ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΓΕΝΙΚΟ ΤΜΗΜΑ
ΤΟΜΕΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ, ΜΑΘΗΜΑΤΙΚΩΝ & ΣΤΑΤΙΣΤΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Αυτοματοποιημένα εργαλεία εξόρυξης βιο-γεωδεδομένων μέσω μεθόδων εξελικτικής κατάτμησης χρονοσειρών

ΘΩΜΑΣ Ι. ΓΚΛΕΖΑΚΟΣ

ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Επιβλέπων:
Καθηγητής Θεόδωρος Τσιλιγκιρίδης

Κωνσταντίνος Γιαλούρης	Λάζαρος Ηλιάδης
Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής

Αθήνα, Ιούλιος 2012

Επταμελής Εξεταστική Επιτροπή

Καθηγητής Θεόδωρος Τσιλιγκιρίδης
Γεωπονικό Πανεπιστήμιο Αθηνών
Τομέας Πληροφορικής, Μαθηματικών & Στατιστικής
Εργαστήριο Πληροφορικής

Αναπληρωτής Καθηγητής Κωνσταντίνος Γιαλούρης
Γεωπονικό Πανεπιστήμιο Αθηνών
Τομέας Πληροφορικής, Μαθηματικών & Στατιστικής
Εργαστήριο Πληροφορικής

Αναπληρωτής Καθηγητής Λάζαρος Ηλιάδης
Δημοκρίτειο Πανεπιστήμιο Θράκης
Τμήμα Δασολογίας και Διαχείρισης Περιβάλλοντος και Φυσικών Πόρων

Καθηγητής Σπυρίδων Κίντζιος
Γεωπονικό Πανεπιστήμιο Αθηνών
Τμήμα Γεωπονικής Βιοτεχνολογίας
Εργαστήριο Φυσιολογίας και Μορφολογίας Φυτών

Καθηγητής Σπυρίδων Λυκοθανάσης
Πανεπιστήμιο Πατρών
Τμήμα Μηχανικών ΗΥ & Πληροφορικής
Τομέας Εφαρμογών και Θεμελιώσεων της Επιστήμης των Υπολογιστών

Καθηγητής Θεμιστοκλής Παναγιωτόπουλος
Πανεπιστήμιο Πειραιά
Τμήμα Πληροφορικής

Καθηγητής Ανδρέας-Γεώργιος Σταφυλοπάτης
Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

στην Ειρήνη,
στον Γιάννη και στο Νώντα

Περιεχόμενα

Δημοσιευμένο ερευνητικό έργο **xv**

Ευχαριστίες **xix**

Περίληψη **xxiii**

Abstract **xxv**

ΠΡΟΛΟΓΟΣ **1**

Κίνητρα και στόχοι της διατριβής 2

Διάταξη της διατριβής 7

1 ΑΝΑΠΑΡΑΣΤΑΣΗ ΧΡΟΝΟΣΕΙΡΩΝ **11**

1.1 Ορισμός και παραδείγματα 12

1.2 Δειγματοληπτικοί αλγόριθμοι 14

1.3 Αλγόριθμοι τμηματικής αναπαράστασης 16

1.4 Τμηματική Γραμμική Αναπαράσταση 21

1.4.1 Αλγόριθμοι διολισθαίνοντος παραθύρου 24

1.4.2 Από-πάνω-προς-τα-κάτω διαδοχική διχοτόμηση 31

1.4.3 Από-κάτω-προς-τα-πάνω διαδοχική συγχώνευση 35

2 ΤΑΞΙΝΟΜΗΤΕΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΧΡΟΝΟΣΕΙΡΕΣ **39**

2.1 Εισαγωγή 39

2.2 Σύντομη ιστορική αναδρομή 41

2.3 Τεχνητά Νευρωνικά Δίκτυα 44

2.3.1 Ορισμοί και χαρακτηριστικά 44

2.3.2 Αρχιτεκτονική ΤΝΔ 47

2.3.3 Πολυεπίπεδα Προσθιόδρομα Perceptrons 52

2.3.4 Αρετές και Περιορισμοί του Αλγορίθμου 53

2.4 Μηχανές Διανυσμάτων Υποστήριξης 54

2.4.1 Διανύσματα Υποστήριξης 55

2.4.2 Ελαστικότητα περιθωρίου υπερ-επιφανειών 58

2.4.3 Μη γραμμική ταξινόμηση 61
2.5 Ανάλυση δεδομένων χρονοσειρών μέσω ταξινομητών ΤΝ 62
2.5.1 Αναπαράσταση χρονοσειρών 62
2.5.2 Προεπεξεργασία χρονοσειρών 65
2.6 Γενετικοί Αλγόριθμοι 68
2.6.1 Ιστορικά Στοιχεία 69
2.6.2 Μεθοδολογία 70
Αρχικοποίηση 71
Επιλογή 71
Αναπαραγωγή 73
Τερματισμός 74
2.6.3 Υποθέσεις επί της προσαρμοστικότητας των ΓΑ 74
2.7 Μειονεκτήματα ταξινομητών 77

3 ΠΛΑΙΣΙΟ ΤΗΣ ΕΡΕΥΝΑΣ **81**
3.1 Περιορισμοί αναπαράστασης και αντιμετώπιση μέ-σω της ΠΕΤ . . . 83
3.2 Μελέτες περίπτωσης και εφαρμογή της ΠΕΤ 85
3.3 Ερευνητικό περιβάλλον 86
3.3.1 Βιοηλεκτρική Δοκιμή Αναγνώρισης BERA 86
3.3.2 Διαχείριση χειμαρρικής επικινδυνότητας 90
3.3.3 Εξελικτικοί αλγόριθμοι στη διαχείριση της χειμαρρικής επικινδυνότητας 91
3.4 Προδιαγραφές του συστήματος 93

4 ΑΝΑΛΥΣΗ ΤΟΥ ΠΡΟΤΥΠΟΥ **97**
4.1 Πρότυπη Εξελικτική Τμηματοποίηση 98
4.1.1 Ορισμός εκπαιδευτών 98
4.1.2 Λειτουργική δομή εκπαιδευτών 99
4.1.3 Καθορισμός τμημάτων 101
4.1.4 Αποτύπωση του σχήματος τμηματοποίησης 102
4.1.5 Σχηματισμός εξελικτικών δεδομένων 104
4.2 Παράδειγμα εργασίας 105

5 ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ **113**
5.1 Δημιουργία Εξελικτικών Δεδομένων 118
5.2 Αξιολόγηση Εκπαιδευτών 121
5.3 Ενδιάμεση γενεά 125

5.4 Πλατφόρμα ανάπτυξης λογισμικού 130
5.5 Ανάπτυξη ταξινομητών και παραμετροποίηση 134
5.5.1 Νευρωνικός ταξινομητής 135
5.5.2 Ταξινομητής Διανυσμάτων Υποστήριξης 137
5.6 Προεπισκόπιση προβλημάτων εφαρμογής 140

6 ΤΑΞΙΝΟΜΗΣΗ ΦΥΤΙΚΩΝ ΙΩΝ 143
6.1 Συμπτωματολογία των ιών TRV και CGMMV 144
6.2 Παραμετροποίηση του συστήματος 147
6.3 Χειρισμός των δεδομένων και αποτελέσματα 149
6.4 Σχολιασμός 155

7 ΠΡΟΒΛΕΨΗ ΧΕΙΜΑΡΡΙΚΗΣ ΕΠΙΚΙΝΔΥΝΟΤΗΤΑΣ 161
7.1 Πλημμυρικά φαινόμενα 162
7.2 Η περιοχή μελέτης 163
7.3 Χειρισμός των δεδομένων 164
7.4 Αποτελέσματα 168
7.5 Σχολιασμός 170

8 ΣΥΜΠΕΡΑΣΜΑΤΑ 173
8.1 Περιβάλλον της έρευνας 173
8.2 Πρότυπη Εξελικτική Τμηματοποίηση 176
8.3 Συμπεράσματα και κριτική αξιολόγηση 176
8.4 Μελλοντική επέκταση 180

Βιβλιογραφία 183

ΠΑΡΑΡΤΗΜΑ 199

Κατάλογος σχημάτων

1.1 *Δειγματοληψία επί της αρχικής χρονοσειράς (a) έχει ως αποτέλεσμα υπερβολική παραμόρφωση (b).* 15

1.2 *Εξομάλυνση χρονοσειράς μέσω κατάτμησης και εξαγωγής του μέσου κάθε τμήματος.* 17

1.3 *Χρονοσειρές και η ΤΓΑ αναπαράστασή τους (A: διαστημική τηλεμετρία, B: Ηλεκτροκαρδιογράφημα) (κατά Keogh et al., 2003).* 22

1.4 *Σωματιδιακή κίνηση Brown (μπλε γραμμή) και η Τμηματική Γραμμική Αναπαράστασή της (κόκκινη γραμμή) μέσω παρεμβολής (κατά Markussen, 2007).* 22

1.5 *Ετήσια παραγωγή ορυζώνων στην περιοχή Zhejiang της Κίνας και η Τμηματική Γραμμική Αναπαράστασή της μέσω παλινδρόμησης [68].* 23

1.6 *Πορεία αναπαράστασης χρονοσειράς βάσει του αλγορίθμου διολισθαίνοντος παραθύρου (κατά Padergnana και Furlani).* 25

1.7 *Σχηματική παράσταση του αλγορίθμου Douglas - Peucker / Ramer.* 33

1.8 *Συμπίεση δεδομένων εικόνας σύμφωνα με τον αλγόριθμο Douglas-Peucker [187].* 34

1.9 *Πορεία αναπαράστασης χρονοσειράς βάσει του αλγορίθμου από-κάτω-προς-τα-πάνω (κατά Padergnana και Furlani).* 35

2.1 *Γραφική αναπαράσταση αρχιτεκτονικής πολυεπίπεδου perceptron τριών επιπέδων.* 48

2.2 *Διαδικασία ενεργοποίησης μεμονωμένου νευρώνα.* 50

2.3 *Συναρτήσεις δραστηριοποίησης ANN.* 51

2.4 *Απλούστευση της διαδικασίας διαχωρισμού κλάσεων σε ανώτερη διάσταση.* 55

2.5 *Η ανάλυση ΜΔΥ επιλέγει τη μοναδική επιφάνεια διαχωρισμού που έχει τέτοια κλίση, ώστε να μεγιστοποιείται το περιθώριο διαχωρισμού μεταξύ των κλάσεων.* 56

2.6 *Εντοπισμός διανυσμάτων υποστήριξης και καθορισμός μέγιστου περιθωρίου μέσω της βέλτισης επιφάνειας διαχωρισμού.* 58

2.7 *Υπο- και υπερ-προσαρμογή δεδομένων από τον ταξινομητή. Και στις δύο περιπτώσεις η ταξινόμηση είναι ανεπαρκής: στην πρώτη λόγω μειωμένης ακρίβειας και στη δεύτερη λόγω κάλυψης μικρού χώρου για τη μια από τις κλάσεις. Η τρίτη περίπτωση περιγράφει ταξινόμηση μειωμένου σφάλματος γενίκευσης.* 59

2.8 *Υλοποίηση γραμμικού διαχωρισμού με αποδοχή ποσοστού σφαλμάτων ταξινόμησης: Το αντίτιμο του σφάλματος υπολογίζεται ως η απόσταση από την επιφάνεια διαχωρισμού, πολλαπλασιασμένη με την παράμετρο κόστους.* . . 60

3.1 *Σχηματική παράσταση ενός τυπικού αισθητήρα BERA. Διακρίνονται (α) το βιολογικό τμήμα το οποίο έχει υποστεί επεξεργασία για την ενσωμάτωση μέσω μεθόδων ηλεκτρο-εισαγωγής αντισωμάτων ή άλλων μορίων, (β) το φυσικοχημικό τμήμα ανίχνευσης με τα ηλεκτρόδια μέτρησης και τον μετατροπέα, (γ) το σύστημα υπολογισμών και αποθήκευσης ([15]).* 87

5.1 *Διάγραμμα ροής της λειτουργίας του συστήματος.* 114
5.2 *Ροή πληροφορίας μεταξύ των βασικών μεθόδων του αλγορίθμου.* 117

6.1 *Καταγραφή της Ακρίβειας των ταξινομητών ΤΝΔ και ΜΔΥ καθ' όλες τις γενεές του προτεινόμενου αλγορίθμου για το πρόβλημα της αναγνώρισης των φυτικών ιών.* 152
6.2 *Μέσες τιμές ακρίβειας ανά γενεά ΤΝΔ (αναγνώριση φυτικών ιών).* 154
6.3 *Μέγιστες τιμές ακρίβειας ανά γενεά ΤΝΔ (αναγνώριση φυτικών ιών).* 155
6.4 *Μέσες τιμές ακρίβειας ανά γενεά ΜΔΥ (αναγνώριση φυτικών ιών).* 156
6.5 *Μέγιστες τιμές ακρίβειας ανά γενεά ΜΔΥ (αναγνώριση φυτικών ιών).* . . . 157
6.6 *Αρχική (α) και εξελικτικά επεξεργασμένη μέσω της μεθόδου ΠΕΤ (β) χρονοσειρά αποκρίσεων του ιού TRV.* 158

7.1 *Πρόβλεψη χειμαρρικής επικινδυνότητας με εφαρμογή ΤΝΔ εκπαιδευμένου με το καλύτερο σχήμα εξελικτικής τμηματοποίησης της 28ης γενεάς και εφαρμοσμένου στο σετ αξιολόγησης. Η γαλάζια συνεχής γραμμή αναπαριστά την πρόβλεψη του συστήματος, ενώ η κόκκινη διακεκομμένη αναπαριστά πραγματικές τιμές.* 170

Κατάλογος Εικόνων

5.1 *Γραφικό Περιβάλλον Διεπαφής Χρήστη του εργαλείου λογισμικού.* 131
6.1 *Πρασινοκίτρινες ομόκεντρες κηλιδώσεις σε φύλλα παιωνίας (Paeonia Sara Bernhardt) που οφείλονται στον ιό του κροταλίσματος του καπνού (κατά Chastagner και Pappu, http://www.apsnet.org).* 144
6.2 *Καστανοκίτρινες τοξοειδείς κηλιδώσεις στη σάρκα με προβολές στην επιφάνεια καρπού πατάτας (Solanum tuberosum) που οφείλονται στον ιό του κροταλίσματος του καπνού (Πηγή: United Nations Economic Commission for Europe, http://www.unece.org).* 145
6.3 *Ο ιός του κροταλίσματος του καπνού (σάρωση από φωτογραφία σε ανάλυση 150 dpi στα 8 bits/pixel greyscale, πηγή: Rothamsted Research, http://www.rothamsted.ac.uk/ppi/links/pplinks/virusems/).* 146
6.4 *Συμπτώματα του ιού της πράσινης ποικιλοχλώρωσης με μωσαϊκό σε φυτά Cucurbitaceae. (a) Σοβαρή προσβολή σε φύλλο Cucumis sativus θερμοκηπίου (b) ήπια προσβολή σε φύλλο Cucumis melo ([131]).* 147
6.5 *Προσβολές CGMMV μέτριας έως σοβαρής έντασης σε φύλλα φυτών θερμοκηπίου (a) χλωρωτική κάκωση επί Chenopodium amaranticolor (b) μωσαϊκό και εξέλκωση σε Citrullus lanatus (c) και (d) έντονη προσβολή Cucumis sativus και Cucumis melo αντίστοιχα (κατά [131]).* 148
6.6 *Πρωτογενή και εξελικτικά δεδομένα και αποτελέσματα δεύτερης γενεάς για την περίπτωση ταυτοποίησης ιών, στο εργαλείο λογισμικού.* 151
7.1 *Γενική άποψη του υδρογραφικού δικτύου της Δημοκρατίας της Κύπρου.* . . 165
7.2 *Δημοκρατία της Κύπρου: Μέσο ετήσιο ύψος βροχής (σε mm) από το 1970 έως το 2000.* 166
7.3 *Πρωτογενή και εξελικτικά δεδομένα και αποτελέσματα για το πρόβλημα της χειμαρρικής επικινδυνότητας στο εργαλείο λογισμικού.* 168

Κατάλογος πινάκων

4.1 Παράδειγμα της δομής εκπαιδευτή 100
4.2 Αποτύπωση του χρωμοσώματος εκπαιδευτή επί πρωτογενούς χρονοσειράς και παραγωγή εξελικτικών δεδομένων σύμφωνα με το τμήμα πυρήνα 111

5.1 Έλεγχος συναρτήσεων ενεργοποίησης για τους νευρώνες του ενδιάμεσου επιπέδου και του επιπέδου εξόδου του ΤΝΔ 138
5.2 Χρόνοι εκπαίδευσης του συστήματος για τις δύο μελέτες περιπτώσεων 142

6.1 Παραμετροποίηση του συστήματος για τη μελέτη περίπτωσης της ταξινόμησης των φυτικών ιών 149
6.2 Δεδομένα εκπαίδευσης, ελέγχου και αξιολόγησης για το πρόβλημα αναγνώρισης των φυτικών ιών 150
6.3 Αποτελέσματα της εφαρμογής του προτεινόμενου εργαλείου στην ταυτοποίηση των φυτικών ιών CGMMV και TRV: Δεδομένα της πρωτογενούς χρονοσειράς έναντι εξελικτικών δεδομένων υπό ΤΝΔ και ΜΔΥ, με την καλύτερη ακρίβεια (Acc) 153

7.1 Δείγμα δεδομένων για το πρόβλημα πρόβλεψης πλημμυρικών κινδύνων: M1-M12:Μήνες του έτους, Fn: επιφάνεια της λεκάνης απορροής, Q_{max}: μέγιστη παροχή ύδατος, P: μέση ετήσια βροχόπτωση, H: απόλυτο υψόμετρο, Jk: απόλυτη κλίση, Qmy: μέση ετήσια παροχή ύδατος. 167
7.2 Αποτελέσματα πρόβλεψης χειμαρρικής επικινδυνότητας: Αρχική χρονοσειρά έναντι εξελικτικών δεδομένων υπό ΤΝΔ και ΜΔΥ, ταξινόμηση με αύξον RMSE 169

Κυριότερα Αρκτικόλεξα

Λατινικά

AAWS	Average Annual Water Supply
AI	Artificial Intelligence
ANN	Artificial Neural Network
APCA	Adaptive Piecewise Constant Approximation
ARCH	Autoregressive Conditional Heteroscedasticity
ARIMA	Auto-Regressive Integrated Moving Average
AZTEC	Amplitude Zone Time Epoch Coding
BBH	Building Block Hypothesis
BERA	Bioelectric Recognition Assay
CGMMV	Cucumber Green Mottle Mosaic Virus
CI	Computational Intelligence
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transformation
EA	Evolutionary Algorithm
ECG	Electrocardiogram
EDA	Exploratory Data Analysis
EDS	Evolutionary Data Set
FEMA	Federal Emergency Management Agency
fMRI	functional Magnetic Resonance Imaging
FP	False Positive sample
GA	Genetic Algorithm
KOSPI	Korean Composite Stock Price Index
LS-SVM	Least Squares - Support Vector Machine
MDL	Minimum Description Length
MLP	Multi-Layer Perceptron
MML	Minimum Message Length
NPV	Negative Prediction Value
PAA	Piecewise Aggregate Approximation
PCA	Principal Component Analysis
PCApp	Piecewise Constant Approximation

PES	Piecewise Evolutionary Segmentation
PIP	Perceptually Important Points
PLR	Piecewise Linear Representation
PPV	Positive Prediction Value
RBF-SVM	Radial Basis Function - Support Vector Machine
RDBMS	Relational Database Management Systems
RMSE	Root Mean Square Error
SAPA	Scan-Along Polygonal Approximation
SAX	Symbolic Aggregate Approximation
SBASS	Segment-Based Approach for Subsequence Search
Sens	Sensitivity
SOM	Self Organizing Maps
Spec	Specificity
SRM	Structured Risk Minimization
SSA	Singular Spectrum Analysis
SSD	Sum Square Difference
SSGE	Scale-Sensitive Gated Experts
SSPSO	Split – Step Particle Swarm Optimization
SVM	Support Vector Machine
TN	True Negative sample
TP	True Positive sample
TRV	Tobacco Rattle Virus
WMRA	Wavelet Multi-Resolution Analysis

Ελληνικά

ΑΚΣ	Ανάλυση Κυρίαρχων Συνιστωσών
ΑΣΣ	Αντιληπτικά Σημαντικά Σημεία
ΓΑ	Γενετικοί Αλγόριθμοι
ΔΑΔ	Διερευνητική Ανάλυση Δεδομένων
ΕΚΕΖ	Εποχική Κωδικοποίηση Εύρους Ζώνης
ΗΚΓ	Ηλεκτροκαρδιογράφημα
ΜΔΥ	Μηχανή Διανυσμάτων Υποστήριξης
ΜΔΥ-ΕΤ	Μηχανή Διανυσμάτων Υποστήριξης Ελαχίστων Τετραγώνων
ΟΣΔΕ	Ολοκληρωμένο Σύστημα Διαχείρισης και Ελέγχου Υπουργείου Αγροτικής Ανάπτυξης και Τροφίμων

ΠΕΤ	Πρότυπη Εξελικτική Τμηματοποίηση
ΠΤΣΠ	Προσαρμοζόμενη Τμηματικά Σταθερή Προσέγγιση
ΣΑΠ	Συμβολική Αθροιστική Προσέγγιση
ΣΔΣΒΔ	Συστήματα Διαχείρισης Σχεσιακών Βάσεων Δεδομένων
ΤΑΠ	Τμηματικά Αθροιστική Προσέγγιση
ΤΝ	Τεχνητή Νοημοσύνη
ΤΝΔ	Τεχνητά Νευρωνικά Δίκτυα
ΤΣΠ	Τμηματικά Σταθερή Προσέγγιση
ΤΤΓΑ	Τμηματικά Γραμμική Αναπαράσταση
ΥΝ	Υπολογιστική Νοημοσύνη

Δημοσιευμένο ερευνητικό έργο

Α. Δημοσιεύσεις Διατριβής

Α1. Δημοσιεύσεις σε Διεθνή Επιστημονικά Περιοδικά:

1. Glezakos, T.J., Tsiligiridis, T.A., Yialouris, C.P., 2012. Piecewise evolutionary segmentation for feature extraction in time series models, Neural Computing and Applications Journal, Springer [Article in Press]

2. Glezakos, T.J., Moschopoulou, G., Tsiligiridis, T.A., Kintzios, S., Yialouris, C.P., 2009. Plant virus identification based on neural networks with evolutionary preprocessing. Journal of Computers and Electronics in Agriculture doi:10.1016/j.compag.2009.09.007

3. Glezakos, T.J., Tsiligiridis, T.A., Iliadis, L.S., Yialouris, C.P., Maris, F.P., Ferentinos, K.P., 2009. Feature Extraction for Time Series Data: an Artificial Neural Network Evolutionary Training Model for the Management of Mountainous Watersheds. Jour nal of NeuroComputing, Neurocomputing 73 (2009) 49–59, doi:10.1016/j.neucom.2008.08.024.

4. Glezakos, TJ., Tsiligiridis, T.A., Yialouris, C.P., 2012. Pesticide residues monitoring using piecewise evolutionary segmentation [Article in preparation]

Α2. Δημοσιεύσεις σε Πρακτικά Διεθνών Συνεδρίων:

1. Glezakos. T.J., Tsiligiridis, T.A., Kintzios, S. and Yialouris, C.P., 2010. Time-series piecewise evolutionary segmentation based on wavelet transformation and Support Vector Machines. In Siddiqi, A.H., Ucan, O.N., Aslan, Z., Oz, H.H., Zontul, M. and G. Erdemir (Eds.) Proceedings of The Fifth International Symposium On Wavelet Applications to World Problems (IWW-2010), 7-8 June, Istanbul, Turkey, ISBN: 978 650 4303 038.

2. Glezakos, T.J., Moschopoulou, G., Tsiligiridis, T.A., Kintzios, S. and C.P. Yialouris, 2008. Assessing the classification accuracy of an evolutionary neural network: the case of plant virus identification, HAICTA 2008, Information Systems & Innovative Technologies in Agriculture, Food and Environment, Athens, Hellas

3. Glezakos, T.J., Tsiligiridis, T.A., Yialouris, C.P., Maris, F., Ferentinos, K.P., 2007. Feature Extraction for Time Series Data: an Artificial Neural Network Evolutionary

Training Model for the Management of Mountainous Watersheds. EANN 2007, 10th International Conference on Engineering Applications of Neural Networks, 29-31 August 2007, Thessaloniki, Hellas.

4. Glezakos, T.J., Maris, F.P., Iliadis, L.S., Tsiligiridis, T.A. and C.P. Yialouris, 2006. Neuro-Genetic Modeling of Torrential Environmental Risk: the Case of the Lakes Volvi and Koroneia. In Tzortzios S., Dalezios N.R. and N. Samaras (Eds), International Conference on: Information Systems in Sustainable Agriculture, Agroenvironment and Food Technology, ISBN: 960-8029-42-2 (set), http://www.epegenorth.gr, pp. 801 – 818, University of Thessaly, Volos, Hellas.

5. Kaloudis, S., Glezakos, T.J., Ferentinos, K.P., Tsiligiridis T.A., and C.P. Yialouris, 2006. Feedforward Neural Network Modeling of Fir Taper in Natural Forests of Greece, International Conference on Sustainable Management and Development of Mountainous and Island Areas, September 29 - October 1 2006, Naxos, Hellas.

6. Glezakos, T.J. and Tsiligiridis T., 2002. Neural Networks for Landscape Applications. In Sideridis, A.B. and C.P. Yialouris, (Eds), 2002, The Impact of ICT in Agriculture, Food and Environment, 1st Conference of Hellenic Association of ICT in Agriculture, Food and Environment, Agriculture University of Athens, Hellas.

Α3. Αναφορές από Τρίτους (citations):

1. Siddiqui, Q.T.M., Hashmi, H.N., Ghumman, A.R., ur Rehman Mughal, H., 2011. Flood Inundation Modeling for a Watershed in the Pothowar Region of Pakistan. Arabian Journal for Science and Engineering, 36(7), pp. 1203-1220, DOI: 10.1007/s 13369-011-0112-2

2. Zhang, Y.D., Wu, L.N., Huo, Y.K., Wang, S.H., 2011 A Novel Global Optimization Method – Genetic Pattern Search. Applied Mechanics and Materials, 44-47, pp. 3240-3244, DOI:10.4028/www.scientific.net/AMM.44-47.3240

3. Rivero, D., Dorado, J., Rabuñal, J., Pazos, A., 2010. Generation and simplification of Artificial Neural Networks by means of Genetic Programming. Neurocomputing, 73(16–18), pp. 3200–3223, DOI: http://dx.doi.org/10.1016/j.neucom.2010.05.010

Β. Άλλο Δημοσιευμένο Ερευνητικό Έργο:

Β1. Διεθνές Περιοδικό

- Pontikakos, C., Sambrakos, M., Glezakos, T.J. and Tsiligiridis T., 2006. Location-based Services: A Framework for an Architecture Design. International Journal of Neural, Parallel & Scientific Computations (NPSC), June, 2006.

Β2. Πρακτικά Διεθνών Συνεδρίων

- Pontikakos, C., Glezakos, T.J., and Tsiligiridis T., 2005. Location-based Services: Architecture Overview. In: Proceedings of the International Congress on Information Technology in Agriculture, Food and Environment (ITAFE'05), October 12-14, 2005, Adana, Turkey.

- Sabrakos, M., Glezakos, T.J. and Tsiligiridis T., 2004. On Evaluative Measurement of Landscape Change. In Vlachopoulou, M., Manthou, V., Iliadis, L., Gertsis, S. and M. Salampasis (Eds.), 2004, HAICTA 2004, Information Systems & Innovative Technologies in Agriculture, Food and Environment, Volume 2, ISBN: 960-287-048-6 (set), http://www.epegenorth.gr, pp. 215 – 219, Aristotle University of Thessalonika, University of Macedonia, Technological Educational Institute, Thessaloniki, Hellas.

- Glezakos, T.J. and Tsiligiridis T., 2003. Environmental Management. An Overview of Landscape Metrics. In Sindir, K.O. (Ed.), 2003, ITAFE '03, International Congress on Information Technology in Agriculture, Food and Environment, pp. 90 – 96, Ege University, Bornova - Izmir, Turkey.

Ευχαριστίες

Χρόνια πριν, δουλεύοντας στο γραφείο που μου είχε παραχωρηθεί στο Εργαστήριο Πληροφορικής του Γεωπονικού Πανεπιστημίου της Αθήνας και έχοντας εκείνη τη στιγμή προβλήματα με τον κώδικα, άκουσα τη φωνή του φίλου - και συνοδοιπόρου για ένα διάστημα σε αυτό το συναρπαστικό ταξίδι - του Γιάννη Φίλη: «Μην ανησυχείς. Θα 'ρθει η στιγμή να γράφεις τις ευχαριστίες στη διατριβή σου και δεν θα το 'χεις καταλάβει!» Στην αρχή γέλασα κοροϊδευτικά. Να, όμως, που όντως έφθασε η στιγμή, Γιάννη! Ας προσπαθήσω, λοιπόν, να βάλω σε μια τάξη τις σκέψεις μου και να ευχαριστήσω όλους όσοι συνέβαλαν στην εκπόνηση αυτής της εργασίας.

Οπωσδήποτε, πολλοί άνθρωποι βοήθησαν, βάζοντας έστω ένα μικρό λιθαράκι στο οικοδόμημα αυτό και η αναφορά σε καθένα όνομα ξεχωριστά ίσως ήταν μια διατριβή από μόνο του. Δεν θα ήταν δυνατόν όμως να μην ξεχωρίσω κάποιες συγκεκριμένες παρουσίες που έπαιξαν ρόλο καταλυτικό. Θα είμαι για πάντα ευνώμων στον καθηγητή κ. Θεόδωρο Α. Τσιλιγκιρίδη, που εμπιστεύθηκε ευθύς εξαρχής τη δυνατότητά μου να φέρω σε πέρας το έργο αυτό και καθοδήγησε επιβλέποντας με ατέρμονη επιστημονική πληρότητα την εργασία μου. Πολύ περισσότερο όμως γιατί ανέχθηκε και ξεπέρασε με μαεστρία όλους τους μικρούς Γολγοθάδες, διδάσκοντάς με καταρχήν την επιμονή στην επιστημονική έρευνα, αλλά και την αξία της καινοτομίας στη δουλειά μου. Η εργασία του είναι μεθοδική, επιστημονική και, στο μέτρο που μου επιτρεπόταν, δανείστηκα μια πλειάδα ιδεών από αυτή. Η μαθηματική προτυποποίηση του εξελικτικού πυρήνα στο σύστημα που αναπτύχθηκε για την παρούσα διατριβή για παράδειγμα, θα ήταν αδύνατη χωρίς την καίρια συμβολή του κ. Τσιλιγκιρίδη. Μακάρι να είχα τη δυνατότητα - ή να μου δοθεί στο μέλλον - να συνεργαστώ περισσότερο και να δανειστώ κι άλλες ιδέες από τη δουλειά του. Σας ευχαριστώ *de profundis* κύριε καθηγητά!

Δεν έχω λόγια για να εκφράσω το πόσο πολύ θέλω να ευχαριστήσω τους καθηγητές Κωνσταντίνο Γιαλούρη και Λάζαρο Ηλιάδη, τόσο για την ανοχή και την υπομονή, όσο και για τις επιστημονικά εμπεριστατωμένες απαντήσεις τους στις ερωτήσεις πάσης φύσης τεχνικού περιεχομένου, με τις οποίες τους πολιορκούσα για καιρό. Ο κ. Γιαλούρης ήταν ο δάσκαλος που μου έδειξε ότι η ιστορία είναι καλή μαζί σου όταν τη γράφεις μόνος σου. Και αυτό υποδεικνύοντάς μου μετ' επιμονής την ανάπτυξη του προτεινόμενου συστήματος, ειδικά της εξελικτικής μεθόδου, με κώδικα που έγραψα ο ίδιος, χωρίς τη χρήση έτοιμων πακέτων λογισμικού. Υποστήριξε δε έμπρακτα την επιμονή του αυτή βοηθώντας με στην πρωτογενή ανάλυση και τη σχεδίαση του εργαλείου λογισμικού. Πιστεύω ακράδαντα ότι δεν θα μπορούσα σε καμιά περίπτωση να φθάσω στο επίπεδο να προτείνω πρόγραμμα τέτοιας πο-

λυπλοκότητας χωρίς την μεστού επιστημονικού και τεχνικού περιεχομένου αρωγή του. Ο κ. Ηλιάδης ήταν αυτός που πρώτος με ενέπνευσε να ασχοληθώ με το αντικείμενο της υπολογιστικής νοημοσύνης, και να το συνδυάσω με την επίλυση προβλημάτων του γεωτεχνικού τομέα, παρέχοντας παράλληλα τα πρωτογενή δεδομένα για το πρόβλημα της πρόβλεψης των πλημμυρικών κινδύνων. Θαυμάζω την επιστημονική του πληρότητα και θα ήθελα κάποτε να πλησιάσω την κατάρτισή του σε θέματα υπολογιστικής νοημοσύνης. Στον καθηγητή Σπυρίδωνα Κίντζιο οφείλω ένα μεγάλο ευχαριστώ για την υπομονή του να εξηγήσει τη μέθοδο BERA και να συνεισφέρει στη διατριβή με τα πρωτογενή δεδομένα αναγνώρισης των φυτικών ιών CGMMV και TRV. Πραγματικά νοιώθω πολύ τυχερός που είχα την καθοδήγηση τέτοιων δασκάλων.

Στο σημείο αυτό θα ήταν παράλειψή μου να μην εκφράσω τις πιο θερμές ευχαριστίες μου στο διευθυντή του Εργαστηρίου Πληροφορικής του Γ.Π.Α., καθηγητή Αλέξανδρο Β. Σιδερίδη που πίστεψε ακράδαντα στη δυνατότητά μου να συνδυάσω τη γεωτεχνική επιστήμη με την πληροφορική. Ο κ. Σιδερίδης δεν ενεπλάκη στο τεχνικό μέρος της διατριβής, αν και είμαι σίγουρος ότι θα είχα εκ μέρους του την καλύτερη επιστημονική υποστήριξη αν του το ζητούσα. Είναι ένας άνθρωπος με όραμα - η ψυχή του Εργαστηρίου - και, παρότι μια αυθεντία πλέον στον τομέα, ήταν πάντα εκεί, προσηνής αλλά και αποτελεσματικός στην επίλυση οποιουδήποτε θέματος προέκυπτε. Κύριε Σιδερίδη, τη στήριξή σας τη χρειαζόμαστε, περισσότερο τώρα, σε αυτούς τους χαλεπούς καιρούς.

Ενα μεγάλο επίσης ευχαριστώ, ίσως το μεγαλύτερο, νοιώθω την υποχρέωση να εκφράσω στον καθηγητή και Υπουργό Αγροτικής Ανάπτυξης Ναπολέοντα Μαραβέγια, ο οποίος ήταν ο πρώτος που με παρότρυνε να αναλάβω αυτή την ευθύνη. Για τη στήριξη και τα καλά σας λόγια ως Πρόεδρος του Δ.Σ του ΕΘ.Ι.ΑΓ.Ε. και ως Ακαδημαϊκός, σας ευχαριστώ κύριε καθηγητά - και όχι κύριε Υπουργέ, γιατί η δεύτερη ιδιότητα φθείρεται φθείροντας, ενώ η πρώτη μένει ανεξίτηλη.

Για τους προϊσταμένους μου στο Υπουργείο Αγροτικής Ανάπτυξης και Τροφίμων, όσες ευχαριστίες και να εκφράσω πάλι θα δείχνουν λίγες. Ο Διευθυντής της Διεύθυνσης Πληροφορικής Γιώργος Αθανασίου και ο Τμηματάρχης Βασίλης Κανελλόπουλος πρέπει να θεωρούνται από τους ανθρώπους που υποστήριξαν με τη μεγαλύτερη ζέση την εκπόνηση αυτής της διατριβής. Είναι στην πραγματικότητα οι άνθρωποι που αγκάλιασαν αυτή την προσπάθειά μου, παρέχοντας κάθε διευκόλυνση στο πλαίσιο της εργασιακής μας σχέσης. Η συμπαράσταση που μου παρείχαν δεν είναι δυνατό να περιγραφεί σε αυτές τις λίγες αράδες και τους είμαι πραγματικά ευγνώμων. Θα ήθελα να ευχαριστήσω τους συναδέλφους μου Μιχάλη Ροδαρέλη, Γιάννη Δασκαλάκη και Γιώργο Μπαναγή για τη στήριξη, ηθική και ουσιαστική που μου παρείχαν. Είμαι βέβαιος ότι αν τους το ζητούσα θα αναλάμβαναν οποιαδήποτε δική μου εργασία προκειμένου να βρω χρόνο για περισσότερη μελέτη. Τα λόγια δεν αρκούν επίσης για τον Κώστα Μπουτεράκο, τον Παναγιώτη Μακρή

και το Γιώργο Λύτρα. Ελπίζω να μου δοθεί η ευκαιρία να επιστρέψω ένα μέρος της αλληλεγγύης που έδειξαν.

Στο συμφοιτητή μου Κώστα Ποντικάκο με τον οποίο ξεκινήσαμε μαζί θα ήθελα να απευθύνω ένα μεγάλο ευχαριστώ για την υπομονή του στους δύσκολους καιρούς. Σε όλους επίσης τους εργαζόμενους στο Εργαστήριο Πληροφορικής, θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες. Μαρία Πλέσσα, η διοργάνωση της υποστήριξης της διατριβής μου θα ήταν αδύνατη χωρίς την πολύτιμη βοήθειά σου. Δημήτρη Κόκκα, Αθηνά Τοκατλίδου και Σπύρο Καλούδη, θα ήταν ένα πολύ μοναχικό Εργαστήριο τα απογεύματα της Τετάρτης και της Πέμπτης χωρίς εσάς. Χριστίνα Παπαδοπούλου, ελπίζω κάποτε να αξιωθώ να σου ξαναφέρω βαθμολογίες φοιτητών.

Στους γονείς μου, τους καλύτερους δασκάλους μου, οφείλω ένα μεγάλο ευχαριστώ, όχι για το ζειν, αλλά για το ευ ζειν.

Για το τέλος άφησα τη γυναίκα μου, την Ειρήνη. Με την αμέριστη συμπαράσταση και υπομονή της όλα αυτά τα χρόνια κατέστη δυνατή η ολοκλήρωση της εργασίας μου. Στις καλές στιγμές, αλλά και στις δύσκολες, η Ειρήνη ήταν πάντα εκεί, δίπλα μου. Για τις διορθώσεις που πρότεινες σε αυτό το κείμενο, αλλά κυρίως για την αγάπη και τη στήριξή σου, την ηθική και την έμπρακτη, σε ευχαριστώ.

Ο Γιάννης και ο Νώντας υπέμειναν την αφοσίωσή μου σε κάτι ξένο προς αυτούς, που πολλές φορές δεν καταλάβαιναν. Παρ' όλ' αυτά, είμαι βέβαιος ότι θα χαρούν όταν μετά από χρόνια ξεφυλλίσουν τη διατριβή και δουν τα ονόματά τους τυπωμένα ανάμεσα στις αράδες της.

Θωμάς Ι. Γκλεζάκος
Αθήνα, Ιούλιος 2012.

Περίληψη

Η διαδικασία προτυποποίησης ποικίλων προβλημάτων έχει σε πλείστες όσες περιπτώσεις βασιστεί στη χρήση ιστορικών πληροφοριών. Η πρακτική αυτή στηρίζεται στην παραδοχή ότι κατανεμημένες μετρήσεις[1] παρελθόντων διαδικασιών είναι ικανές να προσφέρουν σημαντική εισροή δεδομένων για την πιστή αναπαραγωγή των φαινομένων που μελετώνται. Έτσι, η ταυτοποίηση, ο χειρισμός και η προτυποποίηση μη γραμμικών δυναμικών συστημάτων που περιγράφονται από δεδομένα με τη μορφή ακολουθίας τιμών χρονικά κατανεμημένων, απέκτησε ιδιαίτερη σημασία. Πράγματι, η έρευνα σε αυτό το πεδίο στοχεύει κυρίως στην προτυποποίηση του μηχανισμού που είναι υπεύθυνος για την παραγωγή αυτών των δεδομένων. Ένας τυπικός μηχανισμός τέτοιων δεδομένων αποθηκεύει μετρήσεις κατά διαδοχικά και σταθερά χρονικά διαστήματα παράγοντας ακολουθίες τιμών, γνωστές ως χρονοσειρές, κάθε μία από τις οποίες αντιστοιχεί σε δεδομένη κατηγορία ή τιμή.

Ένα σημαντικό πρόβλημα που ανακύπτει κατά την ανάλυση μεγάλων συνόλων δεδομένων χρονοσειρών, τόσο από άποψη διαστατικότητας, όσο και από άποψη μεγέθους, σχετίζεται με την επιλογή ενός αντιπροσωπευτικού υποσυνόλου των αρχικών δεδομένων. Εκ των προτέρων επεξεργασία της χρονοσειράς για την απόκτηση ενός αντιπροσωπευτικού δευτερογενούς υποσυνόλου όχι μόνο μειώνει δραστικά το συνολικό χρόνο επεξεργασίας, αλλά επίσης λειτουργεί ως μια διεργασία ομαλοποίησης της αρχικής πληροφορίας για την απομάκρυνση ανεπιθύμητων μη συστηματικών συνιστωσών[2] που δυσκολεύουν την αναλυτική διαδικασία. Οι περισσότερες παραδοσιακές μέθοδοι προ-επεξεργασίας χρονοσειρών, όπως είναι για παράδειγμα η τμηματοποίηση κατά μήκος του άξονα των χρόνων για ταχεία απόκριση, η μη-γραμμική κανονικοποίηση για να δοθεί έμφαση σε σημαντικά τμήματα της πληροφορίας, η εξαγωγή μέσων όρων για αντιμετώπιση των επιπτώσεων του θορύβου, η μείωση του αριθμού των δειγμάτων για την υλοποίηση αποτελεσματικότερων δικτύων, περιλαμβάνουν στατιστικές μεθόδους, όπως δειγματοληπτικές τεχνικές ή διαδικασίες κινούμενου μέσου, οι οποίες χειρίζονται την αρχική πληροφορία με παράθυρα σταθερού μήκους.

[1] Ο όρος κατανεμημένες αναφέρεται στη χρονική επανάληψη των μετρήσεων, η οποία καθορίζεται από συγκεκριμένο κάθε φορά σταθερό ή μεταβλητό βήμα εκτέλεσης.

[2] Ως μη συστηματικό τμήμα της χρονοσειράς νοείται ο διαταρακτικός όρος που εκφράζει το σφάλμα (θόρυβο) το οφειλόμενο σε αποκλίσεις στις μετρήσεις εξαιτίας αστοχίας του οργάνου μέτρησης ή σε άλλες τυχαίες συνθήκες που επικρατούν σε συγκεκριμένο χρόνο. Οι διαταραχές αυτές είναι σε κάποιες περιπτώσεις στιγμιαίες και εμφανίζονται με τη μορφή παλμών μεγάλου ή μεσαίου εύρους, ενώ σε άλλες περιπτώσεις εισέρχονται σε μεγαλύτερο βάθος στα δομικά στοιχεία της χρονοσειράς επηρεάζοντας την τάση και κατεύθυνσή της σε μεγαλύτερο ή μικρότερο βαθμό.

Στην παρούσα εργασία περιγράφεται ο σχεδιασμός, η ανάπτυξη και η εφαρμογή μιας καινοτόμου μεθόδου ελέγχου του βαθμού διάστασης χρονοσειρών, με τη χρήση εργαλείων υπολογιστικής νοημοσύνης. Ο αλγόριθμος που προτείνεται επιτρέπει την παραγωγή περισσότερο προσαρμοσμένων δευτερογενών δεδομένων, αφού προηγουμένως έχει προ-επεξεργασθεί την αρχική χρονοσειρά με εξελικτικό τρόπο με στόχο τη μείωση της διάστασής της και την παράλληλη διατήρηση της δομής των αρχικών δεδομένων παρά τη μεγάλου εύρους εξομάλυνσή τους. Η όλη διαδικασία υλοποιείται με την ανάπτυξη ενός προσαρμοστικού αναλυτικού εργαλείου εξελικτικής φύσης με τη χρήση των Γενετικών Αλγορίθμων, των Τεχνητών Νευρωνικών Δικτύων και των Μηχανών Διανυσμάτων Υποστήριξης.

Το προτεινόμενο εργαλείο δοκιμάστηκε στη λύση δύο προβλημάτων. Η πρώτη μελέτη περιλαμβάνει την περίπτωση ταυτοποίησης φυτικών ϊών. Είναι γενικά παραδεκτό ότι η ανάλυση χρονοσειρών είναι ιδιαίτερα σημαντική για τη φυτοπαθολογία και την ιολογία, ειδικά όσον αφορά την ταυτοποίηση ιών, η οποία στις περισσότερες περιπτώσεις υλοποιείται μέσω αξιολόγησης τέτοιου είδους δεδομένων. Στην πρώτη αυτή περίπτωση, η οποία είναι ουσιαστικά ένα πρόβλημα ταξινόμησης, δεδομένα παραγόμενα με τη μέθοδο της Βιοηλεκτρικής Αναγνώρισης (Bioelectric Recognition Assay BERA) χρησιμοποιήθηκαν για την ανίχνευση και την τελική ταξινόμηση φυτικών ιών και συγκεκριμένα των ιών του κροταλίσματος του καπνού (TRV: *Tobacco Rattle Virus*) και της πράσινης ποικιλοχλώρωσης με μωσαϊκό της αγγουριάς (CGMMV: *Cucumber Green Mottle Mosaic Virus*). Η μέθοδος εισάγει τη χρήση κατάλληλα προεπεξεργασμένων οργανικών αντιδραστηρίων ως αισθητηρίων στοιχείων. Μετά την αντίδραση με τους εν λόγω βιο-αισθητήρες, καθένας από τους ιούς εκθέτει μοναδιαία πρότυπα αισθητηριακής απόκρισης επί ενός ευρέως φάσματος συγκεντρώσεων, καθιστώντας τις αποκρίσεις αυτές ως συγκεκριμένο χαρακτηριστικό ιδίωμα κάθε ιού. Κάθε τέτοιου είδους υπογραφή είναι ουσιαστικά μια γραφική παράσταση βιο-ηλεκτρικών αποκρίσεων στη μονάδα του χρόνου, η οποία χρησιμοποιείται στην ανίχνευση και ταυτοποίηση εκάστου ιού.

Το δεύτερο πρόβλημα στο οποίο εφαρμόσθηκε η μέθοδος σχετίζεται με τη διαχείριση ορεινών υδατικών αποθεμάτων. Τα δεδομένα εισόδου προέρχονται από το νησί της Κύπρου και περιλαμβάνουν δομικά και δυναμικά στοιχεία στα οποία βασική επίδραση ασκούν τα μηνιαία υδατώδη κατακρημνίσματα. Στην περίπτωση αυτή τα αρχικά δεδομένα, που καλύπτουν ένα μεγάλο χρονικό εύρος, ελήφθησαν από μετεωρολογικές βάσεις δεδομένων βροχόπτωσης που ενημερώνονταν από σταθμούς τοποθετημένους σε λεκάνες απορροής διάσπαρτες σε όλο το υδρογραφικό σύστημα του νησιού. Απώτερος σκοπός της έρευνας αποτελεί η ανάπτυξη ενός συστήματος για τον προσδιορισμό της Μέσης Ετήσιας Παροχής Ύδατος (AAWS: Average Annual Water Supply) σε ετήσια βάση για κάθε ορεινή λεκάνη απορροής.

Abstract

Decision making has in many cases engaged time series historical information. This is often used as an exemplification paradigm, on the grounds that past orderly measurements should be able to give enough input so as to reproduce the phenomenon in question. Thus, the identification, manipulation and modelling of non-linear dynamic systems incorporating time series information has become of crucial importance. In fact, most research on such information seeks to reveal the necessity to uncover the mechanism which is responsible for the production of the data. A typical generator of this kind of record-sets utilizes a sequence of vectors, measured at successive constant time intervals. Each vector either corresponds to a given class or a value, the distribution of which describes the phenomenon in question.

An important problem arising while analyzing large time series data sets, both in dimension and size, relates to the proper selection of a subset of the original features. Preprocessing the time series to obtain a representative meta-data set not only significantly reduces computational time, but also functions as a smoothing technique to weed out possible non systematic portions of the initial information, which may, in an extent, inhibit the analytical process. Conventional methods of time series data preprocessing, such as segmentation along the time axis for fast response, nonlinear normalization to emphasize significant information, averaging samples of the plant virus waves to suppress noise effects, reduction in the number of samples to realize a more compact network, include descriptive statistical methods such as re-sampling techniques or moving average procedures, both of which manipulate the initial information in a fixed width fashion. On the other hand, time series analysis plays an important role for phytopathology and virology, especially as regards to virus identification, which is made possible due to time series assessment.

The design, development and implementation of an innovative method is described in this manuscript, aiming to overcome the limitations posed by the fixed width of the analytical tools. The algorithm allows for the production of effective secondary data, after having preprocessed the original time series information in an evolutionary fashion. Thus, it drastically reduces the size of the raw data table to more compact sets of cases and, at the same time, retaining all the crucial information of the initial time-series. This is achieved by the development of analytical tools of evolutionary adaptive width, propelled by Genetic Algorithms, Artificial Neural Networks and Support Vector Machines.

The proposed methodology was applied for the solution of two problems. In the first case, essentially a classification problem, the Bioelectric Recognition Assay (BERA) method was engaged so as to provide information used in the detection and identification of certain plant viruses, namely the *Tobacco Rattle Virus* (TRV) and the *Cucumber Green Mottle Mosaic Virus* (CGMMV), using appropriately preprocessed reagents as the sensing elements. While reacting to the biosensors, each of the viruses in question exhibit unique patterns of

biosensor responses over specific ranges of concentrations, rendering these responses as a special characteristic for each virus, a real identification signature. Each signature is in essence a graphical curve of bioelectrical responses in the time unit, a time series data set, which should be identified as a characteristic for each virus and effectively classified.

The second problem on which the method was applied relates to the management of water reservoirs. The island of Cyprus was elected as the study area, while the inputs of the problem include structural and dynamic data, in which monthly precipitation particles play a distinct role. In this case, the time series information originated from the historical monthly rainfall data measured at certain watershed stations for a wide temporal period. The issue here was to develop a methodology for the production of evolutionary training/testing data, in order to achieve an effective estimation of the Average Annual Water Supply (AAWS) index on an annual basis, for each mountainous watershed of Cyprus.

ΠΡΟΛΟΓΟΣ

Στο πρώτο αυτό τμήμα της διατριβής θα ήθελα να παρουσιάσω συνοπτικά τις διάφορες συνιστώσες που στο σύνολό τους αποτέλεσαν το κίνητρο για την εκπόνησή της, καθώς επίσης και τους στόχους της. Συνοψίζοντας, επίσης, θα παραθέσω τη συμβολή της πραγματοποιηθείσας έρευνας στο σχετικό επιστημονικό πεδίο.

Στις ημέρες μας, οι συνθήκες υπό τις οποίες η έρευνα και η ανάπτυξη δραστηριοποιούνται προς την κατεύθυνση της βελτίωσης της ποιότητας ζωής του σύγχρονου ανθρώπου έχουν διαφοροποιηθεί σημαντικά. Η συνεχής και αλματώδης ανάπτυξη της σύγχρονης τεχνολογίας, τόσο σε Συστήματα Διαχείρισης Σχεσιακών Βάσεων Δεδομένων (ΣΔΣΒΔ) (RDBMS: Relational Database Management Systems), όσο και σε τεχνολογίες δικτύων υπολογιστών, επέτρεψαν την αύξηση του ρυθμού αποθησαυρισμού βαθμωτών (scalar) και ιστορικών (temporal) δεδομένων, που αντλούνται από διάφορους επιστημονικούς τομείς. Το γεγονός αυτό, σε συνδυασμό με την εξέλιξη των σύγχρονων υπολογιστικών συστημάτων, αναβάθμισε σημαντικά τις δυνατότητες ανάλυσης και επίλυσης προβλημάτων, που τις προηγούμενες δεκαετίες θα ήταν πρακτικά αδύνατον να πραγματοποιηθούν.

Σημαντικό υποσύνολο ιστορικών δεδομένων αποτελούν οι χρονοσειρές οι οποίες στην τυπική τους περίπτωση ορίζονται ως ακολουθίες τιμών κάποιων μεταβλητών, μετρημένων συνήθως σε διαδοχικές χρονικές περιόδους σταθερού βήματος. Πρόκειται για μια εξέχουσα κατηγορία δεδομένων με ευρεία εφαρμογή, τα οποία είναι δυνατόν πλέον εύκολα να εξαχθούν από ένα μεγάλο αριθμό επιστημονικών πηγών. Ο αγροπεριβαλλοντικός τομέας, η ιατρική, η κλινική φαρμακολογία, η βιολογία, η μετεωρολογία, η χρηματιστηριακή οικονομική, η αστρονομία, η φυσιολογία του ζωικού και του φυτικού κυττάρου είναι μερικές μόνο από ένα μακρύ κατάλογο επιστημών στις οποίες τα δεδομένα αυτού του τύπου διαδραματίζουν σημαντικό ρόλο. Μελέτες περιπτώσεων χρονοσειρών θα μπορούσαν για παράδειγμα να αποτελούν οι τιμές κλεισίματος συγκεκριμένων χρηματιστηριακών μετοχών για διάφορες χρονι-

κές περιόδους, η ημερήσια, εβδομαδιαία, μηνιαία ή ετήσια δυναμική ροή ποταμών και η έκπλυση των νιτρικών ενώσεων στη μονάδα του χρόνου μετά την αζωτούχο ανοιξιάτικη λίπανση συγκεκριμένων καλλιεργειών υπό συγκεκριμένες κλιματολογικές, ατμοσφαιρικές και εδαφολογικές συνθήκες ή ακόμη και η σταθερού βήματος απόκριση συγκεκριμένου αντιδραστηρίου σε μετρήσεις που καταγράφονται από αισθητήρες και σχετίζονται με προσβολές από διάφορα παθογόνα.

Κίνητρα και στόχοι της διατριβής

Τα σημαντικότερα προβλήματα που αφορούν σε χρονοσειρές σχετίζονται με την εξόρυξη δεδομένων και την αναγνώριση προτύπων στο διάνυσμα εισόδου. Στην ανάλυση αυτή χρησιμοποιούνται εργαλεία στατιστικής, καθώς επίσης και υπολογιστικής νοημοσύνης. Ενα σύγχρονο σύστημα αυτοματοποιημένης αναγνώρισης προτύπων αποτελείται στην τυπική του μορφή από τρία υπο-συστήματα: της παροχής δεδομένων, της επεξεργασίας και του αποτελέσματος. Το υπο-σύστημα παροχής δεδομένων είναι υπεύθυνο για τη συλλογή των αρχικών δεδομένων και αποτελείται από αισθητήρες καταγραφής ενός φαινομένου ή κάποιας διαδικασίας. Τα δεδομένα συγκεντρώνονται στις συσκευές καταγραφής και προωθούνται στο υποσύστημα επεξεργασίας, όπου κατάλληλα σχεδιασμένοι αλγόριθμοι εξαγωγής χαρακτηριστικών διαμορφώνουν σε πληροφορία τα πρωτογενή δεδομένα των αισθητήρων και τα προσαρμόζουν σε κατάλληλο μαθηματικό πρότυπο μέσω του οποίου επιτυγχάνεται η ταξινόμηση, ή πρόβλεψη, ανάλογα με τις ανάγκες του εκάστοτε προβλήματος. Χαρακτηριστικό γνώρισμα των δεδομένων αυτών είναι η συγκεκριμένη χρονική τους διάταξη, γεγονός που διαφοροποιεί την ανάλυσή τους από την ανάλυση των δεδομένων άλλου τύπου προβλημάτων, τα δεδομένα των οποίων δεν χαρακτηρίζονται από κάποιο φυσικό τύπο κατάταξης. Για παράδειγμα, το ύψος του μισθού ενός υπαλλήλου είναι ανεξάρτητο από τη σειρά με την οποία εμφανίζεται ο εκάστοτε υπάλληλος σε μια υποτιθέμενη καταγραφή των ανθρώπινων διαθεσίμων (HR: Human Resources) ενός οργανισμού ή μιας επιχείρησης. Επίσης, η ανάλυση των χρονοσειρών διαφέρει από την χωρική ανάλυση στην οποία οι παρατηρήσεις σχετίζονται με χωρικά χαρακτηριστικά.

Στις περισσότερες περιπτώσεις, σε ένα πρότυπο χρονοσειρών ισχύει ο νόμος της εγγύτητας, σύμφωνα με τον οποίο πλησιέστερες χρονικά παρατηρήσεις σχετίζονται

περισσότερο μεταξύ τους παρά εάν ήταν πιο απομακρυσμένες χρονικά. Επιπροσθέτως, τα πρότυπα αυτά κάνουν ευρεία χρήση της φυσικής χρονικής μονόδρομης διάταξης, ούτως ώστε παρατηρήσεις για μια δεδομένη περίοδο να είναι δυνατό να εκφρασθούν ως συνάρτηση συγκεκριμένων παρατηρήσεων του παρελθόντος. Εκτός της χαρακτηριστικής τους χρονικής και αριθμητικής διάστασης, τα δεδομένα του τύπου αυτού είναι ουσιαστικά συλλογές παρατηρήσεων με συνεχή πολλές φορές ροή ενημέρωσης. Συνεπώς στις περισσότερες περιπτώσεις, αφενός παρουσιάζουν μεγάλο όγκο τόσο σε εύρος, όσο και σε πληθυσμό και αφετέρου αντιμετωπίζονται αυτούσια ως ενιαίο σύνολο όσον αφορά στη συνέχειά τους. Έτσι οι χρονοσειρές χαρακτηρίζονται συνήθως από διάσταση ανώτερου βαθμού και υψηλά επίπεδα θορύβου. Στην ανάλυση περιλαμβάνονται ποικίλες διαδικασίες που αποσκοπούν στην υποβάθμιση των ανασταλτικών παραγόντων και κατ' επέκταση στην εξαγωγή χρήσιμων συμπερασμάτων για το στατιστικό υπόβαθρο και τα χαρακτηριστικά του μηχανισμού που παράγει τα πρωτογενή χρονικά δεδομένα. Σε αντίθεση με τις παραδοσιακές βάσεις δεδομένων όπου η αναζήτηση ομοιότητας, η ταυτοποίηση και η πρόβλεψη βασίζονται σε ακριβείς όρους, στις χρονοσειρές χρησιμοποιούνται ως επί το πλείστον προσεγγιστικά κριτήρια, που ακολουθούνται από μεθόδους ευρετηρίασης. Στο γενικότερο πλαίσιο της εξόρυξης δεδομένων, το θεμελιώδες πρόβλημα έγκειται στον τρόπο εξομάλυνσης της αρχικής σειράς με τρόπο ώστε να μην υφίσταται σημαντική αποδόμησή της. Όπως συμβαίνει στις περισσότερες περιπτώσεις προβλημάτων που σχετίζονται με την επιστήμη των υπολογιστών, αποτελεσματική λύση προς την κατεύθυνση αυτή προσφέρει η αναπαράσταση των πρωτογενών δεδομένων σε χώρους μικρότερης διάστασης.

Η παρούσα διατριβή εστιάζεται κυρίως στο δεύτερο στάδιο του προαναφερθέντος τυποποιημένου συστήματος αναγνώρισης προτύπων, αυτού της επεξεργασίας της αρχικής χρονοσειράς και της εξαγωγής ουσιωδών χαρακτηριστικών της για την περαιτέρω διευκόλυνση της διαδικασίας ταξινόμησης. Βάσει των προαναφερθέντων, ως κύριος στόχος της παρούσας διδακτορικής διατριβής τέθηκε ο σχεδιασμός, η ανάπτυξη, η υλοποίηση και η αποτίμηση ενός καινοτόμου αυτο-οργανούμενου εξελικτικού ταξινομητή για την επίλυση συγκεκριμένων προβλημάτων που περιλαμβάνουν χρονοσειρές. Επιμέρους στόχους της μελέτης αποτελούν:

α. **Ανάπτυξη μιας νέας μεθόδου αναπαράστασης χρονοσειρών.** Πρώτιστος

στόχος κατά τη φάση του σχεδιασμού της προτεινόμενης μεθόδου τέθηκε η αυξημένη δυνατότητα προσαρμογής του αλγορίθμου στα αρχικά δεδομένα. Η προτεινόμενη μέθοδος προέκυψε λαμβάνοντας υπόψη τα ισχυρά σημεία, αλλά και τις αδυναμίες των μέχρι τώρα ανεπτυγμένων μεθόδων αναπαράστασης, κοινός τόπος και χαρακτηριστικό των οποίων είναι ότι όλες με τον ένα ή τον άλλο τρόπο χρησιμοποιούν παράθυρα και βήματα σταθερού μήκους. Είτε πρόκειται για απλή δειγματοληψία είτε για τις διάφορες μορφές κινούμενου μέσου, η προσέγγιση αυτή είναι χωρίς αμφιβολία αποτελεσματική για διακριτά μη διανυσματικά δεδομένα, αλλά υστερεί έναντι των χρονοσειρών, όπως θα αναλυθεί στα επόμενα. Η Πρότυπη Εξελικτική Τμηματοποίηση (ΠΕΤ) (PES: Piecewise Evolutionary Segmentation) όπως σχεδιάσθηκε στο πλαίσιο της παρούσας μελέτης διευρύνει το φάσμα λειτουργίας της μέσω της ανάπτυξης μιας καινοτόμου εξελικτικής μεθόδου τμηματοποίησης. Με τον όρο αυτό νοείται ότι το εύρος του παραθύρου τμηματοποίησης καθορίζεται εξελικτικά από ειδικά αναπτυγμένο γενετικό αλγόριθμο. Υπό την έννοια αυτή, η προτεινόμενη μέθοδος εφαρμόζεται στην προ-επεξεργαστική φάση των αρχικών δεδομένων, στοχεύοντας στην αποτελεσματική εξαγωγή των χαρακτηριστικών εκείνων που προυσιάζουν τη μεγαλύτερη βαρύτητα στο υπό εξέταση φαινόμενο.

β. **Σχεδίαση του υποσυστήματος ταξινομητών.** Είναι γεγονός ότι καθένα από τα ποικίλα εργαλεία υπολογιστικής νοημοσύνης συγκλίνει γρηγορότερα και αποτελεσματικότερα σε συγκεκριμένα προβλήματα, ενώ αποδεικνύεται λιγότερο κατάλληλο για άλλα. Αναμφίβολα, η ενσωμάτωση περισσότερων του ενός παρόμοιων συναρτησιακών στοιχείων στο τελικό ολοκληρωμένο σύστημα επιφέρει σημαντική διεύρυνση του πεδίου εφαρμογής του, καθιστώντας το ικανό προς επίλυση πολύ μεγαλύτερου αριθμού προβλημάτων. Βεβαίως, στον αντίποδα ελλοχεύει ο κίνδυνος υπερβολικής αύξησης της πολυπλοκότητας αφενός, καθώς επίσης και η αύξηση των απαιτήσεων σε υπολογιστικό κόστος αφετέρου. Κατά τη διάρκεια του σχεδιασμού, εκτός των νευρωνικών δικτύων, ελέγχθηκαν ένα πλήθος ταξινομητών, συμπεριλαμβανομένων των μηχανών διανυσμάτων υποστήριξης, του ταξινομητή Bayes, του δένδρου αποφάσεων και εκείνου των εγγύτερων γειτόνων. Υπό αυτό το πρίσμα, το προτεινόμενο

σύστημα σχεδιάσθηκε έτσι ώστε να περιλαμβάνει τους δύο καλύτερα αποκρινόμενους ταξινομητές, δηλαδή ένα Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ) και μια Μηχανή Διανύσματων Υποστήριξης (ΜΔΥ) εξειδικευμένης δομής. Κατά την εφαρμογή του συστήματος, οι δύο ταξινομητές επιτίθενται στο εκάστοτε πρόβλημα και επιλέγεται ο καλύτερα αποκρινόμενος, δηλαδή αυτός με την μεγαλύτερη ακρίβεια ταξινόμησης ή την καλύτερη πρόβλεψη.

γ. **Σχεδίαση του υποσυστήματος εξελικτικής τμηματοποίησης.** Μια από τις βασικότερες λειτουργίες του προτεινόμενου συστήματος εντοπίζεται στη διαδικασία που είναι υπεύθυνη για την τελική αναπαράσταση της χρονοσειράς. Ο σχεδιασμός του συστήματος επέβαλε την ανάπτυξη μιας εξελικτικής μεθόδου τμηματοποίησης, κατά την οποία τα αρχικά δεδομένα υφίστανται την προ-επεξεργασία ενός ειδικά σχεδιασμένου εξελικτικού προτύπου παραγωγής ολοένα και πιο προσαρμοσμένων συνόλων δευτερογενών δεδομένων. Ο σχεδιασμός των εξελικτικών προτύπων τμηματοποίησης στο χώρο εισόδου υλοποιείται μέσω ειδικά δομημένου Γενετικού Αλγορίθμου (ΓΑ), κάθε γενεά του οποίου συμπληρώνεται από ένα προκαθορισμένο πληθυσμό *χρωμοσωμάτων (εκπαιδευτών)*. Σκοπός ύπαρξης κάθε τέτοιου εκπαιδευτή είναι να αποτυπώνει το χρωμόσωμά του επί της αρχικής χρονοσειράς, διαδικασία που υπαγορεύεται από συγκεκριμένα γονίδια ελέγχου της συμπεριφοράς του. Η αποτύπωση των εξελικτικών προτύπων του αλγορίθμου στην αρχική χρονοσειρά είναι ουσιαστικά μια αναζήτηση στο χώρο εισόδου για την καταλληλότερη δομή του εκπαιδευτή. Μέσω της ρουτίνας αυτής, δημιουργείται μια πλειάδα πιθανών αναπαραστάσεων των αρχικών δεδομένων, κάθε μια από τις οποίες ακολούθως χρησιμοποιείται διαδοχικά στην υπό εποπτεία εκπαίδευση των δύο ενσωματωμένων ταξινομητών. Η απόδοση που σημειώνεται κατά τη φάση ελέγχου κάθε αναπαράστασης ποσοτικοποιείται ως βαθμός καταλληλότητας, ο οποίος αποδίδεται στον αντίστοιχο εκπαιδευτή. Αλλεπάλληλες γενεές συνεχίζουν να σχηματοποιούνται έως ότου επιλεγεί άμεσα ο καταλληλότερος εκπαιδευτής - ήτοι εκείνος που είναι υπεύθυνος για την καλύτερη απόδοση από πλευράς ταξινομητή - και έμμεσα ο καταλληλότερος ταξινομητής.

δ. **Βελτιστοποίηση της αρχιτεκτονικής των ταξινομητών.** Σε όλη ανεξαιρέ-

τως τη σχετική βιβλιογραφία αναφέρεται ότι η αρχιτεκτονική κάθε ταξινομητή αποτελεί έναν από τους κρισιμότερους παράγοντες που επηρεάζει στο μέγιστο βαθμό την αποτελεσματικότητά του. Παρά το γεγονός ότι η επιστημονική κοινότητα έχει ήδη καταθέσει ένα μεγάλο αριθμό προτάσεων σχετικά με τη ρύθμιση των διαφόρων συνιστωσών που καθορίζουν τη δομή των ταξινομητών, ουσιαστικά δεν έχει προτυποποιηθεί μια συγκεκριμένη διαδικασία η οποία να έχει εφαρμογή σε κάθε περίπτωση. Η σχεδίαση του προτεινόμενου συστήματος περιλαμβάνει κανόνες οι οποίοι καθορίζουν την αρχιτεκτονική καθενός από τους εμπλεκόμενους ταξινομητές με βάση τη διάταξη των γονιδίων του αντίστοιχου εκπαιδευτή και του αριθμού των τμημάτων τα οποία αυτός υπαγορεύει, με τρόπο που θα αναλυθεί διεξοδικότερα στα επόμενα. Κατά τη φάση της εφαρμογής του συστήματος μορφοποιούνται βέλτιστοι ταξινομητές αυξημένης απόδοσης.

ε. **Εφαρμογή του συστήματος και μελέτη περιπτώσεων.** Το προτεινόμενο σύστημα σχεδιάσθηκε έτσι ώστε να έχει ευρεία εφαρμογή, τόσο σε προβλήματα ταξινόμησης, όσο και σε προβλήματα παλινδρόμησης/πρόβλεψης. Υπ' αυτή την έννοια, η προτεινόμενη μεθοδολογία εφαρμόσθηκε με ιδιαίτερα ενθαρρυντικά αποτελέσματα στην ανίχνευση φυτικών ιών, καθώς επίσης και στην πρόβλεψη της χειμαρρικής επικινδυνότητας ορεινών υδατικών αποθεμάτων, προβλήματα που επιλέχθηκαν υπό την έννοια ότι αφενός τα διαθέσιμα αρχικά δεδομένα τους παρουσιάζουν υψηλό βαθμό διάστασης και μεγάλο όγκο, ενώ αφετέρου εμπίπτουν στις δύο προαναφερόμενες κατηγορίες προβλημάτων αντίστοιχα.

Κατόπιν τούτων, η βασικότερη συνεισφορά της παρούσας διατριβής στο επιστημονικό πεδίο δραστηριοποίησής της έγκειται στο ότι προτείνει μια προσαρμοστική διαδικασία για την αποτελεσματική αναπαράσταση των χρονοσειρών. Το μεγαλύτερο πρόβλημα κατά τη χρήση αλγορίθμων σταθερού εύρους παραθύρου εντοπίζεται στο γεγονός ότι πιθανά σημαντικά πρότυπα διατρέχουν υψηλό κίνδυνο να υποστούν τεμαχισμό ή να απορριφθούν ολοσχερώς, καθώς ο αλγόριθμος διατρέχει τη χρονοσειρά με χαρακτηριστικά χαμηλό βαθμό προσαρμοστικότητας [89, 61]. Το προτεινόμενο σύστημα αποτελεί αποτελεσματική λύση στο πρόβλημα, καθώς χαρακτηρίζεται από υψηλό βαθμό προσαρμοστικότητας λόγω της εξελικτικής πορείας

του αλγορίθμου, μέσω της οποίας ο κίνδυνος κατακερματισμού ή/και απόρριψης σημαντικών υπαρχόντων προτύπων μειώνεται σημαντικά. Εξάλλου, στην παρούσα εργασία αναπτύσσεται ένα καινοτόμο εργαλείο το οποίο ενσωματώνει παράλληλους ταξινομητές υπολογιστικής νοημοσύνης, η εκπαίδευση των οποίων βελτιώνεται μέσω της εξελικτικής μορφοποίησης δευτερογενών δεδομένων εκπαίδευσης. Τέλος, το προτεινόμενο εργαλείο εφαρμόζεται στην ταυτοποίηση φυτικών ιών και στην πρόβλεψη της χειμαρρικής επικινδυνότητας, καλύπτοντας όλο το εύρος προβλημάτων στους τομείς ταξινόμησης και πρόβλεψης/παλινδρόμησης.

Διάταξη της διατριβής

Η διατριβή, εκτός από το παρόν εισαγωγικό της μέρος, αποτελείται επιπλέον από οκτώ κεφάλαια.

Το αμέσως επόμενο πρώτο κεφάλαιο πραγματεύεται την ανάλυση των χρονοσειρών και αναφέρεται στις διάφορες τεχνικές που έχουν αναπτυχθεί προς την κατεύθυνση της αναπαράστασής τους σε χώρους μικρότερης διάστασης, καθώς επίσης και της μείωσης του ενεχόμενου θορύβου. Στο πλαίσιο αυτό εξετάζονται λεπτομερώς οι διάφορες εκφάνσεις του αλγορίθμου διολισθαίνοντος παραθύρου, ενώ περιλαμβάνεται εκτενής ανάλυση της οικογένειας μεθόδων της τμηματικής γραμμικής αναπαράστασης. Οι δύο αυτές κατηγορίες αλγορίθμων τμηματοποίησης αποτελούν αφετηρία και έμπνευση της προτεινόμενης μεθόδου της Πρότυπης Εξελικτικής Τμηματοποίησης.

Το δεύτερο κεφάλαιο αναφέρεται σε εργαλεία που ανήκουν στον τομέα της υπολογιστικής νοημοσύνης και έχουν χρησιμοποιηθεί στην ανάλυση των χρονοσειρών. Συγκεκριμένα, στο κεφάλαιο αυτό αναλύονται τα Τεχνητά Νευρωνικά Δίκτυα, οι Μηχανές Διανυσμάτων Υποστήριξης και οι Γενετικοί Αλγόριθμοι, με ιδιαίτερη έμφαση ειδικά στην αρχή λειτουργίας τους και στο μαθηματικό πρότυπο που ακολουθούν. Παράλληλα, παρατίθεται σύντομη αναδρομή στη διεθνή βιβλιογραφία σχετικά με την ανάλυση των χρονοσειρών, τις μεθόδους αναπαράστασης και προεπεξεργασίας τους, ενώ επίσης επιχειρείται μια σύντομη ανασκόπηση του προβλήματος της αναπαράστασης χρονοσειρών με τη χρήση εργαλείων τεχνητής νοημοσύνης για την ικανοποίηση αυτών των αναγκών. Τέλος, γίνεται αναφορά στα κυριότερα προβλήματα που απαντώνται κατά τη χρήση ταξινομητών υπολογιστικής

νοημοσύνης και επιχειρείται μια αναδρομή στις ποικίλες τεχνικές που έχουν κατά καιρούς προταθεί για την επίλυσή τους.

Το περιβάλλον της έρευνας που διενεργείται με την παρούσα διατριβή παρουσιάζεται στο κεφάλαιο 3. Συγκεκριμένα, δίνεται το περίγραμμα της προτεινόμενης μεθόδου, καθώς επίσης και των συγκεκριμένων μελετών περίπτωσης χρονοσειρών στις οποίες εφαρμόσθηκε. Ταυτόχρονα, γίνεται εκτενής αναφορά σε παρόμοια προβλήματα και του τρόπου με τον οποίο εργαλεία υπολογιστικής νοημοσύνης υλοποιήθηκαν προς την κατεύθυνση επίλυσής τους. Τέλος, επιχειρείται η κριτική προσέγγιση των διαφόρων τεχνικών ανάλυσης των χρονοσειρών που έχουν παρουσιασθεί μέχρι σήμερα, υπογραμμίζοντας τα κυριότερα μειονεκτήματα που παρουσιάζουν κατά την εφαρμογή τους, ενώ παρατίθενται οι προδιαγραφές του προτεινόμενου συστήματος, καθώς επίσης και η συμβολή του στη βελτίωση της αναπαράστασης των χρονοσειρών.

Στο τέταρτο κεφάλαιο της διατριβής επιχειρείται διεξοδική ανάλυση της προτεινόμενης μεθόδου. Αρχικά τίθενται οι θεωρητικές βάσεις και στη συνέχεια αναλύεται το μαθηματικό πρότυπο που υποστηρίζει τον αλγόριθμο, ενώ γίνεται εκτενής αναφορά των αντικειμένων που αναπτύχθηκαν, της πληροφορίας που παράγουν και τη σκοπιμότητά της, αλλά και των μεταξύ τους σχέσεων.

Το πέμπτο κεφάλαιο αναφέρεται στον τρόπο υλοποίησης του αλγορίθμου, τις βασικές του συνιστώσες και την ανάλυση της ανταλλαγής της πληροφορίας μεταξύ τους. Επίσης στο κεφάλαιο αυτό γίνεται παρουσίαση του γραφικού περιβάλλοντος διεπαφής του εργαλείου λογισμικού όπως αναπτύχθηκε για να υποστηρίξει τη μέθοδο της Πρότυπης Εξελικτικής Τμηματοποίησης.

Οι δύο μελέτες περίπτωσης στις οποίες εφαρμόσθηκε το προτεινόμενο σύστημα αναλύονται στο έκτο και το έβδομο κεφάλαιο. Για κάθε μια από αυτές αναφέρονται διεξοδικά τα υλικά και οι μέθοδοι που χρησιμοποιήθηκαν, ενώ γίνεται εκτενής ανάλυση του τρόπου συγκέντρωσης και ταξινόμησης των πρωτογενών δεδομένων, αλλά και της παραμετροποίησης του συστήματος σε κάθε περίπτωση. Τέλος, δίνονται αναλυτικά τα αποτελέσματα της εφαρμογής του εργαλείου λογισμικού και ο σχολιασμός τους σε κάθε περίπτωση.

Το πόνημα ολοκληρώνεται στο όγδοο κεφάλαιο με τη συνολική παρουσίαση των συμπερασμάτων που εξήχθησαν στο πλαίσιο της διατριβής, σε συνδυασμό με τη συ-

νεισφορά της στο γενικότερο γνωστικό αντικειμένο διερεύνησης. Τέλος, προτείνονται συγκεκριμένες κατευθύνσεις επέκτασης των διεργασιών που αναπτύχθηκαν και θα μπορούσαν να αποτελέσουν αντικείμενο μελλοντικής έρευνας.

Κεφάλαιο 1

ΑΝΑΠΑΡΑΣΤΑΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

Στο κεφάλαιο αυτό παρατίθενται οι διάφορες τεχνικές δειγματοληψίας και αναπαράστασης των χρονοσειρών, ενώ αναλύονται διεξοδικά τα μειονεκτήματα της αναπαράστασης μέσω αλγορίθμων διολισθαίνοντος παραθύρου σταθερού εύρους και γίνεται αναφορά στις μεθόδους μεταβλητού εύρους που αποτελούν μετεξέλιξη των προηγούμενων για την άμβλυνση των ανασταλτικών τους παραγόντων. Ιδιαίτερη έμφαση δίνεται στην οικογένεια των μεθόδων τμηματικής αναπαράστασης, απόρροια της οποίας αποτελεί η προτεινόμενη Πρότυπη Εξελικτική Τμηματοποίηση των χρονοσειρών.

Κατά τη διάρκεια των τελευταίων ετών σημειώνεται κατακόρυφη αύξηση του ενδιαφέροντος σχετικά με προβλήματα που περιλαμβάνουν χρονικές σειρές, οι οποίες χαρακτηρίζονται από υψηλή πολυπλοκότητα λόγω υψηλού βαθμού διάστασης και θορύβου. Σχετική ερευνητική δραστηριότητα έχει επεκταθεί σε ποικίλους τομείς, με χαρακτηριστικότερα παραδείγματα την επεξεργασία είτε ολοκληρωμένων σειρών, είτε τμημάτων τους, με στόχο την ανάπτυξη και τη βελτίωση μεθόδων αναγνώρισης προτύπων για την επίλυση προβλημάτων κατηγοριοποίησης ή πρόβλεψης [61]. Στην πρώτη περίπτωση εμπίπτουν προβλήματα ταυτοποίησης, ανακάλυψης μηχανισμών και κανόνων ή δημιουργίας ομάδων, όπου η χρονοσειρά αποτελεί χαρακτηριστικό γνώρισμα μιας - ή περισσότερων - τάξης αντικειμένων. Στη δεύτερη

περίπτωση η χρονοσειρά είναι συνεχής και η ενημέρωσή της γίνεται συχνά σε πραγματικό χρόνο, το δε ζητούμενο είναι η κατάσταση του συστήματος σε συγκεκριμένο μελλοντικό χρόνο, δεδομένης της ιστορικής του πορείας.

1.1 Ορισμός και παραδείγματα

Με τον όρο χρονοσειρά εννοούμε συνήθως μια ακολουθία, $\{x_t : t = 0, 1, 2, ...\}$ όπου κάθε x_t εκφράζει την κατάσταση ενός συστήματος κατά τη χρονική στιγμή t. Η ακολουθία αυτή εξελίσσεται στον άξονα του χρόνου κατά τυχαίο εν γένει τρόπο, επηρεαζόμενη από διάφορες επικρατούσες εξωτερικές συνθήκες, πρόκειται δε υπό την έννοια αυτή για ένα στοχαστικό σύστημα. Παραδείγματα χρονοσειρών θα μπορούσαν να είναι:

α. Το μήκος των στημόνων x_t των αρσενικών ανθέων συγκεκριμένου φυτού κατά τις περιόδους ανθοφορίας του, με $t = 1, 2, ...$

β. Η υπολειμματικότητα x_t συγκεκριμένων φυτοφαρμάκων σε βάθος χρόνου, με $t \in [0, T]$.

γ. Η ημερήσια παροχή νερού x_t από συγκεκριμένες διατομές ποταμού, με $t = 1, 2, ...$

δ. Η περιεκτικότητα σε σάκχαρα x_t κάποιου καρπού κατά την περίοδο αποθήκευσής του, με $t \in [0, T]$.

ε. Η στη μονάδα του χρόνου - και σε πεπερασμένο χρονικό διάστημα - απόκριση βιο-αισθητήρων για την ταυτοποίηση παθογόνων.

Κατά συνέπεια, οι χρονοσειρές θα μπορούσαν να αφορούν σε διακριτά[1] μεγέθη x_t σε διακριτό χρόνο t (περίπτωση α), διακριτά μεγέθη x_t σε συνεχή χρόνο t (περίπτωση β), συνεχή μεγέθη x_t σε διακριτό χρόνο t (περίπτωση γ) και συνεχή μεγέθη x_t σε συνεχή χρόνο t (περίπτωση δ). Επίσης, θα μπορούσαν να αφορούν σε ένα υποσύνολο μετρήσεων καθεμιά από τις οποίες αντιστοιχεί σε συγκεκριμένη κλάση προερχόμενη από ένα σύνολο δυνατών κλάσεων (περίπτωση ε). Στις περισσότερες περιπτώσεις, τα προβλήματα που σχετίζονται με χρονοσειρές συνίστανται αφενός στην

[1] Οι όροι «διακριτά/συνεχή μεγέθη» και «διακριτός/συνεχής χρόνος» χρησιμοποιούνται εδώ κατ' αντιστοιχία με τους όρους «διακριτές» και «συνεχείς» τυχαίες μεταβλητές

πρόβλεψη μελλοντικών τιμών της ίδιας χρονοσειράς με βάση την ιστορική της πορεία (περίπτωση γ) ή στην πρόβλεψη φαινομένων που σχετίζονται άμεσα με αυτήν (περιπτώσεις α, β και δ). Επίσης υπάρχουν προβλήματα που χρησιμοποιούν δεδομένα χρονοσειρών για να ταυτοποιήσουν ένα φαινόμενο, η κατάσταση του οποίου εξαρτάται από τη χρονοσειρά (περίπτωση ε), όπως είναι για παράδειγμα η περίπτωση ταυτοποίησης παθογόνων ιών που εξαρτάται από δεδομένα βιο-αισθητήρων ή η πρόβλεψη χειμαρρικών φαινομένων που εξαρτάται από μετεωρολογικά δεδομένα χρονοσειρών. Το σύνολο των δυνατών καταστάσεων αποτελεί το χώρο καταστάσεων και συμβολίζεται με S που είναι ένα μονοδιάστατο υποσύνολο του R ή, στη γενική του μορφή, ένα πολυδιάστατο υποσύνολο του R^d. Κατ' αντιστοιχία, το σύνολο τιμών του χρόνου ονομάζεται παραμετρικός χώρος και συμβολίζεται με T, υποσύνολο του R ή του R^k με $k \in N$ στις περιπτώσεις που είναι απαραίτητο το t να παρασταθεί μέσω επιπλέον μεταβλητών πέραν του χρόνου.

Σε όλες σχεδόν τις περιπτώσεις τα δεδομένα που λαμβάνονται για να μορφοποιήσουν κάποια χρονοσειρά, ιδιαίτερα στις βιολογικές επιστήμες, προέρχονται από όργανα μετρήσεων ή αισθητήρες, η όλη δε διαδικασία είναι επιρρεπής στο σφάλμα μέτρησης. Επίσης, τα περισσότερα σύνολα δεδομένων χρονοσειρών χαρακτηρίζονται από υψηλότατα επίπεδα βαθμού διάστασης, ενώ συχνά εμφανίζουν αυξητική (ή φθίνουσα) τάση της μέσης τιμής, ή περιοδικές εναλλαγές μεταξύ αυξητικών και φθινουσών τάσεων, παρουσιάζοντας μια κυκλικά επαναλαμβανόμενη δομή που συνιστά την περιοδικότητα. Τα χαρακτηριστικά αυτά συντιθέμενα προσθετικά δίδουν το γενικευμένο πρότυπο μιας χρονοσειράς:

$$X_t = m_t + s_t + \varepsilon_t,\ t \in R \tag{1.1.1}$$

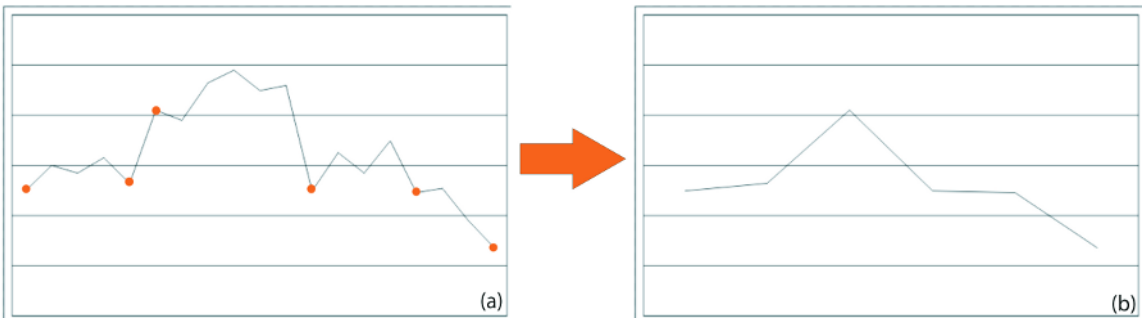
όπου οι συνιστώσες m_t, s_t και ε_t αντιστοιχούν στην τάση, την περιοδικότητα και τον θόρυβο των αρχικών δεδομένων. Τα χαρακτηριστικά των παραγόντων της εξίσωσης 1.1.1 καθιστούν την ανάλυση των δεδομένων μια επίπονη αν όχι δύσκολη έως αδύνατη διαδικασία, επιβάλλοντας την αναπαράσταση των χρονοσειρών σε χώρους μικρότερης διάστασης με παράλληλη εξομάλυνση των αρχικών δεδομένων.

1.2 Δειγματοληπτικοί αλγόριθμοι

Εφόσον δίδεται η τροχιά μιας χρονοσειράς $\{X_t : t \in T\}$ από την αρχική $t = 0$ έως κάποια συγκεκριμένη χρονική στιγμή $t = s$, τότε όλες οι τιμές της μεταξύ των δύο χρονικών αυτών διαστημάτων είναι γνωστές. Είναι λοιπόν φανερό ότι για την πρόβλεψη μιας μελλοντικής τιμής X_{s+h}, $h > 0$, θα πρέπει να ληφθούν υπόψη τα γνωστά αυτά ιστορικά στοιχεία. Για να είναι σε θέση δε η ανάλυση να τα χρησιμοποιήσει για να εξάγει μια αποτελεσματική πρόβλεψη, είναι απαραίτητο όλα τα πιθανοθεωρητικά χαρακτηριστικά της χρονοσειράς να παραμένουν αναλλοίωτα στο χρόνο. Τότε η χρονοσειρά ονομάζεται στάσιμη. Εφόσον ικανοποιούνται αυτές οι προϋποθέσεις, οι συνιστώσες της εξίσωσης 1.1.1 αναλύονται για να επιτευχθεί η προβολή σε μελλοντικό χρόνο.

Στις περιπτώσεις όμως κατά τις οποίες απαιτείται σημαντικό ποσοστό συμπίεσης και ικανοποιητική αποκατάσταση της χρονοσειράς σε δεύτερο χρόνο, ως αποφασιστικής σημασίας παράγοντες αναδεικνύονται κυρίως η επιλογή των κρίσιμων χαρακτηριστικών της χρονοσειράς και η μείωση του βαθμού του ενδογενούς θορύβου των πρωτογενών δεδομένων, η αποτελεσματικότητα των οποίων εξαρτάται άμεσα από τη δυνατότητα αναπαράστασής της σε χώρους μειωμένης διάστασης. Η αναγκαιότητα αυτή αποκτά ιδιαίτερη σημασία στις περιπτώσεις που η χρονοσειρά αποτελεί δομικό στοιχείο ενός συγκεκριμένου φαινομένου, όταν δηλαδή συγκαταλέγεται ανάμεσα στους παράγοντες που επηρεάζουν την έκβασή του ή αποτελεί χαρακτηριστικό μέσω του οποίου είναι δυνατόν το φαινόμενο αυτό να ταυτοποιηθεί ή να προβλεφθεί.

Ποικίλες αναπαραστάσεις ανώτερου επιπέδου για δεδομένα χρονοσειρών έχουν προταθεί, συμπεριλαμβανομένων των μετασχηματισμών Fourier [2, 87] και κυματιδίου [31], ή της μεθόδου συμβολικών απεικονίσεων [3, 48, 148]. Όσον αφορά στις διάφορες τιμές της χρονοσειράς, οι ποικίλες μέθοδοι ανάλυσης είναι δυνατόν να ταξινομηθούν σε δύο μεγάλες κατηγορίες: στις αναλυτικές μεθόδους συχνότητας και σε αυτές του χρόνου. Στην πρώτη κατηγορία ανήκουν η ανάλυση φάσματος και η ανάλυση κυματιδίου, ενώ στη δεύτερη περιλαμβάνονται διάφορες μέθοδοι αυτοσυσχέτισης και παλινδρόμησης [14, 193, 141]. Πιθανώς η απλούστερη αναπαράσταση προκύπτει μετά από εφαρμογή δειγματοληψίας [194, 10] επί της χρονοσειράς. Σύμφωνα με τη μέθοδο αυτή, για την αναπαράσταση χρησιμοποιείται ο λόγος *n/k*, όπου



Σχήμα 1.1: *Δειγματοληψία επί της αρχικής χρονοσειράς (a) έχει ως αποτέλεσμα υπερβολική παραμόρφωση (b).*

n είναι το μήκος χρονοσειράς $\{X_t : t \in T\}$, ενώ k είναι η διάσταση που προκύπτει μετά την υποβάθμισή της. Η εν λόγω μέθοδος παρουσιάζει το μειονέκτημα της υπερβολικής παραμόρφωσης των αρχικών δεδομένων, στις περιπτώσεις όπου ο λόγος n/k πέφτει κάτω από μια ορισμένη τιμή που εξαρτάται από τη φύση των δεδομένων (Εικ. 1.1).

Επιπροσθέτως, σημαντικά βελτιωμένες αναπαραστάσεις προκύπτουν μέσω διατήρησης των σημαντικότερων από τα αρχικά στοιχεία της χρονοσειράς. Ο κυριότερος αλγόριθμος αυτής της κατηγορίας χρησιμοποιήθηκε αρχικά για αντιστοίχιση προτύπων σε εφαρμογές οικονομικού περιεχομένου και αποτελεί βελτίωση προγενέστερης τεχνικής, μέσω της οποίας τα επιλεγμένα στοιχεία χαρακτηρίζονται ως Αντιληπτικά Σημαντικά Σημεία (ΑΣΣ) (PIP: Perceptually Important Points) [42, 62]. Η μέθοδος κατατάσσει τα στοιχεία της χρονοσειράς με φθίνοντα βαθμό σημαντικότητας και στη συνέχεια επιλέγει συγκεκριμένο αριθμό σημαντικών στοιχείων. Διάφορα μέτρα ποσοτικοποίησης της σημαντικότητας έχουν προταθεί και σχετίζονται με:

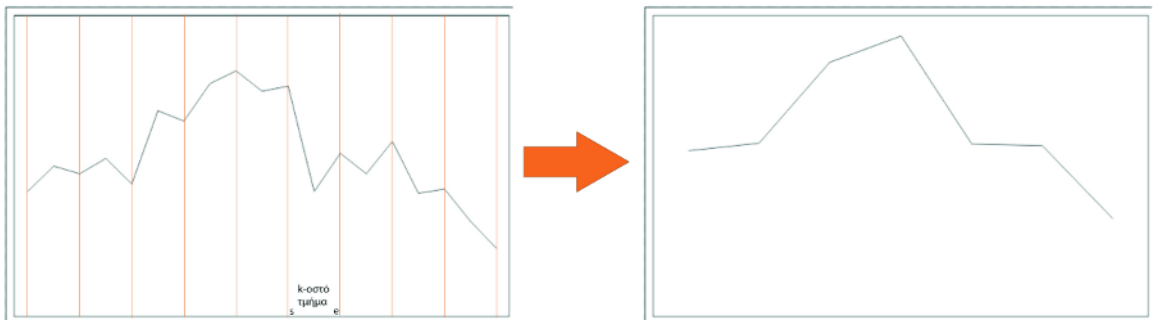
- *την αξιολόγηση της θέσης κάθε στοιχείου στη χρονοσειρά.* Σύμφωνα με τον αλγόριθμο αυτό [42] η θέση επηρεάζει άμεσα τη σημαντικότητα σε βαθμό τέτοιο, ώστε να καθιστά κάποια στοιχεία άμεσα επιλέξιμα στη χρονοσειρά αναπαράστασης. Έτσι, το πρώτο και το τελευταίο στοιχείο αποτελούν τα πλέον σημαντικά και κατατάσσονται πρώτα στην ταξινομημένη κατά βαθμό σημαντικότητας νέα χρονοσειρά. Ως τρίτο σημαντικότερο ορίζεται το στοιχείο με τη μεγαλύτερη ευκλείδεια απόσταση από τα δύο πρώτα, ενώ ως τέταρτο εκείνο

με τη μεγαλύτερη απόσταση από την ευθεία που ενώνει τα δυο παρακείμενα προς αυτό ΑΣΣ. Η διαδικασία συνεχίζεται έως ότου όλα τα στοιχεία της αρχικής χρονοσειράς έχουν αξιολογηθεί ή ένας προκαθορισμένος αριθμός ΑΣΣ ξεπεραστεί.

- *την εύρεση σημαντικών κορυφών και κοιλοτήτων.* Τα στοιχεία αυτά θεωρούνται ως σημεία ελέγχου μέσω των οποίων είναι δυνατή η πλήρης αναπαράσταση της αρχικής πληροφορίας. Οι αλγόριθμοι που χρησιμοποιούνται κάνουν χρήση προτύπων ορόσημου (landmark models) σε συνδυασμό με μέτρα ομοιότητας (similarity measures) [112], ή δικτυωτού (lattice models) [119].
- *τον καθορισμό σημαντικών ακρότατων ελάχιστων ή μέγιστων.* Μέσω της εν λόγω τεχνικής απορρίπτονται λιγότερο σημαντικές διακυμάνσεις και στη συνέχεια η αρχική πληροφορία συμπιέζεται, ώστε να διατηρεί συγκεκριμένα σημαντικά ακρότατα [158]. Ο αλγόριθμος χρησιμοποιεί μια παράμετρο $R\ (>1)$ που ορίζεται ως ο βαθμός συμπίεσης, ο οποίος καθορίζει το ποσοστό των στοιχείων της αρχικής πληροφορίας που συμπεριλαμβάνονται στην αναπαράστασή της. Δεδομένων των δεικτών *i* και *j*, ένα σημείο x_t, όπου $i \leq t \leq j$, μιας χρονοσειράς X θα θεωρείται ως σημαντικό ελάχιστο (ή μέγιστο) εάν και μόνο εάν το x_t είναι το ελάχιστο (ή το μέγιστο) μεταξύ των σημείων $x_i \ldots x_j$, καθώς $x_i/x_t \geq R$ και $x_j/x_t \geq R$. Κατά συνέπεια, καθώς το R αυξάνει, λιγότερα στοιχεία πρωτογενούς πληροφορίας λαμβάνουν μέρος στη μορφοποίηση της αναπαράστασης. Ο αλγόριθμος ταυτοποιεί τα σημαντικότερα σημεία της ακολουθίας με βάση την τοπική πληροφορία κάθε δημιουργούμενου τμήματός της και τελικά επιστρέφει την τιμή και τον πίνακα ευρετηρίασής τους, έχοντας συμπεριλάβει σε αυτά τα σημεία αρχής και τέλους.

1.3 Αλγόριθμοι τμηματικής αναπαράστασης

Σε γενικές γραμμές, οι διάφορες μέθοδοι αναπαράστασης των χρονοσειρών που έχουν κατά καιρούς προταθεί θα μπορούσαν να ταξινομηθούν σε δυο μεγάλες κατηγορίες, ανάλογα με την κατάτμηση ή μη των πρωτογενών δεδομένων. Η κατάτμηση ουσιαστικά προκύπτει από την ανάγκη βελτίωσης των απλούστερων αλγορίθμων αναπαράστασης, ένεκα συγκεκριμένων προβλημάτων που ανακύπτουν κατά τη χρήση τους. Για παράδειγμα, στην προαναφερθείσα περίπτωση της δειγματοληψίας



Σχήμα 1.2: *Εξομάλυνση χρονοσειράς μέσω κατάτμησης και εξαγωγής του μέσου κάθε τμήματος.*

του σχήματος 1.1, σημαντική βελτίωση στην εξομάλυνση των πρωτογενών δεδομένων επέρχεται εάν η σειρά κατατμηθεί και στη συνέχεια εξαχθεί ο μέσος όρος κάθε τμήματος για την αναπαράσταση [8].

Είναι πολύ συχνή η εμφάνιση μεθόδων τμηματοποίησης στην αναπαράσταση των χρονοσειρών. Στο πλαίσιο της εξόρυξης δεδομένων, οι διάφορες αυτές τεχνικές αποσκοπούν στην υποστήριξη:

1. εκτέλεσης ταχέων αναζητήσεων ομοιότητας (fast similarity search) σε ογκώδεις βάσεις δεδομένων χρονοσειρών

2. καθορισμού προτύπων μέτρων απόστασης (novel distance measures), συμπεριλαμβανομένων ερωτημάτων ασάφειας (fuzzy queries) [166], ερωτήσεων πολλαπλών επιλύσεων (multi-resolution queries) [180, 113], δυναμικής χρονικής παραμόρφωσης (DTW: Dynamic Time Warping) [143] και ερωτημάτων με τη χρήση βαρών (weighted queries) [90]

3. της αποτελεσματικότητας προτύπων αλγορίθμων ταυτοποίησης (classification) μέσω προ-επεξεργασίας των δεδομένων και δημιουργίας συστάδων [90]

4. προτυποποίησης σημαντικών σημείων και ανάπτυξης μεθόδων εντοπισμού χαρακτηριστικών σημείων διαφοροποίησης [11]

Σε γενικές γραμμές, το πρόβλημα σχεδιασμού ενός σχήματος τμηματοποίησης για δεδομένη χρονοσειρά είναι δυνατόν να ορισθεί έτσι ώστε στην τελική ζητούμενη αναπαράσταση:

α. να χρησιμοποιείται συγκεκριμένος προκαθορισμένος αριθμός τμημάτων,

β. το μέγιστο σφάλμα για κάθε τμήμα να μην υπερβαίνει μια προκαθορισμένη τιμή κατωφλίου,

γ. το συνδυασμένο σφάλμα όλων των τμημάτων να είναι μικρότερο μιας προκαθορισμένης τιμής κατωφλίου.

Χαρακτηριστικό παράδειγμα αποτελεί η μέθοδος γνωστή ως Τμηματικά Αθροιστική Προσέγγιση (ΤΑΠ) (PAA: Piecewise Aggregate Approximation) [88] που αναφέρεται στη βιβλιογραφία και ως Τμηματικά Σταθερή Προσέγγιση (ΤΣΠ) (Piecewise Constant Approximation) [181]. Σύμφωνα με τη μέθοδο αυτή, ο βαθμός διάστασης μιας χρονοσειράς $\{X_t : t \in T\}$ δεδομένου μήκους n μειώνεται μέσω αντιστοίχισής της σε νέα ακολουθία $\{\bar{X}_t : t \in T\}$, μήκους w, όπου τυπικά $w \ll n$. Το k-οστό στοιχείο της νέας ακολουθίας $\bar{X}$ δίδεται από τη σχέση:

$$\bar{x_k} = \frac{w}{n} \sum_{i=\frac{n}{w}(k-1)+1}^{\frac{n}{w}k} x_i \tag{1.3.1}$$

Η σχέση 1.3.1 δίνεται και και ως:

$$\bar{x_k} = \frac{1}{e_k - s_k + 1} \sum_{i=s_k}^{e_k} x_i \tag{1.3.2}$$

όπου s_k, e_k είναι αντίστοιχα οι δείκτες του αρχικού και τελικού στοιχείου τα οποία ανήκουν στο k-οστό τμήμα της αρχικής χρονοσειράς (Εικ. 1.2).

Σύμφωνα με τη μέθοδο ΤΑΠ, όπως περιγράφεται από τις σχέσεις 1.3.1 και 1.3.2, για να επιτευχθεί η μείωση του βαθμού διάστασής της από n σε w, η αρχική χρονοσειρά διαμερίζεται σε w ίσα τμήματα, για καθένα από τα οποία εξάγεται ο μέσος όρος των τιμών της. Η νέα υπο-ακολουθία που προκύπτει κατ' αυτό τον τρόπο αποτελεί μια αναπαράσταση των αρχικών δεδομένων μειωμένου βαθμού διάστασης. Επέκταση αυτής της μεθόδου αποτελεί η Προσαρμοζόμενη Τμηματικά Σταθερή Προσέγγιση (ΠΤΣΠ) (APCA: Adaptive Piecewise Constant Approximation), κατά την οποία τα τμήματα δεν είναι σταθερού μήκους, αλλά προσαρμοζόμενα κάθε φορά στο σχήμα της χρονοσειράς, ενώ παρόμοια είναι η μέθοδος αναπαράστασης που βα-

σίζεται στην υπογραφή του κάθε τμήματος. Πέραν της χρήσης του μέσου όρου για την αναπαράσταση των τμημάτων, ποικίλες άλλες μέθοδοι έχουν προταθεί, συμπεριλαμβανομένου του αλγορίθμου υπολογισμού του αθροίσματος της παραλλακτικότητας ανά τμήμα, καθώς επίσης και της αντιστοίχισης κάθε στοιχείου του τμήματος σε συγκεκριμένο δυφίο[2] [61].

Σημαντική επίσης τεχνική που μετέρχεται μεθόδους τμηματοποίησης της αρχικής χρονοσειράς συνιστά ο αλγόριθμος μετατροπής του αριθμητικού σε συμβολικό τύπο, με αποτέλεσμα να προκύπτουν συμβολοσειρές [67, 128, 123]. Η μέθοδος αυτή περιλαμβάνει αρχικά την κατάτμηση της χρονοσειράς σε τμήματα, σε καθένα από τα οποία στη συνέχεια αντιστοιχίζεται ένα σύμβολο. Στην κατηγορία αυτή εμπίπτει και η μέθοδος της Συμβολικής Αθροιστικής Προσέγγισης (ΣΑΠ) (SAX: Symbolic Aggregate Approximation), σύμφωνα με την οποία χρησιμοποιείται μια ενδιάμεση αναπαράσταση μεταξύ της αρχικής χρονοσειράς και της τελικής συμβολοσειράς [116, 120, 70]. Ο αλγόριθμος ΣΑΠ ξεκινά μετασχηματίζοντας τα αρχικά δεδομένα μέσω της μεθόδου ΤΑΠ και στη συνέχεια αντιστοιχίζει μια διακεκριμένη σειρά συμβόλων σε κάθε προκύπτουσα τμηματική αναπαράσταση ώστε να δημιουργήσει συμβολοσειρά μειωμένου βαθμού διάστασης. Αφού ολοκληρωθεί η ενδιάμεση αναπαράσταση, ο αλγόριθμος ΣΑΠ την τεμαχίζει σε α ισοδύναμες περιοχές, καθορίζοντας ένα διατεταγμένο σύνολο σημείων τεμαχισμού $B = \beta_1, \ldots, \beta_{\alpha-1}$ έτσι ώστε η περιοχή υπό καμπύλη Gauss $N(0,1)$ από το σημείο β_i έως το $\beta_{i+1} = 1/\alpha$ (τα σημεία β_0 και β_a θεωρούνται ως $-\infty$ και $+\infty$ αντίστοιχα).

Αφού καθορισθούν οι περιοχές κατάτμησης αρχίζει η διακριτοποίηση των τεμαχίων που προκύπτουν. Ολοι οι συντελεστές της ΤΑΠ που τοποθετούνται κάτω από το μικρότερο σημείο κατάτμησης αποτυπώνονται στο σύμβολο «a». Οι συντελεστές που είναι μεγαλύτεροι ή ίσοι με το δεύτερο τη τάξει σημείο κατάτμησης αποτυπώνονται στο σύμβολο «b» κ.ο.κ. Η αλληλοσύνδεση συμβόλων που αντιπροσωπεύουν μια υπο-ακολουθία αποτελεί μια «λέξη». Με άλλα λόγια, μια υπο-ακολουθία X μήκους n αποτελεί μια «λέξη» $\hat{X}=\hat{x_1}, \ldots, \hat{x_w}$ ως ακολούθως: Έστω α_i το i-οστό στοιχείο του αλφάβητου, π.χ. $alpha_1 = a$ και $alpha_2 = b$. Τότε η αποτύπωση από μια αναπαράσταση $\bar{X}$ της ΤΑΠ σε μια λέξη $\hat{X}$ προκύπτει μέσω της συνθήκης:

[2]Συντομευμένη μορφή του σύμπλοκου όρου «δυαδικό ψηφίο» που παράγεται με σύμμειξη των δύο συνθετικών του: δυ(αδικό) + (ψη)φίο = δυφίο ως απόδοση του αγγλικού συμμείγματος bi(nary) + (digi)t = bit

$$\hat{x_i} = \alpha_j, \qquad \text{εάν και μόνο εάν} \qquad \beta_{j-1} \leq \bar{x}_i \leq \beta_j \tag{1.3.3}$$

Σύμφωνα με τον αλγόριθμο ΣΑΠ, αρχικά ο χώρος κατανομής χωρίζεται σε ισοδύναμες περιοχές, σε καθεμιά από τις οποίες αντιστοιχίζεται ένα σύμβολο. Η παραγόμενη μετασχηματισμένη χρονοσειρά τελικά αναπαρίσταται από μια νέα συμβολοσειρά, αφού καθορισθούν δυο κρίσιμες παράμετροι: το μήκος της υπο-ακολουθίας και το μέγεθος της αλφαβήτου που χρησιμοποιείται ως δειγματικός χώρος συμβόλων. Πέραν της χρήσης του μέσου των τμημάτων για τη δημιουργία των συμβόλων, άλλες μέθοδοι χρησιμοποιούν το βαθμό μεταβλητότητάς τους για τη δημιουργία ενός κατηγορικού αλφάβητου [62], με κατηγορίες οι οποίες προκύπτουν από:

α. την τάση της μεταβολής σε ισχυρά/ασθενώς αυξητική, σταθερή ή ισχυρά/ασθενώς ελαττούμενη,

β. την κλίση της μεταβολής σε ανοδική, καθοδική και επίπεδη,

γ. τη μεταβολή του λόγου συνεχόμενων αριθμητικών τιμών.

Άλλες μέθοδοι μετατροπής μιας χρονοσειράς σε συμβολοσειρά περιλαμβάνουν:

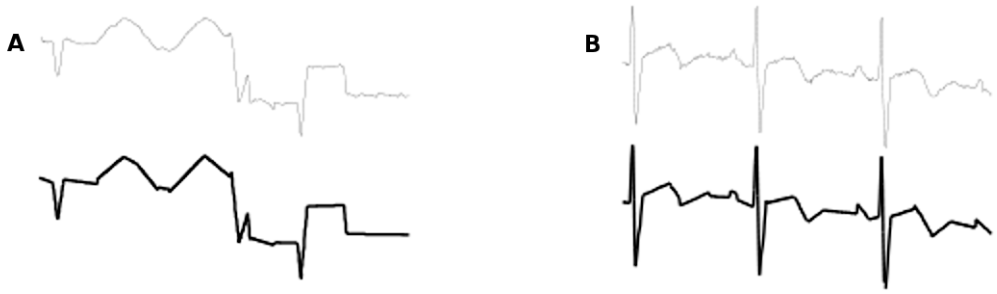
- την ταυτοποίηση κάθε δημιουργούμενου τμήματος με μια λέξη-κλειδί η οποία προέρχεται από συγκεκριμένο λεξικό όρων (codebook) [123],
- τις διαδικασίες διακριτοποίησης χωρίς επίβλεψη [132], οι οποίες βασίζονται στην αξιολόγηση της ποιότητας κάθε τμήματος και τη διατήρηση της χρονικής συνιστώσας για κάθε στοιχείο της πρωτογενούς πληροφορίας,
- τις διαδικασίες δημιουργίας ομάδων και την αντιστοίχισή τους σε σύμβολα. Σύμφωνα με τη μέθοδο της εξόρυξης πολλαπλών επιπέδων αφαίρεσης, τα σύμβολα που χρησιμοποιούνται καθορίζονται από τις συστάδες που δημιουργούνται από απλά περιγραφικά και ανώτερου βαθμού στατιστικά χαρακτηριστικά κάθε τμήματος, όπως συντελεστές παλινδρόμησης, μέσο τυπικό τετραγωνικό σφάλμα και ιστογράμματα των υπολοίπων της παλινδρόμησης.

1.4 Τμηματική Γραμμική Αναπαράσταση

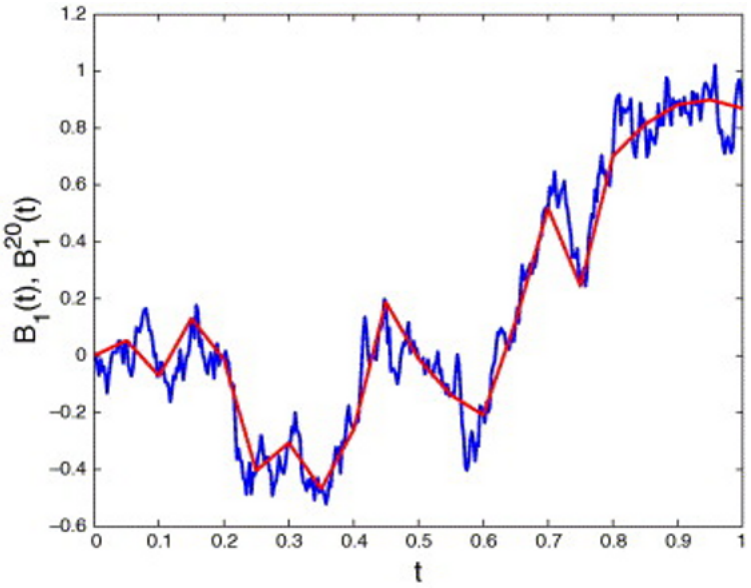
Αποτελεσματική εξομάλυνση πρωτογενών δεδομένων χρονοσειρών με παράλληλη διατήρηση της δομής τους επιτυγχάνεται μέσω της κατά προσέγγιση αναπαράστασής τους με ευθύγραμμα τμήματα. Η τεχνική αυτή, στην οποία ουσιαστικά περιλαμβάνεται μια μεγάλη ομάδα αλγορίθμων υπό τη γενική ονομασία Τμηματική Γραμμική Αναπαράσταση (ΤΓΑ) (PLR: Piecewise Linear Representation), αποτελεί ίσως την πιο συχνά χρησιμοποιούμενη ομάδα μεθόδων τμηματοποίησης [69, 96, 34, 90, 115, 111, 113, 92, 139, 143, 198, 166, 167, 180, 92, 170]. Σύμφωνα με τον Keogh [89], κάθε αλγόριθμος ο οποίος δέχεται στην είσοδο μια χρονοσειρά και αποδίδει στην έξοδο μια τμηματική γραμμική αναπαράστασή της, εμπίπτει στην κατηγορία των αλγορίθμων τμηματοποίησης. Συνοπτικά, ως ΤΓΑ ορίζεται η κατά προσέγγιση αναπαράσταση μιας χρονοσειράς T, μήκους n, με K ευθύγραμμα τμήματα (Εικ. 1.3). Σύμφωνα με τη μέθοδο αυτή, προσεγγιστική ευθεία για το τμήμα $(p_i,\ldots,p_j)$ είναι εκείνη που διέρχεται από τα σημεία p_i, και p_j, κατά συνέπεια η εν λόγω μέθοδος τείνει να ταυτίζει το τελικό με το αρχικό σημείο δύο διαδοχικών τμημάτων της χρονοσειράς, αποδίδοντας μια συνεχή προσεγγιστική γραμμή. Αρχικά δημιουργεί μια πρόχειρη προσέγγιση του n χρησιμοποιώντας $n/2$ τμήματα, συγχωνεύοντας δε στη συνέχεια επαναληπτικά τα πιο επουσιώδη από αυτά, έως ότου καταλήξει στον απαιτούμενο προκαθορισμένο αριθμό τμημάτων. Η πολιτική του αλγορίθμου όσον αφορά στη συγχώνευση έγκειται στη σύγκριση του κόστους συγχώνευσης όμορων τμημάτων. Η εύρεση της προσεγγιστικής γραμμής την οποία η ΤΓΑ θα χρησιμοποιήσει στην αναπαράσταση είναι δυνατόν να ευρεθεί με δύο τουλάχιστον τρόπους (Εικόνες 1.4 και 1.5):

1. Τμηματική Γραμμική Παρεμβολή (Linear Interpolation): Δεδομένου του σχεδιασμένου τμήματος $T[a:b]$ της χρονοσειράς T στα σημεία t_a και t_b, η προσεγγιστική γραμμή ορίζεται από τα σημεία t_a και t_b, είναι δε δυνατόν να ληφθεί σε συνεχή χρόνο

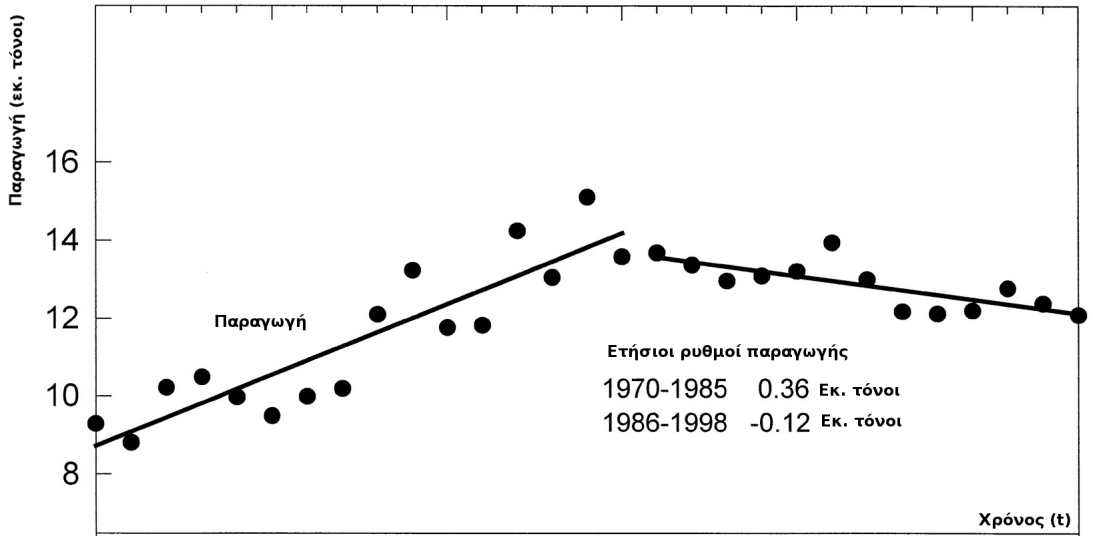
2. Τμηματική Γραμμική Παλινδρόμηση (Linear Regression): Δεδομένου του σχεδιασμένου τμήματος $T[a:b]$ της χρονοσειράς T στα σημεία t_a και t_b, η προσεγγιστική γραμμή προκύπτει με τη μέθοδο των ελαχίστων τετραγώνων (least squares)



Σχήμα 1.3: *Χρονοσειρές και η ΤΓΑ αναπαράστασή τους (Α: διαστημική τηλεμετρία, Β: Ηλεκτροκαρδιογράφημα) (κατά Keogh et al., 2003).*



Σχήμα 1.4: *Σωματιδιακή κίνηση Brown (μπλε γραμμή) και η Τμηματική Γραμμική Αναπαράστασή της (κόκκινη γραμμή) μέσω παρεμβολής (κατά Markussen, 2007).*



Σχήμα 1.5: *Ετήσια παραγωγή ορυζώνων στην περιοχή Zhejiang της Κίνας και η Τμηματική Γραμμική Αναπαράστασή της μέσω παλινδρόμησης [68].*

Η μέθοδος της γραμμικής παρεμβολής παράγει αναπαραστάσεις στις οποίες ταυτίζεται το τελικό σημείο του προηγούμενου τμήματος με το αρχικό σημείο του επόμενου, με αποτέλεσμα η τελική τμηματική αναπαράσταση να παρουσιάζει μεγαλύτερο ποσοστό εξομάλυνσης, σε αντίθεση με την τμηματική γραμμική παλινδρόμηση που συνήθως δημιουργεί μη συνεχείς αναπαραστάσεις στις περισσότερες περιπτώσεις. Η αισθητική υπεροχή της γραμμικής παρεμβολής σε συνδυασμό με τις χαμηλές απαιτήσεις της σε υπολογιστική ισχύ την καθιστούν δημοφιλή επιλογή σε εφαρμογές επεξεργασίας γραφικών μέσω υπολογιστών συστημάτων. Παρ' όλ' αυτά, η ποιότητα της αναπαράστασης είναι σε γενικές γραμμές υποδεέστερη εκείνης που παρέχεται από την προσέγγιση της γραμμικής παλινδρόμησης.

Επειδή ο αριθμός των τμημάτων που δημιουργούνται είναι τυπικά πολύ μικρότερος του αριθμού των στοιχείων της αρχικής χρονοσειράς, η αναπαράσταση αυτού του είδους όχι μόνο καθιστά αποτελεσματικότερη την αποθήκευση και τη μετάδοση των δεδομένων, αλλά επίσης επιτρέπει ακριβέστερους υπολογισμούς και ταξινομήσεις. Οι διάφοροι αλγόριθμοι οι οποίοι ανήκουν στην ΤΓΑ εμφανίζονται στη διεθνή βιβλιογραφία υπό ποικίλες ονομασίες, οι δε υλοποιήσεις τους είναι δυνατό να κα-

ταταγούν στις εξής κατηγορίες:

α. Αλγόριθμοι διολισθαίνοντος παραθύρου (Sliding Window Algorithms): Τα τμήματα αυξάνουν σε μήκος έως ότου αυτό ξεπεράσει ένα προκαθορισμένο όριο σφάλματος. Η διαδικασία κατάτμησης συνεχίζεται με τον ίδιο τρόπο, αφού το επόμενο σημείο της χρονοσειράς έχει επισημανθεί ως το αρχικό σημείο του επόμενου σχηματιζόμενου τμήματος.

β. Από πάνω προς τα κάτω διαδοχική διχοτόμηση (Top-Down Algorithm): Η χρονοσειρά διχοτομείται επαναληπτικά έως ότου ικανοποιηθεί συγκεκριμένο κριτήριο περαίωσης.

γ. Από κάτω προς τα πάνω διαδοχική συγχώνευση (Bottom-Up Algorithm): Ξεκινώντας από τη λεπτομερέστερη αναπαράσταση, τα τμήματα που έχουν δημιουργηθεί συγχωνεύονται διαδοχικά, έως ότου ικανοποιηθεί συγκεκριμένο κριτήριο περαίωσης.

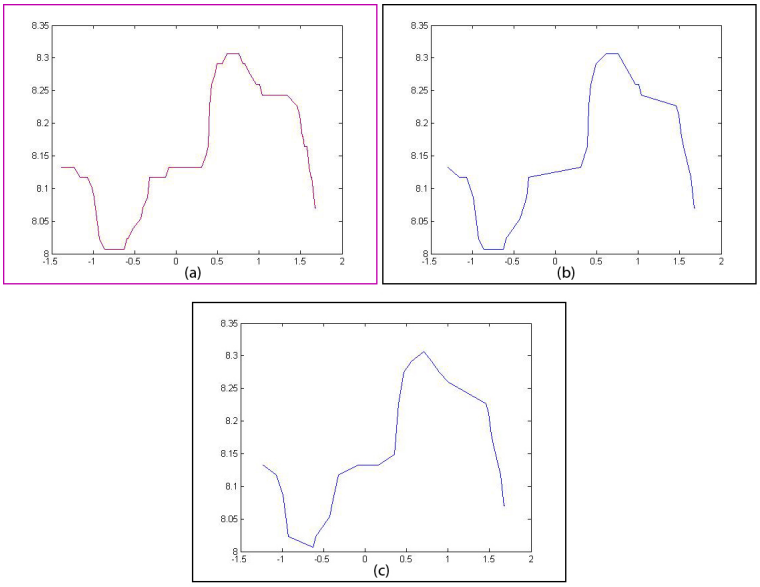
Όλοι οι αλγόριθμοι τμηματοποίησης απαιτούν την ενσωμάτωση μιας μεθόδου αποτίμησης όσον αφορά στην ποιότητα προσαρμογής κάθε σχεδιαζόμενου τμήματος. Κοινά μέτρα προς την κατεύθυνση αυτή, σε συνδυασμό με τη γραμμική παλινδρόμηση, είναι:

- το άθροισμα τετραγώνων ή το σφάλμα υπολοίπου, το οποίο υπολογίζεται ως το άθροισμα των τετραγωνισμένων διαφορών μεταξύ των αρχικών σημείων και των αντίστοιχών τους σημείων της καλύτερα προσαρμοσμένης γραμμής.

- η απόσταση μεταξύ της καλύτερα προσαρμοσμένης γραμμής και του πλέον απομακρυσμένου σημείου των πρωτογενών δεδομένων.

1.4.1 Αλγόριθμοι διολισθαίνοντος παραθύρου

Η κατάτμηση αποτελεί ένα πολύ σημαντικό εργαλείο αναπαράστασης των χρονοσειρών, χρησιμοποιείται δε είτε σαν μέθοδος προ-επεξεργασίας σε διάφορα προβλήματα εξόρυξης δεδομένων, είτε ως τεχνική ανάλυσης τάσεων, είτε ακόμα και ως τεχνική διακριτοποίησης. Για παράδειγμα, μια απλή μέθοδος διακριτοποίησης



Σχήμα 1.6: *Πορεία αναπαράστασης χρονοσειράς βάσει του αλγορίθμου διολισθαίνοντος παραθύρου (κατά Padergnana και Furlani).*

περιλαμβάνει αλγόριθμο που υλοποιεί παράθυρο σταθερού εύρους για την τμηματοποίηση της χρονοσειράς σε υπο-ακολουθίες και στη συνέχεια τη χρήση των πρωτόλειων πρότυπων σχημάτων τα οποία διαμορφώνονται για την αναπαράστασή της [40]. Η διαδικασία αυτή εξαρτάται πρωτίστως από το μέγεθος του παραθύρου κατάτμησης. Ωστόσο, η χρήση παραθύρου σταθερού εύρους αποτελεί υπεραπλουστευμένη μέθοδο αναπαράστασης, δεδομένου ότι στις περισσότερες περιπτώσεις πρότυπα ουσιώδους σημασίας

- υφίστανται στη χρονοσειρά σε διαφορετικά μεγέθη,
- περιλαμβάνουν ποικίλα δομικά της συστατικά και
- η θέση τους στον άξονα του χρόνου δεν είναι εκ των προτέρων γνωστή.

Συνεπώς, κατά την κατάτμηση μέσω αλγορίθμων σταθερού μήκους παραθύρου, παρατηρούνται τουλάχιστον τρία σοβαρά μειονεκτήματα:

1. Η κατάτμηση των δεδομένων κατά τον άξονα του χρόνου παρουσιάζει μικρότερη πιθανότητα να συμπεριλάβει στα διαμορφωθέντα τμήματα ικανοποιητικό ποσοστό προτύπων.

2. Αυξάνεται σε υπερβολικό βαθμό η πιθανότητα κατάτμησης και απώλειας ουσιωδών προτύπων, εφόσον ο αλγόριθμος κατακερματίζει τα αρχικά δεδομένα με σταθερό βήμα.

3. Ο αλγόριθμος αδυνατεί να ανακαλύψει σημαντικά χρονικά σημεία της χρονοσειράς, ιδιαίτερα στις περιπτώσεις όπου αυτά αποτελούν σημεία έναρξης ή λήξης ενός ουσιώδους προτύπου.

Δεδομένου ότι τα προαναφερόμενα μειονεκτήματα αποδείχθηκαν ιδιαίτερης σημασίας ως προς την ποιότητα της αναπαράστασης, η έρευνα εγκατέλειψε τις μεθόδους παραθύρου σταθερού εύρους και κατευθύνθηκε προς την εξεύρεση δυναμικών λύσεων. Οι τεχνικές που προτάθηκαν ενσωματώνουν ευέλικτες διαδικασίες ταυτοποίησης συγκεκριμένων σημαντικών χρονικών σημείων, χρησιμοποιώντας αλγορίθμους μεταβλητού εύρους παραθύρου κατάτμησης.

Προς την κατεύθυνση αυτή, οι συχνότερα χρησιμοποιούμενες μέθοδοι περιλαμβάνουν τη μέθοδο των Αντιληπτικά Σημαντικών Σημείων (ΑΣΣ) ή την ανίχνευση αξιοσημείωτων γεγονότων σε καίρια χρονικά σημεία της χρονοσειράς, τη μέθοδο του Ελαχίστου Μήκους Μηνύματος (MML: Minimum Message Length) και τη μέθοδο του Ελάχιστου Μήκους Περιγραφής (MDL: Minimum Description Length) [127]. Επίσης, η μέθοδος της Ανάλυσης των Κυρίαρχων Συνιστωσών (ΑΚΣ) (PCA: Principal Component Analysis) έχει προταθεί ως ικανοποιητική λύση στο θέμα της δυναμικής κατάτμησης, είτε αυτούσια [16, 110] είτε σε συνδυασμό με ασαφή λογική [145] και παρόμοιες τεχνικές. Άλλες συνδυαστικές μέθοδοι δυναμικής κατάτμησης περιλαμβάνουν αλγορίθμους που λειτουργούν σε δύο φάσεις, εκ των οποίων η πρώτη χαρακτηρίζεται συνήθως από την εφαρμογή μεθόδων για την επίτευξη ενός πρωτόλειου μορφώματος κατάτμησης. Το πρόχειρο αυτό σχέδιο λειτουργεί κυρίως ως οδηγός για τη δεύτερη φάση του αλγορίθμου, κατά την οποία συγκεκριμενοποιούνται τα ασαφή όρια των τμημάτων.

Μια ακόμη μεγάλη κατηγορία δυναμικών κατατμητικών μεθόδων σχετίζεται με την αξιολόγηση και την ταυτοποίηση σημαντικών σημείων της ακολουθίας τα οποία λειτουργούν ως σημεία διαφοροποίησης, με την έννοια ότι στα χρονικά αυτά στάδια παρατηρείται σημαντική αλλαγή της συμπεριφοράς της χρονοσειράς [174]. Τυπική λύση στο πρόβλημα εντοπισμού τέτοιων σημείων αποτελεί καταρχήν ο υπολογισμός του αριθμού τους και στη συνέχεια η διακρίβωση της θέσης τους, ενώ ακολουθεί η

εύρεση συναρτήσεων προσαρμογής των διαστημάτων μεταξύ διαδοχικών σημείων διαφοροποίησης. Ο Chu [40] προτείνει μια διαδικασία κατάτμησης διολισθαίνοντος δοκιμαστικού παραθύρου βασισμένη σε μη στάσιμη (non-stationary) ανίχνευση στατιστικών μέτρων διακύμανσης (fluctuation statistics) και εντοπισμού των σημείων διαφοροποίησης. Σε άλλες περιπτώσεις [61], κατάλληλα τροποποιημένα πρότυπα προσαρμόζονται στα σχεδιαζόμενα τμήματα και στη συνέχεια χρησιμοποιούνται κριτήρια πιθανοφάνειας (likelihood criteria) για την αξιολόγηση του ενδεχομένου περαιτέρω κατάτμησης. Επίσης, μέθοδοι προς την κατεύθυνση της προτυποποίησης του μηχανισμού που διέπει τη διαφοροποίηση της συμπεριφοράς σε συγκεκριμένα σημεία της χρονοσειράς περιλαμβάνουν τον προσδιορισμό του αριθμού των υπο-διαδικασιών και της δυναμικής καθεμιάς από αυτές. Σε άλλες περιπτώσεις, η μέθοδος του δυναμικού προγραμματισμού χρησιμοποιείται για τον καθορισμό του συνολικού αριθμού των διαστημάτων, τον εντοπισμό της τοποθεσίας τους και την τάξη του προτύπου που αντιστοιχεί σε κάθε διάστημα [111].

Το πρόβλημα της κατάτμησης μιας χρονοσειράς έχει επίσης μελετηθεί με βάση τη διερευνητική ανάλυση δεδομένων (ΔΑΔ) (EDA: Exploratory Data Analysis) και εξόρυξης δεδομένων, αλλά και της εύρεσης κυκλικής περιοδικότητας των τμημάτων. Στην πρώτη περίπτωση χρησιμοποιείται η μέθοδος των Scale-Sensitive Gated Experts (SSGE) για την κατάτμηση πολύπλοκων μη-γραμμικών ακολουθιών σε ομάδες απλούστερων [183], ενώ η χρήση της μεθόδου του ψευδούς εγγύτερου γείτονα (false nearest neighbour) συμβάλλει στον προσδιορισμό του κατάλληλου βαθμού διάστασης και καθυστέρησης [155]. Σε άλλες περιπτώσεις επιχειρείται εκ των προτέρων εξόρυξη δεδομένων με διάφορες μεθόδους, σε συνδυασμό με τμηματοποίηση, ούτως ώστε να ευρεθεί σύγχρονη ή ασύγχρονη περιοδικότητα σταθερού μήκους [190].

Κοινός τόπος και κυριότερο χαρακτηριστικό των προαναφερθέντων δυναμικών μεθόδων είναι ότι αδυνατούν να προσεγγίσουν αποτελεσματικά το πρόβλημα της κατάτμησης της χρονοσειράς, όταν ουσιώδη πρότυπα είναι άγνωστα εκ των προτέρων. Είναι συνεπώς απαραίτητο να ταυτοποιηθούν *a-priori* συγκεκριμένα πρότυπα, ούτως ώστε να μορφοποιηθεί ένας πολύπλευρος χώρος εξόρυξης δεδομένων. Η μορφοποίηση συγκεκριμένων κανόνων για κάθε περίπτωση, οι οποίοι διέπουν την εξαγωγή χαρακτηριστικών και την αναπαράσταση των δεδομένων, αυξάνει

την αποτελεσματικότητα των δυναμικών αλγορίθμων κατάτμησης [22]. Στα περισσότερα βεβαίως προβλήματα, ιδιαίτερα βιολογικής φύσης, η *a-priori* ταυτοποίηση συγκεκριμένων προτύπων είναι πολύ δύσκολη, ενώ η σύνταξη κανόνων εξαρτάται από την προηγούμενη γνώση του προβλήματος και γίνεται πολύ δύσκολη. Οι περιπτώσεις αυτές συνήθως συνιστούν ουσιαστικά προβλήματα βελτιστοποίησης, για τα οποία οι εξελικτικοί αλγόριθμοι αποτελούν τις καταλληλότερες μεθόδους προσέγγισης [43, 36, 107].

Σε γενικές γραμμές, το αποτέλεσμα της επεξεργασίας μέσω του διολισθαίνοντος παραθύρου είναι η τμηματοποίηση των αρχικών δεδομένων και η εξαγωγή μιας αντιπροσωπευτικής αξίας για κάθε τμήμα. Τα τμήματα δημιουργούνται διαδοχικά, το ένα μετά το άλλο, το δε εύρος τους είναι προοδευτικά αυξανόμενο και καθορίζεται από μια τιμή κατωφλίου. Η συγκεκριμένη αυτή τιμή αποτελεί βασική παράμετρο του αλγορίθμου και τίθεται από το χρήστη πριν την εκκίνησή του. Καθ' όλη τη διάρκεια της προοδευτικής αύξησης του εύρους ενός εκάστου τμήματος, το παραγόμενο σφάλμα συγκρίνεται με την τιμή αυτή και, όταν αυξηθεί σε επίπεδα μεγαλύτερα της τιμής κατωφλίου, ο αλγόριθμος σταματά την αύξηση του τρέχοντος τμήματος, επισημαίνει το προηγούμενο σημείο ως το τελικό του τμήματος, ενώ το τρέχον σημείο ως το αρχικό του επόμενου. Η διαδικασία επαναλαμβάνεται από την αρχή έως ότου όλα τα σημεία της χρονοσειράς εξαντληθούν και δημιουργηθεί με τον τρόπο αυτό μια γραμμική αναπαράσταση των πρωτογενών δεδομένων. Ο γενικευμένος[3] ψευδο-κώδικας του αλγορίθμου διολισθαίνοντος παραθύρου έχει ως εξής:

Listing 1.1: Ψευδοκώδικας Διολισθαίνοντος Παραθύρου (κατά Keogh κ.α. 2003)

```
1 Algorithm Seg_TS = Sliding_window(T, max_error)
2 anchor = 1;
3 WHILE not finished segmentation process:
4     i = 2;
5     WHILE calculate_error(T[anchor:anchor + i]) < max_error
6         i = i+1;
7     END WHILE;
8     Seg_TS = concat(Seg_TS, create_segment(T[anchor:anchor+(i-1)]));
```

[3] Ο όρος χρησιμοποιείται για να δηλώσει ότι ο ψευδοκώδικας που ακολουθεί περιγράφει σε αδρές γραμμές όλες τις υλοποιήσεις του αλγορίθμου

```
9       anchor = anchor + i
10  END WHILE
```

Ο αλγόριθμος διολισθαίνοντος παραθύρου είναι ιδιαίτερα ελκυστικός εξαιτίας της μεγάλης ευκολίας στην κατανόηση και την εφαρμογή του, αλλά και εξαιτίας της σειριακής (online)[4] φύσης του [88]. Βασικότερη παράμετρος στην περίπτωση αυτή είναι το βήμα με το οποίο γίνεται ο έλεγχος του μεγέθους κάθε τμήματος. Ποικίλες παραλλαγές και βελτιώσεις του βασικού σχήματος του αλγορίθμου έχουν προταθεί, με βάση τον καθορισμό αυτής της παραμέτρου και σε συνδυασμό με διαφόρων τύπων δεδομένα και προβλήματα. Συγκεκριμένα, αναφέρεται ότι είναι δυνατή η επιτάχυνση του αλγορίθμου με στατιστικά μη σημαντικές επιδράσεις στην έξοδο, βελτιστοποιώντας το μέγεθος του βήματος τμηματοποίησης [149]. Δεδομένου του γεγονότος ότι το σφάλμα του υπολοίπου βαίνει μονότονα μη-μειούμενο με την προσθήκη νέων σημείων δεδομένων, δεν είναι απαραίτητος ο έλεγχος του βήματος από το δεύτερο σημείο και μετά. Ικανοποιητική επιτάχυνση του αλγορίθμου με ελάχιστες απώλειες στην ποιότητα αναπαράστασης επιτυγχάνεται εάν τεθεί η αρχική τιμή του βήματος στη μέση τιμή εύρους των προηγούμενων τμημάτων. Στην περίπτωση που το σφάλμα παραμένει μικρότερο της τιμής κατωφλίου, ο αλγόριθμος προχωρεί στην αύξηση του βήματος σύμφωνα με την κλασσική μέθοδο. Σε αντίθετη περίπτωση, η τιμή του μειώνεται έως ότου επιτευχθεί η μείωση του σφάλματος σε επίπεδα μικρότερα του κατωφλίου [61]. Η μέθοδος αυτή είναι αποτελεσματική εάν το μέσο μήκος των τμημάτων είναι μεγάλο συγκρινόμενο με την τυπική τους απόκλιση.

Παρά το γεγονός ότι στις περισσότερες περιπτώσεις προβλημάτων οι συνθήκες υπό τις οποίες ο αλγόριθμος παρουσιάζει μέγιστη αποτελεσματικότητα δεν έχουν τεκμηριωθεί πλήρως, πιθανώς εξαιτίας του γεγονότος ότι έχει τεθεί υπό δοκιμή κυρίως σε δεδομένα υψηλού σχετικά θορύβου (βάσεις δεδομένων χρηματιστηριακού ενδιαφέροντος), στην έρευνα [49] το πρόβλημα αναλύεται διεξοδικά, όπως επίσης και η δυσκολία της καθολικής του εφαρμογής. Πράγματι, η ποιότητα της παραγόμενης αναπαράστασης που προκύπτει από την εφαρμογή του αλγορίθμου διολισθαίνοντος παραθύρου μειώνεται στις περιπτώσεις κατά τις οποίες τίθενται υπό ανά-

[4] Στην επιστήμη των υπολογιστών ως online αλγόριθμος ορίζεται αυτός που δέχεται τα δεδομένα εισαγωγής σειριακά και κατά δέσμες. Κατά συνέπεια, ο αλγόριθμος καλείται να λάβει αποφάσεις μετά από κάθε ερέθισμα στην εισαγωγή, χωρίς να γνωρίζει εκ των προτέρων όλα τα δεδομένα.

λυση δεδομένα χρονοσειρών ιδιαίτερα μεγάλου εύρους. Το πρόβλημα γίνεται εντονότερο κατά την ανάλυση δεδομένων που προέρχονται από αποτελέσματα πειραμάτων βιολογικών ή γεωτεχνικών επιστημών, όπου μεγαλύτερη σημασία από τις μεμονωμένες τιμές της χρονοσειράς έχει η εξεύρεση των μεταξύ τους σχέσεων, καθώς επίσης και η αναγνώριση και εξαγωγή προτύπων που υποκρύπτονται στα αρχικά δεδομένα. Επίσης, σε πραγματικά προβλήματα που παρουσιάζουν ποικίλα επίπεδα θορύβου στη χρονοσειρά, το πλήθος των τμημάτων που σχηματοποιούνται από τη μέθοδο είναι στις περισσότερες περιπτώσεις υπερβολικό.

Διάφορες παραλλαγές του αλγορίθμου έχουν προταθεί για την διεύρυνση του πεδίου εφαρμογής του, με κυριότερη τη μέθοδο τμηματοποίησης που βασίζεται στη μονοτονικότητα των σχεδιαζόμενων τμημάτων υπό την ονομασία Τμηματική Προσέγγιση για Αναζητήσεις Υπο-Ακολουθιών (SBASS: Segment-Based Approach for Subsequence Search) [142]. Σύμφωνα με τη μέθοδο αυτή, τα τμήματα που σχηματίζονται απαρτίζονται από σημεία δεδομένων του τύπου $x_1 \leq x_2 \leq ... \leq x_n$ ή $x_1 \geq x_2 \geq ... \geq x_n$. Βασική προϋπόθεση για τη συμμετοχή ενός σημείου της αρχικής χρονοσειράς στο δημιουργούμενο τμήμα είναι να διατηρεί τη μονοτονικότητα (αύξουσα ή φθίνουσα) των προηγούμενων τιμών του συγκεκριμένου τμήματος. Κάθε επόμενο σημείο της χρονοσειράς δοκιμάζεται και, στις περιπτώσεις που η μονοτονικότητα στο τρέχον σημείο αλλάζει φορά, το τμήμα περατώνεται στο προηγούμενο σημείο και ένα καινούριο τμήμα ξεκινάει στο σημείο μεταβολής. Η εν λόγω μέθοδος καθόρισε μια αρκετά ικανοποιητική αναπαράσταση για τις συγκεκριμένες χρονοσειρές στις οποίες εφαρμόσθηκε.

Οι διάφορες παραλλαγές του αλγορίθμου είναι ιδιαίτερα δημοφιλείς στην ιατρική κοινότητα, ως μεθοδολογία συμπίεσης δεδομένων και εξόρυξης προτύπων από αυτά. Το ηλεκτροκαρδιογράφημα (ΗΚΓ) (ECG: Electrocardiogram) αποτελεί μια γραφική αναπαράσταση της καρδιακής λειτουργίας, μέσω της οποίας παρέχονται απαραίτητες πληροφορίες στους καρδιολόγους. Μια τυπική συσκευή ΗΚΓ παράγει μεγάλο όγκο δεδομένων που σε πολλές περιπτώσεις μπορεί να φθάσουν στα επίπεδα των 60 - 480 Kb/min ανά αισθητήρα [138], ενώ συνήθως περιλαμβάνει περισσότερους του ενός αισθητήρες. Επιπροσθέτως, απαιτείται να προβλέπονται ανωμαλίες της καρδιακής λειτουργίας κατά τη διάρκεια ροής των δεδομένων σε πραγματικό χρόνο, δηλαδή το ΗΚΓ και η αναγνώριση του εν λόγω σήματος είναι

μια κατά κύριο λόγο σειριακή διαδικασία. Συνεπώς, υπάρχει η ανάγκη αφενός αποτελεσματικής συμπίεσης για τον αποθησαυρισμό των σημάτων [59] και αφετέρου αποτελεσματικής αναδόμησης της συμπιεσμένης πληροφορίας σε δεύτερο χρόνο, ενώ τέλος, απαιτείται αποτελεσματική αναπαράσταση και εξαγωγή προτύπων από το αρχικό σήμα [151]. Διάφορες παραλλαγές του αλγορίθμου διολισθαίνοντος παραθύρου έχουν προταθεί προς την κατεύθυνση της συμπίεσης της αρχικής πληροφορίας και αναδόμησής της [86]. Για τη μείωση του βαθμού διάστασης και την αναγνώριση προτύπων κυρίαρχες μέθοδοι που εμπλέκουν τεχνικές διολισθαίνοντος παραθύρου είναι ο αλγόριθμος του αθροίσματος της διαφοράς τετραγώνων (SSD: Sum Square Difference) [41] και ο αλγόριθμος εποχικής κωδικοποίησης εύρους ζώνης (EKEZ) (AZTEC: Amplitude Zone Time Epoch Coding) [101], ο οποίος επιτυγχάνει μεγάλη μείωση του βαθμού διάστασης αλλά παρουσιάζει προβλήματα κατά τη φάση αναδόμησης της αρχικής πληροφορίας [162]. Τέλος, η μέθοδος της πολυγωνικής προσέγγισης και ειδικότερα η παραλλαγή της σάρωσης (SAPA: Scan-Along Polygonal Approximation) είναι μια από τις πλέον τεκμηριωμένες μεθόδους κατάτμησης χρονοσειρών [86, 164, 17, 160, 161, 117, 147, 18, 78, 192, 197, 100, 144, 29] η οποία χρησιμοποιεί τεχνικές γραμμικής αναπαράστασης για τη συμπίεση και απόδοση των αρχικών δεδομένων.

1.4.2 Από-πάνω-προς-τα-κάτω διαδοχική διχοτόμηση

Η λειτουργία των αλγορίθμων αυτών έγκειται σε μια μέθοδο εξέτασης κάθε δυνατής τμηματοποίησης της χρονοσειράς και τελικά διαχωρισμού της στα καταλληλότερα σημεία. Στη συνέχεια, τα τμήματα που δημιουργούνται αξιολογούνται ως προς το επίπεδο του σφάλματος προσέγγισης που σημειώνουν. Εάν αυτό είναι μεγαλύτερο από μια προκαθορισμένη τιμή κατωφλίου, ο αλγόριθμος συνεχίζει να διαχωρίζει τα τμήματα της χρονοσειράς, έως ότου όλα τα σφάλματα προσέγγισης μειωθούν κάτω από το συγκεκριμένο κατώφλι. Ο γενικευμένος ψευδο-κώδικας του αλγορίθμου από-πάνω-προς-τα-κάτω αποδίδεται ως εξής:

Listing 1.2: Ψευδοκώδικας από-πάνω-προς-τα-κάτω (κατά Keogh κ.α. 2003)

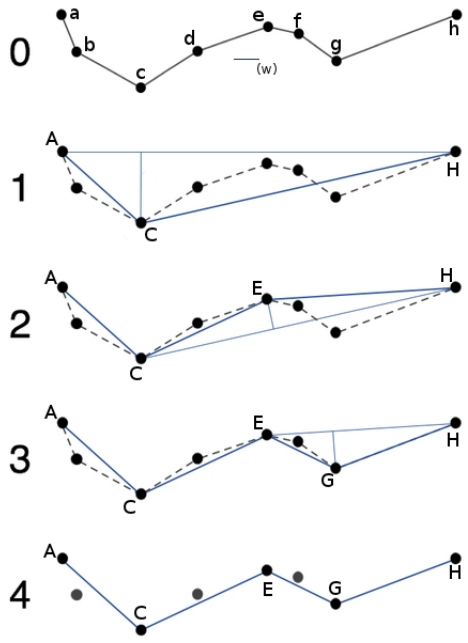
```
1 Algorithm Seg_TS = Top_Down(T, max_error)
2 best_so_far = inf;
3 for i = 2 to (length(T) - 2)
4 improvement_in_approximation = improvement_splitting_here(T,i);
```

```
5           if improvement_in_approximation < best_so_far
6                   breakpoint = i;
7                   best_so_far = improvement_in_approximation;
8           end;
9   end;
10  if calculate_error(T[1:breakpoint]) > max_error
11          Seg_TS = Top_Down(T[1:breakpoint]);
12  end;
13  if calculate_error(T[breakpoint+1:length(T)]) > max_error
14          Seg_TS = Top_Down(T[breakpoint+1:length(T)]);
15  end;
```

Αρχικά ο αλγόριθμος αναζητεί την άριστη θέση διαμερισμού (γραμμή 3), ενώ στη συνέχεια (γραμμές 10 και 13 αντίστοιχα) ξεκινούν οι ρουτίνες αναδρομικού τεμαχισμού του αριστερού και δεξιού τμήματος, εφόσον αυτό είναι απαραίτητο. Ο εν λόγω αλγόριθμος έχει χρησιμοποιηθεί για να υποστηρίξει την αναπαράσταση δεδομένων χρονοσειρών σε ποικίλα αφηρημένα επίπεδα[5] [113] ή για την εκτέλεση προσεγγιστικών αναζητήσεων [166]. Στις εργασίες αυτές μια χρονοσειρά αναπαρίσταται μέσω ενός συνόλου διακεκριμένων απλών και σύνθετων αντικειμένων. Τα πρώτα συνίστανται από ομογενείς υπο-ακολουθίες των αρχικών δεδομένων, ενώ τα δεύτερα από τη σύνθεση πολλών απλών αντικειμένων μέσω συγκεκριμένων τελεστών. Τα απλά αντικείμενα καθορίζονται σε διάφορα επίπεδα αφαίρεσης με χαμηλότερο το σύνολο των αρχικών σημείων της χρονοσειράς, αμέσως μετά το επίπεδο των χαρακτηριστικών της και τέλος το σημασιολογικό ή εννοιολογικό (semantic) της επίπεδο.

Ο αλγόριθμος τυγχάνει ευρύτατης εφαρμογής σε ποικίλα επιστημονικά πεδία. Βασικός του σκοπός είναι η αποτελεσματική μείωση των σημείων δεδομένης καμπύλης η οποία αποτελείται από ευθύγραμμα τμήματα. Η αρχική εκδοχή του αλγορίθμου προτάθηκε σε αναφορά με την επεξεργασία εικόνας το 1972 από τον Urs

[5] Ως *αφαίρεση* στην επιστήμη των υπολογιστών νοείται η διαδικασία κατά την οποία δεδομένα ή/και προγράμματα αναπαρίστανται με βάση το νόημά τους (semantics) ενώ οι εφαρμοστικές τους λεπτομέρειες τίθενται στο περιθώριο, έτσι ώστε η ανάπτυξη ενός συστήματος να εστιάζεται στις σημαντικότερες συνιστώσες του. Είναι δυνατόν να καθορίζονται ποικίλα επίπεδα αφαίρεσης και μέσω αυτών να παρουσιάζονται διαφορετικές απόψεις του προβλήματος. Χαμηλότερα επίπεδα αφαίρεσης μπορεί να σχετίζονται με το υλικό (hardware) επί του οποίου τρέχει το πρόγραμμα, ενώ υψηλότερα επίπεδα με τη λογική του ίδιου του προγράμματος.

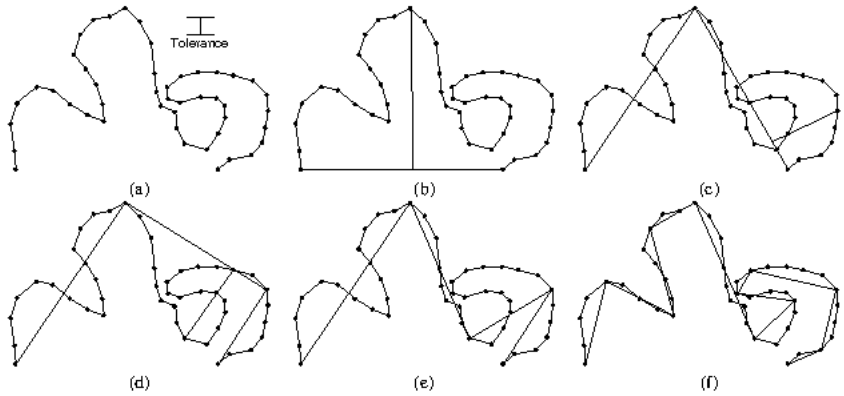


Σχήμα 1.7: *Σχηματική παράσταση του αλγορίθμου Douglas - Peucker / Ramer.*

Ramer [153], ενώ ένα χρόνο αργότερα, σε ανεξάρτητη έρευνα σχετικά με τη συμπίεση δεδομένων χαρτογραφίας, εισάγεται από τους David Douglas και Thomas Peucker [53] [6]. Παράδειγμα συμπίεσης εικόνας σύμφωνα με τον αλγόριθμο Douglas-Peucker δίδεται στα σχήματα 1.7 και 1.8.

Κατά την αρχικοποίησή τους αμφότεροι οι αλγόριθμοι δέχονται στην εισαγωγή ένα σύνολο διατεταγμένων σημείων και μια σταθερά που παίζει το ρόλο ελάχιστης απόστασης. Η πρώτη προσεγγιστική αναπαράσταση που ορίζεται από τον αλγόριθμο είναι το ευθύγραμμο τμήμα που ενώνει τα σημεία αρχής και τέλους των αρχικών δεδομένων. Στη συνέχεια ο αλγόριθμος βρίσκει το σημείο με τη μεγαλύτερη απόσταση από το ευθύγραμμο αυτό τμήμα. Εάν η απόσταση αυτή είναι μικρότερη από την προκαθορισμένη τιμή κατωφλίου, τότε τερματίζεται και αποδίδει το ευθύγραμμο τμήμα ως αναπαράσταση. Εάν η απόσταση είναι μεγαλύτερη, τότε επαναλαμβάνει τη διαδικασία για τα σημεία που περιβάλλουν τα σχηματιζόμενα δύο νέα ευθύγραμμα τμήματα και τερματίζεται όταν καλυφθούν όλα τα σημεία του αρχι-

[6] Στον τομέα της συμπίεσης εικόνας είναι γνωστός με τις ονομασίες αλγόριθμος Ramer–Douglas–Peucker ή επαναληπτικός αλγόριθμος καταλληλότητας τελικού σημείου ή ακόμη αλγόριθμος κατάτμησης-και-συγχώνευσης

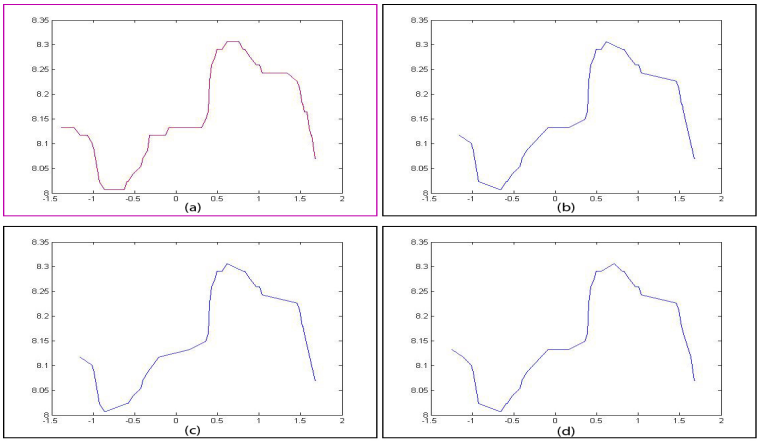


Σχήμα 1.8: *Συμπίεση δεδομένων εικόνας σύμφωνα με τον αλγόριθμο Douglas-Peucker [187].*

κού διατεταγμένου συνόλου. Η τελική αναπαράσταση αποτελείται από τα σημεία και μόνο εκείνα που παρουσιάζουν απόσταση από τα σχηματιζόμενα τμήματα μεγαλύτερη της προκαθορισμένης. Για παράδειγμα, στο Σχήμα 1.7 αρχικά δίδεται η ακολουθία σημείων a, b, c, d, e, f, g, h και η ελάχιστη απόσταση (w) του αλγορίθμου (Σειρά 0). Στην πρώτη φάση ο αλγόριθμος στοχοποιεί το σημείο c ως το πιο απομακρυσμένο σημείο από την πρώτη ακατέργαστη αναπαράσταση (ευθύγραμμο τμήμα AH). Επειδή η απόσταση του c από το AH είναι μεγαλύτερη της τιμής κατωφλίου, το c θεωρείται πλέον σημείο της τελικής αναπαράστασης. Ταυτόχρονα, ορίζονται τα τμήματα AC και CH (Σειρά 1). Μεταξύ των a και c βρίσκεται το σημείο b του οποίου η απόσταση από το AC είναι μικρότερη της τιμής κατωφλίου και συνεπώς απορρίπτεται. Μεταξύ των σημείων c και h βρίσκονται τα σημεία d, e, f και g, με πιο απομακρυσμένο από το τμήμα CH το σημείο e, η απόσταση του οποίου από το CH είναι μεγαλύτερη της τιμής κατωφλίου (Σειρά 2). Συνεπώς το σημείο e θα συμπεριληφθεί στην τελική αναπαράσταση. Ταυτόχρονα ορίζονται τα ευθύγραμμα τμήματα CE και EH για τα οποία ακολουθείται η ίδια διαδικασία, έως ότου όλα τα σημεία της αρχικής σειράς εξαντληθούν. Συνεπώς ο αλγόριθμος εφαρμόζεται σε όλα τα σχεδιασμένα τμήματα στα οποία παρατηρείται σφάλμα πέραν κάποιου προκαθορισμένου κατωφλίου [143]. Η τελική αναπαράσταση A, C, E, G, H προκύπτει από την εφαρμογή του αλγορίθμου τοπικά σε ήδη σχεδιασμένα τμήματα, υπό την προϋπόθεση ότι αυτά παρουσιάζουν σχετικά υψηλό σφάλμα, υψηλότερο από την τιμή κατωφλίου που καθορίζεται κατά την αρχικοποίηση. Πρόκειται για μια γενικότερη

κατηγορία τμηματοποίησης η οποία όταν εφαρμόζεται σε δεδομένα χρονοσειρών αποτυπώνει όλες τις περιοχές υψηλού και χαμηλού επιπέδου, ενώ τα καταγεγραμμένα ακρότατα σημεία χρησιμοποιούνται για να καθορίσουν μια αδρή τμηματοποίηση. Όπως είναι φυσικό, όσο μικρότερη είναι η προκαθορισμένη τιμή κατωφλίου, τόσο πιστότερη θα είναι η αναπαράσταση, αλλά και τόσο μικρότερη η συμπίεση.

1.4.3 Από-κάτω-προς-τα-πάνω διαδοχική συγχώνευση



Σχήμα 1.9: *Πορεία αναπαράστασης χρονοσειράς βάσει του αλγορίθμου από-κάτω-προς-τα-πάνω (κατά Padergnana και Furlani).*

Επίσης και στην περίπτωση του αλγορίθμου αυτού, κατά την αρχικοποίηση δίδεται ένα μέτρο απόστασης το οποίο χρησιμοποιείται ως τιμή κατωφλίου η οποία αντιστοιχεί στο κόστος συγχώνευσης όμορων τμημάτων. Η μέθοδος αρχικά ορίζει την πρώτη τμηματοποίηση χρησιμοποιώντας *n/2* τμήματα για την προσέγγιση μιας χρονοσειράς αποτελούμενης από *n* σημεία δεδομένων. Στη συνέχεια υπολογίζεται το κόστος συγχώνευσης κάθε ζεύγους όμορων τμημάτων συγκρίνοντας για παράδειγμα το άθροισμα τετραγώνων των τμημάτων αυτών με την προκαθορισμένη τιμή κατωφλίου. Ο αλγόριθμος προχωρά με τη συγχώνευση των ζευγών χαμηλότερου κόστους έως ότου εκπληρωθεί κάποιο κριτήριο περαίωσης.

Από τη στιγμή που υλοποιείται η συγχώνευση δύο τμημάτων σε ένα μεγαλύτερο, ακολουθεί ο υπολογισμός του κόστους συγχώνευσης αυτού του μεγαλύτερου τμήματος με το προηγούμενο και το επόμενο τμήμα για να αξιολογηθεί η περίπτωση εκ νέου συγχώνευσης. Ο γενικευμένος ψευδο-κώδικας του αλγορίθμου αποδίδεται ως

εξής:

Listing 1.3: Ψευδοκώδικας από-κάτω-προς-τα-πάνω (κατά Keogh κ.α. 2003)

```
1  Algorithm Seg_TS = Bottom_Up(T, max_error)
2  for i = 1:2:length(T)
3          Seg_TS = concat(Seg_TS, create_segment(T[i:i+1]));
4  end;
5  for i = 1:length(Seg_TS)-1
6          merge_cost(i) = calculate_error([merge(Seg_TS(i), Seg_TS(i+1)
               )]);
7  end
8  while min(merge_cost) < max_error
9          index = min(merge_cost)
10         Seg_TS(index) = merge(Seg_TS(index), Seg_TS(index+1));
11         delete (Seg_TS(index+1));
12         merge_cost(index) = calculate_error(merge(Seg_TS(index),
               Seg_TS(index+1)));
13         merge_cost(index-1) = calculate_error(merge(Seg_TS(index-1),
               Seg_TS(index)));
14 end;
```

Εκκινώντας ο αλγόριθμος σχεδιάζει μια πρώτη πρόχειρη τμηματοποίηση (γραμμή 2), ενώ στη συνέχεια δημιουργείται η ρουτίνα υπολογισμού του κόστους συγχώνευσης των τμημάτων (γραμμή 5). Το ελάχιστο κόστος χρησιμοποιείται ως μέτρο καταλληλότητας για τη συγχώνευση ενός ζεύγους (γραμμή 9). Δισδιάστατες και τρισδιάστατες παραλλαγές του αλγορίθμου απαντώνται συχνά στην επιστήμη των υπολογιστικών γραφικών υπό τη γενικότερη ονομασία «μέθοδοι υποδιαίρεσης» (decimation methods) [39], ενώ εκτεταμένη χρήση του σημειώνεται στην εξόρυξη δεδομένων χρονοσειρών [90, 92, 89].

Ποικίλες είναι οι ερευνητικές προσπάθειες για εμπλουτισμό, επέκταση και βελτίωση της ΤΓΑ. Σε αυτές περιλαμβάνονται η προσαρμογή δυνατοτήτων βελτιστοποίησης προβλημάτων [69], η αξιολόγηση της σημαντικότητας των τμημάτων που προτείνονται με την χρήση πίνακα βαρών [90], ή ακόμη η ανατροφοδότηση συνάφειας (relevance feedback) από το χρήστη [91].

Η πρόσφατη αναβάθμιση που σημειώνεται στα εργαλεία εξελικτικού προγραμ-

ματισμού και τεχνητής νοημοσύνης, στα οποία θα αναφερθούμε αμέσως μετά, παράλληλα με την ολοένα και αυξανόμενη ανάγκη για ανάπτυξη αποτελεσματικότερων μεθόδων αναπαράστασης δεδομένων χρονοσειρών, αποτελούν τα βασικότερα κίνητρα της παρούσας έρευνας, στην οποία η μέθοδος ΤΓΑ αναβαθμίζεται μέσω ορισμένης εξελικτικής διαδικασίας παραγωγής δευτερογενών δεδομένων εκπαίδευσης συγκεκριμένων ταξινομητών, ώστε να αποτελέσει μια πρότυπη μέθοδο προ-επεξεργασίας χρονοσειρών.

Κεφάλαιο 2

ΤΑΞΙΝΟΜΗΤΕΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΧΡΟΝΟΣΕΙΡΕΣ

Στο κεφάλαιο αυτό παρουσιάζονται τα κυριότερα εργαλεία τεχνητής νοημοσύνης που αποτελούν τους ταξινομητές, αλλά και τη συνάρτηση αξιολόγησης του γενετικού αλγορίθμου του προτεινόμενου εργαλείου. Δίνεται έμφαση κυρίως στον τρόπο λειτουργίας τους - και στη μαθηματική προτυποποίησή της - έτσι ώστε να αποτελέσει ένα θεωρητικό υπόβαθρο για τη μελέτη του συστήματος, η οποία θα ακολουθήσει στα επόμενα κεφάλαια.

2.1 Εισαγωγή

Η σημασία της ανάλυσης των χρονοσειρών έχει ήδη υπογραμμισθεί ως κρίσιμης σημασίας σε ποικίλα ερευνητικά πεδία από την οικονομετρία και τα οικονομικά, ως τη βιολογία, την κλινική φαρμακολογία, τη μετεωρολογία, την υδρολογία και υδραυλική, τη δασοκομία, τη φυτική και ζωική παραγωγή και τη διαχείριση του περιβάλλοντος. Πρόκειται για το σύνολο των διαδικασιών μέσω των οποίων είναι δυνατός ο εντοπισμός και η αποκάλυψη των τάσεων και των προτύπων που υποκρύπτονται - αλλά και είναι δημιουργός συνιστώσα των πρωτογενών δεδομένων. Παραδοσιακά, το συχνότερα χρησιμοποιούμενο εργαλείο σε αυτού του είδους την ανάλυση είναι η στατιστική, κυρίως μέσω της στατιστικής διαστρωμάτωσης, της

ανάλυσης διασποράς / παλινδρόμησης, καθώς και των αυτοπαλινδρομικών προτύπων ARCH (Autoregressive Conditional Heteroscedasticity) [20, 13, 189]. Παράλληλα με τις προσπάθειες αυτές, τα τελευταία χρόνια παρατηρείται μεγάλη ώθηση στο πεδίο της υπολογιστικής νοημοσύνης η οποία περιλαμβάνει, εκτός των άλλων, επίσης και την ανάπτυξη ποικίλων αναλυτικών εργαλείων, κυρίως υπό εποπτεία ταξινομητών. Σύμφωνα με τη βιβλιογραφία, πρότυπα τεχνητών νευρωνικών δικτύων, μηχανών διανυσμάτων υποστήριξης και γενετικών αλγορίθμων έχουν για μια ακόμη φορά προσελκύσει το ενδιαφέρον των μελετητών, κυρίως εξαιτίας του γεγονότος ότι σήμερα υπάρχει διαθέσιμη κατάλλλη τεχνογνωσία και επαρκές από άποψη ισχύος και ποιότητας υλικό για να προσομοιάσουν πολύπλοκες διεργασίες και μηχανισμούς. Πολλές περιπτώσεις Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ) (ANN: Artificial Neural Networks), ασαφούς λογικής και εξελικτικών αλγορίθμων, παράλληλα με συστήματα Bayes, Εγγύτερων Γειτόνων (Nearest Neigbours) ή Δένδρων Αποφάσεων (Decision Trees) συγκαταλέγονται σε μια μακρά λίστα μεθοδολογιών αναπτυγμένων ειδικά για την ανάλυση χρονοσειρών [33, 35, 85, 94, 102, 114, 134, 154].

Η υπολογιστκή νοημοσύνη προκύπτει ένεκα της αναζήτησης του ανθρώπου για τα αίτια της λογικής και της αιτίασης, αλλά και των μηχανισμών της κίνησης των διαφόρων τμημάτων του σώματος, προσπάθειας που χρονολογείται από το λυκαυγές της ιστορίας του. Από τα πρώτα ακόμη χρόνια της συνάθροισης σε κοινωνίες, η περιέργεια αυτή αποτέλεσε την κινητήρια δύναμη για την ανάπτυξη της επιστήμης των νευρώνων (Neuroscience), η οποία απέδωσε θαυμαστά αποτελέσματα στην κατανόηση της λειτουργίας των φυσιολογικών μηχανισμών και των φυσικοχημικών νόμων που τους διέπουν. Συνετέλεσε δε αποφασιστικά στη συνειδητοποίηση ότι ο ανθρώπινος εγκέφαλος είναι ο ίδιος ένας εκπληκτικών δυνατοτήτων βιολογικός μηχανισμός παράλληλης και κατανεμημένης επεξεργασίας χαρακτηριστικής πολυπλοκότητας και απαράμιλλης τελειότητας, την απόδοση και τις επιδόσεις του οποίου κανένας ηλεκτρονικός υπολογιστής δεν έχει καν πλησιάσει μέχρι τις ημέρες μας. Ακόμη, πως κάθε ζωντανό κύτταρο φέρει ενσωματωμένες στον πυρήνα του όλες τις απαραίτητες γενετικές πληροφορίες για τη δημιουργία ενός πλήρους ανθρώπινου οργανισμού και πως, ταυτόχρονα, το γενετικό υλικό αποτελεί ένα σύστημα ελέγχου αποτελούμενο από χιλιάδες επιμέρους υποσυστήματα που υπαγορεύουν τα χαρακτηριστικά και τη συμπεριφορά των έμβιων όντων που θα προκύψουν στις επόμενες γενεές.

2.2 Σύντομη ιστορική αναδρομή

Σύμφωνα με τον Ray Kurzweil [104], η μετα-βιομηχανική εποχή δεν θα τροφοδοτείται από καύσιμα, αλλά από ένα καινοφανές είδος διαθέσιμου πόρου που ονομάζεται Τεχνητή Νοημοσύνη[1] (ΤΝ) (AI: Artificial Intelligence). Ο όρος αναφέρεται στον κλάδο της επιστήμης των υπολογιστών που ασχολείται με τη σχεδίαση, την ανάπτυξη, την υλοποίηση και εφαρμογή συστημάτων τα οποία έχουν ως πρότυπο την ανθρώπινη λειτουργία της νόησης και επιδεικνύουν στοιχειώδη ευφυία υπό την έννοια της εκμάθησης, της προσαρμοστικότητας, της γενίκευσης στην εξαγωγή συμπερασμάτων, καθώς επίσης και στην επίλυση διαφόρων τύπων προβλημάτων. Αποτελεί το σημείο τομής πολλών επιστημονικών πεδίων, από την επιστήμη των υπολογιστών, μέχρι τη νευρολογία, τη φιλοσοφία και τη γλωσσολογία. Η ΤΝ, ανάλογα με το είδος ευφυίας που προσπαθεί να προσομοιάσει, είναι δυνατό να διακριθεί σε δύο μεγάλες κατηγορίες. Στις περιπτώσεις λογισμικού το οποίο αναπτύσσεται για την επίλυση αυστηρά ενός τύπου προβλήματος και μόνο, οπότε αναφερόμαστε στην ασθενή ΤΝ. Αντίθετα, ισχυρή ΤΝ είναι εκείνη που στοχεύει στην προσομοίωση της πραγματικής ευφυίας, με πολλαπλά επίπεδα γενίκευσης και εξαγωγής συμπερασμάτων για μεγάλο εύρος περιπτώσεων.

Κεντρικό δόγμα της ΤΝ είναι ότι ο άνθρωπος, ως νοήμον είδος, είναι σε θέση να κατασκευάσει τεχνητά συστήματα με χαρακτηριστικά νόησης. Βεβαίως, στο σημείο αυτό πρέπει να αναφερθεί ένας πολύ σημαντικός ανασταλτικός παράγοντας: παρά το γεγονός ότι ο Homo sapiens βρίσκεται στην κορυφή όσον αφορά στα επίπεδα νοημοσύνης σε σχέση με όλα τα υπόλοιπα έμβια είδη του πλανήτη, ωστόσο ο εγκέφαλός του δεν είναι σε θέση εκτέλεσης αυτο-διάγνωσης. Δεν έχει δηλαδή τη δυνατότητα αποτύπωσης, χαρτογράφησης και αποκωδικοποίησης των ίδιων των λειτουργιών του και των μεταξύ τους σχέσεων, ώστε να προκύπτει ευανάγνωστο διάγραμμα της συνολικής του λειτουργίας και του τρόπου επίτευξης των στόχων του. Παρά το γεγονός αυτό, οι προσπάθειες ανάπτυξης «αυτόματων», μηχανών δηλαδή που «σκέπτονται», είναι τόσο παλιά, όσο και η ανθρώπινη μυθολογία - μυθιστορία. Οι ελληνικοί μύθοι οι σχετικοί με τη δραστηριότητα του «μηχανικού» των θεών Ήφαιστου, ο οποίος ήταν υπεύθυνος για την κατασκευή μηχανικών υπηρετών, καθώς επίσης και

[1] Ο όρος τείνει να αντικατασταθεί τα τελευταία χρόνια από τον όρο Υπολογιστική Νοημοσύνη (ΥΝ) (CI: Computational Intelligence)

το «χάλκινο ανθρωποειδές» Τάλως, παραπέμπουν απευθείας στην έννοια ευφυών ρομπότ. Κατά τον 5ο π.Χ. αιώνα ο Αριστοτέλης εισάγει τη συλλογιστική λογική και την εφαρμόζει στη διδασκαλία του. Κατά το διάστημα του 13ου έως το 16ο αι. μ.Χ. μια σειρά μηχανικών καινοτομιών κάνουν την εμφάνισή τους, από την επινόηση του Γουτεμβέργιου, έως την επανάσταση της πρώτης χρονομετρικής μηχανής και των ποικίλων πειραματικών μηχανικών κατασκευών που προέκυψαν[2]. Κατά το 17ο αι. λόγιοι όπως ο Descartes, ο Pascal και ο Leibniz επεκτείνουν περαιτέρω τη σχέση ανθρώπου - μηχανής[3], ενώ το 18ο αι. ο Gottlob Frege θεμελιώνει τη σύγχρονη μαθηματική λογική.

Στις αρχές του 20ου αιώνα οι Bertrand Russell και Alfred North Whitehead εκδίδουν το μνημειώδες έργο τους «Principia Mathematica»[4] μέσω του οποίου η θεωρία του Frege επεκτείνεται περαιτέρω, αποκτώντας στέρεες βάσεις, ενώ τριάντα χρόνια αργότερα οι McCulloch και Pitts [122] θέτουν τα θεμέλια για τη θεωρία η οποία διέπει τα Τεχνητά Νευρωνικά Δικτυα[5]. Την ίδια περίοδο ο Turing [178] προτείνει μεθόδους μηχανικού ελέγχου της νοήμονος συμπεριφοράς, ενώ ο Shannon [165] προτυποποιεί τον αλγόριθμο μέσω του οποίου ένα υπολογιστικό σύστημα μπορεί να παίξει σκάκι. Κατά τη δεκαετία 1955 - 1965 ο John McCarthy [121] επινοεί τη γλώσσα προγραμματισμού Lisp και ο Marvin Minsky [124] εισάγει νέους αλγορίθμους αυτοδιδαχής. Κατά την ίδια δεκαετία εργασίες αποδεικνύουν τις δυνατότητες υπολογιστικών συστημάτων έναντι ελέγχων νοημοσύνης (IQ tests) ή κατανόησης της καθομιλουμένης, ενώ δέκα χρόνια αργότερα επινοείται η γλώσσα προγραμματισμού Prolog από τον Alain Colmerauer και ο Minsky [125] προτείνει μεθόδους αναπαράστασης της γνώσης. Τέλος, στα μέσα της δεκαετίας του '80 η μορφοποίηση του αλγορίθμου της οπισθόδρομης μετάδοσης του σφάλματος και των δικτύων Hopfield δίνουν νέα ώθηση στην ανάπτυξη εργαλείων ΤΝ. Ταυτόχρονα, γενετικοί αλγόριθμοι και άλλες συναφείς τεχνολογίες αρχίζουν να κάνουν την εμφάνισή τους ως δράσεις εξελικτικού

[2]Βλέπε για παράδειγμα τον βαδίζοντα λέοντα του Da Vinci κ.λπ.

[3]Σχετικά νωρίς κατά το 17ο αι. ο Descartes προτείνει τη θεωρία ότι τα σώματα των έμβιων όντων είναι σύμπλοκες μηχανές ανώτερης πολυπλοκότητας, μια σκέψη η οποία αποτέλεσε επανάσταση για την επιστημονική κοινότητα της εποχής, προκαλώντας μεγάλο ενδιαφέρον. Επίσης, το 1642 ο Pascal κατασκευάζει και ανακοινώνει την πρώτη αναλογική υπολογιστική μηχανή, ενώ ο Leibniz τη βελτιώνει με δυνατότητες πολλαπλασιασμού και διαίρεσης.

[4]Το έργο εκδόθηκε για πρώτη φορά σε τρεις τόμους κατά τα έτη 1910, 1912 και 1913.

[5]Το άρθρο τους με τίτλο «A Logical Calculus of the Ideas Immanent in Nervous Activity» αποτελεί αναφορά για τους ερευνητές ακόμη και μέχρι τις ημέρες μας.

προγραμματισμού.

Στις ημέρες μας, η ΤΝ περιλαμβάνει ένα πλήθος εργαλείων και εμφανίζει διαφορετικές κατευθύνσεις ως προς την έρευνα. Η εκμάθηση και η απόκτηση γνώσης βασίζεται σε εμπειρικά δεδομένα, ενώ έχουν περιγραφεί μέθοδοι εκμάθησης τόσο υπό όσο και άνευ εποπτείας. Μέχρι τώρα ο άνθρωπος έχει επιτύχει μεγάλα άλματα στον τομέα της ταχύτητας επεξεργασίας και αποθηκευτικής μνήμης των συστημάτων ΤΝ που αναπτύσσει, ωστόσο οι αλγόριθμοι αυτοί είναι προς το παρόν καταφανώς κατώτεροι της δικής του νοημοσύνης. Παρ' ολ' αυτά το κενό όλο και μειώνεται, καθώς οι μέθοδοι εκμάθησης μηχανής όλο και βελτιώνονται. Οπως συμπεραίνει και ο Kurzweil, δεν μπορούμε να ισχυριστούμε το ίδιο για την ανθρώπινη νόηση. Συνεπώς το ερώτημα που εγείρεται είναι τελικά η διαφορά θα μειωθεί; Και ίσως ακόμη περισσότερο προκλητικά, είναι δυνατόν μια οντότητα με νόηση να ξεπεράσει το δημιουργό της; Αν και οι γνώμες διϊστανται, απαντήσεις επί του θέματος δεν είναι δυνατό να δοθούν στον παρόντα χρόνο, ούτε στο πλαίσιο της διατριβής αυτής. Το σίγουρο συμπέρασμα είναι ότι μετά από μακρές περιόδους ύφεσης και ανάπτυξης, η ΤΝ βρίσκεται στην παρούσα φάση σε μια περίοδο έντονης κινητικότητας με εφαρμογή σε πάρα πολλά προβλήματα του πραγματικού κόσμου και η μελέτη της δομής και των προτύπων της προσφέρει ευκαιρίες για πιθανή βελτίωση. Οι διάφορες συνιστώσες της υπολογιστικής νοημοσύνης είναι δυνατό να διακριθούν σε:

- Ταξινομητές ΤΝ. Πρόκειται για μια μεγάλη ομάδα εργαλείων υπολογιστικής νοημοσύνης στην οποία περιλαμβάνονται Τεχνητά Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, δίκτυα Bayes και k-εγγύτερων γειτόνων

- Συστήματα ασαφούς λογικής, τα οποία αποτελούν τεχνικές λήψης απόφασης υπό αβεβαιότητα. Βασίζονται στη δόμηση κανόνων που περιγράφουν μη-αυστηρά διαχωριζόμενες καταστάσεις με βαρύτητα που λαμβάνεται υπόψη κατά περίσταση.

- Εξελικτική υπολογιστική. Περιλαμβάνει γενετικούς αλγορίθμους και εξελικτικό προγραμματισμό, προκύπτει δε από τη μελέτη της εξέλιξης των έμβιων όντων και χαρακτηρίζεται από έννοιες της βιολογίας, όπως προσαρμοστικότητα, κληρονομικότητα, μετάλλαξη και επιλογή. Οι σχετιζόμενοι αλγόριθμοι αναπαριστούν τις λύσεις του εκάστοτε προβλήματος ως ένα σύνολο χρωμο-

σωμάτων που εξελίσσεται μέσω γενεών, η δε εξέλιξη χαρακτηρίζεται από τον Δαρβινιανό κανόνα της επιβίωσης του πλέον προσαρμοσμένου.

Στα επόμενα θα επιχειρηθεί η ανάλυση των ταξινομητών ΤΝ και των γενετικών αλγορίθμων που αποτελούν αντικείμενο της παρούσας διατριβής.

2.3 Τεχνητά Νευρωνικά Δίκτυα

Η ανάπτυξη των Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ), έχει τη βάση της στις μελέτες τις σχετικές με τη λειτουργία του ανθρώπινου νευρικού συστήματος και του εγκεφάλου ειδικότερα. Από την αρχή έγινε φανερό ότι ο εγκέφαλος είναι ένα πολύπλοκο, μη γραμμικό σύστημα παράλληλης επεξεργασίας πληροφοριών, το οποίο εκτελεί τις υπολογιστικές του διεργασίες με εντελώς διαφορετικό τρόπο από αυτόν που χρησιμοποιεί ένας τυπικός ηλεκτρονικός υπολογιστής. Παρότι ο τελευταίος παρουσιάζει απαράμιλλη ταχύτητα όσον αφορά στην εκτέλεση απλών μαθηματικών πράξεων, ο εγκέφαλος των έμβιων όντων παρουσιάζει το χαρακτηριστικό να μπορεί να οργανώνει κατά τέτοιον τρόπο τα δομικά λειτουργικά του στοιχεία, γνωστά ως νευρώνες, ώστε να επιτυγχάνει θαυμαστά αποτελέσματα στους τομείς της αναγνώρισης προτύπων, της αντίληψης του περιβάλλοντος χώρου, της αναπαραγωγής της συσσωρευμένης γνώσης και του ελέγχου της κίνησης, σε ταχύτητες απλησίαστες ακόμη και για τους ισχυρότερους υπολογιστές.

2.3.1 Ορισμοί και χαρακτηριστικά

Ένα Τεχνητό Νευρωνικό Δίκτυο είναι στην πιο γενικευμένη του έννοια μια κατασκευή σχεδιασμένη έτσι ώστε να προτυποποιεί τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος εκτελεί ορισμένες τουλάχιστον από τις λειτουργίες του. Το δίκτυο συνήθως υλοποιείται με τη χρήση ηλεκτρονικών μερών, ή προσομοιάζεται μέσω λογισμικού σε ηλεκτρονικό υπολογιστή. Σύμφωνα με τον Haykin [75],

«…ένα νευρωνικό δίκτυο είναι ένας μαζικά διατεταγμένος επεξεργαστής παράλληλης λειτουργίας, κατασκευασμένος από απλές επεξεργαστικές μονάδες που ενέχει την ιδιότητα αποθήκευσης εμπειρικής μνήμης, την οποία δύναται να επαναφέρει προς χρήση …».

Το όλο σύστημα προσομοιάζει το βιολογικό εγκέφαλο κατά δύο έννοιες:

1. η γνώση αποκτάται από τον περιβάλλοντα χώρο μέσω μιας διαδικασίας εκμάθησης

2. μέσο αποθήκευσης της αποκτώμενης γνώσης αποτελούν οι διασυνδέσεις μεταξύ των νευρώνων, γνωστές ως συναπτικά βάρη.

Η διαδικασία που οδηγεί στην εκμάθηση βασίζεται πάντα στον λεγόμενο αλγόριθμο εκμάθησης, σκοπός του οποίου είναι η μεταβολή των συναπτικών βαρών του νευρωνικού δικτύου με δομημένο τρόπο, ούτως ώστε να επιτευχθεί το επιθυμητό αποτέλεσμα. Η μεταβολή των συναπτικών βαρών αποτελεί την παραδοσιακή πλέον μέθοδο εκπαίδευσης ενός νευρωνικού δικτύου, αλλά είναι πιθανό να χρησιμοποιηθούν και άλλες μέθοδοι, όπως για παράδειγμα η αυτόματη μεταβολή της τοπολογίας από το ίδιο το δίκτυο, μέσω του οποίου προσομοιάζεται η διαδικασία θανάτου και ανάπτυξης νέων νευρώνων στον ανθρώπινο εγκέφαλο.

Ένα νευρωνικό δίκτυο αποτελείται από μαζικά αλληλοσυνδεδεμένους τεχνητούς νευρώνες οι οποίοι είναι οργανωμένοι σε επάλληλα στρώματα ή επίπεδα (layers). Συνήθως υπάρχει ένα επίπεδο εισόδου που μπορεί να περιλαμβάνει ή όχι επεξεργαστικές δυνατότητες και το οποίο μεταφέρει το διάνυσμα εισόδου σε ένα ή περισσότερα κρυφά επίπεδα, όπου γίνεται η μη γραμμική επεξεργασία των πληροφοριών. Στη συνέχεια η πληροφορία διατίθεται στο τελικό επίπεδο εξόδου, το οποίο ενέχει επεξεργαστικής δυνατότητας και παρέχει τα αποτελέσματα του δικτύου στο εξωτερικό του περιβάλλον. Ο αριθμός των τεχνητών νευρώνων που αποτελούν ένα νευρωνικό δίκτυο κυμαίνεται από μερικούς έως κάποιες χιλιάδες, σε συνάρτηση και με το πρόβλημα που πρέπει να επιλυθεί. Τέλος, τα νευρωνικά δίκτυα αναφέρονται στη βιβλιογραφία επίσης ως νευρο-υπολογιστές (neurocomputers), δίκτυα σύνδεσης (connectionist networks) ή παράλληλα κατανεμημένοι επεξεργαστές (parallel distributed processors). Στη διατριβή αυτή χρησιμοποιείται κυρίως ο όρος «τεχνητά νευρωνικά δίκτυα».

Είναι προφανές ότι τα νευρωνικά δίκτυα οφείλουν την επεξεργαστική τους ισχύ καταρχήν στη μαζική και παράλληλα δομημένη αρχιτεκτονική τους, καθώς επίσης και στη δυνατότητά τους να εκπαιδεύονται και να εφαρμόζουν αυτή τη γνώση σε μια διαδικασία γενίκευσης, ούτως ώστε να την εφαρμόζουν σε παρόμοια προβλήματα. Η χρήση των νευρωνικών δικτύων παρουσιάζει τις εξής χρήσιμες ιδιότητες και χαρακτηριστικά κατά Haykin [75]:

- Μη γραμμική λειτουργία. Ένα νευρωνικό δίκτυο που αποτελείται από μη γραμμικούς αλληλοσυνδεδεμένους νευρώνες, είναι από κατασκευής μη γραμμικό, ιδιότητα διανεμημένη εξ' ολοκλήρου στο δίκτυο. Πρόκειται για μια πολύ σπουδαία ιδιότητα, ειδικά στις περιπτώσεις κατά τις οποίες ο μηχανισμός που διέπει το φαινόμενο που θέλουμε να μελετήσουμε ή το πρόβλημα που θέλουμε να επιλύσουμε μέσω του νευρωνικού δικτύου είναι μη γραμμικός.

- Αντιστοιχία εισόδου - εξόδου. Τα νευρωνικά δίκτυα έχουν τη δυνατότητα να αντιστοιχούν σε ένα διάνυσμα εισόδου, ένα επιθυμητό διάνυσμα εξόδου. Η αντιστοιχία αυτή επιτυγχάνεται μέσω μιας διαδικασίας εκπαίδευσης που στοχεύει στην ελαχιστοποίηση του σφάλματος που ορίζεται εδώ ως η διαφορά μεταξύ του διανύσματος εισόδου και του επιθυμητού διανύσματος. Η δυνατότητα αυτή χρησιμοποιείται κατά κόρον σε περιπτώσεις ταξινόμησης προτύπων (pattern classification), όπου το ζητούμενο είναι να αποδοθεί μια προαποφασισμένη κατηγορία σε ένα σήμα εισόδου το οποίο αντιπροσωπεύει ένα φυσικό αντικείμενο ή γεγονός. Αυτή η ιδιότητα είναι ιδιαίτερα χρήσιμη στις περισσότερες περιβαλλοντικές εφαρμογές, όπως θα δούμε στα επόμενα κεφάλαια.

- Προσαρμοστικότητα. Τα νευρωνικά δίκτυα έχουν ενσωματωμένη την ιδιότητα να προσαρμόζουν τα συναπτικά βάρη μεταξύ των νευρώνων τους σε μεταβολές του περιβάλλοντος. Ένα δίκτυο εκπαιδευμένο να λειτουργεί σε ένα συγκεκριμένο περιβάλλον, μπορεί πολύ εύκολα να επανεκπαιδευθεί, ώστε να αποδίδει το ίδιο καλά σε μικρές αλλαγές στις λειτουργικές συνθήκες του περιβάλλοντός του. Σε άλλες πάλι περιπτώσεις, ένα νευρωνικό δίκτυο είναι δυνατόν να σχεδιαστεί κατά τέτοιο τρόπο ώστε να μεταβάλει αυτόματα τα συναπτικά του βάρη σε πραγματικό χρόνο, ώστε να λειτουργεί σε ένα μη στατικό περιβάλλον, ένα περιβάλλον δηλαδή το οποίο μεταβάλλεται με την πάροδο του χρόνου. Τέτοια περιβάλλοντα απαντώνται σε πολλές εφαρμογές που εμπλέκουν αναγνώριση και ταξινόμηση προτύπων, επεξεργασία σημάτων και εφαρμογές ελέγχου, στα οποία η συμβολή των νευρωνικών δικτύων είναι σημαντική όπως επίσης θα δούμε στα επόμενα κεφάλαια.

- Τεκμηρίωση των αποκρίσεων. Ιδιαίτερα στις περιπτώσεις της αναγνώρισης

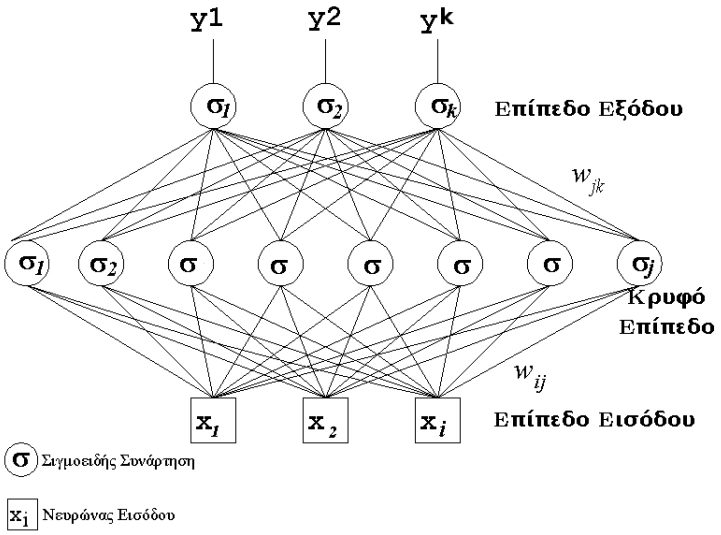
και ταξινόμησης προτύπων, ένα νευρωνικό δίκτυο είναι δυνατόν να κατασκευαστεί κατά τέτοιο τρόπο ώστε όχι μόνο να προτείνει την επιλογή κάποιου προτύπου, αλλά επίσης και να παρέχει πληροφορίες σχετικές με τα επίπεδα ανοχής αυτής της απόφασης. Η ιδιότητα αυτή μπορεί να χρησιμοποιηθεί για την απόρριψη αμφιλεγόμενων ή διφορούμενων προτύπων, βελτιώνοντας σημαντικά την απόδοση του δικτύου.

- Ανοχή λειτουργικών σφαλμάτων. Τα νευρωνικά δίκτυα, ειδικά αυτά που υλοποιούνται με βάση μηχανικά μέρη, παρουσιάζουν την πολύ σημαντική ιδιότητα της βαθμιαίας υποβάθμισης της απόδοσής τους σε αντιδιαστολή με την απότομη παύση λειτουργίας, όταν συμβαίνουν κατασκευαστικά ή λειτουργικά σφάλματα. Στις περιπτώσεις κατά τις οποίες ένας νευρώνας ή οι όμοροί του αναστείλουν τη λειτουργία τους εξαιτίας κατασκευαστικών ή λειτουργικών ελαττωμάτων, η ζημιά θα πρέπει να είναι αρκετά εκτεταμένη πριν το δίκτυο αναστείλει παντελώς τη λειτουργία του. Αυτό οφείλεται κατά ένα μέρος στην παράλληλη επεξεργαστική λειτουργία του νευρωνικού δικτύου και κατά ένα άλλο στην κατανεμημένη αποθηκευμένη στο δίκτυο γνώση.

- Αναλυτική και σχεδιαστική ομοιομορφία. Τα νευρωνικά δίκτυα παρουσιάζουν ομογενοποίηση σε ότι αφορά στη λειτουργία τους ως επεξεργαστές πληροφορίας. Οσον αφορά στην εφαρμογή τέτοιων δικτύων σε διάφορους τομείς, η χρησιμοποιούμενη ορολογία είναι η ίδια. Για παράδειγμα, οι νευρώνες κατά τον ένα ή τον άλλο τρόπο αντιπροσωπεύουν ένα συστατικό κοινό σε όλα τα νευρωνικά δίκτυα. Η ομοιομορφία αυτή έχει ως αποτέλεσμα τη δυνατότητα εφαρμογής θεωριών ή εκπαιδευτικών αλγορίθμων που είναι κατασκευασμένοι για διαφορετικές εφαρμογές.

2.3.2 Αρχιτεκτονική ΤΝΔ

Στη γενικευμένη τους μορφή τα ΤΝΔ αποτελούνται από ένα σύνολο μαζικά αλληλοσυνδεδεμένων και αλληλοεξαρτώμενων επεξεργαστικών μονάδων οι οποίες, κατ' αντιστοιχία με τα βιολογικά δίκτυα, ονομάζονται νευρώνες. Οι κόμβοι αυτοί επικοινωνούν μεταξύ τους αποστέλλοντας σήματα ο ένας στον άλλο μέσω ενός δικτύου πολυάριθμων σταθμισμένων συνδέσεων. Στη βασική αρχιτεκτονική ενός τυπικού δικτύου περιλαμβάνονται:



Σχήμα 2.1: *Γραφική αναπαράσταση αρχιτεκτονικής πολυεπίπεδου perceptron τριών επιπέδων.*

- Ένα σύνολο επεξεργαστικών μονάδων που ονομάζονται νευρώνες.
- Η κατάσταση ενεργοποίησης για κάθε νευρώνα, ισοδύναμη με την έξοδό του
- Συνδέσεις μεταξύ των νευρώνων. Τυπικά, κάθε σύνδεση αντιπροσωπεύεται από ένα στάθμισμα το οποίο συμβολίζει την επίδραση της εξόδου ενός νευρώνα στην είσοδο ενός άλλου.
- Ο κανόνας δρομολόγησης (propagation rule) μέσω του οποίου υπολογίζεται η αποτελεσματική είσοδος σε ένα νευρώνα από τις επί μέρους εισόδους του.
- Η συνάρτηση ενεργοποίησης μέσω της οποίας υπολογίζονται τα νέα επίπεδα ενεργοποίησης με βάση την αποτελεσματική είσοδο και την τρέχουσα ενεργοποίηση.
- Η εξωτερική επίδραση, καλούμενη πόλωση (bias, offset, threshold) για κάθε νευρώνα.

- Ο κανόνας εκπαίδευσης, ο οποίος ουσιαστικά συνιστά τη μέθοδο για τη συλλογή, εκμάθηση και αποθήκευση των γνώσεων. Η εκπαίδευση των ΤΝΔ, διακρίνεται σε εκπαίδευση υπο- και άνευ- εποπτείας, ανάλογα με το εάν τα δεδομένα εξόδου γίνονται γνωστά κατά τη διάρκεια της εκμάθησης ή όχι.

- Το περιβάλλον λειτουργίας του συστήματος μέσα στο οποίο αυτό αναμένεται να παρέχονται σήματα εισόδου - εξόδου και, αν είναι απαραίτητο, σήματα σφάλματος.

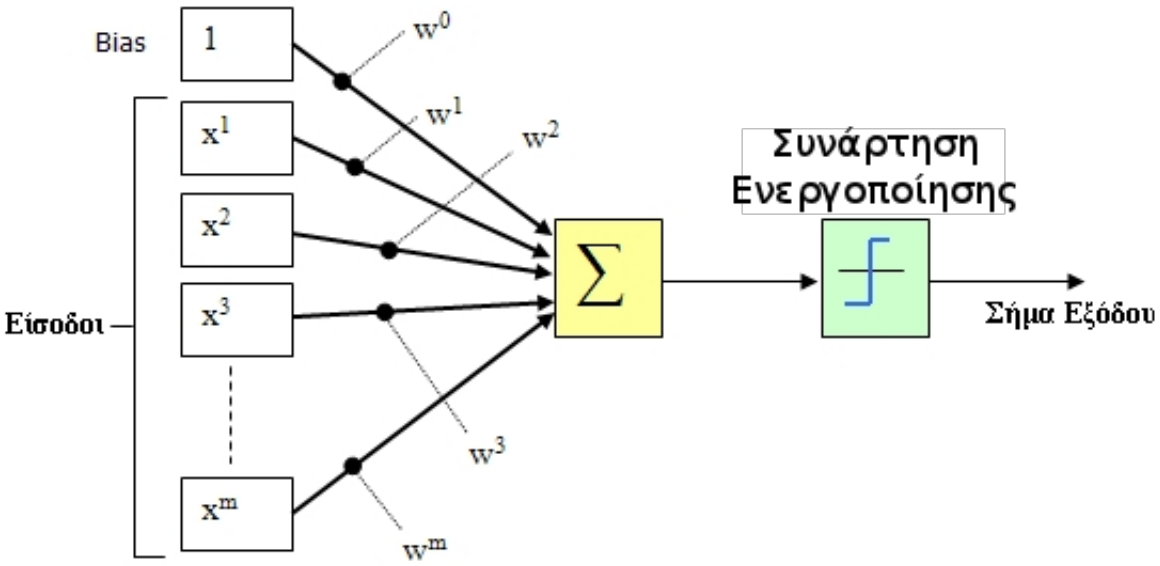
Κάθε επεξεργαστική μονάδα εκτελεί μια σχετικά απλή λειτουργία (Σχήμα 2.2). Είναι εξειδικευμένη στο να λαμβάνει σήματα εισόδου από τις όμορές της μονάδες - ή από εξωτερικές πηγές - και να τα χρησιμοποιεί για να υπολογίζει ένα σήμα εξόδου το οποίο δρομολογείται σε άλλες μονάδες. Μια δεύτερη επίσης λειτουργία κάθε επεξεργαστικής μονάδας αποτελεί η προσαρμογή των συναπτικών βαρών. Ο σχεδιασμός του συστήματος είναι τέτοιος ώστε να εξασφαλίζεται παράλληλη επεξεργασία, με την έννοια ότι πολλές επεξεργαστικές μονάδες μπορούν να εκτελούν τους υπολογισμούς τους κατά την ίδια χρονική περίοδο. Οι επεξεργαστικές μονάδες των νευρωνικών δικτύων είναι κατανεμημένες κατά ομάδες σε

α. *μονάδες εισόδου,* μέσω των οποίων επιτυγχάνεται η λήψη του σήματος από το εξωτερικό περιβάλλον του δικτύου,

β. *μονάδες εξόδου,* οι οποίες χρησιμεύουν για να αποδίδουν το τελικό αποτέλεσμα των επεξεργασιών του δικτύου στο εξωτερικό του περιβάλλον και

γ. *κρυφές μονάδες,* οι οποίες παραμένουν αθέατες από τον εξωτερικό παρατηρητή και των οποίων τα σήματα εισόδου και εξόδου παραμένουν στο εσωτερικό του δικτύου.

Κατά τη διάρκεια της λειτουργίας τους, οι μονάδες επεξεργασίας είναι δυνατόν να ενημερώνονται σύγχρονα ή ασύγχρονα, ανάλογα με τον τρόπο ενημέρωσης της δραστηριοποίησής τους. Στη σύγχρονη ενημέρωση, όλες οι μονάδες ενημερώνονται ταυτόχρονα, ενώ στην ασύγχρονη, κάθε μονάδα έχει μια συγκεκριμένη πιθανότητα να ενημερωθεί σε συγκεκριμένο χρόνο, και μόνο μια μονάδα επεξεργασίας είναι σε θέση να το κάνει αυτό στη μονάδα του χρόνου.



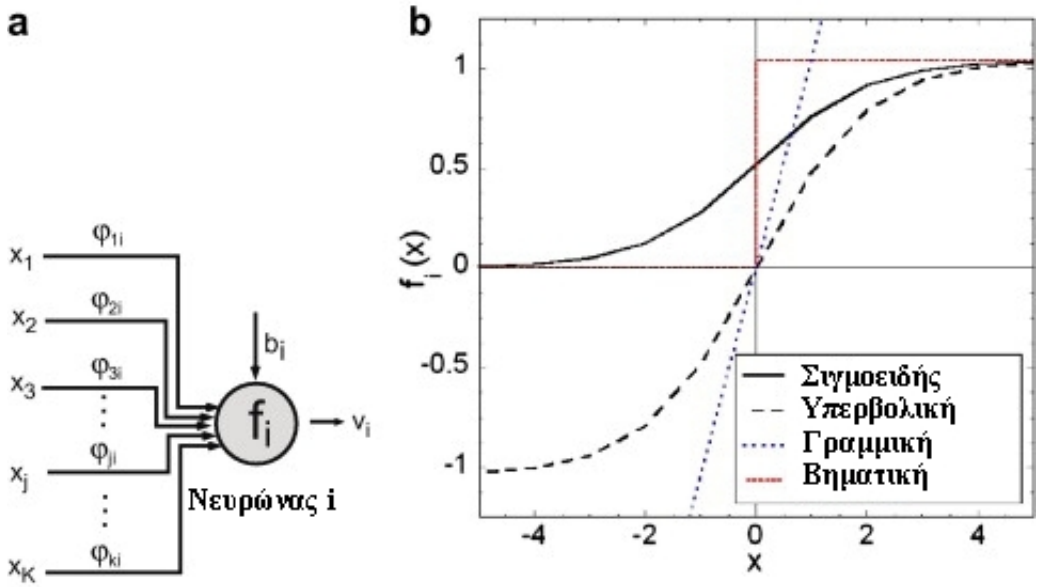
Σχήμα 2.2: *Διαδικασία ενεργοποίησης μεμονωμένου νευρώνα.*

Στις περισσότερες περιπτώσεις, κάθε μονάδα επεξεργασίας συμβάλλει αθροιστικά με την έξοδό της στην είσοδο κάθε όμορής της συνδεδεμένης μονάδας. Η συνολική είσοδος ενός νευρώνα, με άλλα λόγια το φέρον τοπικό πεδίο (induced local field) για μια συγκεκριμένη χρονική στιγμή - ή για μια συγκεκριμένη επανάληψη - είναι το σταθμισμένο άθροισμα όλων των ξεχωριστών εξόδων των συνδεδεμένων με αυτόν το νευρώνα μονάδων, προσαυξημένο κατά έναν όρο που αποτελεί την πόλωση. Η πόλωση παίζει το ρόλο του κατωφλίου για να αποφασισθεί η ενεργοποίηση ή όχι του νευρώνα

Επίσης χρησιμοποιείται ένας κανόνας, μέσω του οποίου υπολογίζεται η επίδραση της συνολικής εισόδου στη δραστηριοποίηση ενός νευρώνα. Για το σκοπό αυτό χρησιμοποιείται μια συνάρτηση ή οποία παράγει μια νέα τιμή της δραστηριοποίησης του νευρώνα με βάση τη συνολική είσοδο και την τρέχουσα ενεργοποίησή του την προηγούμενη χρονική στιγμή (ή επανάληψη)

Συνεπώς, κατά τη λειτουργία ενός νευρωνικού δικτύου, μπορούμε να διακρίνουμε τρία βασικά στοιχεία στο επίπεδο του νευρώνα.

α. *Πίνακας συναπτικών βαρών* . Πρόκειται για ένα σύνολο τιμών μέσω των οποίων συνδέονται οι νευρώνες μεταξύ τους κατ' αντιστοιχία με τις βιολογικές εγκε-



Σχήμα 2.3: *Συναρτήσεις δραστηριοποίησης ANN.*

φαλικές συνάψεις. Κάθε τέτοια τιμή πολλαπλασιάζεται υπό προϋποθέσεις με κάποιο σήμα στην είσοδο μιας σύναψης μεταξύ δύο νευρώνων. Αντίθετα με τη βιολογική σύναψη, το συναπτικό βάρος ενός τεχνητού νευρωνικού δικτύου είναι δυνατό να περιλαμβάνει τόσο θετικές, όσο και αρνητικές τιμές.

β. *Σύστημα άθροισης των σημάτων εισόδου,* τα οποία είναι σταθμισμένα με τα αντίστοιχα συναπτικά βάρη και τέλος

γ. *Συναρτήσεις δραστηριοποίησης,* για τον περιορισμό της έντασης της εξόδου του νευρώνα.

Οι συναρτήσεις δραστηριοποίησης που χρησιμοποιούνται είναι διαφόρων ειδών και κυμαίνονται από την απλή βηματική συνάρτηση, έως τη σιγμοειδή συνάρτηση, ανάλογα με το προς επίλυση πρόβλημα. Η πιο συχνά χρησιμοποιούμενη συνάρτηση ενεργοποίησης είναι η σιγμοειδής και ειδικότερα η λογιστική συνάρτηση του τύπου

$$\phi(u) = [1 + exp(-\alpha u)]^{-1} \tag{2.3.1}$$

όπου α είναι η παράμετρος κλίσης της σιγμοειδούς καμπύλης. Μεταβάλλοντας την παράμετρο αυτή μπορούμε να επιτύχουμε σιγμοειδείς καμπύλες διαφορετικής μορ-

φής. Στην οριακή περίπτωση, καθώς η κλίση τείνει στο άπειρο, η καμπύλη τείνει να εξομοιωθεί με τη βηματική συνάρτηση κατωφλίου (Εικ. 2.3). Σε αντίθεση με τη συνάρτηση κατωφλίου που λαμβάνει τις τιμές 0 ή 1, η σιγμοειδής μπορεί να λάβει ένα συνεχές πλήθος τιμών στο διάστημα [0,1], είναι δε δυνατόν να διαφοροποιηθεί, γεγονός που την καθιστά ιδιαίτερα σημαντική για τη θεωρία των νευρωνικών δικτύων.

2.3.3 Πολυεπίπεδα Προσθιόδρομα Perceptrons

Κεντρική θέση στη λύση του προβλήματος της αποτύπωσης μη γραμμικών σχέσεων δεδομένων από υπολογιστικές μηχανές παραλληλίας κατέχει η θεώρηση ότι η μεταβολή των συναπτικών βαρών των νευρώνων ενός επιπέδου του δικτύου θα έπρεπε να εξαρτάται από το σφάλμα που παρουσιάζεται στους νευρώνες του αμέσως επόμενου επιπέδου. Ο τρόπος με τον οποίο οι νευρώνες ενός επιπέδου ενημερώνονται για το σφάλμα που παρουσιάζεται στους νευρώνες του αμέσως επόμενου επιπέδου, ουσιαστικά δηλαδή η «προς τα πίσω» μετακίνηση του σφάλματος, απέδωσε την ονομασία «οπισθόδρομη διάδοση του σφάλματος» (back propagation of error) στον αλγόριθμο. Στην τυπική του μορφή αυτό το πρότυπο δικτύου αποτελείται από επίπεδα επεξεργαστικών μονάδων δια μέσου των οποίων ρέει το σήμα εισόδου, ένα επίπεδο τη φορά, με μονοκατευθυντικό τρόπο. Υπάρχει το επίπεδο εισόδου το οποίο περιλαμβάνει στη γενική του μορφή νευρώνες εισόδου χωρίς επεξεργαστικές δυνατότητες. Στη συνέχεια το σήμα εισόδου διαπερνά ένα ή περισσότερα ενδιάμεσα κρυφά επίπεδα, όπου υφίσταται επεξεργασία πριν μετατραπεί σε σήμα εξόδου και περάσει από το επίπεδο εξόδου προς το εξωτερικό του δικτύου περιβάλλον. Ο όρος «πολυεπίπεδα perceptrons» οφείλεται στην ύπαρξη περισσότερων του ενός επιπέδων σε αυτά τα νευρωνικά δίκτυα, τα οποία έχουν εφαρμοσθεί με επιτυχία για την επίλυση μερικών πολύ δύσκολων προβλημάτων - για το λόγο αυτό το πολυπεπίπεδο perceptron (MLP: Multi-Layer Perceptron) αποτελεί ένα από τους διαδεδομένους τύπους νευρωνικού δικτύου. Στη μεγάλη επιτυχία του προτύπου αυτού έχει χωρίς αμφιβολία συμβάλλει ο αλγόριθμος εκπαίδευσής του, ένας πολύ αποτελεσματικός τρόπος εκμάθησης που προσδίδει στο δίκτυο μεγάλη ευελιξία, σταθερότητα και υψηλό βαθμό απόδοσης και ακρίβειας, ο οποίος βασίζεται στον κανόνα μάθησης με την ονομασία «διόρθωση σφάλματος» (error correction learning rule) [75]. Ο αλγόριθμος ανήκει στην κατηγορία των αλγορίθμων εκπαίδευσης με εποπτεία, με

την έννοια ότι κατά τη διάρκεια της εκπαιδευτικής διαδικασίας παρουσιάζονται στο δίκτυο τόσο τα διανύσματα εισόδου, όσο και τα επιθυμητά διανύσματα εξόδου, ενώ η ελαχιστοποίηση του σφάλματος του δικτύου βασίζεται στον υπολογισμό του σφάλματος εξόδου ως η διαφοροποίηση της απόκρισης του συστήματος από την επιθυμητή και τη χρήση αυτής της διαφοράς ως «παραδειγματισμό» του δικτύου.

Κάθε νευρώνας ή επεξεργαστική μονάδα λαμβάνει σταθμισμένα σήματα εισόδου τα οποία αθροίζει. Το άθροισμα αυτό συγκρίνεται με μια προκαθορισμένη τιμή κατωφλίου για να αποφασισθεί η ενεργοποίηση του νευρώνα. Ο ενεργοποιημένος νευρώνας αποδίδει ένα σήμα εξόδου που με τη σειρά του αποτελεί την είσοδο στον νευρώνα κάποιου άλλου επιπέδου για να συνεχισθεί η διαδικασία, ή αποδίδει ένα αποτέλεσμα προς το εξωτερικό περιβάλλον, εφόσον ανήκει στο επίπεδο εξόδου του δικτύου. Η τοπολογία του δικτύου είναι τέτοια ώστε η ροή της πληροφορίας να είναι μονοκατευθυντική ανάμεσα στα διάφορα επίπεδα πριν το αποτέλεσμα των υπολογισμών αποδοθεί από το τελικό επίπεδο εξόδου. Η παρουσία του κρυφού - ή των κρυφών - επιπέδου(ων) αποτελεί μια «ενδυνάμωση» του δικτύου με την έννοια ότι επιτρέπει την αποκάλυψη μη γραμμικών σχέσεων υψηλού βαθμού διάστασης μεταξύ των διανυσμάτων των σχετικών με τα δεδομένα εισόδου. Ετσι, τόσο ο αριθμός των επιπέδων όσο και ο αριθμός των νευρώνων που περιέχονται σε κάποιο επίπεδο, σχετίζονται και καθορίζουν την πολυπλοκότητα της υποκείμενης συνάρτησης.

2.3.4 Αρετές και Περιορισμοί του Αλγορίθμου

Ο αλγόριθμος οπισθόδρομης μετάδοσης του σφάλματος αποτελεί τον πλέον δημοφιλή τρόπο εκπαίδευσης των πολυεπίπεδων προσθιόδρομων perceptrons λόγω του ότι προσφέρει ευκολία στη χρήση και στον υπολογισμό των τοπικών μεταβλητών αφενός, αλλά και διότι επιτρέπει την εκτέλεση στοχαστικής βαθμωτής κατάβασης στο χώρο των βαρών με αποτέλεσμα να επιτυγχάνεται προοδευτική ενημέρωσή τους, καθώς τα πρότυπα εισόδου διαδέχονται το ένα το άλλο. Πρόκειται για ένα παράδειγμα μαζικής αλληλοσύνδεσης που βασίζεται σε τοπικούς υπολογισμούς για την αποκάλυψη της λύσης σε ένα πρόβλημα. Ο περιορισμός αυτός αναφέρεται ως «τοπικός»' με την έννοια ότι συμβαίνει στο επίπεδο της επεξεργαστικής μονάδας του δικτύου και επηρεάζεται αυστηρά μόνο από τις επεξεργαστικές μονάδες που βρίσκονται σε φυσική επαφή με αυτήν. Η «τοπικότητα» αυτή αφενός προωθεί τη χρήση παράλληλης επεξεργασίας κατά τη λειτουργία του νευρωνικού δικτύου, αφε-

τέρου επιτρέπει τη δημιουργία συστημάτων ανεκτικών σε σφάλματα (fault-tolerant systems).

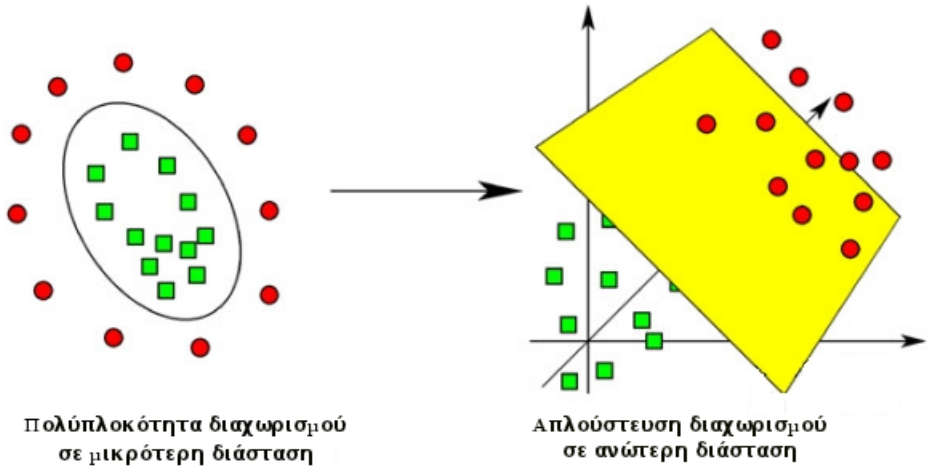
Αντίθετα, η λειτουργία τους δε φαίνεται βιολογικά αληθοφανής. Συγκεκριμένο συναπτικό βάρος μεταξύ δύο νευρώνων ενός δικτύου μπορεί να μεταπίπτει από διεγερτικό σε ανασταλτικό κατά την πορεία των εκπαιδευτικών επαναλήψεων, ενώ αντίθετα τα βιολογικά συναπτικά βάρη στις περισσότερες περιπτώσεις εμφανίζονται αποκλειστικά και μόνο με μια από τις δυο αυτές μορφές. Επίσης, τα τεχνητά νευρωνικά δίκτυα δεν παρουσιάζουν το φαινόμενο της σφαιρικής επικοινωνίας[6] η οποία έχει σαν αποτέλεσμα τη μάθηση, την προσοχή και τη διέγερση του ανθρώπινου εγκεφάλου. Τέλος, ο αλγόριθμος εκπαίδευσης του τεχνητού μοντέλου εμπεριέχει εκπαίδευση με τη μορφή μιας υπό εποπτεία διαδικασίας, η οποία δεν έχει σχέση με τις βιολογικές διεργασίες σε όλες τους τις εκφάνσεις.

Παρ' όλες αυτές τις βιολογικές αναφορικές ανακρίβειες, το γεγονός παραμένει ότι ο αλγόριθμος οπισθόδρομης μετάδοσης του σφάλματος για τα πολυεπίπεδα προσθιόδρομα perceptrons έχει επιτρέψει τη δημιουργία συστημάτων με πολυποίκιλες εφαρμογές σε μεγάλο εύρος διαφορετικών πεδίων, συμπεριλαμβανομένης και της εξομοίωσης νευρο-βιολογικών φαινομένων.

2.4 Μηχανές Διανυσμάτων Υποστήριξης

Στη στατιστική και την επιστήμη των υπολογιστών ως Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) νοούνται οι μη-πιθανοτικοί (non probabilistic) δυαδικοί γραμμικοί ταξινομητές (binary linear classifiers) στους οποίους ενσωματώνονται μέθοδοι εκπαίδευσης υπο εποπτεία και χρησιμοποιούνται στην ανάλυση δεδομένων και στην αναγνώριση προτύπων, τόσο σε προβλήματα ταξινόμησης, όσο και σε προβλήματα πρόβλεψης. Πολύ συχνά, τα δεδομένα του προβλήματος ορίζονται σε χώρο πεπερασμένης διάστασης (finite dimensional space) όπου ο γραμμικός τους διαχωρισμός είναι δύσκολος έως πολλές φορές αδύνατος. Στις περιπτώσεις αυτές, τα πρότυπα ΜΔΥ αναπαριστούν τα αρχικά δείγματα σε χώρους ανώτερου βαθμού διάστασης, όπου ο γραμμικός διαχωρισμός είναι ευκολότερος, επιτυγχάνεται δε μέσω υπερεπιφανειών (hyperplanes), η θέση των οποίων καθορίζεται κάθε φορά από το εκπαι-

[6] Η *σφαιρική επικοινωνία* εμφανίζεται στα βιολογικά δίκτυα με τη μορφή της επίδρασης συγκεκριμένων ορμονών και συνδυασμών τους



Σχήμα 2.4: *Απλούστευση της διαδικασίας διαχωρισμού κλάσεων σε ανώτερη διάσταση.*

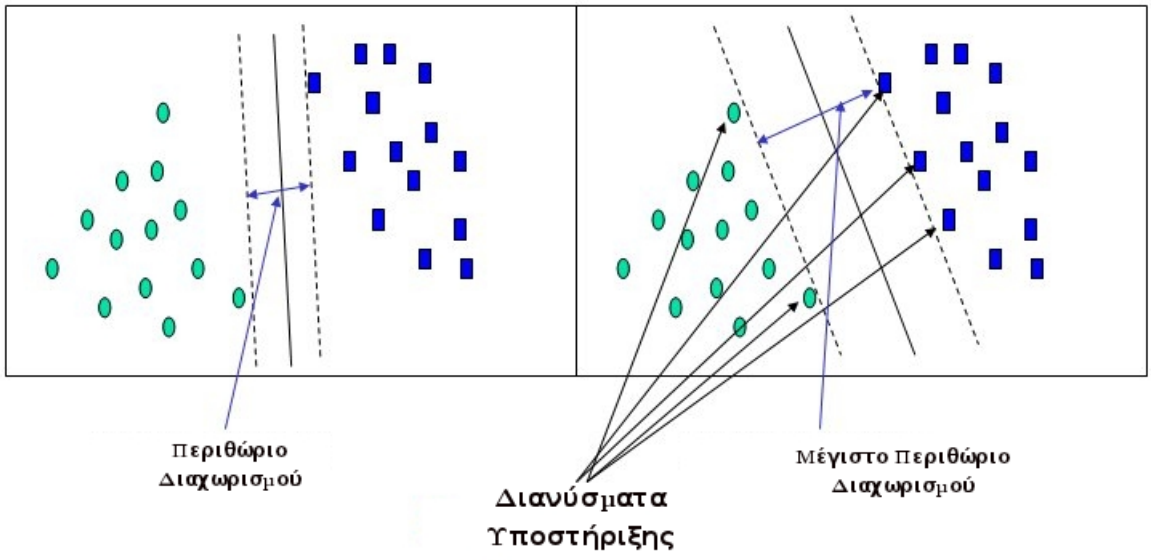
δευτικό σύνολο δεδομένων.

Ο χώρος που ορίζεται από τις υπερ-επιφάνειες αυτές ονομάζεται ωφέλιμο περιθώριο (functional margin), και έχει μέγεθος το οποίο βαίνει αντιστρόφως ανάλογα του μεγέθους του σφάλματος γενίκευσης του ταξινομητή. Σύμφωνα με τη θεωρία μεγίστου περιθωρίου ταξινόμησης (maximum margin classification theory) [21] ισχυρότερη γενίκευση στην ταξινόμηση επιτυγχάνεται από τις υπερ-επιφάνειες που ορίζονται από τα εγγύτερα σημεία του συνόλου δεδομένων, δημιουργώντας με τον τρόπο αυτό το μέγιστο χώρο μεταξύ των κλάσεων.

Η απαίτηση για μέγιστο περιθώριο περιλαμβάνει δύο πλεονεκτήματα. Αφενός προτείνει μοναδιαία λύση για προβλήματα που είναι δυνατό να ταξινομηθούν γραμμικά και αφετέρου προσφέρει υψηλότερη ανεκτικότητα έναντι του θορύβου που πιθανόν ενέχεται στα αρχικά δεδομένα. Βασιζόμενα στην αρχή της δομημένης ελαχιστοποίησης του κινδύνου (SRM: Structured Risk Minimization principle), τα πρότυπα ΜΔΥ σχεδιάζονται με βάση την ελαχιστοποίηση του άνω ορίου του σφάλματος γενίκευσης.

2.4.1 Διανύσματα Υποστήριξης

Έστω ένα εκπαιδευτικό σύνολο δεδομένων $\mathcal{D}$ του τύπου



Σχήμα 2.5: *Η ανάλυση ΜΔΥ επιλέγει τη μοναδική επιφάνεια διαχωρισμού που έχει τέτοια κλίση, ώστε να μεγιστοποιείται το περιθώριο διαχωρισμού μεταξύ των κλάσεων.*

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p, \;\; y_i \in \{-1, \; 1\}\}_{i=1}^{n}$$

αποτελούμενο από (α) υποδείγματα[7] $\mathbf{x}_i$ και (β) κλάσεις[8] y_i. Τότε το δυαδικό πρόβλημα ταξινόμησης συνίσταται στην εύρεση επιφανειών που παρουσιάζουν το μέγιστο μεταξύ τους περιθώριο κατά το διαχωρισμό των σημείων που ανήκουν στην κλάση $y_i = 1$ από αυτά που ανήκουν στην $y_i = -1$. Μια οποιαδήποτε επιφάνεια είναι δυνατό να γραφεί ως το σύνολο των $\mathbf{x}$ που ικανοποιούν τη σχέση:

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \tag{2.4.1}$$

όπου το σύμβολο $\cdot$ δηλώνει εσωτερικό γινόμενο και $\mathbf{w}$ το κάθετο στην επιφάνεια διάνυσμα.

Η παράμετρος $\frac{b}{||w||}$ καθορίζει τη μετάθεση της επιφάνειας από την αρχή των αξόνων. Το ζητούμενο είναι η κατάλληλη επιλογή των $\mathbf{w}$ και b ώστε να μεγιστοποιηθεί το όριο, η απόσταση δηλαδή μεταξύ των παράλληλων διαχωριστικών επιφα-

[7] Οι όροι που απαντώνται στη σύγχρονη βιβλιογραφία είναι features (χαρακτηριστικά), patterns (πρότυπα) ή examples (παραδείγματα)

[8] επίσης καλούμενων ετικετών (labels)

νειών, οι οποίες είναι δυνατό να γραφούν ως:

$$\mathbf{w} \cdot \mathbf{x} - b = \begin{cases} 1, & \text{για την πρώτη κλάση} \\ -1, & \text{για τη δεύτερη κλάση} \end{cases} \tag{2.4.2}$$

Στην περίπτωση που οι κλάσεις είναι γραμμικά διαχωριστέες, οι δύο επιφάνειες είναι δυνατό να επιλεγούν με τρόπο ώστε να μην υπάρχουν σημεία υποδειγμάτων μεταξύ τους και στη συνέχεια μεγιστοποιείται η μεταξύ τους απόσταση. Επειδή η απόσταση αυτή καθορίζεται από το μέγεθος $\frac{2}{||w||}$, η μεγιστοποίηση της μεταξύ τους απόστασης προκύπτει από την ελαχιστοποίηση του $||w||$. Ο δεύτερος όρος του προβλήματος, δηλαδή η απαίτηση ο χώρος μεταξύ των δυο επιφανειών να μην περιλαμβάνει κανένα αρχικό υπόδειγμα, ικανοποιείται από την εισαγωγή ενός νέου περιορισμού. Για κάθε i πρέπει να ισχύει:

$$\mathbf{w} \cdot \mathbf{x}_i - b \begin{cases} \geq 1, & \text{για τα σημεία που ανήκουν στην πρώτη κλάση} \\ \leq -1, & \text{για τα σημεία που ανήκουν στη δεύτερη κλάση} \end{cases} \tag{2.4.3}$$

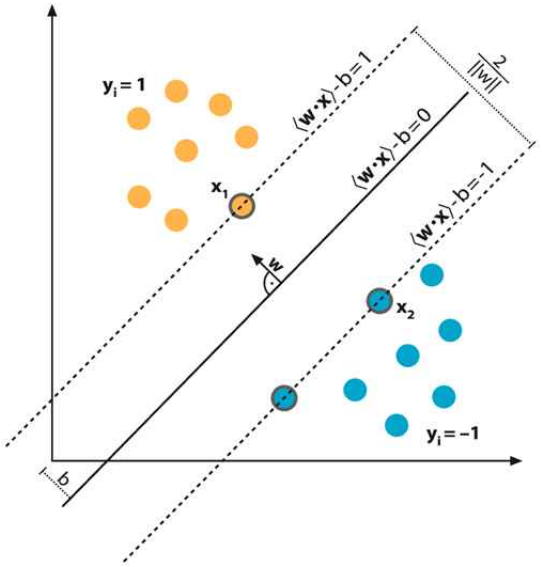
η οποία μπορεί να γραφεί και ως:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \;\; \forall 1 \leq i \leq n \tag{2.4.4}$$

εφόσον $y_i \in \{-1,\ 1\}$

Συνοψίζοντας, το πρόβλημα μεγιστοποίησης του ορίου εκφράζεται μέσω της ελαχιστοποίησης του μεγέθους $||\mathbf{w}||$ υπό την προϋποθεση ότι (subject to) $y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \;\; \forall 1 \leq i \leq n$. Σύμφωνα με τον ορισμό αυτό, η μεγιστοποίηση εξαρτάται από το μέγεθος $||\mathbf{w}||$, το μέτρο δηλαδή του διανύσματος $\mathbf{w}$, το οποίο μπορεί να αντικατασταθεί με $\frac{1}{2}||w||^2$ χωρίς να επηρεαστεί η λύση. Έτσι η βελτιστοποίηση μεταπίπτει στο δευτεροβάθμιο προγραμματισμό (quadratic programming optimization) κατά τον οποίο αναζητώνται σημεία που καθορίζουν τις διαχωρίζουσες επιφάνειες ως εξής:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{2.4.5}$$



Σχήμα 2.6: *Εντοπισμός διανυσμάτων υποστήριξης και καθορισμός μέγιστου περιθωρίου μέσω της βέλτισης επιφάνειας διαχωρισμού.*

Τα οριζόμενα $\mathbf{x}_i$ είναι ακριβώς τα διανύσματα υποστήριξης (support vectors), τα οποία καθορίζουν τις διαχωριστικές επιφάνειες και ικανοποιούν τη συνθήκη $y_i(\mathbf{w} \cdot \mathbf{x}_i - b) = 1$.

2.4.2 Ελαστικότητα περιθωρίου υπερ-επιφανειών

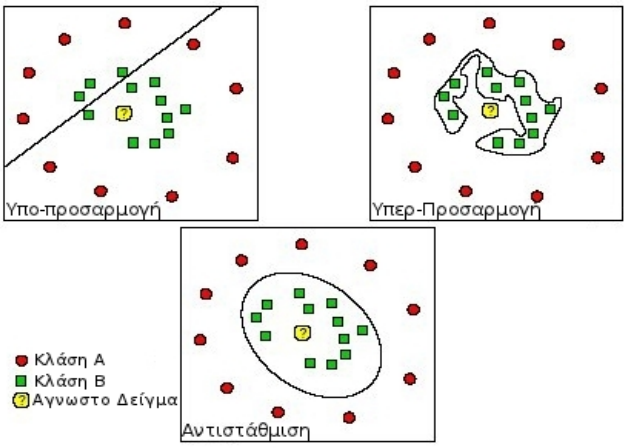
Οι ΜΔΥ που χρησιμοποιούνται σήμερα οφείλουν το μειωμένο σφάλμα γενίκευσης που παρουσιάζουν σε μια τροποποιημένη θεώρηση του ορισμού του μέγιστου περιθωρίου, μέσω του οποίου επιτρέπεται ένα ποσοστό λανθασμένων ταξινομήσεων. Ο καθορισμός των διανυσμάτων υποστήριξης, όπως θεωρήθηκε μέχρι τώρα, περιλαμβάνει τον περιορισμό να υπάρχει μηδενικό σφάλμα ταξινόμησης, δηλαδή να μην υπάρχει κανένα υπόδειγμα στο περιθώριο μεταξύ των δύο επιφανειών διαχωρισμού. Στις περιπτώσεις που δεν υπάρχει δυνατότητα καθαρού διαχωρισμού των δύο κλάσεων, η θεωρία του ελαστικού περιθωρίου (soft margin) [45] επιτρέπει σε ένα μικρό ποσοστό υποδειγμάτων αμφότερων των κλάσεων να εισέρχονται στο περιθώριο διαχωρισμού, ενώ παράλληλα αυτό μεγιστοποιείται μεταξύ των δύο κλάσεων. Το ποσοστό του επιτρεπόμενου σφάλματος ταξινόμησης νοείται μέσω του καθορισμού συγκεκριμένων μεταβλητών «χαλάρωσης» (slack variables) ξ_i, οι οποίες

επιτρέπουν σε ένα υπόδειγμα (pattern) είτε να εμπίπτει εντός του περιθωρίου σφάλματος, είτε να αποτυγχάνει τελείως κατά την ταξινόμηση. Οι εν λόγω μεταβλητές εισάγονται ως εξής:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \quad 1 \leq i \leq n \tag{2.4.6}$$

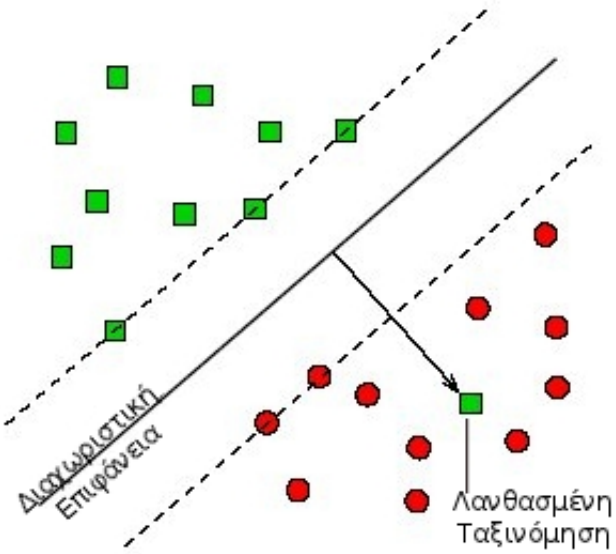
Η αντικειμενική συνάρτηση στην περίπτωση αυτή αυξάνεται με την προσθήκη ενός μεγέθους για τη διαχείριση (δηλαδή, την αύξηση ή τη μείωση ανάλογα) του ποσοστού των μη μηδενικών ξ_i. Στις περιπτώσεις που η συνάρτηση διαχείρισης των ξ_i είναι γραμμική, το πρόβλημα βελτιστοποίησης γράφεται:

$$\min_{w,b,\xi} \left\{ \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i \right\} \tag{2.4.7}$$



Σχήμα 2.7: *Υπο- και υπερ-προσαρμογή δεδομένων από τον ταξινομητή. Και στις δύο περιπτώσεις η ταξινόμηση είναι ανεπαρκής: στην πρώτη λόγω μειωμένης ακρίβειας και στη δεύτερη λόγω κάλυψης μικρού χώρου για τη μια από τις κλάσεις. Η τρίτη περίπτωση περιγράφει ταξινόμηση μειωμένου σφάλματος γενίκευσης.*

υπό την προϋπόθεση της 2.4.6. το άθροισμα των ξ_i αποτελεί το όριο των σφαλμάτων ταξινόμησης (misclassified instances). Τα εκπαιδευτικά διανύσματα αποτυπώνονται σε ένα χώρο μεγαλύτερης διάστασης, ο οποίος χρησιμοποιείται ως βάση επί της οποίας η ΜΔΥ καθορίζει την επιφάνεια του γραμμικού διαχωρισμού, εντός του μέγιστου περιθωρίου. Ο δεύτερος όρος της αντικειμενικής συνάρτησης καθορίζει το



Σχήμα 2.8: *Υλοποίηση γραμμικού διαχωρισμού με αποδοχή ποσοστού σφαλμάτων ταξινόμησης: Το αντίτιμο του σφάλματος υπολογίζεται ως η απόσταση από την επιφάνεια διαχωρισμού, πολλαπλασιασμένη με την παράμετρο κόστους.*

μέγεθος του σφάλματος ταξινόμησης για το οποίο η σταθερά C αποτελεί καθοριστική παράμετρο (penalty parameter). Η εισαγωγή ποσοστού ανοχής όσον αφορά στα σφάλματα ταξινόμησης (misclassifications), μειώνει την πιθανότητα του ενδεχομένου εξέλιξης ενός αυστηρού κατά την εκπαίδευση ταξινομητή σε υπερ-προσαρμογέα (over-fitter).

Η σταθερά $C > 0$ αντιπροσωπεύει ακριβώς αυτό το ρυθμιστή του περιθωρίου μεταξύ των κλάσεων. Κατά συνέπεια, η βελτιστοποίηση μεταπίπτει σε ένα είδος αντιστάθμισης του μεγέθους του ωφέλιμου περιθωρίου μεταξύ των επιφανειών ($\frac{1}{2}||w||^2$) και του μεγέθους του επιτρεπόμενου ποσοστού εσφαλμένων ταξινομήσεων ($C\sum_{i=1}^{n}\xi_i$): όταν το μέγεθος του περιθωρίου αυξάνει, το μέγεθος του ποσοστού μειώνεται και αντίστροφα, έτσι ώστε να ελαχιστοποιείται η 2.4.7.

Το σημαντικότερο πλεονέκτημα της χρήσης γραμμικών συναρτήσεων διαχείρισης των ξ_i είναι ότι, στην περίπτωση δυαδικών προβλημάτων ταξινόμησης, οι μεταβλητές αυτές δεν υπεισέρχονται πλέον στην εξίσωση βελτιστοποίησης, αφήνοντας μόνο τη σταθερά C ως άνω όριο καθορισμού του μέγιστου περιθωρίου.

2.4.3 Μη γραμμική ταξινόμηση

Στις περιπτώσεις που οι δύο κλάσεις είναι μη-γραμμικά διαχωριστέες, χρησιμοποιούνται μετασχηματισμοί, με σκοπό τη δημιουργία μη-γραμμικών ταξινομητών. Με τη χρήση συναρτήσεων πυρήνα [21] εισάγεται η έννοια των υπερ-επιφανειών στον καθορισμό του μέγιστου ωφέλιμου περιθωρίου διαχωρισμού. Ο αλγόριθμος που προκύπτει ακολουθεί τα πρότυπα της σχέσης 2.4.1, με τη διαφορά ότι το εσωτερικό γινόμενο αντικαθίσταται από μια μη-γραμμική συνάρτηση πυρήνα, η οποία επιτρέπει στον αλγόριθμο να προσαρμόσει την υπερ-επιφάνεια μέγιστου περιθωρίου σε ένα μετασχηματισμένο χώρο (transformed feature space). Ο μετασχηματισμός αυτός στις περισσότερες περιπτώσεις είναι μη-γραμμικός και ο μετασχηματισμένος χώρος παρουσιάζει υψηλό βαθμό διάστασης στον οποίο ο διαχωρισμός των κλάσεων είναι ευκολότερος.

Το εύρος των συναρτήσεων πυρήνα καθορίζεται εκ των προτέρων (a priori) από το χρήστη, οι δε συχνότερα χρησιμοποιούμενοι πυρήνες είναι πολυωνυμικοί ή συναρτήσεων ακτινωτής βάσης:

- Πολυωνυμικός (ομοιογενής): $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$
- Πολυωνυμικός (ανομοιογενής): $k(\mathbf{x}_i, \mathbf{x}_j) = (a + b(\mathbf{x}_i \cdot \mathbf{x}_j))^d$. Μια ειδική περίπτωση αυτού του πυρήνα για $a = 0,\ b = d = 1$ αποτελεί ο γραμμικός
- Συνάρτησης ακτινωτής βάσης: $k(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2),\ \ \gamma > 0$

Η έννοια του πυρήνα σχετίζεται με το μετασχηματισμό $\phi(\mathbf{x}_i)$ με βάση τη σχέση $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. Κατ' αντιστοιχία, στο μετασχηματισμένο χώρο η σχέση 2.4.5 γίνεται $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\mathbf{x}_i)$.

Στην πράξη, η αποτελεσματικότητα οποιασδήποτε ΜΔΥ εξαρτάται πρώτιστα από τις παραμέτρους d και γ του πυρήνα, καθώς επίσης και τη σταθερά C που εκφράζει το ελαστικό ωφέλιμο περιθώριο. Για τα περισσότερα προβλήματα, κατάλληλοι συνδυασμοί παραμέτρων ευρίσκονται κατόπιν εφαρμογής αναζήτησης πλέγματος (grid search) με ακολουθίες εκθετικά αυξανόμενων τιμών των εν λόγω παραμέτρων. Κατόπιν κάθε συνδυασμός συγκρίνεται με τους υπόλοιπους μέσω ποικίλων τύπων διαδικασιών cross-validation [79] και επιλέγονται είτε οι καλύτεροι συνδυασμοί, είτε μέσες τιμές από συγκεκριμένο αριθμό των καλύτερων από αυτούς.

2.5 Ανάλυση δεδομένων χρονοσειρών μέσω ταξινομητών ΤΝ

Ενας σημαντικός αριθμός ερευνητικών εργασιών αφιερώνονται στη μελέτη της ανάπτυξης και προτυποποίησης της εκπαίδευσης ταξινομητών μέσω μεθόδων μετα-ευρετικής[9] και στην εφαρμογή τέτοιων εκπαιδευμένων προτύπων είτε για την επιτυχή και αποτελεσματική αναπαράσταση δεδομένων χρονοσειρών, είτε για την προεπεξεργασία τους.

2.5.1 Αναπαράσταση χρονοσειρών

Στο πλαίσιο αυτό, πρότυπα ΤΝΔ αναφέρονται ως αποτελεσματικές μέθοδοι ανάλυσης χρονοσειρών [1, 55, 136, 150, 157, 169]. Σε άλλες περιπτώσεις χρησιμοποιούνται τροποποιημένα υβριδικά πρότυπα που περιλαμβάνουν συνδυασμούς πολυεπίπεδων perceptrons με μεθόδους Ολοκληρωμένων Αυτοπαλινδρομικών Προτύπων Κινούμενου Μέσου (ARIMA: Auto Regressive Integrated Moving Average) [52, 188] ΤΝΔ ανάδρομης δομής [5] ή αυτό-καθοριζόμενα ΤΝΔ [71]. Εναλλακτικές προσεγγίσεις αφορούν σε γενετικούς αλγόριθμους ή συνδυασμούς εξελικτικού προγραμματισμού με ΤΝΔ [82, 27, 44, 63].

Αν και η χρήση των ΤΝΔ ως εναλλακτικών αναλυτικών προτύπων έχει εδραιωθεί τα τελευταία χρόνια, τα εγγενή προβλήματά τους που συνοψίζονται στην υπερπροσαρμογή της εκπαίδευσης (training over-fitting), στην παγίδευση της επιδιωκόμενης λύσης σε τοπικά ελάχιστα καθώς και στις δυσκολίες που απαντώνται στον καθορισμό της βέλτιστης αρχιτεκτονικής, οδήγησαν πολλούς μελετητές στην αναζήτηση εναλλακτικών ταξινομητών. Τα πρότυπα ΜΔΥ αναφέρονται ως ικανοποι-

[9] Στην επιστήμη της πληροφορικής ο όρος μετα-ευρετική (meta-heuristic) αναφέρεται σε μια υπολογιστική μέθοδο μέσω της οποίας ένα πρόβλημα βελτιστοποιείται με την επαναληπτική αναβάθμιση μιας υποψήφιας λύσης. Στις περισσότερες περιπτώσεις κατά την εφαρμογή μεθόδων μετα-ευρετικής, η εν λόγω αναβάθμιση επιτυγχάνεται με την αναζήτηση του αλγορίθμου για τη λύση εκείνη που προσομοιάζει περισσότερο προς ένα μέτρο ποιότητας, το οποίο συνήθως έχει τεθεί εκ των προτέρων. Παρά το γεγονός ότι ο αλγόριθμος έχει τη δυνατότητα να εκτελεί αναζητήσεις σε ιδιαίτερα ευρείς δειγματικούς χώρους πιθανών λύσεων, εντούτοις δεν εγγυάται την ανεύρεση της βέλτιστης σε κάθε περίπτωση. Η τελική επιτυχία του εξαρτάται από τις δυνατότητες αναζήτησης, οι οποίες με τη σειρά τους εξαρτώνται από την εκπαίδευση που έχει λάβει ο αλγόριθμος για τη σχέση εισόδου - εξόδου του προβλήματος. Υπό την έννοια αυτή, πολλές μέθοδοι μετα-ευρετικής συχνά υλοποιούν κάποιο είδος στοχαστικής βελτιστοποίησης στο αντίστοιχο πρόβλημα το οποίο επεξεργάζονται.

ητική εναλλακτική των ΤΝΔ λύση, καθώς παρουσιάζουν μεγαλύτερες δυνατότητες γενίκευσης ενώ επιτυγχάνουν καλύτερα αποτελέσματα, ειδικά στις περιπτώσεις που συνδυάζονται με δεδομένα τα οποία έχουν υποστεί προ-επεξεργασία μέσω εξελικτικών μεθόδων [81, 118]. Τα ΜΔΥ παρουσιάζουν αξιοσημείωτες δυνατότητες σε προβλήματα ταξινόμησης και πρόβλεψης των χρονοσειρών.

Σήμερα αρκετοί μελετητές χρησιμοποιούν τροποποιημένα υποδείγματα ΜΔΥ όπως για παράδειγμα είναι τα πρότυπα ΜΔΥ Ελαχίστων Τετραγώνων (ΜΔΥ-ΕΤ) (LS-SVM: Least Squares - Support Vector Machines) κανονικού [28] καθώς και ανατροφοδοτούμενου [173] τύπου, το ΜΔΥ ελάχιστης παραλλακτικότητας κλάσης (minimum class variance SVM) [106], το ΜΔΥ Συνάρτησης Ακτινωτής Βάσης (ΜΔΥ-ΣΑΒ) (RBF-SVM: Radial Basis Function - SVM) σε συνδυασμό με τα ΤΝΔ [46, 54], το κύριο ρεύμα έρευνας εστιάζει σε κανονικά ΜΔΥ τα οποία ως επί το πλείστον χρησιμοποιούν είτε τον πυρήνα της ΣΑΒ είτε τον πολυωνυμικό πυρήνα τόσο για προβλήματα πρόβλεψης όσο και ταξινόμησης. Οι εφαρμογές τους είναι πολυάριθμες. Το 2009 παρουσιάστηκε μέθοδος [108] που χρησιμοποιούσε το πρότυπο ΜΔΥ-ΣΑΒ το οποίο τροφοδοτήθηκε με δεδομένα Λειτουργικής Απεικόνισης Μαγνητικού Συντονισμού (fMRI: functional Magnetic Resonance Imaging) με σκοπό την αυτοματοποιημένη κατηγοριοποίηση των ανθρώπινων σκέψεων. Οι Ceccarelli και Maratea πρότειναν το 2009 το συνδυασμό ανάλυσης χρονοσειρών με εικονικό γενετικό κώδικα ώστε να μορφοποιηθεί ένα αποτελεσματικό σχήμα κωδικοποίησης γονιδιακών ακολουθιών χρήσιμο σε διάφορα πεδία της γενετικής [30]. Το σύστημα αναπτύχθηκε βάσει ΜΔΥ-ΣΑΒ και εφαρμόσθηκε με επιτυχία στην ταξινόμηση του *Caenorhabditis elegans*.[10] Παραμένοντας στη μελέτη των προβλημάτων ταξινόμησης, ο Kim (2003) διερεύνησε τη χρήση των υποδειγμάτων των ΜΔΥ στην ανάπτυξη ενός έξυπνου συστήματος [95], το οποίο εφάρμοσε στην πρόβλεψη της τάσης χρηματιστηριακών δεικτών συγκρίνοντας τους δύο κυριότερους πυρήνες τον ΣΑΒ και τον πολυωνυμικό, καθώς επίσης και την απόδοση του όλου συστήματος που ανέπτυξε έναντι απλών ΤΝΔ. Από τα δεδομένα που χρησιμοποιήθηκαν στην έρευνα εξήχθησαν συγκεκριμένοι τεχνικοί δείκτες οι οποίοι στη συνέχεια χρησιμοποιήθηκαν ως είσοδοι

[10] Ο όρος αναφέρεται σε ελεύθερο στη φύση, διάφανο νηματώδη του εδάφους, μήκους περίπου 1mm, ο οποίος τρέφεται σαπροφυτικά. Η έρευνα επί της μοριακής και φυσιολογικής εξέλιξης του εν λόγω νηματώδους ξεκίνησαν κατά τη δεκαετία του '70 και συνεχίζεται αμείωτη μέχρι τις ημέρες μας, καθιστώντας τον συχνά ως πρότυπο σύγκρισης.

στο σύστημα. Στην έξοδο αποτιμάται η τάση του ημερήσιου Κορεάτικου χρηματιστηριακού δείκτη (KOSPI: Korean Composite Stock Price Index). Εάν το σύστημα προβλέπει άνοδο του δείκτη στην έξοδο αντιστοιχίζεται η τιμή 1, ενώ σε αντίθετη περίπτωση η τιμή 0. Τα αποτελέσματα της εργασίας κατέδειξαν την υπεροχή του πυρήνα και επομένως του συνολικού συστήματος ΜΔΥ έναντι του ΤΝΔ.

Συγκρίσεις εμφανίζονται και στις περιπτώσεις προβλημάτων πρόβλεψης. Οι Tay και Cao πριν μια δεκαετία περίπου ανέπτυξαν μια ΜΔΥ για την πρόβλεψη σε χρηματοοικονομικές χρονοσειρές [175]. Το σύστημα αναπτύχθηκε βάσει του πυρήνα ΣΑΒ ο οποίος σε όλες τις μετρήσεις απέδωσε καλύτερα από τον πολυωνυμικό. Επίσης και στην περίπτωση αυτή υπάρχει σύγκριση ΜΔΥ και ΤΝΔ με τα αποτελέσματα της πρώτης να υπερτερούν σαφώς έναντι της χρήσης του δεύτερου. Η πρόβλεψη του φορτίου του ηλεκτρικού ρεύματος έχει μελετηθεί αρκετά με βάση μηχανές πυρήνα. Οι Wu κ.α. (2009) χρησιμοποίησαν γενετικούς αλγορίθμους για να καθορίσουν την άριστη παραμετροποίηση και τη συνάρτηση πυρήνα ΜΔΥ, η οποία στη συνέχεια εφαρμόσθηκε για να προβλέψει το μέγιστο ημερήσιο ηλεκτρικό φορτίο για τον Ιανουάριο του 1999 [185]. Το εκπαιδευτικό σύνολο δεδομένων κάλυπτε την περίοδο 1997 - 1999 και αποτελείτο από τιμές ηλεκτρικού φορτίου ανά μισή ώρα, τη μέση ημερήσια θερμοκρασία και τις δημόσιες αργίες για το διάστημα αυτό. Τα χρωμοσώματα του γενετικού αλγορίθμου περιλάμβαναν γονίδια για τη συνάρτηση του πυρήνα, το βαθμό του πολυωνύμου για τον πολυωνυμικό πυρήνα, καθώς επίσης και τη διακύμανση και τη σταθερά του εκπαιδευτικού σφάλματος για τον πυρήνα ΣΑΒ. Η έρευνα απέδειξε ότι για το συγκεκριμένο πρόβλημα ο πολυωνυμικός πυρήνας βαθμού 4.55 και σταθεράς 192.85 είναι ο αποτελεσματικότερος.

Με το ίδιο πρόβλημα ασχολήθηκε και ο Hong [77] αναπτύσσοντας ένα πρότυπο ΜΔΥ Παλινδρόμησης (ΜΔΥΠ) με πυρήνα ΣΑΒ και αυτοματοποιημένης παραμετροποίησης μέσω ανοσοποιητικού αλγορίθμου (immune algorithm). Στη μελέτη ως εκπαιδευτικό σύνολο χρησιμοποιούνται δεδομένα ηλεκτρικού φορτίου της Ταϊβάν που καλύπτουν εικοσαετή χρονική περίοδο, από το 1981 έως το 2000. Τα πειραματικά αποτελέσματα έδειξαν ότι το προτεινόμενο μοντέλο ΜΔΥΠ ήταν αποτελεσματικότερο των υπόλοιπων ταξινομητών με τους οποίους συγκρίθηκε αναφορικά με την ακρίβεια πρόβλεψης.

Οι Lu και Wang (2005) διερεύνησαν τη δυνατότητα κατασκευής μιας παραλλα-

γής ΜΔΥ για την πρόβλεψη του επιπέδου αέριων περιβαλλοντικών ρυπαντών [118] μέσω δεδομένων χρονοσειρών που προήλθαν από σχετική βάση δεδομένων του σταθμού περιβαλλοντικής παρακολούθησης Causeway Bay του Hong Kong. Το σύστημα αναπτύχθηκε σε περιβάλλον Matlab χρησιμοποιώντας έναν αλγόριθμο ακολουθίας για την ελαχιστοποίηση του σφάλματος και περιελάμβανε μια ΜΔΥ με πυρήνα Gauss. Τα αποτελέσματα της έρευνας επιβεβαίωσαν την υπόθεση ότι η προσέγγιση ΜΔΥ πλεονεκτεί έναντι των ΤΝΔ, προσφέροντας γενικότερα καλύτερη πρόβλεψη, ενώ παράλληλα παρουσιάζουν ευκολότερη παραμετροποίηση.

Επιπροσθέτως, αναφέρεται ότι η προ-επεξεργασία δεδομένων χρονοσειρών βάσει γενετικών αλγορίθμων βελτιώνει την ικανότητα πρόβλεψης των ΜΔΥ. Οι Huang και Wu (2008) εφάρμοσαν μια τεχνική προεπεξεργασίας μέσω γενετικών αλγορίθμων για την εξαγωγή κατάλληλων χαρακτηριστικών από τα αρχικά δεδομένα, με σκοπό τη χρήση τους στην εκπαίδευση ΜΔΥ [81]. Η έρευνα αποσκοπούσε στο σχεδιασμό ενός υβριδικού μοντέλου ΜΔΥ για την πρόβλεψη των τάσεων του Ασιατικού χρηματιστηριακού δείκτη και του δείκτη G7, χρησιμοποίησε δε ως ιστορικά δεδομένα εκπαίδευσης τις τιμές των δεικτών αυτών από τον Ιανουάριο του 2003 έως το Δεκέμβριο του 2005. Τα αποτελέσματα καταδεικνύουν τη βελτίωση που επήλθε μετά την εξελικτική εξαγωγή χαρακτηριστικών από την αρχική χρονοσειρά.

2.5.2 Προεπεξεργασία χρονοσειρών

Ποικίλες είναι οι εργασίες που αναφέρονται στην προ-επεξεργασία χρονοσειρών ως προπαρασκευαστικό στάδιο για την κατάλληλη εκπαίδευση εργαλείων εξόρυξης δεδομένων. Ο Palmer ανέπτυξε το 2005 ένα σύστημα εκπαίδευσης μηχανής βασισμένο σε ΤΝΔ πρόβλεψης της τουριστικής κίνησης με τη χρήση χρονοσειρών [140]. Η έρευνα αφορούσε στις Βαλεαρίδες Νήσους (Balearic Islands) για μια χρονική περίοδο δεκατεσσάρων συνεχόμενων ετών, από το 1986 έως το 2000. Οι ερευνητές εφάρμοσαν μια καινοτόμο προ-επεξεργασία των αρχικών χρονοσειρών μέσω λογαριθμικής μετατροπής για την ανάδειξη των τάσεων, την ελαχιστοποίηση του θορύβου, την αποκάλυψη συσχετίσεων και την εξομάλυνση της κατανομής των αρχικών μεταβλητών, διαδικασία που είχε ως αποτέλεσμα την αύξηση της ακρίβειας του ΤΝΔ. Την ίδια χρονιά μελετήθηκε η επίδραση της ελαχιστοποίησης της εποχικότητας και της τάσης των χρονοσειρών στην απόδοση των ΤΝΔ [196]. Τα αποτελέσματα της έρευνας έδειξαν ότι τα ΤΝΔ δεν είναι σε θέση απευθείας να προτυποποι-

ήσουν την εποχικότητα των συγκεκριμένων χρονοσειρών. Η ανάπτυξη κατάλληλης διαδικασίας προ-επεξεργασίας των πρωτογενών δεδομένων αναβαθμίζει την αποτελεσματικότητα του νευρωνικού ταξινομητή.

Οι μελέτες επί της επίδρασης της προ-επεξεργασίας στην αποτελεσματικότητα διαφόρων ταξινομητών συμπεριλαμβάνει ποικίλες μεθόδους με σημαντικά πολλές φορές αποτελέσματα. Το 2006, χρησιμοποιήθηκε ο Διακριτός Μετασχηματισμός Κυματιδίου (DWT: Discrete Wavelet Transformation) και η διαμέριση των δεδομένων για την προ-επεξεργασία πληροφορίας σχετικής με τη μηνιαία απορροή που καταγράφηκε στη λεκάνη απορροής του Tirso της Σαρδηνίας [26]. Τα δεδομένα που προέκυψαν χρησιμοποιήθηκαν στην εκπαίδευση ενός ΤΝΔ για την πρόβλεψη της απορροής κατά ένα μήνα. Τα αποτελέσματα της έρευνας ήταν ενθαρρυντικά σχετικά με τη σύγκριση των προ-επεξεργασμένων δεδομένων έναντι της αρχικής πρωτόλειας χρονοσειράς. Την ίδια χρονιά, διεξήχθη έρευνα σχετικά με τη χρησιμότητα της ομαδοποίησης των χρονοσειρών [168]. Η έρευνα προέκυψε ένεκα της εμφάνισης αποτελεσμάτων που συνηγορούσαν υπέρ της άποψης ότι η εν λόγω προεπεξεργασία δεν έχει νόημα και την αντικρούει εισάγοντας την έννοια της αναπτυσσόμενης προ-επεξεργασίας. Τα αποτελέσματα της έρευνας δείχνουν ότι η ομαδοποίηση συγκεκριμένων χρονοσειρών δίνει καλά αποτελέσματα, υπό την προϋπόθεση ότι έχει προηγηθεί αποτελεσματική προ-επεξεργασία των πρωτογενών δεδομένων. Πρόσφατα, μελετήθηκε η επίδραση τριών διαφορετικών προ-επεξεργαστικών τεχνικών στην απόδοση ενός νευρωνικού ταξινομητή [186]. Συγκεκριμένα, μέσω προ-επεξεργασίας χρονοσειρών με τις μεθόδους του κινούμενου μέσου, της φασματικής ανάλυσης ιδιαζόντων τιμών (SSA: Singular Spectrum Analysis) και της κυματιδιακής πολυ-κλιμακωτής ανάλυσης (WMRA: Wavelet Multi-Resolution Analysis) προέκυψαν δευτερογενή δεδομένα προκειμένου να χρησιμοποιηθούν στην εκπαίδευση ΤΝΔ, η σύγκριση των αποδόσεων του οποίου ανέδειξε την υπεροχή της μεθόδου του κινούμενου μέσου. Στα βασικότερα συμπεράσματα της έρευνας συγκαταλέγεται η σπουδαιότητα της προ-επεξεργασίας της χρονοσειράς ως μέσο βελτιστοποίησης της απόδοσης του ταξινομητή.

Η ανάλυση δεδομένων χρονοσειρών και η εξαγωγή χαρακτηριστικών από αυτές πολύ συχνά μετέρχεται προτύπων παλινδρόμησης, όπως είναι τα πρότυπα του αυτο-συσχετιζόμενου και του κινούμενου μέσου. Ποικίλες περιπτώσεις νευρωνικών

ταξινομητών [12, 19, 58, 152, 159, 182], ασαφούς λογικής [129, 56, 105, 57] και εξελικτικών αλγορίθμων [72], παράλληλα με κλασσικούς ταξινομητές, όπως ο ταξινομητής Bayes, οι μέθοδοι του εγγύτερου γείτονα και του δένδρου αποφάσεων [130, 137, 184, 9] συγκαταλέγονται σε ένα μεγάλο κατάλογο τεχνικών αντιμετώπισης της μη γραμμικότητας και του υψηλού βαθμού διάστασης των χρονοσειρών. Τέλος, ο σχεδιασμός και η προτυποποίηση συστημάτων συνδυασμού διαφόρων ταξινομητών έχει μελετηθεί κατά τη διάρκεια των τελευταίων χρόνων ως εναλλακτική λύση σε παρόμοια προβλήματα. Στην τελευταία περίπτωση, ποικίλες τεχνικές ταξινόμησης συνδυάζονται σε μεμονωμένα συστήματα εκπαιδευμένα για την αντιμετώπιση προβλημάτων αναγνώρισης προτύπων. Οι συνδυαστικές αυτές μέθοδοι είναι δυνατόν σε κάποιες περιπτώσεις να προσφέρουν υλοποιήσιμες λύσεις, ιδιαίτερα για προβλήματα πολύπλοκων χρονοσειρών όπου συστήματα μονού ταξινομητή αδυνατούν να ανταποκριθούν επαρκώς [6].

Σχετικά με την εφαρμογή εργαλείων εκπαίδευσης μηχανής στη μικροβιολογία, λίγες είναι οι περιπτώσεις που αντιστοιχούν σε φυτικούς ιούς. Μια περίπτωση είναι η μελέτη και η ανάπτυξη ενός συστήματος εξελικτικού ΤΝΔ για τον ευφυή έλεγχο των μικροβιακών διεργασιών που αναπτύσσονται κατά την αποθήκευση καρπών [133]. Η μελέτη συνδυάζει τη σχετική υγρασία με την υδατική απώλεια και την αύξηση των μυκητιακών προσβολών στη μονάδα του χρόνου. Σε πρώτη φάση αναγνωρίσθηκαν δυναμικοί συνδυασμοί των υδατικών απωλειών και του ποσοστού μυκητιακής ανάπτυξης στους καρπούς υπό την επήρρεια της σχετικής υγρασίας του περιβάλλοντος του χώρου αποθήκευσης. Τα δεδομένα εισήχθησαν σε κατάλληλα δομημένο ΤΝΔ μέσω του οποίου ταξινομήθηκαν τα επίπεδα της σχετικής υγρασίας και στη συνέχεια υποδείχθηκαν τα άριστα επίπεδα αυτής μέσω Γενετικών Αλγορίθμων (ΓΑ).

Σε αντιδιαστολή, πολυάριθμες μελέτες εκμεταλλεύονται εργαλεία υπολογιστικής νοημοσύνης για την επίλυση προβλημάτων ταυτοποίησης ζωικών ιών. Ενα παράδειγμα τέτοιας εφαρμογής είναι η χρήση ΤΝΔ για την ταυτοποίηση της κίρρωσης του ήπατος ασθενών με χρόνια ηπατίτιδα τύπου C [74]. Το δίκτυο εκπαιδεύθηκε με παραμέτρους του ιού και του ξενιστή, ώστε να είναι σε θέση να αναγνωρίζει την ύπαρξη του ιού στους ασθενείς. Τα αποτελέσματα δείχνουν ότι το πρότυπο ΤΝΔ ξεπέρασε σε ακρίβεια τη λογιστική παλινδρόμηση φθάνοντας σε τιμές ευαισθησίας

το 92% και ειδικότητας το 98.9%. Η προγνωστική αξία των θετικών και αρνητικών αποτελεσμάτων ήταν 95% και 97% αντίστοιχα, καθιστώντας το ΤΝΔ έναν ιδιαίτερα αξιόλογο ταξινομητή.

Προς τα τέλη του περασμένου αιώνα η υπολογιστική νοημοσύνη εδραίωσε μεταξύ των μελετητών τις δυνατότητές της ως χρήσιμου εργαλείου για μικροβιολογική και βιοχημική έρευνα. Τεχνικές εκπαίδευσης μηχανής [135] χρησιμοποιήθηκαν με σκοπό την προτυποποίηση των διαδρομών που ακολουθούν οι πρωτεΐνες προς τον τελικό ενδοκυτταρικό τους προορισμό[11]. Οι ερευνητές χρησιμοποίησαν συγκεκριμένο λογισμικό (SignalIp) βασισμένο στην τεχνολογία ΤΝΔ για τον ασφαλέστερο προσδιορισμό του γνωστότερου sorting signal, αυτού του εκκριτικού πεπτιδίου. Πιο πρόσφατα, μελετήθηκε η περίπτωση της χρήσης ΤΝΔ στην παρακολούθηση της γρίππης τύπου Α [51]. Συγκεκριμένα, το δίκτυο εκπαιδεύθηκε έτσι ώστε να αναγνωρίζει πρότυπα εικόνων φθορισμού του ιού, επιδεικνύοντας τελικά υψηλές κλινικές τιμές ευαισθησίας και ειδικότητας της τάξης του 95% και 92% αντίστοιχα.

Επίσης, ποικίλες μορφές ΤΝΔ έχουν χρησιμοποιηθεί στην ανάπτυξη εργαλείων λογισμικού για την ιατρική παθολογία, προς την κατεύθυνση της αντιμετώπισης επιδημιών ή την αναγνώριση ατόμων πιο ευαίσθητων στις λοιμώξεις. Δίκτυα Bayes αναφέρονται ως αποτελεσματικά στην αναγνώριση δειγμάτων φυσιολογικών και προσβεβλημένων ιστών ή στη διάκριση της προσβολής από παθογόνο ιό ή μύκητα σε ενδιάμεσα ή τελικά στάδια της λοίμωξης [64], ενώ στα πρώιμα στάδια απαιτείται προ-επεξεργασία των χρονοσειρών για την αύξηση της διακριτικής ικανότητας της μεθόδου.

2.6 Γενετικοί Αλγόριθμοι

Οι γενετικοί αλγόριθμοι (ΓΑ) (GA: Genetic Algorithms) αποτελούν μια κατηγορία ευρετικών μεθόδων με βάση την κτηθείσα εμπειρία, που λειτουργούν μιμούμενοι τη φυσική εξελικτική διαδικασία. Πρόκειται δηλαδή για μια υποπερίπτωση της κατηγορίας των Εξελικτικών Αλγορίθμων (EA: Evolutionary Algorithms - EA) η οποία περιλαμβάνει μεθόδους για την επίλυση κυρίως προβλημάτων βελτιστοποίησης (optimization), συγκεντρώνοντας γνώση βάσει εμπειριών με τη χρήση τεχνι-

[11] Οι πρωτεΐνες ακολουθούν συγκεκριμένες διαδρομές στον οργανισμό γνωστές στη διεθνή βιβλιογραφία με τον όρο *sorting signals*

κών εμπνευσμένων από τη θεωρία της Φυσικής Επιλογής[12], όπως κληρονομικότητα (inheritance), επιλογή (selection), μετάλλαξη (mutation) και διασταύρωση (cross-over) ή ανα–συνδυασμός (recombination).

2.6.1 Ιστορικά Στοιχεία

Εξομοιώσεις της διαδικασίας της φυσικής εξέλιξης με τη χρήση υπολογιστών άρχισαν να κάνουν την εμφάνισή τους ήδη από τη δεκαετία του 1950 στο Πανεπιστήμιο του Princeton, ενώ κατά την επόμενη δεκαετία έγιναν πιο συχνές και καταγράφηκαν σε διάφορα συγγράμματα και δοκίμια. Μάλιστα κατά την περίοδο αυτή οι εξομοιώσεις που καταγράφονται περιλαμβάνουν όλα τα απαραίτητα στοιχεία γενετικής εξέλιξης που συνθέτουν τους πλέον εξελιγμένους σημερινούς ΓΑ, συμπεριλαμβανομένων και των παραμέτρων του ανα–συνδυασμού και της μετάλλαξης [126].

Παρά την επιτυχία που σημείωσαν οι πρώιμες αυτές προσπάθειες, έπρεπε να παρέλθει περίπου μια δεκαετία και να μεσολαβήσει η εργασία του John Holland στο Πανεπιστήμιο του Michigan και η έκδοση του βιβλίου του «Adaptation in Natural and Artificial Systems» [76] πριν οι ΓΑ αποκτήσουν τη φήμη - που ακόμη και σήμερα τους συνοδεύει - ως αποτελεσματικές τεχνικές επίλυσης διαφόρων ειδών προβλημάτων. Παρότι ο Holland παρουσίασε ένα κανονιστικό πλαίσιο για την πρόβλεψη του βαθμού «ποιότητας» της επόμενης γενεάς του πληθυσμού του αλγορίθμου (Θεώρημα του Holland - Holland's Schema Theorem), η έρευνα επάνω στους ΓΑ παρέμεινε σε θεωρητικό επίπεδο μέχρι περίπου τα μέσα της δεκαετίας του '80, περίοδο κατά την οποία έλαβε χώρα το Πρώτο Διεθνές Συνέδριο για τους Γενετικούς Αλγορίθμους στο

[12]Το 1858, ο φυσιοδίφης, βιολόγος και γεωλόγος Charles Darwin στο έργο του «On the Origin of Species by means of Natural Selection», όρισε την Φυσική Επιλογή ως τη διαδικασία εξέλιξης των ειδών. Μέσω αυτής, οργανισμοί που είναι «καλύτερα» προσαρμοσμένοι στον περιβάλλοντα χώρο ή ενδιαίτημα παράγουν περισσότερους απογόνους από εκείνους τους οργανισμούς που παρουσιάζουν μικρότερο βαθμό «προσαρμοστικότητας». Ένεκα του συγκριτικού πολυάριθμου των απογόνων, το γονίωμα των γονέων αποκτά μεγαλύτερες πιθανότητες να επιβιώσει μέσω αυτών αφενός, αλλά και αφετέρου να «προσαρμοστεί» ακόμη καλύτερα στο φυσικό περιβάλλον μέσω διαδικασιών γενετικής επιλογής και εξέλιξης. Αν λοιπόν ένα κληρονομήσιμο γνώρισμα προσφέρει προσαρμοστικό πλεονέκτημα στο φορέα του, αυτός (είτε γιατί επιβιώνει περισσότερο, είτε γιατί επιλέγεται περισσότερο από τα άτομα του άλλου φύλου, σε σύγκριση με όσους δεν το φέρουν) αφήνει περισσότερους απογόνους με αποτέλεσμα να το μεταβιβάζει με αυξημένη συχνότητα στα άτομα της επόμενης γενιάς. Με τον τρόπο αυτό συσσωρεύονται από γενιά σε γενιά τα ευνοϊκά για την επιβίωση γνωρίσματα, γεγονός που πιθανόν να οδηγήσει βαθμιαία στη δημιουργία ενός νέου «καλύτερα προσαρμοσμένου» πληθυσμού.

Πίτσμπουργκ της Πενσυλβανίας. Αμέσως μετά και παράλληλα με την σημαντική αύξηση της διαθέσιμης υπολογιστικής ισχύος, αυξήθηκε το ακαδημαϊκό και βιομηχανικό ενδιαφέρον, με αποτέλεσμα ένα μεγάλο ποσοστό επιχειρήσεων του ιδιωτικού τομέα να χρηματοδοτούν προγράμματα προς την κατεύθυνση της εφαρμογής των αποτελεσμάτων της μέχρι τότε έρευνας επί των ΓΑ. Στις ημέρες μας, οι ΓΑ θεωρούνται πλέον μια εδραιωμένη τεχνική προς την κατεύθυνση της επίλυσης όχι μόνο προβλημάτων βελτιστοποίησης, αλλά επίσης και πρόβλεψης ή ταξινόμησης, βρίσκοντας εφαρμογή στους τομείς της βιο-πληροφορικής (bioinformatics), στην φυλογενετική (phylogenetics), στην επιστήμη των υπολογιστών (computational science), στη μηχανική (engineering), στα οικονομικά (economics), στη χημεία (chemistry), στη βιομηχανία / μεταποίηση (manufacturing), στη φυσική (physics) και αλλού.

2.6.2 Μεθοδολογία

Για το σχεδιασμό ενός τυπικού γενετικού αλγορίθμου απαιτούνται:

1. Η κατασκευή μιας γενετικής αναπαράστασης / απεικόνισης του δειγματικού χώρου των δυνητικών λύσεων (solution domain) και

2. Ο σχηματισμός της συνάρτησης προσαρμοστικότητας, μέσω της οποίας επιτυγχάνεται η αξιολόγηση των λύσεων αυτών

Στον πυρήνα της ιδέας γύρω από τον γενετικό αλγόριθμο έγκειται η εξέλιξη, υπό το πρίσμα της ανέλιξης ενός αρχικού πληθυσμού πιθανών λύσεων ενός προβλήματος ανά διαδοχικές γενεές, έως ότου βρεθεί μια λύση η οποία να ικανοποιεί κάποιες αρχικές συνθήκες. Υπ' αυτή την έννοια, ένας ΓΑ αρχικοποιείται από ένα πληθυσμό ατόμων τα οποία ονομάζονται χρωμοσώματα (ή γενότυπος του γονιώματος) σε καθένα από τα οποία έχει κωδικοποιηθεί μια δυνητική υποψήφια λύση ενός συγκεκριμένου προβλήματος. Παραδοσιακά, τέτοιες λύσεις απεικονίζονται ως δυαδικές συμβολοσειρές (binary strings), αν και διαφορετικές θεωρήσεις είναι επίσης υπό έρευνα. Βασικό χαρακτηριστικό το οποίο καθιστά χρήσιμες τέτοιες απεικονίσεις είναι ότι τα μέρη από τα οποία αποτελείται κάθε άτομο (λύση) και τα οποία καλούνται γονίδια (genes) είναι σταθερά σε πλήθος, επιτρέποντας απλές διαδικασίες επιλογής και ανα–συνδυασμού.

Η εξέλιξη του αρχικού τυχαιοποιημένου πληθυσμού προχωρεί μέσω διαδοχικών γενεών. Σε κάθε γενεά υπολογίζεται ο «βαθμός προσαρμοστικότητας» («fitness

score») για κάθε άτομο που ανήκει σε αυτήν, αποτελεί δε μια ποσοτικοποίηση της εγγύτητας της συγκεκριμένης λύσης προς την ιδεατή λύση του προβλήματος. Ο υπολογισμός του βαθμού προσαρμοστικότητας κάθε χρωμοσώματος επιτυγχάνεται με την εφαρμογή της συνάρτησης προσαρμοστικότητας (fitness function) η οποία ποσοτικοποιεί την «ποιότητα» της κάθε φορά αναπαριστάμενης λύσης.

Στη συνέχεια, οι «καλύτερες» λύσεις (με την έννοια: τα άτομα με το μεγαλύτερο βαθμό προσαρμοστικότητας) συγκεντρώνονται στοχαστικά και «ζευγαρώνουν» (με την έννοια: τροποποιούνται με συγκεκριμένες πιθανότητες επιλογής, ανα–συνδυασμού και μετάλλαξης) για να σχηματοποιήσουν μια νέα γενεά η οποία θα χρησιμοποιηθεί στην επόμενη επανάληψη (iteration) του αλγορίθμου.

Ο αλγόριθμος προχωρεί σε διακεκριμένες φάσεις, ως εξής:

Αρχικοποίηση

Στην αρχή, ένα πλήθος χρωμοσωμάτων / λύσεων παράγονται κατά τέτοιο τρόπο ώστε το γονίωμά τους να αποτελείται από τυχαία επιλεγμένα, δυαδικού τύπου γονίδια. Ο αριθμός των χρωμοσωμάτων της αρχικής γενεάς ποικίλλει και εξαρτάται από το πρόβλημα, περιορίζεται δε όσο μειώνεται ο απαιτούμενος χρόνος σύγκλισης του αλγορίθμου, αλλά παραμένει ικανός ώστε να καλύπτει το σύνολο των δυνητικών λύσεων (search plane). Περιστασιακά, είναι δυνατόν η αρχική γενεά να μην προκύπτει από πλήρη τυχαιοποίηση, αλλά να εμπλουτίζεται με περιπτώσεις που πιθανόν να οδηγούν σε άριστες λύσεις.

Επιλογή

Με τον όρο «επιλογή» υπονοείται η διαδικασία κατά την οποία δύο χρωμοσώματα του προηγούμενου πληθυσμού προκρίνονται ως κατάλληλα για να αναπαραχθούν και να αποδώσουν τους απογόνους τους στην επόμενη γενεά. Κατά τη διάρκεια κάθε διαδοχικής γενεάς, ένα μέρος του πληθυσμού επιλέγεται με συγκεκριμένα κριτήρια τα οποία βασίζονται στην προσαρμοστικότητα του κάθε ατόμου κατά τρόπο ανάλογο. Έχουν αναπτυχθεί ποικίλες μέθοδοι επιλογής των «καλύτερων» χρωμοσωμάτων. Κάποιες από αυτές, υπολογίζουν την απόδοση ενός εκάστου χρωμοσώματος και επιλέγουν εκείνα που εμφανίζουν το μεγαλύτερο βαθμό προσαρμοστικότητας. Άλλες, για να αποφύγουν το χρονικό κόστος που παρουσιάζουν οι πρώτες, αξιολογούν όσον αφορά την προσαρμοστικότητα μόνο ένα τυχαίο δείγμα του αρχικού πληθυσμού. Όλες οι μέθοδοι επιλογής είναι στοχαστικές και λαμβά-

νουν υπόψη το γεγονός ότι η παραλλακτικότητα κάθε επόμενης γενεάς θα πρέπει να συντηρείται σε υψηλά επίπεδα, έτσι ώστε να μη «χάνονται» πιθανές άριστες λύσεις. Για το λόγο αυτό μεριμνούν ώστε ένα μικρό ποσοστό λύσεων μικρότερης προσαρμοστικότητας να «περνούν» επίσης στη φάση της αναπαραγωγής. Δύο είναι οι επικρατέστερες μέθοδοι επιλογής: η Μέθοδος της Ρουλέτας και η Μέθοδος του Διαγωνισμού:

- *Μέθοδος της Ρουλέτας (Roulette Wheel Method):* επίσης γνωστή και ως μέθοδος της αναλογικής προσαρμοστικότητας (fitness proportionate selection), η μέθοδος αυτή χρησιμοποιεί τη συνάρτηση προσαρμοστικότητας για να αποδώσει ένα βαθμό προσαρμοστικότητας f_i στο χρωμόσωμα i της τρέχουσας γενεάς, ο οποίος χρησιμοποιείται με τρόπο ώστε να αντιστοιχιστεί μια πιθανότητα επιλογής (selection probability) p_i σε αυτό το χρωμόσωμα, ίση με:

$$p_i = \frac{f_i}{\sum_{j=1}^{n} f_j} \tag{2.6.1}$$

  όπου n είναι το πλήθος των χρωμοσωμάτων της συγκεκριμένης γενεάς. Παρατηρούμε εδώ ότι σε κάθε χρωμόσωμα αντιστοιχίζεται μια πιθανότητα επιλογής ανάλογη της προσαρμοστικότητάς του. Αυτό θα μπορούσε να παρομοιασθεί με μια ρουλέτα τα τμήματα της οποίας είναι ανισομερή, με επιφάνεια που «μεγαλώνει» ανάλογα με την προσαρμοστικότητα του αντίστοιχου χρωμοσώματος. Συνεπώς είναι πιο πιθανό η μπίλια (με την έννοια: η επιλογή) να σταματήσει στις μεγαλύτερες επιφάνειες (με την έννοια: πιο προσαρμοσμένα χρωμοσώματα). Έτσι, ενώ υποψήφιες λύσεις μεγαλύτερης προσαρμοστικότητας είναι πιο απίθανο να εξαφανιστούν, ωστόσο υπάρχει (έστω και μια) μικρή πιθανότητα ότι αυτό θα συμβεί. Αντιθέτως, υπάρχει μια μικρή επίσης πιθανότητα να επιλεγούν λύσεις μικρότερης προσαρμοστικότητας, λόγω του ότι τέτοιες λύσεις πιθανόν να κρύβουν στο γονίωμά τους ενδιαφέροντες συνδυασμούς.

- *Μέθοδος του Διαγωνισμού (Tournament Method):* Σύμφωνα με αυτήν, κάθε φορά επιλέγεται τυχαία με επανάθεση ένα μέρος του αρχικού πληθυσμού χρωμοσωμάτων και από αυτό προκρίνεται προς τη φάση της αναπαραγωγής το

χρωμόσωμα με το μεγαλύτερο βαθμό προσαρμοστικότητας. Η διαδικασία αυτή συνεχίζεται έως ότου επιλεγεί για αναπαραγωγή ικανός αριθμός χρωμοσωμάτων.

Ψευδοκώδικας της μεθόδου του διαγωνισμού:

Επανέλαβε μέχρι να γεμίσει η λίμνη αναπαραγωγής:

Επέλεξε k τυχαία χρωμοσώματα του αρχικού πληθυσμού

Από αυτά επέλεξε το καλύτερο με πιθανότητα p

Από τα υπόλοιπα επέλεξε το δεύτερο καλύτερο με πιθανότητα $p(1-p)$

Από τα υπόλοιπα επέλεξε το τρίτο καλύτερο με πιθανότητα $p(1-p)^2$

(...κ.ο.κ όσες φορές τεθεί)

Επανάθεσε το δείγμα των k χρωμοσωμάτων στον αρχικό πληθυσμό

Το εύρος του υποσυνόλου (k) που κάθε φορά αποτελεί το πεδίο του διαγωνισμού επηρεάζει την πίεση που ασκεί ο αλγόριθμος στην τελική επιλογή από άποψη ελιτίστικης προσαρμοστικότητας, υπό την έννοια ότι όσο μεγαλώνει το εύρος του υποσυνόλου, τόσο μειώνεται η πιθανότητα να επιλεγούν χρωμοσώματα μικρότερης προσαρμοστικότητας. Υπό ντετερμινιστικές[13] συνθήκες, όταν τεθεί $p = 1$ ο αλγόριθμος της μεθόδου του διαγωνισμού επιλέγει πάντα το «καλύτερο» χρωμόσωμα σε κάθε διαγωνισμό, ενώ όταν τεθεί $k = 1$ η μέθοδος μεταπίπτει σε τυχαία επιλογή.

Αναπαραγωγή

Με βάση το βαθμό προσαρμοστικότητας, κατά τη φάση της επιλογής προκρίνεται ένας αριθμός χρωμοσωμάτων τα οποία θα αποτελέσουν τη γενετική λίμνη (genetic pool) από όπου θα προκύψει η επόμενη γενεά του αλγορίθμου. Από αυτήν επιλέγονται κάθε φορά με (ή χωρίς) επανάθεση δυο χρωμοσώματα που αντιπροσωπεύουν τους γονείς που θα αναπαραχθούν. Η αναπαραγωγή επιτυγχάνεται με δυο τρόπους, σε κάθε έναν από τους οποίους έχει τεθεί μια πιθανότητα ενεργοποίησης:

α. *Αυτούσια επιλογή* , κατά την οποία ένας από τους δυο γονείς (ή σε κάποιες περιπτώσεις και οι δύο) μεταφέρεται αυτούσιος στην επόμενη γενεά

[13] Ντετερμινισμός (determinism) είναι μια φιλοσοφική προσέγγιση που αναλύει δράσεις με βάση το αίτιο και το αιτιατό. Σύμφωνα με την εν λόγω θεωρία, δοθέντων συγκεκριμένων συνθηκών, το αποτέλεσμα ενός συστήματος δεν είναι δυνατό να αλλάξει εάν δεν αλλάξουν και οι συνθήκες αυτές.

β. *Ανα–συνδυασμός* των γονέων, όπου τμήμα του ενός γονέα συνδυάζεται με το συμπληρωματικό του από τον άλλο γονέα για να μορφοποιηθεί το χρωμόσωμα που θα αποτελέσει τον απόγονο

γ. *Μετάλλαξη,* κατά την οποία ένα (ή πιο σπάνια περισσότερα) δυαδικό γονίδιο που ανήκει στο γονίωμα των απογόνων μεταπίπτει με συγκεκριμένη προκαθορισμένη πιθανότητα σε αντίθετο στροβιλισμό (spin) (από 0 σε 1 και αντίστροφα)

Τερματισμός

Η εξελικτική διαδικασία συνεχίζεται έως ότου ικανοποιηθεί κάποια από τις συνθήκες τερματισμού που έχουν προκαθορισθεί κατά τη φάση της αρχικοποίησης του αλγορίθμου. Κοινοί τόποι τερματισμού αποτελούν

- Η περίπτωση εύρεσης ιδανικής λύσης, δηλαδή του χρωμοσώματος που ικανοποιεί τα βέλτιστα χαρακτηριστικά που έχουν προκαθορισθεί κατά τη φάση της αρχικοποίησης
- Η ολοκλήρωση ενός συγκεκριμένου προκαθορισμένου αριθμού γενεών
- Η υπέρβαση ενός προκαθορισμένου ορίου σε υπολογιστικό χρόνο
- Η περίπτωση κατά την οποία ο βαθμός προσαρμοστικότητας της γενετικής λίμνης σταθεροποιείται σε κάποιο όριο (φθάνει σε «πλατώ») και δε βελτιώνεται περαιτέρω. Οι μηχανισμοί ελέγχου στην περίπτωση αυτή είναι ιδιαίτερα δαπανηροί, αλλά προσφέρουν προφανή πλεονεκτήματα στον αλγόριθμο
- Συνδυασμός δύο ή περισσότερων από τα ανωτέρω

2.6.3 Υποθέσεις επί της προσαρμοστικότητας των ΓΑ

Κατά τη φυσική επιλογή, πληθυσμοί ατόμων ανταγωνίζονται για επιβίωση και αναπαραγωγή. Κατά συνέπεια, σχετικά πιο προσαρμοσμένα άτομα επιβιώνουν επί μακρόν παράγοντας περισσότερους απογόνους. Μετά την πάροδο εύλογου χρονικού διαστήματος, καλύτερα προσαρμοσμένοι απόγονοι συναθροίζονται σε μεγαλύτερες πυκνότητες στον αρχικό πληθυσμό με αποτέλεσμα η μέση προσαρμοστικότητά του να αυξάνει. Σύμφωνα με τη Δαρβινική θεωρία, αυτή η διαδικασία της

φυσικής επιλογής είναι ουσιαστικά η γενεσιουργός αιτία της εξέλιξης όλων των μορφών ζωής όπως τη γνωρίζουμε σήμερα στον πλανήτη μας. Η προσομοίωση της διαδικασίας αυτής σε τεχνικό επίπεδο περιλαμβάνει άτομα του πληθυσμού ως υποψήφιες λύσεις σε κάποιο πρόβλημα, με το βαθμό προσαρμοστικότητας κάθε μιας να αποτελεί ουσιαστικά το «μέτρο» επιβίωσής της.

Ο γενετικός αλγόριθμος, όπως αναφέρθηκε στα προηγούμενα, εμφανίζεται ως μια διαδικασία εξερεύνησης του χώρου των δυνατών λύσεων για την εύρεση της βέλτιστης, ωστόσο, ενώ παρουσιάζει απλότητα στην εφαρμογή του, μεγάλες δυσκολίες ανακύπτουν όταν πρέπει να κατανοηθεί και να καταγραφεί ο τρόπος με τον οποίο επιτυγχάνονται εντυπωσιακά μερικές φορές αποτελέσματα, κατά την εφαρμογή σε πρακτικά προβλήματα της καθημερινής ζωής. Σύμφωνα με τον Holland [76] η προσαρμογή των ΓΑ οφείλεται στη δράση των δημιουργηθέντων προτύπων μέσα στο γενότυπο του χρωμοσώματος. Σύμφωνα με αυτή την ανάλυση, ο ΓΑ είναι περισσότερο μια διαδικασία εξερεύνησης των χαρακτηριστικών του παρά του ίδιου του γενοτύπου. Ένα «χαρακτηριστικό» καθορίζεται από ένα *σχήμα*[14], δηλαδή μια σειρά γονιδίων με καθορισμένες σταθερές τιμές σε συγκεκριμένες θέσεις. Με αυτές τις προϋποθέσεις, αντικειμενικός σκοπός του ΓΑ πλέον είναι να αυξήσει την «πυκνότητα» των σχημάτων υψηλής προσαρμοστικότητας στον πληθυσμό.

Υπό αυτή την έννοια, ο τρόπος λειτουργίας των ΓΑ έχει αποτελέσει αντικείμενο ενδελεχούς μελέτης και έχουν διατυπωθεί ποικίλες αντικρουόμενες απόψεις. Κάποιοι μελετητές καταλήγουν στο συμπέρασμα ότι οι ΓΑ διασπούν ένα αρχικά πολύπλοκο (ανώτερου βαθμού) πρόβλημα σε πολυάριθμα δομικά στοιχεία (Building Block Hypothesis - BBH) υψηλής συνάφειας όσον αφορά στη λύση του αρχικού προβλήματος, ο συνδυασμός και ανα–συνδυασμός των οποίων τελικά οδηγεί στην εύ-

[14]Ως σχήμα (schema) ορίζεται ένα πρότυπο το οποίο ταυτοποιεί ένα χρωμοσωματικό υποσύνολο με ομοιότητες σε συγκεκριμένες θέσεις γονιδίων. Για παράδειγμα, ας θεωρήσουμε τα χρωμοσώματα 100011 και 111001. Αυτά παρουσιάζουν το σχήμα 1**0*1 το οποίο πιθανόν να έχει αποδειχθεί χρήσιμο σε κάποιο σημείο στο γενετικό αλγόριθμο. Το συγκεκριμένο σχήμα περιγράφει το σύνολο όλων των χρωμοσωμάτων με εύρος 6 γονιδίων, τα οποία έχουν 1 στην πρώτη και έκτη θέση και 0 στην τέταρτη, ενώ οι θέσεις 2, 3 και 5 μπορούν να καταλαμβάνονται είτε από 0 είτε από 1. Οι θέσεις που περιλαμβάνουν σταθερά γονίδια ονομάζονται *χαρακτηριστικές (defining)*, ενώ οι συνδυασμοί τιμών των θέσεων με αστερίσκους παράγουν τους αλληλόμορφους (alleles) του σχήματος. Ως *τάξη (order)* ορίζεται ο αριθμός των χαρακτηριστικών θέσεων του σχήματος (3 στο παράδειγμα), ενώ ως *χαρακτηριστικό εύρος (defining length)* ορίζεται η απόσταση μεταξύ της πρώτης και της τελευταίας χαρακτηριστικής θέσης (5 στο παράδειγμα) του. Ως *βαθμός προσαρμοστικότητας* του σχήματος ορίζεται η μέση προσαρμοστικότητα των αλληλομόρφων του.

ρεση της βέλτιστης δυνατής λύσης [60]. Άλλοι διαφωνούν, υποστηρίζοντας ότι ο τρόπος ανα–συνδυασμού των δομικών στοιχείων είναι τυχαίος και εξαρτάται κάθε φορά από το πρόβλημα [176].

Σύμφωνα με την υπόθεση των δομικών στοιχείων (BBH), μετά από δειγματοληψία επί της γενετικής λίμνης προκύπτει ένας αριθμός περιορισμένου μήκους και χαμηλού βαθμού σχημάτων (schemata) τα οποία ανα-συνδυάζονται κατάλληλα ούτως ώστε να παράξουν χρωμοσώματα όλο και υψηλότερου βαθμού προσαρμοστικότητας [76]. Η απλότητα των σχημάτων είναι βασικά υπεύθυνη για τη μείωση της πολυπλοκότητας, αντικαθιστώντας την αναζήτηση της λύσης ανάμεσα σε πολύπλοκα χρωμοσώματα με τον ανα–συνδυασμό τμημάτων χρωμοσωματικών προτύπων που αναδείχθηκαν ως καταλληλότερα σε προηγούμενες δοκιμές (γενεές) του αλγορίθμου.

Ως παραμετροποίηση του ΓΑ νοείται ο καθορισμός των άριστων τιμών των παραμέτρων του, δηλαδή του άριστου σημείου στο οποίο πρέπει να τείνει ο αλγόριθμος, του τρόπου επιλογής των απογόνων, της πιθανότητας ανα–συνδυασμού και της πιθανότητας μετάλλαξης. Οι παράμετροι αυτές επιδρούν σημαντικά στην απόδοση του αλγορίθμου, εν τούτοις δεν υπάρχει μια ενιαία αποδεκτή μέθοδος για την εύρεση των άριστων κάθε φορά τιμών. Αντίθετα, στις περισσότερες περιπτώσεις, η έρευνα προχωρά με τη μέθοδο δοκιμής/λάθους (trial and error) και εξαρτάται κάθε φορά από τα δεδομένα του προβλήματος. Ο καθορισμός άριστης πιθανότητας ανα–συνδυασμού επηρεάζει την εμφάνιση νέων δομικών στοιχείων, άρα εμπλουτίζει την παραλλακτικότητα της γενετικής λίμνης η οποία μειώνεται όσο η πιθανότητα ανα–συνδυασμού μειώνεται. Υψηλές (πάνω από την άριστη) τιμές της παραμέτρου αυτής δημιουργούν σχάση των βασικών δομικών μονάδων με αποτέλεσμα να δημιουργείται ανάσχεση της σύγκλισης του αλγορίθμου και δημιουργία πολλών πεδίων (plateaus) τοπικών ελαχίστων (ή μεγίστων) στα οποία ο αλγόριθμος είναι δυνατόν να παγιδευτεί.

Όσον αφορά στη μετάλλαξη, αυτή στοχεύει στη δημιουργία νέων δυναμικών και καινοτόμων δομικών στοιχείων με τα οποία εμπλουτίζεται το χρωμόσωμα, με συνέπεια την αύξηση της παραλλακτικότητας. Ως συχνότητα αλληλομορφίας (allele frequency) ορίζεται το ποσοστό των αντιγράφων ενός δομικού στοιχείου που χαρακτηρίζεται από συγκεκριμένη γενετική παραλλακτικότητα. Είναι δηλαδή ο λόγος

του αριθμού των όμοιων προς τους συνολικούς αλληλόμορφους ενός γενετικού τόπου (locus) ενός πληθυσμού. Η συχνότητα αλληλομορφίας απεικονίζει τη γενετική παραλλακτικότητα των ειδών και των ατόμων μέσα στα είδη. Οι αλληλόμορφοι των απογόνων αποτελούν υποσύνολο (δείγμα) των αλληλομόρφων των γονέων, ενώ το ενδεχόμενο επιβίωσης αποφασίζεται τυχαία. Η πιθανότητα μετάλλαξης επηρεάζει στο μέγιστο βαθμό το φαινόμενο της γενετικής παρέκκλισης (genetic drift) μέσω της οποίας διαταράσσεται η ισορροπία με τη μεταβολή της συχνότητας αλληλομορφίας (συχνότητα εμφάνισης της παραλλαγής ενός γονιδίου-αλληλομόρφου), εξαιτίας τυχαίας δειγματοληψίας (random sampling). Συνεπώς, η πιθανότητα μετάλλαξης, όταν λαμβάνει υπερβολικά υψηλές ή χαμηλές τιμές, προκαλεί γενετική παρέκκλιση επηρεάζοντας τη συχνότητα αλληλομορφίας και συχνά αποτελεί αιτία κατάρρευσης της γενετικής παραλλακτικότητας του πληθυσμού.

2.7 Μειονεκτήματα ταξινομητών

Τα διάφορα πρότυπα ταξινομητών υπολογιστικής νοημοσύνης όπως αυτά έχουν μέχρι τώρα προταθεί στη βιβλιογραφία παρουσιάζουν συγκεκριμένα μειονεκτήματα, τα οποία κατά κύριο λόγο εξαρτώνται από τα προς ανάλυση δεδομένα και σε πολλές περιπτώσεις αποτελούν σημαντικό ανασταλτικό της λειτουργίας τους παράγοντα.

- *Δειγματοληψία στο διάνυσμα εισόδου.* Η αποτελεσματικότητα της εκπαίδευσης οποιουδήποτε ταξινομητή εξαρτάται από τα δεδομένα, δηλαδή τα χαρακτηριστικά του διανύσματος εισόδου, τόσο όσον αφορά στο βαθμό διάστασης, όσο και στον θόρυβο των αρχικών δεδομένων. Σε κάθε περίπτωση, η εξομάλυνση του διανύσματος εισόδου ισοδυναμεί με τη μορφοποίηση ενός υποσυνόλου μέσω κατάλληλης επιλογής χαρακτηριστικών ικανών να αντικαταστήσουν χωρίς απώλειες τα αρχικά δεδομένα, διατηρώντας σημαντικό ποσοστό της πρωτογενούς πληροφορίας. Με άλλα λόγια, το ζητούμενο είναι το επιλεγμένο υποσύνολο να διατηρεί πρακτικά αναλλοίωτη τη συμμετοχή του στην αντιστοίχιση των χαρακτηριστικών επί των κλάσεων, προσεγγίζοντας όσο το δυνατόν αυτή του πρωτόλειου συνόλου δεδομένων. Το ζήτημα αυτό, το οποίο αποδεικνύεται ιδιαίτερα σημαντικό για προβλήματα που περιλαμβάνουν χρονοσειρές, αντιμετωπίζεται συνήθως σε προγενέστερο της εκπαίδευσης των ταξινομητών στάδιο, ως μια διαδικασία προ-επεξεργασίας των δεδομένων,

με χρήση συγκεκριμένης ανεξάρτητης μεθόδου, όπως είναι για παράδειγμα η μέθοδος ΑΚΣ. Το πρόβλημα που ανακύπτει στις περιπτώσεις αυτές έγκειται στο ότι η διαδικασία προ-επεξεργασίας των δεδομένων, η οποία στην πραγματικότητα επεμβαίνει στο επίπεδο εισόδου του ταξινομητή, δε συμμετέχει καθ' οιονδήποτε τρόπο στον καθορισμό της τελικής του μορφής. Συνεπώς, ο καθορισμός της αρχιτεκτονικής του εκάστοτε ταξινομητή όσον αφορά στα υπόλοιπα επίπεδα της δομής του ολοκληρώνεται ανεξάρτητα από τη διαδικασία μορφοποίησης του επιπέδου εισόδου του, με αποτέλεσμα σε πλείστες περιπτώσεις να προκύπτουν ανισόρροπα πρότυπα μέτριας δομής και εκπαιδευτικής χωρητικότητας που χαρακτηρίζονται από υποβαθμισμένη απόδοση.

- *Βέλτιστη παραμετροποίηση.* Η αναγνώριση της δομής των διαφόρων ταξινομητών υπολογιστικής νοημοσύνης αποτελεί σημαντικότατη παράμετρο, η βέλτιστη ρύθμιση της οποίας είναι δυνατό να σηματοδοτεί τη διαφορά μεταξύ αποδοτικών και μη ταξινομητών.

  - *Τεχνητά Νευρωνικά Δίκτυα:* Ως αρχιτεκτονική ενός νευρωνικού ταξινομητή νοείται κατά κύριο λόγο η επιλογή του αριθμού των κρυφών επιπέδων και των νευρώνων ανά επίπεδο του δικτύου, ο τρόπος διασύνδεσης των νευρώνων διαφορετικών επιπέδων, η συνάρτηση ενεργοποίησης και η κωδικοποίηση της εξόδου, η παραμετροποίηση των οποίων επηρεάζει άμεσα το είδος (υπο- ή υπερ-εκπαίδευση) και τη χρονική διάρκεια της εκπαίδευσης. Παρότι αυτές οι παράμετροι έχουν κατά καιρούς αντιμετωπιστεί μέσω εμπειρικών θεωρήσεων, η ρύθμισή τους επιτυγχάνεται κατά περίπτωση και κυρίως μέσω της διαδικασίας αποτίμησης δοκιμής-λάθους (trial and error evaluation procedure). Σημειώνεται ότι εξ όσων γνωρίζουμε δεν αναφέρεται στη βιβλιογραφία συγκεκριμένη ολοκληρωμένη μεθοδολογία βελτιστοποίησής τους.
  - *Μηχανές Διανυσμάτων Υποστήριξης:* Οι ΜΔΥ - ή μηχανές πυρήνα - αναφέρονται σε μια νέα σχετικά ομάδα αλγορίθμων που βασίζονται στη θεώρηση ότι αυξάνοντας το βαθμό διάστασης διευκολύνεται η εύρεση διαχωριστικής υπερ-επιφάνειας μεταξύ των δεδομένων εισόδου. Πρόκειται για εξαιρετικά ισχυρούς μη γραμμικούς ταξινομητές, ικανούς να αποδώσουν σε προβλήματα υψηλότατου βαθμού πολυπλοκότητας με αποτέλε-

σμα να χαρακτηρίζονται από υψηλές απαιτήσεις σε υπολογιστική ισχύ και αποθηκευτική ικανότητα για την εκπαίδευση και την αξιολόγηση, ενώ παράλληλα παρουσιάζουν σχετική ευπάθεια στην ύπαρξη εκτροπών (outliers) στα δεδομένα, τα οποία προκαλούν άμεση υποβάθμιση της ικανότητάς τους προς γενίκευση. Τέλος, η απόδοση των ΜΔΥ βαίνει αντιστρόφως ανάλογη προς το συνολικό αριθμό χαρακτηριστικών του διανύσματος εισόδου, αλλά και προς το συνολικό αριθμό των εκπαιδευτικών προτύπων.

– *Γενετικοί Αλγόριθμοι:* Οι ΓΑ αποτελούν μια σχετικά πρόσφατη κλάση αλγορίθμων αναζήτησης και βελτιστοποίησης, που βασίζονται στην αρχή της εξέλιξης των ειδών. Ένας ΓΑ επιτίθεται σε κάποιο πρόβλημα σχηματοποιώντας μια ομάδα πιθανών λύσεων και πραγματοποιεί αναζήτηση για τη βέλτιστη λύση μέσω γενεών του αρχικού πληθυσμού οι οποίες προκύπτουν μετά από προσομοιωμένες διαδικασίες ανα–συνδυασμού, αντιγραφής και μετάλλαξης των ατόμων του πληθυσμού τους. Τα άτομα του πληθυσμού κάθε γενεάς σχηματοποιούνται μέσω κωδικοποίησης με συμβολοσειρές, με αποτέλεσμα η λειτουργία των ΓΑ να εξαρτάται αποκλειστικά και μόνο από την κωδικοποίηση ενός συνόλου τιμών που είναι δυνατό να λάβουν οι μεταβλητές του προβλήματος και όχι από τις μεταβλητές αυτές καθ'αυτές. Στην απλούστερη των περιπτώσεων, κάθε λύση αναπαρίσταται μέσω δυαδικής συμβολοσειράς καθορισμένου μήκους, αν και έχουν προταθεί ποικίλες κωδικοποιήσεις από δειγματικές συμβολοσειρές έως συμβολοσειρές πραγματικών αριθμών. Συνεπώς η εύρεση της καταλληλότερης κωδικοποίησης αποτελεί κρίσιμο βήμα για την επιτυχία του αλγορίθμου. Κάθε γενεά προωθεί στην επόμενη τα πλέον προσαρμοσμένα στοιχεία της δια μέσου ορισμένης αντικειμενικής συνάρτησης[15], η οποία επίσης αποτελεί σημαντικό στοιχείο του αλγορίθμου, καθορίζοντας ουσιαστικά το συνολικό πλαίσιο δράσης του. Οι ΓΑ είναι πολύ ευαίσθητοι και υφίστανται απότομη κατάρρευση στις περιπτώσεις κατά τις οποίες δεν ικανοποιούνται πλήρως οι περιορισμοί στο πεδίο ορισμού της αντικειμενικής τους συνάρτησης.

[15]Επίσης στη βιβλιογραφία η αντικειμενική συνάρτηση (objective function) αναφέρεται και ως συνάρτηση ικανότητας, αξιολόγησης ή καταλληλότητας (fitness ή evaluating function)

- *Προσαρμογή στα δεδομένα:* Η επιτυχία στη διακριτική ικανότητα κάθε ταξινομητή εξαρτάται από τη δυνατότητά του να προσαρμόζεται καλά στα δεδομένα του προβλήματος. Η προσαρμογή αυτή βεβαίως εξαρτάται πρωτίστως από την ποιότητα των δεδομένων όπως αναλύθηκε προηγουμένως, αλλά επίσης εξαρτάται και από τις δυνατότητες του ταξινομητή. Επειδή δεν είναι δυνατό να ορισθεί εκ των προτέρων το σημείο του γενικού ελαχίστου (global minimum) η εκπαίδευση των ταξινομητών είναι στην ουσία μια εξερεύνηση της επιφάνειας του σφάλματος (error surface). Εκκινώντας από μια αρχική, τυχαιοποιημένη ρύθμιση του πίνακα βαρών και των τιμών κατωφλίου, τουτέστιν από ένα τυχαίο σημείο στην επιφάνεια του σφάλματος, οι αλγόριθμοι προοδευτικά αναζητούν το γενικό ελάχιστο της επιφάνειας αυτής,[16] η οποία είναι στις περισσότερες περιπτώσεις πολύπλοκη και χαρακτηρίζεται από μια πλειάδα σημείων παγίδευσης[17] που δυσκολεύουν τη δραπέτευση του αλγορίθμου, συνιστώντας ένα από τα μεγαλύτερα προβλήματα στους αλγορίθμους παράλληλης επεξεργασίας.

[16] δηλαδή τη βέλτιστη λύση στο πρόβλημα.

[17] Ως *σημεία παγίδευσης* του αλγορίθμου νοούνται περιοχές τοπικών ελάχιστων, δηλαδή κυρτώσεις ή απότομες σχισμές στην επιφάνεια του σφάλματος, οι οποίες κείνται χαμηλότερα από τις περιβάλλουσες αυτών περιοχές, αλλά σε σύγκριση με το γενικό ελάχιστο βρίσκονται πάντα σε υψηλότερα επίπεδα.

Κεφάλαιο 3
ΠΛΑΙΣΙΟ ΤΗΣ ΕΡΕΥΝΑΣ

Στο σημείο αυτό παρουσιάζεται το γενικότερο περιβάλλον της παρούσης έρευνας, το πλαίσιο στο οποίο αναπτύχθηκε η μέθοδος της Πρότυπης Εξελικτικής Τμηματοποίησης, καθώς επίσης και το εργαλείο λογισμικού που την υλοποιεί. Συγκεκριμένα αναφέρονται βασικές μέθοδοι αναπαράστασης των χρονοσειρών και τα κυριότερα προβλήματα που ανακύπτουν κατά τη χρήση τους, ενώ περιγράφεται η συμβολή της προτεινόμενης μεθόδου στην επίλυσή τους. Γίνεται μια πρώτη αναφορά στην εφαρμογή της σε δύο διαφορετικές μελέτες περίπτωσης, ενώ παράλληλα περιγράφεται το ερευνητικό περιβάλλον με την έννοια της προγενέστερης έρευνας σε παρόμοια προβλήματα. Τέλος, δίνονται οι βασικές προδιαγραφές του συστήματος που αναπτύχθηκε και οι κυριότεροι στόχοι που τίθενται.

Η αποτελεσματική εξαγωγή χαρακτηριστικών από δεδομένα χρονοσειρών πραγματικών αριθμών - και τελικά η αναπαράστασή τους σε δειγματικούς χώρους χαμηλότερης διάστασης - αποτελεί το θεμέλιο λίθο για την επιτυχή επίλυση ποικίλων προβλημάτων ταξινόμησης ή πρόβλεψης και το επίκεντρο της παρούσας μελέτης. Ως απόρροια ποικίλων τροποποιημένων μεθόδων τμηματοποίησης των χρονοσειρών, η προτεινόμενη μέθοδος διατηρεί ανέπαφη την κρίσιμη πληροφορία, ενώ αποκαλύπτει και απορρίπτει μη συστημικά συστατικά της χρονοσειράς. Με τον

τρόπο αυτό βελτιστοποιείται η διακριτική ικανότητα του μηχανισμού ταξινόμησης ο οποίος προκύπτει μετά την αξιολόγηση ενός πλήθους υποψήφιων λύσεων και την επιλογή του πιο αποδοτικού. Οι εν λόγω λύσεις προκύπτουν ως αποτέλεσμα μιας εξελικτικής διαδικασίας του μηχανισμού τμηματοποίησης μέσω της οποίας η διαθέσιμη γενετική δεξαμενή πληρούται με πιθανά κατάλληλα γενετικά δομικά στοιχεία, ο γενετικός συνδυασμός των οποίων έχει ως αποτέλεσμα τελικά τη μορφοποίηση ενός άριστου συνόλου δεδομένων. Βάσεις δεδομένων όπου διατηρούνται ιστορικά και χρονικά στοιχεία αποτελούν αντικείμενα - στόχους του αναπτυγμένου συστήματος, είτε η ανάλυση των ζητούμενων χρονοσειρών αντιστοιχεί σε προβλήματα ταξινόμησης, είτε σε προβλήματα πρόβλεψης. Συγκεκριμένα, προβλήματα βιολογικού, φυσιολογικού και περιβαλλοντικού ενδιαφέροντος, όπου περιλαμβάνονται τοπογραφικά, μετεωρολογικά, κλιματολογικά και άλλα δεδομένα προερχόμενα από αισθητήρες, τα οποία χαρακτηρίζονται από τη χρονική διάσταση των τιμών με τη μορφή χρονοσειρών, θα ωφεληθούν σημαντικά από την αναπαράσταση της αρχικής πληροφορίας σε μειωμένη διάσταση. Επίσης, η εξελικτική παραγωγή σειράς δευτερογενών δεδομένων εξειδικευμένων στην εκπαίδευση συγκεκριμένων ταξινομητών, έχει θετική επίδραση στην απόδοσή τους και μορφοποιεί ισχυρά αντικείμενα αυξημένης διακριτικής ικανότητας είτε απαιτείται ταξινόμηση, είτε πρόβλεψη μελλοντικής κατάστασης ενός συγκεκριμένου φαινομένου.

Η ανάλυση χρονοσειρών στοχεύει στην επίλυση προβλημάτων που βασικά εμπίπτουν στις δύο προαναφερόμενες κατηγορίες, αμφότερες των οποίων αποσκοπούν στην κατανόηση των κινητηρίων δυνάμεων που δημιούργησαν τα πρωτογενή δεδομένα. Στις περισσότερες περιπτώσεις, δυο βασικοί παράγοντες αποτελούν τροχοπέδη: η διαστατικότητα του διανύσματος εισόδου και η ύπαρξη θορύβου. Συνεπώς, κατά την ανάλυση τα δεδομένα που εμπλέκονται θεωρούνται ως αποτελούμενα από δυο συστατικά: ένα συστημικό αναγνωρίσιμο διάνυσμα και μη συστημικό τυχαίο θόρυβο με τη μορφή συστηματικού σφάλματος [23]. Οι διαδικασίες προ-επεξεργασίας της χρονοσειράς στοχεύουν στην ελαχιστοποίηση του συστημικού σφάλματος και την υποβάθμιση του βαθμού διάστασης μέσω της εφαρμογής κάποιας τεχνικής εξομάλυνσης. Η προτεινόμενη μέθοδος της Πρότυπης Εξελικτικής Τμηματοποίησης (ΠΕΤ) (PES: Piecewise Evolutionary Segmentation) αξιοποιεί μια εξελικτική μέθοδο τμηματοποποίησης που οδηγεί στην παραγωγή ενός μεγάλου αριθμού δευτερογενών δεδομένων για τη βέλτιστη εκπαίδευση του ταξινομητή.

3.1 Περιορισμοί αναπαράστασης και αντιμετώπιση μέ-σω της ΠΕΤ

Οι μελετητές που ασχολούνται με τη διερευνητική ανάλυση δεδομένων χρονοσειρών έχουν πρόσφατα στρέψει την προσοχή τους στην εκμετάλλευση ενός μεγάλου αριθμού εργαλείων τεχνητής νοημοσύνης, όπως ΤΝΔ και ΓΑ, η άριστη παραμετροποίηση των οποίων σε πολλές περιπτώσεις απαιτεί τη χρήση μεθόδων προεπεξεργασίας της αρχικής πληροφορίας για τον καθορισμό του διανύσματος εισόδου [114]. Το πρώτιστο βήμα στην ανάλυση αυτού του είδους είναι η εξομάλυνση του σφάλματος, διαδικασία που υλοποιείται με τη χρήση ποικίλων τεχνικών με σκοπό την πρόκληση αμοιβαίας κατάργησης των μη συστημικών συστατικών που εμπεριέχονται στις παρατηρήσεις. Η εφαρμογή τέτοιου είδους διηθητικών τεχνικών είναι δυνατόν να απομακρύνουν τυχαία παραλλακτικότητα και να αποκαλύψουν τάσεις και κυκλικά συστατικά (cyclic components) της σειράς, υπό την προϋπόθεση ότι η παραμετροποίησή τους έχει διεξαχθεί με τρόπο κατάλληλο [50]. Οι πλέον κοινές μέθοδοι εξομάλυνσης περιλαμβάνουν κάποιο είδος δειγματοληψίας, καθώς επίσης και ποικίλες εκθετικές μεθόδους με την εφαρμογή αλγορίθμων απλών ή κινούμενων μέσων για την αντιμετώπιση ποικίλων τάσεων στα αρχικά δεδομένα.

Σύμφωνα με την τεχνική του κινούμενου μέσου, κάθε στοιχείο της χρονοσειράς αντικαθίσταται είτε από τον απλό είτε από το σταθμικό μέσο (weighted average) *n* στοιχείων, όπου το μέγεθος *n* αντιστοιχεί στο πλάτος του κινούμενου παραθύρου [23, 179]. Όσον αφορά στη διαδικασία δειγματοληψίας, η χρονοσειρά μετατρέπεται σε νέα σειρά αποτελούμενη από το *n-οστό* στοιχείο του παραθύρου καθώς αυτό διολισθαίνει επί της αρχικής, καθιστώντας στην περίπτωση αυτή το μέγεθος *n* ως το ρυθμό δειγματοληψίας των αρχικών δεδομένων. Φυσικά, η έννοια του διαμέσου μπορεί να αντικαταστήσει το μέσο στην πρώτη περίπτωση, εάν ενδιαφερόμαστε για ανθεκτικότητα έναντι της επίδρασης ακρότατων τιμών (outliers). Πάντως η χρήση διαμέσων παράγει σε κάθε περίπτωση πιο ακανόνιστες γραμμές απ' ότι οι αντίστοιχοι μέσοι. Σε αμφότερες τις περιπτώσεις το εύρος του κινούμενου παραθύρου και ο ρυθμός δειγματοληψίας παίζουν το σημαντικότερο ρόλο όχι μόνο για την ακρίβεια της διαδικασίας εξομάλυνσης, αλλά επίσης και για τη συνολική ποσότητα πληροφορίας που θα απορριφθεί. Εάν αυτό ορισθεί υπερβολικά υψηλό ή χαμηλό,

η αναπαράσταση της αρχικής χρονοσειράς υποβαθμίζεται δραστικά εξαιτίας σημαντικής απώλειας ζωτικής πληροφορίας και ανεπαρκούς μείωσης του θορύβου αντίστοιχα. Μια άλλη παράμετρος που δημιουργεί προβλήματα στην ποιότητα αναπαράστασης των αρχικών δεδομένων έγκειται στη διαδικασία παραμετροποίησης των ποικίλων αλγορίθμων προ-επεξεργασίας της αρχικής πληροφορίας, η οποία γίνεται κατά βάση με στατικό τρόπο με διαδικασίες δηλαδή δοκιμής και λάθους, με αποτέλεσμα να μορφοποιούνται τελικά μέθοδοι εξαιρετικά αυστηρές (strict) όσον αφορά στην αξιολόγηση της πληροφορίας και την προσαρμογή στο υποκείμενο σύνολο δεδομένων.

Τα κυριότερα προβλήματα που προκύπτουν κατά την εφαρμογή του αλγορίθμου διολισθαίνοντος παραθύρου είναι αφενός η αδυναμία του να προβλέψει αποτελεσματικά ποικίλες καταστάσεις και αφετέρου το γεγονός ότι δέχεται ένα μικρό ποσοστό της πληροφορίας κάθε φορά. Το γεγονός αυτό έχει ως αποτέλεσμα να μην υπάρχει πλήρης επισκόπηση ολόκληρης της σειράς των αρχικών δεδομένων, με αποτέλεσμα στις περισσότερες περιπτώσεις να προκύπτει χαμηλής ποιότητας αναπαράσταση. Εναλλακτικοί αλγόριθμοι[1] παρουσιάζουν καλύτερη ποιότητα αναπαράστασης, αλλά είναι κατά βάση αλγόριθμοι δέσμης και απαιτούν τη σάρωση ολόκληρης της χρονοσειράς, πρακτικά ανεφάρμοστη στις περισσότερες περιπτώσεις εξόρυξης δεδομένων μεγάλου όγκου ή συνεχούς ενημέρωσης.

Συνεπώς, δεδομένων των ανεπαρκειών που περιγράφηκαν παραπάνω και εντοπίζοντας το πρόβλημα της ανάλυσης αφενός στην ποιότητα της αναπαράστασης, καθώς επίσης και στην παραμετροποίηση των εργαλείων προ-επεξεργασίας, η παρούσα εργασία εισάγει την έννοια της πρότυπης εξελικτικής τμηματοποίησης. Πρόκειται για μια μέθοδο κατά την οποία ο αλγόριθμος τμηματοποίησης διαπερνά την αρχική χρονοσειρά εκμεταλλευόμενος μια δυναμική εξελικτική διαδικασία τμηματοποίησης, η οποία προσαρμόζει το εύρος του διολισθαίνοντος παραθύρου επί του υποκείμενου συνόλου των αρχικών χρονοσειριακών δεδομένων. Τα παραγόμενα δεδομένα χρησιμοποιούνται ως επιμορφωτικό υλικό έναντι ποικίλων ταξινομητών υπό εκπαίδευση και αξιολογούνται βάσει της απόδοσης των τελευταίων ώστε να προκριθούν τα καταλληλότερα. Μέσω της μεθόδου αυτής σχηματοποιούνται ισχυροί ταξινομητές, άριστα παραμετροποιημένοι και προσαρμοσμένοι κατάλληλα στα

[1] όπως για παράδειγμα ο αλγόριθμος από-πάνω-προς-τα-κάτω, καθώς επίσης και ο από-κάτω-προς-τα-πάνω

δεδομένα, ενώ παράλληλα βελτιστοποιείται η αναπαράσταση της αρχικής χρονοσειράς τόσο όσον αφορά στην εξόρυξη των πλέον κρίσιμων χαρακτηριστικών, όσο και στην πιστότητα των δεδομένων που προκύπτουν.

3.2 Μελέτες περίπτωσης και εφαρμογή της ΠΕΤ

Στην παρούσα διατριβή μελετάται το πρόβλημα της αναπαράστασης των χρονοσειρών μέσω της ανάπτυξης νέων προτύπων εξελικτικής εξαγωγής χαρακτηριστικών από τα πρωτογενή δεδομένα, ενώ περαιτέρω η μεθοδολογία που αναπτύχθηκε εφαρμόζεται σε δύο πραγματικά προβλήματα. Τα κριτήρια που ακολουθήθηκαν για την επιλογή των μελετών αυτών ήταν κατ' αρχήν η φύση του προβλήματος το οποίο ήταν επιθυμητό να άπτεται του γεωτεχνικού χώρου. Σε δεύτερο επίπεδο, οι περιπτώσεις που μελετήθηκαν θα έπρεπε όχι μόνο να περιλαμβάνουν δεδομένα χρονοσειρών, αλλά επίσης τα δεδομένα αυτά να είναι υψηλού βαθμού διάστασης, ώστε να δοκιμασθεί η αποτελεσματικότητα της μεθόδου αφενός στη δραστική μείωση αυτής και αφετέρου στην ποιότητα της τελικής αναπαράστασης των δεδομένων. Τέλος, επιλέχθηκε ένα πρόβλημα ταξινόμησης και ένα πρόβλεψης της πορείας ενός φαινομένου, με σκοπό το σύστημα να δοκιμασθεί σε όλο το εύρος του φάσματος των τυπικών προβλημάτων που συναντώνται.

Στην πρώτη περίπτωση ανήκει το πρόβλημα της ταξινόμησης φυτικών ιών, κατά το οποίο τα δεδομένα προέρχονται από την αντίδραση βιοαισθητήρων με συγκεκριμένους παθογόνους μικροοργανισμούς. Το αποτέλεσμα της αντίδρασης μετράται ως διαφορά ηλεκτρικής τάσης στη μονάδα του χρόνου, είναι μια πεπερασμένη χρονοσειρά, χαρακτηριστική για κάθε ιό, εναπόκειται δε σε ειδικούς η αναγνώριση του ιού από τις χαρακτηριστικές της καμπύλες αναπαράστης. Στη δεύτερη περίπτωση αξιολογείται η χειμαρρική επικινδυνότητα ορεινών περιοχών με την ανάπτυξη συστήματος πρόβλεψης συγκεκριμένου περιβαλλοντικού δείκτη που την επηρεάζει. Τα δεδομένα που συγκεντρώθηκαν αφορούν σε δομικά στοιχεία διαφόρων λεκανών απορροής, καθώς επίσης και σε μεταβλητές περιβάλλοντος, από τις οποίες το ύψος της μηνιαίας βροχόπτωσης θεωρείται ο πλέον σημαντικός παράγοντας. Για το λόγο αυτό δεδομένα μηνιαίας βροχόπτωσης που ελήφθησαν ως χρονοσειρές και καλύπτουν μια ευρεία χρονική περίοδο, αποτέλεσαν το διάνυσμα εισόδου της ΠΕΤ.

3.3 Ερευνητικό περιβάλλον

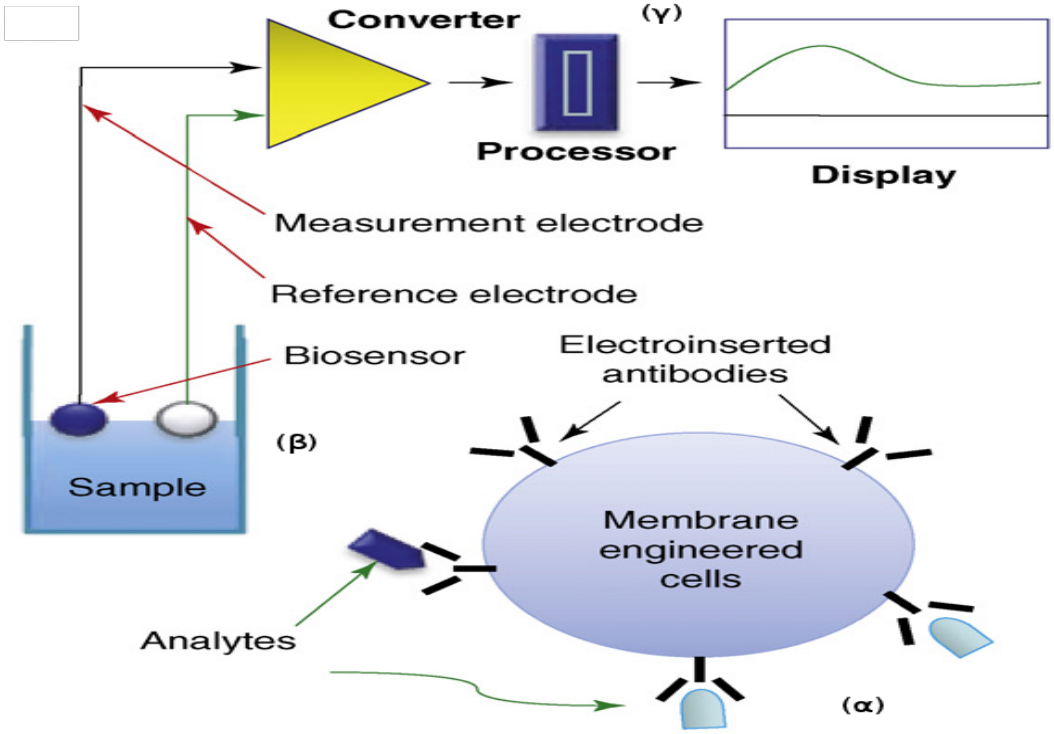
Σχεδιάζοντας τον πυρήνα του συστήματος, σχεδόν αμέσως εμφανίσθηκε η αδήριτη ανάγκη ενός συνδυασμού λειτουργιών. Αφενός στο κεντρικό αυτό σύστημα αφικνούνται δευτερογενή εξελικτικά δεδομένα προερχόμενα από συγκεκριμένο χρωμόσωμα του περιφερειακού γενετικού αλγορίθμου, στο οποίο απαιτείται να αντιστοιχισθεί μια ποσοτικοποιημένη αξιολόγηση. Αφετέρου, αφού το σύστημα φθάσει στο τελικό συμπέρασμα περί του τρόπου αναπαράστασης των αρχικών δεδομένων, είναι απαραίτητο να επιλεγεί ο πιο προσαρμοσμένος στα δεδομένα ταξινομητής. Συνεπώς, οι ταξινομητές που υλοποιούν τον κεντρικό πυρήνα υλοποιούν διττή λειτουργία: όχι μόνο αποδίδουν συγκεκριμένο βαθμό προσαρμοστικότητας σε κάθε εκπαιδευτή, αλλά επίσης διαγκωνίζονται για την επικράτηση και την ανάληψη της ταξινόμησης από έναν εξ' αυτών σε κάθε πρόβλημα στο οποίο επιτίθεται το σύστημα. Κατόπιν εμπεριστατωμένης μελέτης στη διεθνή βιβλιογραφία, προκρίθηκε η χρήση ΤΝΔ και ΜΔΥ, ως των πλέον κοινώς χρησιμοποιούμενων ταξινομητών για τα προβλήματα που αντιμετωπίσθηκαν. Ενδεικτικά ακολουθεί επισκόπηση της μεθοδολογίας παραγωγής δεδομένων ταυτοποίησης φυτικών ιών μέσω της μεθόδου BERA, καθώς επίσης και ανασκόπηση της υλοποίησης ΤΝΔ και γενετικών αλγορίθμων στη διαχείριση υδατικών διαθεσίμων.

3.3.1 Βιοηλεκτρική Δοκιμή Αναγνώρισης BERA

Η μέθοδος της βιοηλεκτρικής δοκιμής αναγνώρισης (BERA: Bioelectric Recognition Assay) αποτελεί ένα σύστημα μέτρησης της μεταβολής του ηλεκτρικού δυναμικού σε επίπεδο ιστού, με σκοπό την ανίχνευση ιών ή άλλων βιοενεργών ουσιών. Πρόκειται για μια καινοτόμο μέθοδο, η οποία έχει αναπτυχθεί τα τελευταία χρόνια [99, 98] και βασίζεται στην παρακολούθηση και καταμέτρηση των μεταβολών των ηλεκτρικών ιδιοτήτων ομάδας κυττάρων κατάλληλα ακινητοποιημένων εντός πήγματος, έτσι ώστε να διατηρούνται οι φυσιολογικές κυτταρικές λειτουργίες κατά τη διάρκεια της αλληλεπίδρασης με τα υπό ανίχνευση αντικείμενα. Το σύστημα των ακινητοποιημένων κυττάρων μέσω του οποίου γίνεται η διάγνωση, αποτελεί ουσιαστικά ένα πλέγμα βιο-αισθητήρων.

Κατά τη διάρκεια των τελευταίων ετών σημειώνεται σημαντική αύξηση στην εφαρμογή διαγνωστικών συστημάτων που χρησιμοποιούν βιο-αισθητήρες, η λει-

τουργία των οποίων, στην πλειονότητά τους, βασίζεται στην έμμεση μέτρηση βιοχημικών προτύπων ενζύμων ή μορίων αντισωμάτων. Με τον όρο «βιο-αισθητήρας» αποδίδονται συσκευές οι οποίες σε αδρές γραμμές μετατρέπουν κάποια βιολογική διαδικασία σε μετρήσιμο, ηλεκτρικό συνήθως, σήμα. Υπό μια άλλη έννοια, πρόκειται για συσκευές μέσω των οποίων καθίσταται δυνατή η ανίχνευση μιας χημικής ένωσης και η μετατροπή της αντίδρασής της με γνωστές ενώσεις σε ηλεκτρικό σήμα, στη μονάδα του χρόνου.



Σχήμα 3.1: *Σχηματική παράσταση ενός τυπικού αισθητήρα BERA. Διακρίνονται (α) το βιολογικό τμήμα το οποίο έχει υποστεί επεξεργασία για την ενσωμάτωση μέσω μεθόδων ηλεκτροεισαγωγής αντισωμάτων ή άλλων μορίων, (β) το φυσικοχημικό τμήμα ανίχνευσης με τα ηλεκτρόδια μέτρησης και τον μετατροπέα, (γ) το σύστημα υπολογισμών και αποθήκευσης ([15]).*

Στην τυπική του μορφή (Εικ. 3.1), ένας βιο-αισθητήρας της μεθόδου βιοηλεκτρικής δοκιμής αναγνώρισης διακρίνεται σε τρία μέρη [99]:

- το βιολογικό τμήμα, στο οποίο περιλαμβάνονται όλα τα βιολογικά συστατικά του αισθητήρα όπως κύτταρα, ένζυμα ή αντισώματα

- το φυσικοχημικό τμήμα ανίχνευσης. Στο τμήμα αυτό επιτυγχάνεται η μετατροπή της αντίδρασης του βιολογικού τμήματος του αισθητήρα με το υπό εξέταση οργανικό δείγμα σε ποσοτικοποιημένο σήμα στη μονάδα του χρόνου

- το τμήμα υπολογισμών και αποθήκευσης όπου διενεργείται η αποθήκευση και η αξιολόγηση του σήματος που παράγεται από το τμήμα ανίχνευσης

Οι διάφοροι τύποι των βιο-αισθητήρων καθορίζονται ουσιαστικά από τον τύπο του βιολογικού τους τμήματος, ανάλογα με το είδος του κυττάρου που χρησιμοποιείται. Υπό το πρίσμα αυτό διακρίνονται δύο είδη βιο-αισθητήρων, οι μικροβιακοί και οι αισθητήρες ιστών η ολοκληρωμένων κυττάρων. Στην πρώτη περίπτωση χρησιμοποιούνται ζύμες ή βακτήρια με στόχο την παραγωγή ενζύμου, ενώ στη δεύτερη μεμονωμένα κύτταρα φυτικών ή ζωικών ιστών. Σε αμφότερες τις περιπτώσεις τα κύτταρα ακινητοποιούνται σε εξειδικευμένο υλικό για τη δημιουργία ευνοϊκού κυτταρικού περιβάλλοντος, όπως άγαρ, ή αλγινικό ασβέστιο το οποίο πλεονεκτεί έναντι του πρώτου όσον αφορά στην ποιότητα του περιβάλλοντος που δημιουργεί [99].

Συσκευές που εκμεταλλεύονται ζώντα κύτταρα προς την κατεύθυνση ανίχνευσης διαφόρων βιοενεργών ουσιών πλεονεκτούν κυρίως λόγω

- της ταχύτητας εφαρμογής, καθώς η διάρκεια της ανάλυσης δεν ξεπερνά τα έξι λεπτά ανά δείγμα στην περίπτωση των παθογόνων ιών

- της υψηλής ευαισθησίας

- της αυξημένης αποδοτικότητας, εφόσον παρέχει τη δυνατότητα πραγματοποίησης έως και 1.000 αναλύσεων ανά ημέρα για κάθε τεχνικό

- του χαμηλού σχετικά κόστους ανάλυσης και

- των δυνατοτήτων του εξοπλισμού για μεταφορά και εγκατάσταση σε οποιαδήποτε περιοχή.

Η λειτουργία του βιο-αισθητήρα έγκειται στη δημιουργία χημικής αντίδρασης του προς εξέταση δείγματος με το κυτταρικό σύνολο το οποίο περιέχει. Το γεγονός ότι η αντίδραση αυτή διενεργείται υπό φυσικές βιολογικές συνθήκες αυξάνει τη σταθερότητά του, αλλά και την ακρίβεια της μέτρησης. Χαρακτηριστική ιδιότητα

για κάθε βιο-αισθητήρα αποτελεί η ευαισθησία του έναντι του βιολογικού παράγοντα τον οποίο καλείται να ανιχνεύσει και η οποία εξαρτάται από το είδος των χρησιμοποιούμενων κυττάρων, καθώς επίσης και από αυτόν τον ίδιο τον υπό εξέταση παράγοντα, ιδιότητα που καθορίζει την καταλληλότητα ανά περίπτωση. Κάθε τύπος αισθητήρα έχει συγκεκριμένη διάρκεια ζωής, η οποία καθορίζεται από τη διάρκεια ζωής των κυττάρων του, αλλά και τις συνθήκες αποθήκευσης, όπως η θερμοκρασία του περιβάλλοντος και το pH, τα οποία πρέπει να κυμαίνονται εντός εύρους συγκεκριμένων ορίων για κάθε τύπο αισθητήρα. Επίσης κατά τη διάρκεια της αποθήκευσης, κρίσιμη θεωρείται η τροφοδοσία των κυττάρων του αισθητήρα με θρεπτικά στοιχεία.

Πρόκειται για μια καινοτόμο μέθοδο ποσοτικοποίησης, με τη μορφή ηλεκτρικής απόκρισης, της αντίδρασης καλλιέργειας κυττάρων ακινητοποιημένων σε κατάλληλο πήγμα με ποικίλες βιοενεργές ουσίες, οι οποίες προσκολλώνται στο αντιδραστήριο και επηρεάζουν τη φυσιολογία του. Η μέθοδος έχει εφαρμοσθεί με επιτυχία στη διενέργεια γρήγορων και φθηνών δοκιμών στην περίπτωση ιών φυτικού και ζωικού κυττάρου, καθώς επίσης και στην ανίχνευση της υπολειμματικότητας διαφόρων φυτοφαρμάκων.

Οι βιοαισθητήρες της μεθόδου εξελίχθηκαν διαδοχικά μέσα από μια σειρά βελτιώσεων στην πέμπτη γενεά, η οποία περιλαμβάνει εξαιρετικά μικροσκοπικούς αισθητήρες που χαρακτηρίζονται από υψηλή ταχύτητα παραγωγής και χαμηλό κόστος, ενώ η διάρκεια της δοκιμής έχει μειωθεί σημαντικά[2]. Οι αισθητήρες έκτης γενεάς περιλαμβάνουν εξειδικευμένα αντισώματα στις μεμβράνες τους.

Τα σημαντικότερα προβλήματα στα οποία έχει εφαρμοσθεί η μέθοδος BERA εντοπίζονται στην ανίχνευση φυτικών ιών και στην υπολειμματικότητα φυτοφαρμάκων. Στην παρούσα εργασία, δεδομένα προερχόμενα από βιο-αισθητήρες BERA για την ανίχνευση φυτικών ιών χρησιμοποιήθηκαν ως δεδομένα εισόδου της προτεινόμενης μεθόδου Πρότυπης Εξελικτικής Τμηματοποίησης για εξαγωγή των κρισιμότερων χαρακτηριστικών και στην συνέχεια την ταυτοποίηση καθενός από τους ιούς CGMMV και TRV. Οι εν λόγω ιοί επιδεικνύουν μοναδιαία πρότυπα κατά την αντίδρασή τους με τα εξειδικευμένα κύτταρα των βιο-αισθητήρων, τα οποία είναι

[2] Στις αρχικές γενεές μια τυπική δοκιμή αναγνώρισης διαρκούσε περίπου σαράντα δευτερόλεπτα, ενώ η διάρκεια του απαιτούμενου χρόνου μειώθηκε σε μόλις δύο δευτερόλεπτα στις νεώτερες γενεές βιοαισθητήρων

χαρακτηριστικά για κάθε ιό και τον ταυτοποιούν. Με άλλα λόγια, κατά την αντίδραση σύμφωνα με τη μέθοδο BERA, σε κάθε ιό αντιστοιχεί μια χαρακτηριστική χρονοσειρά, η οποία μπορεί να λάβει τη μορφή μιας χαρακτηριστικής καμπύλης - ψηφιακής υπογραφής, όπου οι μονάδες στον άξονα τετμημένων αντιστοιχούν στο χρόνο, ενώ ο άξονας τεταγμένων αντιπροσωπεύει μετρήσεις διαφοράς δυναμικού του βιο-αισθητήρα.

3.3.2 Διαχείριση χειμαρρικής επικινδυνότητας

Σε γενικές γραμμές είναι τεκμηριωμένη η χαρακτηριστική ανοχή που παρουσιάζουν τα ΤΝΔ στο σφάλμα ταξινόμησης ή πρόβλεψης (fault tolerance), καθώς επίσης και η αυξημένη ικανότητα επίλυσης μη γραμμικών προβλημάτων, γεγονός το οποίο προσελκύει μεγάλο αριθμό ερευνητών που ασχολούνται με τη διαχείριση υδάτινων πόρων. Η ανάγκη μετάβασης σε ισχυρότερα ερευνητικά εργαλεία εν πολλοίς υπαγορεύεται και από τις δραματικές κλιματικές αλλαγές που συμβαίνουν τα τελευταία χρόνια και θεωρούνται υπεύθυνες για σημαντικές καταστροφές σε όλο τον κόσμο. Στο [19] οι Bodri και Cermak ανέπτυξαν μια μέθοδο ΤΝΔ πρόβλεψης πλημμυρικών καταστάσεων στη Μοράβια, στο ανατολικό τμήμα της Δημοκρατίας της Τσεχίας. Χρησιμοποιήθηκαν δεδομένα επαναλαμβανόμενων πλημμύρων της περιοχής η οποία είχε αντιμετωπίσει πολύ σημαντικό πρόβλημα κατά τη διάρκεια των σημαντικών Ευρωπαϊκών πλημμύρων του 1997. Οι Toth κ.α. [177] σύγκριναν την ακρίβεια των βραχυπρόθεσμων προβλέψεων βροχόπτωσης που σημείωσαν διάφορες τεχνικές ανάλυσης χρονοσειρών, χρησιμοποιώντας μόνο ιστορικά στοιχεία ύψους βροχής ως εισαγωγή. Στη μελέτη τους οι ερευνητές συμπεριέλαβαν πρότυπα της γραμμικής στοχαστικής μεθόδου του ολισθαίνοντος μέσου (moving average), ΤΝΔ και τη μη παραμετρική μέθοδο του εγγύτερου γείτονα (nearest neighbour), συμπεραίνοντας ότι η ανάλυση της χρονοσειράς με τη μέθοδο ΤΝΔ παρείχε σημαντικές βελτιώσεις στην ακρίβεια της πρόβλεψης πλημμύρων. Οι Wei κ.α. [182] ανέπτυξαν ένα παρόμοιο σύστημα για την Κίνα, με τη διαφορά ότι στην περίπτωση αυτή το πρότυπο παρέχει τη δυνατότητα χειρισμού δεδομένων χρονοσειρών με αρκετά σφάλματα.

Οι Ni και Xue [134] επίσης δραστηριοποιήθηκαν προς την κατεύθυνση της πρόβλεψης πλημμυρικών φαινομένων στην κοίτη και τις όχθες του ποταμού Yangtze της Κίνας, περιοχή η οποία έχει υποστεί σημαντική περιβαλλοντική υποβάθμιση εξαι-

τίας της εντατικοποιημένης ανθρώπινης δραστηριοποίησης τα τελευταία χρόνια. Το μοντέλο που σχεδίασαν χρησιμοποιούσε ένα RBF ΤΝΔ και εφαρμόσθηκε αφενός στην ακριβή πρόβλεψη της επικινδυνότητας και αφετέρου στην κατάταξη των περιοχών σε πέντε επίπεδα ασφάλειας. Οι Filho και dos Santos σύγκριναν τις αποδόσεις ΤΝΔ και πολυ-παραγοντικών μοντέλων auto – regression [58]. Το σύστημα που ανέπτυξαν τροφοδοτείται με δεδομένα χρονοσειρών για την πρόβλεψη της ροής του ρεύματος στη λεκάνη απορροής του ποταμού Tamanduatei στην περιοχή του Sao Paulo της Βραζιλίας. Η έρευνα καταδεικνύει την υπεροχή του ΤΝΔ έναντι της μεθόδου auto – regression όσον αφορά στην πρόβλεψη ακαριαίων πλημμύρων (flash flood) στην περιοχή. Οι Harpham και Dawson [73] μελέτησαν τη διακύμανση στο σφάλμα της δοκιμής (test set error) μεταξύ έξι διαφορετικών RBF ΤΝΔ. Οι δοκιμές διενεργήθηκαν σε ποικίλες χρονοσειρές πρόβλεψης πλημμύρων στις περιοχές των ποταμών Amber και Mole στο Ηνωμένο Βασίλειο. Οι Kerh και Lee [94] μελέτησαν μια προσέγγιση ΤΝΔ για την πρόβλεψη πλημμυρικής αποφόρτισης στα κατάντη του ποταμού Kaoping. Η είσοδος στο σύστημα προήλθε από μια σειρά σταθμών στα ανάντη του ποταμού, ενώ δεν υπήρχε καμιά πληροφορία στην περιοχή της πλημμύρας. Η εν λόγω έρευνα επιβεβαίωσε την υπόθεση ότι το πρότυπο ΤΝΔ που αναπτύχθηκε αποδίδει καλύτερα από την παραδοσιακή μέθοδο Muskingum που χρησιμοποιείται ευρέως για την καταγραφή της πλημμυρικής πορείας σε φυσικά κανάλια και ποταμούς. Παρόμοια μοντέλα έχουν αναπτυχθεί από διάφορους άλλους ερευνητές [152, 159] προς την κατεύθυνση της πρόβλεψης της χειμαρρικής επικινδυνότητας σε ποικίλες λεκάνες απορροής σε όλο τον κόσμο, συμπεραίνοντας ότι η μέθοδος ΤΝΔ είναι πολλά υποσχόμενη. Οι Jain και Kumar [85] προτείνουν ένα υβριδικό μοντέλο συνδυάζοντας συμβατικές στατιστικές μεθόδους με μοντέλα ΤΝΔ για την πρόβλεψη πλημμύρων με βάση μηνιαίες χρονοσειρές ροής του ποταμού Colorado στην περιοχή Lees Ferry των Η.Π.Α.

3.3.3 Εξελικτικοί αλγόριθμοι στη διαχείριση της χειμαρρικής επικινδυνότητας

Πολυάριθμες είναι οι αναφορές που σχετίζονται με τη χρήση γενετικών αλγορίθμων στη διαχείριση υδατικών διαθεσίμων γενικότερα, καθώς επίσης και ειδικότερα στην πρόβλεψη της χειμαρρικής επικινδυνότητας. Οι Cai κ.α. [24], συνδύασαν ένα ειδικά σχεδιασμένο ΓΑ με γραμμικό προγραμματισμό ώστε να προταθούν απο-

τελεσματικές λύσεις σε προβλήματα σχετικά με υδατικά διαθέσιμα μεγάλης κλίμακας. Η έρευνά τους αποσκοπεί στην απλοποίηση των σχετικών μεταβλητών του προβλήματος και στη μετατροπή του σε γραμμικό. Στο [37] οι Cheng κ.α. προτείνουν το συνδυασμό ασαφών συστημάτων με ΓΑ προς την κατεύθυνση της επίλυσης του προβλήματος βαθμονόμησης των μοντέλων βροχόπτωσης – απορροής στον ταμιευτήρα Shuangpai στην Κίνα. Στο προτεινόμενο μοντέλο, ο ΓΑ χρησιμοποιείται για τη βαθμονόμηση τόσο της υδατικής ισορροπίας, όσο και στη δρομολόγηση της απορροής. Σε νέα του εργασία [38] ο ερευνητής τροποποιεί το σχετικό ΓΑ με σκοπό την επιτυχή αντιμετώπιση του προβλήματος αναγνώρισης των καλύτερων συμπεριφορών του συστήματος κατά τη διάρκεια της βαθμονόμησης. Οι Agrawal και Singh χρησιμοποίησαν τους ίδιους αλγόριθμους για την ανάπτυξη και βελτιστοποίηση ενός μοντέλου πρόβλεψης της απορροής [4] που εφαρμόσθηκε στη λεκάνη απορροής Kashinagar του ποταμού Vamsadhara κοντά στην πόλη Orissa της Ινδίας. Στην περίπτωση αυτή οι ΓΑ χρησιμοποιούνται για τον υπολογισμό των παραμέτρων του μοντέλου και για την αριστοποίηση διαφόρων λειτουργιών. Βασισμένοι στην πεποίθηση ότι η προτυποποίηση των λεκανών απορροής είναι απαραίτητη για τη διαχείριση των υδατικών διαθεσίμων και ότι απαιτεί κατάλληλη περιγραφή της χωρικής διακύμανσης της βροχόπτωσης, οι Chang κ.α. προτείνουν ένα πρότυπο ΓΑ [33] για τον προσδιορισμό των παραμέτρων ασαφών συναρτήσεων στοιχείων (fuzzy membership functions) οι οποίες αντιπροσωπεύουν τοποθεσίες χωρίς αρχεία βροχόπτωσης. Τα αποτελέσματα της εργασίας δείχνουν ξεκάθαρη μείωση του σφάλματος υπολογισμού (estimated error) στην περίπτωση που χρησιμοποιείται ΓΑ.

Η εκπαίδευση των ΤΝΔ με εξελικτικά παραγόμενα σύνολα δεδομένων έχει μελετηθεί αρκετά σε σχέση με προβλήματα διαχείρισης υδατικών αποθεμάτων. Οι Srinivasulu και Jain συνέκριναν τρεις διαφορετικές τεχνικές εκπαίδευσης με δεδομένα για μοντέλα βροχόπτωσης – απορροής, ένα από τα οποία ενσωμάτωνε ΓΑ για την μορφοποίηση των εκπαιδευτικών ζευγών [171]. Μοντέλο αυτό-οργανωμένων χαρτών (SOM: Self Organizing Maps) ήταν υπεύθυνο για την κατηγοριοποίηση του χώρου εισόδου / εξόδου πριν από την ανάπτυξη ΤΝΔ για κάθε κατηγορία. Τα αποτελέσματα δείχνουν την υπεροχή της τεχνικής με ΓΑ έναντι της κλασσικής εκπαιδευτικής διαδικασίας. Περίπου κατά τη διάρκεια της ίδιας περιόδου, οι Anctil κ.α. [7] πρότειναν ένα πρότυπο ΤΝΔ αναβαθμισμένης ικανότητας πρόβλεψης μέσω της αριστοποίησης των χρονοσειρών μέσης ημερήσιας βροχόπτωσης, η οποία υλοποιήθηκε

με την εφαρμογή συγκεκριμένου ΓΑ. Πιο πρόσφατα, ο Chau ανέπτυξε ένα πρότυπο Split – Step Particle Swarm Optimization (SSPSO) για την εκπαίδευση πολυεπίπεδων perceptrons που χρησιμοποιήθηκαν για την πρόβλεψη σε πραγματικό χρόνο των επιπέδων στάθμης του ποταμού Shing Mun στο Hong Kong [35]. Τα αποτελέσματα δείχνουν βελτιωμένα επίπεδα ακρίβειας στη σύγκριση με κλασσικά ΤΝΔ. Παράλληλα, οι Kerachian και Karamouz [93] ανέπτυξαν μια στοχαστική τεχνική επίλυσης αντιφάσεων βασισμένη σε ΓΑ. Η μέθοδος, συνδυαζόμενη με σύστημα προσομοίωσης της ποιότητας των νερών μέσω ανάλυσης χρονοσειρών, προτυποποίησε με επιτυχία τη λειτουργία του ταμιευτήρα και τον καταμερισμό του φορτίου αποβλήτων του ποταμού Ghomrud στο κεντρικό Ιράν. Τέλος, οι Damle και Yalcin περιγράφουν μια καινοτόμο προσέγγιση για υπερχείλιση ποταμών βασισμένη σε εξόρυξη δεδομένων χρονοσειρών [47]. Οι διαδικασίες που περιγράφονται χρησιμοποιούν, μεταξύ άλλων, ΓΑ για την ανεύρεση βέλτιστων προτύπων συστάδων (optimal pattern clusters) στο υπό μελέτη σύνολο δεδομένων. Η εν λόγω μέθοδος χρησιμοποιήθηκε με επιτυχία στο σταθμό μέτρησης βροχόπτωσης St. Louis του ποταμού Mississippi των Η.Π.Α.

3.4 Προδιαγραφές του συστήματος

Με βάση τις ανάγκες που δημιουργούνται για αποτελεσματικότερη αναπαράσταση των πρωτογενών δεδομένων χρονοσειρών, αλλά και την ανάπτυξη μιας μεθόδου με όσο το δυνατόν ευρύτερη εφαρμογή, το εργαλείο λογισμικού που προτείνεται στην παρούσα διατριβή αναπτύχθηκε με βάση συγκεκριμένες προδιαγραφές, ώστε να επιτυγχάνεται:

- *Αποτελεσματική εξαγωγή χαρακτηριστικών από το διάνυσμα εισόδου.* Ο εξελικτικός αλγόριθμος που αναπτύχθηκε ως βάση του προτεινόμενου εργαλείου εφαρμόζεται στην προ-επεξεργαστική φάση των δεδομένων. Βασικότερος στόχος στη φάση αυτή είναι ο εντοπισμός των κυριότερων χαρακτηριστικών προτύπων της χρονοσειράς. Ο αλγόριθμος προσαρμόζεται στα αρχικά δεδομένα δημιουργώντας, μετά την παρέλευση συγκεκριμένου αριθμού εξελικτικών γενεών, την πλέον αποτελεσματική αναπαράσταση της αρχικής χρονοσειράς.

- *Βέλτιστη παραμετροποίηση βάσει της εξαγωγής χαρακτηριστικών.* Κατά τη διάρκεια εφαρμογής του αλγορίθμου, ένα μεγάλο πλήθος δεδομένων σχημα-

τοποιείται και δοκιμάζεται από τους ενσωματωμένους ταξινομητές του συστήματος, η αρχιτεκτονική των οποίων εξαρτάται κάθε φορά από το σχήμα τμηματοποίησης του αντίστοιχου εκπαιδευτή. Με τον τρόπο αυτό αντιμετωπίζεται αποτελεσματικά το πρόβλημα του καθορισμού της βέλτιστης παραμετροποίησης των ταξινομητών.

- *Εξομάλυνση των αρχικών δεδομένων.* Λόγω του γεγονότος ότι οι δοκιμές που εκτελεί ο αλγόριθμος κατά τη διάρκεια της εξελικτικής διαδικασίας ποσοτικοποιούνται στη φάση του ελέγχου, δηλαδή σε δεδομένα τα οποία ο εκάστοτε ταξινομητής δεν έχει ξανασυναντήσει κατά τη διάρκεια της εκπαίδευσης, σε συνδυασμό με την πολιτική επιλογής των καταλληλότερων σχημάτων αναπαράστασης σε κάθε γενεά, αυξάνουν το βαθμό προσαρμοστικότητας του συστήματος έναντι των αρχικών δεδομένων, αυξάνοντας χαρακτηριστικά τις δυνατότητες γενίκευσής του.

- *Αποτελεσματικό χειρισμό ακρότατων τιμών.* Η εφαρμογή του εξελικτικού αλγορίθμου σχηματίζει ζώνες σημαντικότητας μέσω της τμηματοποίησης που επιτυγχάνεται, με αποτέλεσμα να μειώνεται η επίδραση τυχόν ακρότατων τιμών της αρχικής χρονοσειράς στο τελικό αποτέλεσμα.

- *Απρόσκοπτη συνεργασία των υπο-συστημάτων.* Οι ενσωματωμένοι ταξινομητές του συστήματος, οι οποίοι αναλαμβάνουν το έργο της εξαγωγής του τελικού αποτελέσματος, αποτελούν ουσιαστικά την αντικειμενική συνάρτηση του εξελικτικού αλγορίθμου από τον οποίο προκύπτει η εξαγωγή των χαρακτηριστικών της χρονοσειράς. Τα δύο υπο-συστήματα συνεργάζονται κατ' αυτό τον τρόπο χωρίς προβλήματα για τη μορφοποίηση του καταλληλότερου σχήματος τμηματοποίησης.

- *Δυνατότητα χειρισμού μεγάλου όγκου δεδομένων.* Ως όγκος δεδομένων για το τελικό παραγόμενο σύστημα νοείται, εκτός από το συνολικό αριθμό των δειγμάτων, επίσης και το πλήθος των μεταβλητών κάθε δείγματος. Η αναγνώριση των προτύπων που διενεργείται στην αρχική χρονοσειρά έχει ως αποτέλεσμα τη δραστική μείωση του βαθμού διάστασης χωρίς παράλληλη απώλεια ζωτικής πληροφορίας. Υπό την έννοια αυτή το προτεινόμενο εργαλείο είναι σε

θέση πρακτικά να επεξεργαστεί δεδομένα χρονοσειρών, ανεξαρτήτως εύρους και πλήθους δειγμάτων.

- *Προσαρμογή σε ποικίλους τύπους προβλημάτων.* Η ενσωμάτωση περισσοτέρων του ενός ταξινομητών στο τελικό σύστημα και η επιλογή αυτού που αποδίδει καλύτερα ανά περίπτωση διευρύνει το εύρος εφαρμογής του προτεινόμενου εργαλείου.

- *Συγκράτηση λειτουργικού κόστους του συστήματος.* Ένεκα της λειτουργίας του έναντι μεγάλου όγκου δεδομένων και πολύπλοκης αυτοματοποιημένης παραμετροποίησης του εξελικτικού αλγορίθμου, όπως και των ενσωματωμένων ταξινομητών, το κόστος επεξεργασίας από πλευράς υπολογιστικής ισχύος είναι σχετικά υψηλό κατά τη διάρκεια της εξελικτικής εκπαίδευσης. Μετά την ολοκλήρωσή της όμως οι υπολογιστικές απαιτήσεις μειώνονται δραστικά, κυρίως λόγω της αποτελεσματικής εξομάλυνσης και αναπαράστασης των αρχικών δεδομένων.

Στη συνέχεια θα παρουσιασθεί λεπτομερειακά το πρότυπο του αλγορίθμου που αναπτύχθηκε, καθώς επίσης και η εφαρμογή του στις επιλεγμένες μελέτες περίπτωσης μαζί με τα αποτελέσματα που εξήχθησαν από την εφαρμογή αυτή.

Κεφάλαιο 4
ΑΝΑΛΥΣΗ ΤΟΥ ΠΡΟΤΥΠΟΥ

Στο κεφάλαιο αυτό τίθενται οι θεωρητικές βάσεις και αναλύεται διεξοδικά το μαθηματικό πρότυπο στο οποίο στηρίζεται ο προτεινόμενος αλγόριθμος της Πρότυπης Εξελικτικής Τμηματοποίησης των χρονοσειρών.

Στο γενικότερο πλαίσιο της εξόρυξης δεδομένων από χρονοσειρές, σημαντικό ρόλο παίζει η εξομάλυνση των πρωτογενών δεδομένων και η αναπαράστασή τους σε χώρους μικρότερης διάστασης, στο μέτρο βεβαίως κατά το οποίο δεν επέρχεται σημαντική απώλεια πληροφορίας. Η μεθοδολογία που αναπτύσσεται στην παρούσα διατριβή μετέρχεται εξελικτικής διαδικασίας, μέσω της οποίας επιτυγχάνεται αποτελεσματική αναπαράσταση των πρωτογενών χρονοσειριακών δεδομένων. Πρόκειται ουσιαστικά για μια μέθοδο προ-επεξεργασίας της χρονοσειράς, κατά την οποία πλήθος δευτερογενών δεδομένων προκύπτει από τα πρωτογενή με εξελικτικό τρόπο. Η διαδικασία αυτή υλοποιείται μέσω ειδικά δομημένου γενετικού αλγορίθμου, οι γενεές του οποίου αποτελούνται από πληθυσμούς χρωμοσωμάτων[1] όλο και πιο προσαρμοσμένων στα πρωτογενή δεδομένα.

[1] Από το σημείο αυτό και μετέπειτα, τα «χρωμοσώματα» που αποτελούν τον πληθυσμό των εξελικτικών γενεών του γενετικού αλγορίθμου θα αναφέρονται ως «εκπαιδευτές»

Στην αρχική γενεά παράγεται πληθυσμός εκπαιδευτών το μέγεθος του οποίου είναι σταθερό για όλες τις επόμενες γενεές του αλγορίθμου και καθοριζόμενο από το χρήστη κατά την αρχικοποίησή του. Κάθε εκπαιδευτής αποτελείται από τυχαία επιλεγμένα δυαδικά γονίδια, ο αριθμός των οποίων εξαρτάται από τον αριθμό των δεδομένων της αρχικής χρονοσειράς και χρησιμοποιείται για τη δημιουργία σχήματος τμηματοποίησης επί των πρωτογενών χρονοσειριακών δεδομένων. Ουσιαστικά μέσω του εκπαιδευτή καθορίζεται μια πιθανή αναπαράσταση της αρχικής χρονοσειράς, η αξιολόγηση της οποίας προσδιορίζεται μέσω της απόδοσης των ταξινομητών του συστήματος. Χρησιμοποιώντας τις μεθόδους της επιλογής, του ανασυνδυασμού και της μετάλλαξης, οι επόμενες γενεές του αλγορίθμου εποικίζονται με εκπαιδευτές συνεχώς αυξανόμενης προσαρμοστικότητας, με την έννοια της βελτιωμένης απόδοσης εκ μέρους των ταξινομητών.

Στη συνέχεια θα παρουσιασθεί το πρότυπο αποτύπωσης του σχήματος τμηματοποίησης επί των αρχικών δεδομένων.

4.1 Πρότυπη Εξελικτική Τμηματοποίηση

Για την ανάλυση του προτύπου τμηματοποίησης της πρωτογενούς χρονοσειράς, είναι απαραίτητη η εισαγωγή της έννοιας των εκπαιδευτών, των οποίων η δομή και η λειτουργία καθορίζεται στα επόμενα. Παράλληλα αναλύεται η μέθοδος αποτύπωσης του φέροντος σχήματος τμηματοποίησης επί του διανύσματος εισόδου, καθώς επίσης και η ενσωμάτωση της εξόδου για την παραγωγή των εξελικτικών δεδομένων.

4.1.1 Ορισμός εκπαιδευτών

Ας υποθέσουμε ότι για συγκεκριμένο πρόβλημα:

1. ο πίνακας εισόδου $\mathbf{X} = [x_{mn}]\,,\ x_{mn} \in \mathbb{R}, m = 1, 2, \cdots, M,\ \ n = 1, 2, \cdots, N$ σχηματοποιείται μέσω ενός συνόλου *M* εγγραφών δεδομένων τα οποία προέρχονται από κάποια χρονοσειρά. Στην περίπτωση αυτή, καθε εγγραφή αποτελείται από *N* χρονοσειριακά στοιχεία και συγκεκριμένα:

$$\mathbf{X} = [\mathbf{X}_m]_{m=1(1)M} = [x_{m1}, x_{m2}, \cdots, x_{mN}]_{m=1(1)M} \tag{4.1.1}$$

2. Κατ' αντιστοιχία, ο πίνακας εξόδου $\mathbf{Y} = [y_{mp}]$, $y_{mp} \in \{0, 1\},\ m = 1, 2, \cdots, M,\ \ p = 1, 2, \cdots, P$ θα περιγράφεται απο το ίδιο πλήθος *M* εγγραφών, κάθε μια από τις οποίες θα αποτελείται από διαφορετικό πλήθος *P* προτύπων και συγκεκριμένα:

$$\mathbf{Y} = [\mathbf{Y}_m]_{m=1(1)M} = [y_{m1}, y_{m2}, \cdots, y_{mP}]_{m=1(1)M} \tag{4.1.2}$$

3. Ο πίνακας εκπαιδευτών $\mathbf{\Omega} = [\omega_{qn}],\ \ \omega_{qn} \in \{0, 1\},\ \ q = 1, 2, \cdots, Q,\ \ n = 1, 2, \cdots, N+2$ περιγράφεται από *Q* το πλήθος εγγραφών (εκπαιδευτών), κάθε μια από τις οποίες αποτελείται από *N*+2 στοιχεία και συγκεκριμένα:

$$\mathbf{\Omega} = [\mathbf{\Omega}_q]_{q=1(1)Q} = \left[\omega_{q1}, \omega_{q2}, \cdots, \omega_{qN} \vdots \omega_{qN+1}, \omega_{qN+2}\right]_{q=1(1)Q} \tag{4.1.3}$$

Ορισμός 4.1.1. *Κάθε εκπαιδευτής*[2] $\mathbf{\Omega}_q = (\omega_{q1}, \omega_{q2}, \cdots, \omega_{qN} \vdots \omega_{qN+1}, \omega_{qN+2}),\ \ q = 1, 2, \cdots, Q$, *θεωρείται ως ένα χρωμόσωμα, το οποίο αρχικά σχηματοποιείται από δυαδικά γονίδια επιλεγμένα με τυχαίο τρόπο. Κάθε τέτοιο χρωμόσωμα αποτελείται από δύο τμήματα, το* τμήμα δραστηριοποίησης (activation block) *και το* τμήμα πυρήνα (core block) *(Πίνακας 4.1).*

4.1.2 Λειτουργική δομή εκπαιδευτών

Όπως θα δειχθεί στα επόμενα, το τμήμα πυρήνα λειτουργεί ως ο συμπεριφορικός μηχανισμός του εκπαιδευτή. Με τον όρο αυτό εννοείται ότι στο τμήμα πυρήνα έχει ενσωματωθεί ένα σύνολο κανόνων το οποίο διέπει και κατά κάποιο τρόπο υπαγορεύει την ενέργεια που θα εκτελεσθεί εκ μέρους του λοιπού τμήματος του χρωμοσώματος, δηλαδή του τμήματος δραστηριοποίησης, έναντι των πρωτογενών δεδομένων τα οποία κάθε φορά χειρίζεται ο συγκεκριμένος εκπαιδευτής.

Ορισμός 4.1.2. *Τα N το πλήθος δυαδικά γονίδια του τμήματος δραστηριοποίησης καθενός από τους εκπαιδευτές* $\mathbf{\Omega}_q,\ \ q = 1, 2, \cdots, Q$ *ορίζουν τμήμα* $\mathbf{\Omega}_q^k,\ \ k \in \mathbb{N}$, *ως ακολούθως:*

---

[2] Η σημειογραφία $\vdots$ διαφοροποιεί το τμήμα δραστηριοποίησης από το τμήμα πυρήνα

Πίνακας 4.1: Παράδειγμα της δομής εκπαιδευτή

Γραμμή	Τμήμα Δραστηριοποίησης										Τμήμα Πυρήνα
1	1	0	0	1	1	0	0	1	0	1	1 0
2	Ω^3			Ω^1	Ω^3			Ω^2		Ω^1	
3	44	32	17	8	8	12	1	18	30	48	
4	X^3			X^1	X^3			X^2		X^1	

$$\mathbf{\Omega}_q^1 = \underbrace{1}_{1-element}, \mathbf{\Omega}_q^2 = \underbrace{10}_{2-elements}, \mathbf{\Omega}_q^3 = \underbrace{100}_{3-elements}, \cdots, \mathbf{\Omega}_q^k = \underbrace{100\cdots0}_{k-elements} \tag{4.1.4}$$

ως ένα *k-υποσύνολο* δυαδικών στοιχείων δηλαδή, το πρώτο εκ των οποίων είναι 1, ενώ ταυτόχρονα όλα τα υπόλοιπα είναι 0.

Είναι απαραίτητο στο σημείο αυτό να τονιστεί ότι εφόσον η πρώτη τιμή της χρονοσειράς σηματοδοτεί την έναρξη του φαινομένου, το πρώτο γονίδιο κάθε εκπαιδευτή τίθεται πάντα ως 1. Κατ' αντιστοιχία, το τελευταίο γονίδιο κάθε εκπαιδευτή επίσης τίθεται ως 1 για να αποδώσει την περάτωση του φαινομένου. Η διαφοροποίηση αυτή συμμορφώνεται απόλυτα με τον αλγόριθμο ΑΣΣ, καθώς επίσης και τον αλγόριθμο Douglas-Peucker, κατά τους οποίους τα δύο ακραία στοιχεία της χρονοσειράς θεωρούνται ως τα σημαντικότερα και κατατάσσονται πρώτα στον ταξινομημένο κατά βαθμό σημαντικότητας κατάλογο των αντιληπτικά σημαντικών σημείων για την αναπαράσταση. Στον πίνακα 4.1 διακρίνεται το τμήμα δραστηριοποίησης και το τμήμα πυρήνα ενός τυπικού εκπαιδευτή, όπως αυτός παράγεται από το γενετικό αλγόριθμο (γραμμή 1). Το τμήμα δραστηριοποίησης διαχωρίζεται περαιτέρω σε πέντε τμήματα (γραμμή 2), μέσω των οποίων επιτυγχάνεται η επεξεργασία της εγγραφής μιας δεκαμελούς χρονοσειράς (γραμμή 3). Το αποτέλεσμα είναι ένα σχήμα τμηματοποίησης αποτελούμενο από πέντε διαδοχικά τμήματα (γραμμή 4), σε καθένα από τα οποία εφαρμόζεται ένα στατιστικό μέτρο που υποδεικνύεται από το τμήμα πυρήνα.

4.1.3 Καθορισμός τμημάτων

Κατά τον ίδιο τρόπο, κάθε μια από τις *M* εγγραφές $\mathbf{X}_m = (x_{m1}, x_{m2}, \cdots, x_{mN})$ του πίνακα $\mathbf{X} = [x_{mn}]$, $m = 1, 2, \cdots, M, \quad n = 1, 2, \cdots, N$ διαχωρίζεται σε ίσο αριθμό χρονοσειριακών τμημάτων, ενώ κάθε $\mathbf{X}_m^k; k \in \mathbb{N}$ καθορίζεται από το αντίστοιχο εκπαιδευτικό τμήμα $\mathbf{\Omega}_q^k$. Για παράδειγμα, παραβλέποντας χάριν ευκολίας το δείκτη *m*, το τμήμα δραστηριοποίησης του χρωμοσώματος που εμφαίνεται στον Πίνακα 4.1 αποτελείται από τα ακόλουθα πέντε εκπαιδευτικά τμήματα:

$\mathbf{\Omega}^3\ \mathbf{\Omega}^1\ \mathbf{\Omega}^3\ \mathbf{\Omega}^2\ \mathbf{\Omega}^1$

Τα αντίστοιχα πέντε διαδοχικά τμήματα που θα σχεδιασθούν στην αρχική εγγραφή της πρωτογενούς χρονοσειράς του ίδιου πίνακα, δίνονται από το σχήμα:

$\mathbf{X}^3\ \mathbf{X}^1\ \mathbf{X}^3\ \mathbf{X}^2\ \mathbf{X}^1$

ή, το οποίο είναι ισοδύναμο, από τις ομάδες:

[44 32 17] , [8] , [8 12 1] , [18 30] , [48].

Για να συμπληρωθεί η σημειογραφία του προτύπου, είναι απαραίτητο να συμπεριληφθεί ο αριθμός των σχεδιαζόμενων τμημάτων, έστω *r*, της ως άνω καθορισμένης εγγραφής, έστω *m*, καθώς επίσης και ο αριθμός των στοιχείων μέσα σε κάθε τμήμα, έστω $k_j\ ; j = 1(1)r$. Κατ' αυτό τον τρόπο, παριστώντας τα τμήματα που ορίσθηκαν προηγουμένως με $\mathbf{\Omega}_m^{k_1}\ \mathbf{\Omega}_m^{k_2}\ \mathbf{\Omega}_m^{k_3}\ \mathbf{\Omega}_m^{k_4}\ \mathbf{\Omega}_m^{k_5}$ και $\mathbf{X}_m^{k_1}\ \mathbf{X}_m^{k_2}\ \mathbf{X}_m^{k_3}\ \mathbf{X}_m^{k_4}\ \mathbf{X}_m^{k_5}$ αντίστοιχα, μπορούμε εύκολα να επαληθεύσουμε ότι ο αριθμός των τμημάτων που καθορίζονται στην εγγραφή *m* είναι *r*=5, με $k_j\ ; j = 1(1)5$ και $k_1 = 3,\ k_2 = 1,\ k_3 = 3,\ k_4 = 2,$ και $k_5 = 1$.

Ο φαινότυπος ενός εκπαιδευτή ο οποίος αποτελείται από δύο μέρη, το μέρος δραστηριοποίησης τμηματοποιημένο όπως καθορίζεται στα προηγούμενα, και το μέρος πυρήνα, έχει ως ακολούθως:

$$\begin{aligned} \mathbf{\Omega} \;=\; [\mathbf{\Omega}_q]_{q=1(1)Q} \;&=\; [\omega_{q1}, \omega_{q2}, \cdots, \omega_{qN} \vdots \omega_{qN+1}, \omega_{qN+2}]_{q=1(1)Q} \;= \\ &=\; \left[\mathbf{\Omega}_q^{k_1}\ \mathbf{\Omega}_q^{k_2},\ \cdots, \mathbf{\Omega}_q^{k_r} \vdots \omega_{qN+1}, \omega_{qN+2}\right]_{q=1(1)Q} \end{aligned} \tag{4.1.5}$$

όπου $\sum_{j=1}^{r} k_j = N,\ \forall\ k_j,\ N,\ r,\ q\ \in \mathbb{N}$.

4.1.4 Αποτύπωση του σχήματος τμηματοποίησης

Υπό την έννοια αυτή, κάθε εκπαιδευτής $\mathbf{\Omega}_q = (\omega_{q1}, \omega_{q2}, \cdots, \omega_{qN} \vdots \omega_{qN+1}, \omega_{qN+2})$, συνδυάζεται με τον πίνακα εισόδου X, μέσω του τελεστή $\circ$ ώστε να παραχθεί ο πίνακας $\mathbf{\Xi}^q = [\xi^q_{mk}] \quad \forall m = 1, 2, \cdots, M, \;\; k = 1, 2, \cdots, k_r, \;\; q = 1, 2, \cdots, Q$, όπου $M,\ k_r,\ Q \in \mathbb{N}$, ως ακολούθως:

$$
\begin{array}{l}
\mathbf{\Xi}^q \leftarrow \mathbf{X} \circ \mathbf{\Omega}_q : \; [\xi^q_{mk}] = [x_{mk}] \circ \mathbf{\Omega}_q = \\
(x_{m1}, x_{m2}, \cdots, x_{mN})_{m=1(1)M} \circ (\omega_{q1}, \omega_{q2}, \cdots, \omega_{qN} \vdots \omega_{qN+1}, \omega_{qN+2}) = \\
\left[\mathbf{X}^{k_1}_m \mathbf{X}^{k_2}_m \cdots \mathbf{X}^{k_r}_m\right]_{m=1(1)M} \circ \left[\mathbf{\Omega}^{k_1}_q \, \mathbf{\Omega}^{k_2}_q \;\; \cdots \mathbf{\Omega}^{k_r}_q \vdots \omega_{qN+1}, \omega_{qN+2}\right] = \\
\left[\; \mathbf{X}^{k_1}_m \circ \mathbf{\Omega}^{k_1}_q \quad \mathbf{X}^{k_2}_m \circ \mathbf{\Omega}^{k_2}_q \quad \cdots \; \mathbf{X}^{k_r}_m \circ \mathbf{\Omega}^{k_r}_q \vdots \omega_{qN+1}, \omega_{qN+2} \;\right]_{m=1(1)M}
\end{array}
\tag{4.1.6}
$$

όπου $\sum_{j=1}^{r} k_j = N \;\; \forall \;\; k_j,\ N,\ r,\ q \;\in \mathbb{N}.$

Ορισμός 4.1.3. *Ο πίνακας εισόδου* $\mathbf{X} = [x_{mn}]$ *μετουσιώνεται σύμφωνα με τις συνδυασμένες τιμές του τμήματος πυρήνα κάθε εκπαιδευτή* $(\omega_{qN+1}, \omega_{qN+2})$ *και του σχήματος τμηματοποίησης* $\mathbf{\Omega}^k_m \; ; k \in \mathbb{N}$*, στον εξελικτικό πίνακα δεδομένων* $\mathbf{\Xi}^q = [\xi^q_{mk}]\,,\ m = 1, 2, \cdots, M, \;\; k = 1, 2, \cdots, k_r,$

όπου $\sum_{j=1}^{r} k_j = N,\ \forall\, k_j,\ N,\ r,\ q \in \mathbb{N}$*, ως ακολούθως:*

$$
\mathbf{\Xi}^q \leftarrow \mathbf{X} \circ \mathbf{\Omega}_q : \; [\xi^q_{mk}] = \begin{bmatrix} \xi^q_{11} & \xi^q_{12} & \cdots & \xi^q_{1k_r} \\ \xi^q_{21} & \xi^q_{22} & \cdots & \xi^q_{2k_r} \\ \vdots & \vdots & \vdots & \vdots \\ \xi^q_{M1} & \xi^q_{M2} & \cdots & \xi^q_{Mk_r} \end{bmatrix}
\tag{4.1.7}
$$

το οποίο ισχύει $\forall q = 1, 2, \cdots, Q$ *όπου,*

$$\xi_{mk}^{q} = \begin{cases} \sum_{k=1}^{k_r} x_{mk}\omega_{mk} & \text{εάν} \quad (\omega_{qN+1}, \omega_{qN+2}) = (0,0) \\ \sum_{k=1}^{k_r} x_{mk}\omega_{mk}/k_r & \text{εάν} \quad (\omega_{qN+1}, \omega_{qN+2}) = (1,1) \\ \text{Διάμεσος } \mathbf{X}_m^{k_r} & \text{εάν} \quad (\omega_{qN+1}, \omega_{qN+2}) = (0,1) \\ \text{Μέγιστο } \mathbf{X}_m^{k_r} - \text{Ελάχιστο } \mathbf{X}_m^{k_r} & \text{εάν} \quad (\omega_{qN+1}, \omega_{qN+2}) = (1,0) \end{cases} \tag{4.1.8}$$

Οι οδηγίες που φθάνουν στο τμήμα δραστηριοποίησης του εκπαιδευτή είναι ένα πρωτόκολλο λογικής συμπεριφοράς έναντι των πρωτογενών δεδομένων, για τα οποία τέσσερις πολιτικές χειρισμού έχουν ουσιαστικά επιλεγεί. «Απόρριψη-Ολων-Των-0», είναι μια απευθείας δειγματοληπτική διαδικασία σύμφωνα με την οποία ένα ποσοστό της αρχικής πληροφορίας απορρίπτεται, εκ προοιμίου τεκμαίροντας ότι δε σχετίζεται με το υπό μελέτη πρόβλημα και αντιστοιχεί σε θόρυβο, ο οποίος αποτελεί τροχοπέδη προς την ανεύρεση της άριστης λύσης. Τα υπόλοιπα σενάρια συμπεριφοράς περιλαμβάνουν περισσότερο ή λιγότερο αυστηρούς δειγματολήπτες. «Πρώτο-Ένα-Τελευταίο-0 μέσος», «Πρώτο-Ένα-Τελευταίο-0 διάμεσος» και «Πρώτο-Ένα-Τελευταίο-0 ελάχιστο-μέγιστο», σχηματίζουν συστοιχίες δεδομένων επί της πρωτογενούς χρονοσειράς και αποδίδουν μοναδική τιμή σε καθεμιά. Η τιμή αυτή δεν αποτελεί μια αναπαράσταση της συστοιχίας των δεδομένων μόνο, αλλά επίσης και ένα είδος «μνήμης» για το σύστημα, το οποίο με τον τρόπο αυτό λαμβάνει υπόψη ένα ποσοστό της πληροφορίας που χάνεται και δεν την απορρίπτει πάραυτα. Η μόνη παράμετρος που ποικίλλει στις τρεις προαναφερόμενες συμπεριφορές του εκπαιδευτή είναι η ουσία της μνήμης αυτής. Στην πρώτη περίπτωση, κάθε συστοιχία αντιστοιχίζεται με το μέσο όρο των δεδομένων της, ενώ στη δεύτερη η συστοιχία διαχωρίζεται σε δύο τμήματα και αναπαρίσταται με το ακριβές μέσο της. Τέλος, με την τελευταία περίπτωση ο αλγόριθμος είναι σε θέση να υπολογίσει το ουσιαστικό εύρος κάθε σχεδιασμένης συστοιχίας.

Υπ' αυτό το πρίσμα, όταν τα γονίδια του μηχανισμού πυρήνα είναι 00, ο εκπαιδευτής αντιστοιχεί στη δειγματοληπτική διαδικασία «Απόρριψη-Ολων-Των-0». Στην περίπτωση αυτή, η χρονοσειρά θα στερηθεί από τις τιμές της για τις οποίες τα αντίστοιχα γονίδια του εκπαιδευτή είναι 0. Εάν τα γονίδια του μηχανισμού πυρήνα είναι 11, τότε ο εκπαιδευτής αντιστοιχεί στο μηχανισμό συστοιχίας «Πρώτο-Ένα-Τελευταίο-0 μέσος», περίπτωση κατά την οποία ο εκπαιδευτής εξάγει το μέσο όρο

κάθε τμήματος της αρχικής χρονοσειράς τα στοιχεία της οποίας καθορίζονται από το πρώτο 1 και το τελευταίο 0 των γονιδίων του. Στην περίπτωση κατά την οποία ο μηχανισμός πυρήνα είναι 01, τότε ο εκπαιδευτής συμπεριφέρεται ως «Πρώτο-Ένα-Τελευταίο-0 διάμεσος» και επιστρέφει το διάμεσο κάθε τμήματος όπως ορίσθηκε προηγουμένως. Τέλος, εάν ο μηχανισμός πυρήνα είναι 10, τότε ο εκπαιδευτής επιστρέφει τη διαφορά της μέγιστης από την ελάχιστη τιμή κάθε τμήματος της αρχικής χρονοσειράς. Με άλλα λόγια, εάν τα γονίδια πυρήνα είναι αμφότερα μηδέν, τότε μόνο το πρώτο στοιχείο κάθε τμήματος επιστρέφεται στο εξελικτικό σύνολο δεδομένων. Εάν είναι αμφότερα μονάδες, τότε επιστρέφεται η μέση τιμή κάθε τμήματος. Στην περίπτωση κατά την οποία τα γονίδια πυρήνα είναι (0,1) τότε επιστρέφεται ο διάμεσος κάθε τμήματος, ενώ όταν είναι (1,0) τότε, για κάθε τμήμα, επιστρέφεται η διαφορά του στοιχείου με τη μεγαλύτερη τιμή από εκείνο με τη μικρότερη, υπό την προϋπόθεση ότι οι τιμές έχουν καταταχθεί κατάλληλα για τις δύο αυτές τελευταίες περιπτώσεις.

4.1.5 Σχηματισμός εξελικτικών δεδομένων

Με τον τρόπο αυτό, ο αλγόριθμος παράγει έναν αριθμό δευτερογενών εξελικτικών συνόλων δεδομένων (EDS: Evolutionary Data Sets) ίσο με το συνολικό αριθμό των εκπαιδευτών, τα οποία αποτελούν τον πίνακα $\boldsymbol{\Xi} = [\boldsymbol{\Xi}^q]_{q=1(1)Q}$. Κάθε παραγόμενο σύνολο δεδομένων αποτελείται από ίσο αριθμό εγγραφών όπως και το αρχικό πρωτογενές σύνολο, αλλά είναι μειωμένου εύρους, σύμφωνα με το σχήμα τμηματοποίησης του εκάστοτε εκπαιδευτή:

$$\boldsymbol{\Xi} = [\boldsymbol{\Xi}^q]_{q=1(1)Q} \leftarrow [\boldsymbol{X} \circ \boldsymbol{\Omega}_q]_{q=1(1)Q} = \left[\begin{array}{c} \left[\begin{array}{cccc} \xi_{11}^1 & \xi_{12}^1 & \cdots & \xi_{1r_1}^1 \\ \xi_{21}^1 & \xi_{22}^1 & \cdots & \xi_{2r_1}^1 \\ \vdots & \vdots & \vdots & \vdots \\ \xi_{M1}^1 & \xi_{M2}^1 & \cdots & \xi_{Mr_1}^1 \end{array} \right] \\ \left[\begin{array}{cccc} \xi_{11}^2 & \xi_{12}^2 & \cdots & \xi_{1r_2}^2 \\ \xi_{21}^2 & \xi_{22}^2 & \cdots & \xi_{2r_2}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \xi_{M1}^2 & \xi_{m2}^2 & \cdots & \xi_{Mr_2}^2 \end{array} \right] \\ \cdots\cdots\cdots\cdots\cdots \\ \left[\begin{array}{cccc} \xi_{11}^Q & \xi_{12}^Q & \cdots & \xi_{1r_M}^Q \\ \xi_{21}^Q & \xi_{22}^Q & \cdots & \xi_{2r_M}^Q \\ \vdots & \vdots & \vdots & \vdots \\ \xi_{M1}^Q & \xi_{M2}^Q & \cdots & \xi_{Mr_M}^Q \end{array} \right] \end{array} \right] \tag{4.1.9}$$

Μετά την αποτύπωση του γονιώματος κάθε εκπαιδευτή στο εκπαιδευτικό και ελεγκτικό σύνολο της αρχικής χρονοσειράς ακολουθεί η «φάση επανένωσης», κατά την οποία κάθε «γενετικά» παραγόμενη εκπαιδευτική σειρά θα ενσωματώσει εκ νέου τα δομικά και δυναμικά δεδομένα από τα οποία αρχικά είχε αποχωρισθεί, ώστε να εξελιχθεί γενετικά. Έτσι, τα τελικά παραγόμενα σύνολα δεδομένων (EDS), εκείνα που χρησιμοποιούνται στην εκπαίδευση των ταξινομητών, προκύπτουν αφού ο πίνακας εξόδου $\mathbf{Y}$ προσαρτηθεί σε κάθε $\boldsymbol{\Xi}^q$ κατά σειρά. Στην επόμενη φάση τα δευτερογενή αυτά δεδομένα τροφοδοτούνται στο ΤΝΔ για αξιολόγηση.

4.2 Παράδειγμα εργασίας

Στο σημείο αυτό θα παρατεθεί ένα παράδειγμα εργασίας το οποίο εμπίπτει στη γενικότερη κατηγορία προβλημάτων δυαδικής ταξινόμησης στα οποία περιλαμβάνονται δεδομένα χρονοσειρών. Συγκεκριμένα, ας θεωρήσουμε διδιάστατο πίνακα εγγραφών χρονοσειράς, κάθε μια από τις οποίες αποτελείται από δέκα στοιχεία και αντιστοιχεί σε μια από δύο κλάσεις, έστω *Κλάση 1* και *Κλάση 2*. Ως ζητούμενο τίθεται η επιτυχής ταξινόμηση οποιασδήποτε άγνωστης δεκαμελούς χρονοσειριακής

εγγραφής σε μια από τις δύο δεδομένες κλάσεις. Με άλλα λόγια ζητείται η επιτυχής αντιστοίχηση κάθε άγνωστης ψηφιακής υπογραφής σε μια από τις δύο γνωστές κατηγορίες αντικειμένων.

Ας υποθέσουμε ότι το πρωτογενές σύνολο δεδομένων δίνεται από τον ακόλουθο πίνακα:

$$\left[\begin{array}{cccccccccc} 3 & 8 & 7 & 8 & 2 & 1 & 4 & 5 & 9 & 6 & \text{Κλάση 1} \\ 2 & 9 & 4 & 3 & 5 & 2 & 6 & 3 & 8 & 7 & \text{Κλάση 1} \\ 7 & 6 & 3 & 9 & 2 & 1 & 5 & 4 & 6 & 8 & \text{Κλάση 2} \\ 5 & 4 & 5 & 6 & 3 & 4 & 2 & 1 & 7 & 8 & \text{Κλάση 1} \\ 6 & 3 & 2 & 4 & 7 & 5 & 8 & 9 & 1 & 4 & \text{Κλάση 1} \\ 4 & 7 & 6 & 5 & 1 & 8 & 3 & 2 & 9 & 5 & \text{Κλάση 2} \\ 9 & 5 & 8 & 1 & 4 & 6 & 7 & 8 & 4 & 2 & \text{Κλάση 2} \\ 8 & 2 & 9 & 7 & 6 & 3 & 1 & 5 & 3 & 1 & \text{Κλάση 2} \end{array}\right]$$

Δημιουργώντας ένα δυαδικό πίνακα εξόδου με τη χρήση των αντιστοιχίσεων *Κλάση 1* $\longrightarrow 0$ και *Κλάση 2* $\longrightarrow 1$, καθίσταται δυνατός ο διαχωρισμός του πρωτογενούς συνόλου δεδομένων στους πίνακες εισόδου $\mathbf{X}$ και εξόδου $\mathbf{Y}$, ως ακολούθως:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} = \begin{bmatrix} 3 & 8 & 7 & 8 & 2 & 1 & 4 & 5 & 9 & 6 \\ 2 & 9 & 4 & 3 & 5 & 2 & 6 & 3 & 8 & 7 \\ 7 & 6 & 3 & 9 & 2 & 1 & 5 & 4 & 6 & 8 \\ 5 & 4 & 5 & 6 & 3 & 4 & 2 & 1 & 7 & 8 \\ 6 & 3 & 2 & 4 & 7 & 5 & 8 & 9 & 1 & 4 \\ 4 & 7 & 6 & 5 & 1 & 8 & 3 & 2 & 9 & 5 \\ 9 & 5 & 8 & 1 & 4 & 6 & 7 & 8 & 4 & 2 \\ 8 & 2 & 9 & 7 & 6 & 3 & 1 & 5 & 3 & 1 \end{bmatrix} \text{ και } \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Επίσης, ας υποθέσουμε ότι ο αλγόριθμος σε κάθε διαδοχική γενεά του παράγει έξι εκπαιδευτές. Συνεπώς ο πίνακας των εκπαιδευτών $\mathbf{\Omega}$ θα αποτελείται από έξι

εγγραφές (*εκπαιδευτές*), καθένας από τους οποίους θα αποτελείται από 12 (=*10+2*) στοιχεία, συγκεκριμένα:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_1 \\ \mathbf{\Omega}_2 \\ \mathbf{\Omega}_3 \\ \mathbf{\Omega}_4 \\ \mathbf{\Omega}_5 \\ \mathbf{\Omega}_6 \end{bmatrix} = \left[\begin{array}{cccccccccc:cc} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

Συνεπώς, κάθε εκπαιδευτής αποτελείται από τα ακόλουθα σχήματα τμηματοποίησης και τα αντίστοιχα τμήματα πυρήνα:

$$\begin{array}{lcll} \mathbf{\Omega}_1 & = & \mathbf{\Omega}_1^3\mathbf{\Omega}_1^1\mathbf{\Omega}_1^3\mathbf{\Omega}_1^2\mathbf{\Omega}_1^1 & \vdots \quad 0 \quad 0 \\ \mathbf{\Omega}_2 & = & \mathbf{\Omega}_2^1\mathbf{\Omega}_2^2\mathbf{\Omega}_2^4\mathbf{\Omega}_2^3 & \vdots \quad 1 \quad 1 \\ \mathbf{\Omega}_3 & = & \mathbf{\Omega}_3^2\mathbf{\Omega}_3^3\mathbf{\Omega}_3^3\mathbf{\Omega}_3^1\mathbf{\Omega}_3^1 & \vdots \quad 1 \quad 0 \\ \mathbf{\Omega}_4 & = & \mathbf{\Omega}_4^1\mathbf{\Omega}_4^4\mathbf{\Omega}_4^3\mathbf{\Omega}_4^1\mathbf{\Omega}_4^1 & \vdots \quad 0 \quad 1 \\ \mathbf{\Omega}_5 & = & \mathbf{\Omega}_5^6\mathbf{\Omega}_5^1\mathbf{\Omega}_5^3 & \vdots \quad 1 \quad 1 \\ \mathbf{\Omega}_6 & = & \mathbf{\Omega}_6^2\mathbf{\Omega}_6^1\mathbf{\Omega}_6^3\mathbf{\Omega}_6^4 & \vdots \quad 1 \quad 0 \end{array} \tag{4.2.1}$$

Καθένας από τους εκπαιδευτές της 4.2.1 θα δημιουργήσει ένα αντίστοιχο εξελικτικό σύνολο δεδομένων, αποτυπώνοντας το σχήμα τμηματοποίησης το οποίο φέρεται στο τμήμα δραστηριοποίησής του επί του πρωτογενούς ανεπεξέργαστου συνόλου δεδομένων. Για παράδειγμα, η εφαρμογή του $\mathbf{\Omega}_1$ επί του πρωτογενούς συνόλου δεδομένων δίδει[3]:

[3]Στην αναφερόμενη αποτύπωση του σχήματος τμηματοποίησης που φέρει ο εκάστοτε εκπαιδευτής επί του πρωτογενούς συνόλου δεδομένων, ο τελεστής '$\circ$' καθορίζει τον πολλαπλασιασμό των δύο πινάκων στοιχείο με στοιχείο. Η συμπεριφορά που περιγράφεται στο συγκεκριμένο παράδειγμα υπαγορεύεται από το τμήμα πυρήνα του εκπαιδευτή (*0,0*).

$$\boldsymbol{\Xi}^1 = \mathbf{X} \circ \boldsymbol{\Omega}_1 = \left[\mathbf{X}_m^3\, \mathbf{X}_m^1 \mathbf{X}_m^3\, \mathbf{X}_m^2\, \mathbf{X}_m^1\right]_{m=1(1)8} \circ \left[\boldsymbol{\Omega}_1^3\, \boldsymbol{\Omega}_1^1\, \boldsymbol{\Omega}_1^3\, \boldsymbol{\Omega}_1^2\, \boldsymbol{\Omega}_1^1\right] =$$
$$= \left[\mathbf{X}_m^3 \circ \boldsymbol{\Omega}_1^3\;\; \mathbf{X}_m^1 \circ \boldsymbol{\Omega}_1^1\;\; \mathbf{X}_m^3 \circ \boldsymbol{\Omega}_1^3\;\; \mathbf{X}_m^2 \circ \boldsymbol{\Omega}_1^2\;\; \mathbf{X}_m^1 \circ \boldsymbol{\Omega}_1^1\right]_{m=1(1)8}$$

Κατά τον τρόπο αυτό

$$\boldsymbol{\Xi}^1 = \begin{bmatrix} 3 & 8 & 2 & 5 & 6 \\ 2 & 3 & 5 & 3 & 7 \\ 7 & 9 & 2 & 4 & 8 \\ 5 & 6 & 3 & 1 & 8 \\ 6 & 4 & 7 & 9 & 4 \\ 4 & 5 & 1 & 2 & 5 \\ 9 & 1 & 4 & 8 & 2 \\ 8 & 7 & 6 & 5 & 1 \end{bmatrix} \quad \Rightarrow \quad EDS_1 = \begin{bmatrix} 3 & 8 & 2 & 5 & 6 & 0 \\ 2 & 3 & 5 & 3 & 7 & 0 \\ 7 & 9 & 2 & 4 & 8 & 1 \\ 5 & 6 & 3 & 1 & 8 & 0 \\ 6 & 4 & 7 & 9 & 4 & 0 \\ 4 & 5 & 1 & 2 & 5 & 1 \\ 9 & 1 & 4 & 8 & 2 & 1 \\ 8 & 7 & 6 & 5 & 1 & 1 \end{bmatrix}$$

όπου το παραγόμενο εξελικτικό σύνολο δεδομένων EDS_1 προκύπτει αφού ο πίνακας εξόδου Y προσαρτηθεί στο $\boldsymbol{\Xi}^1$ κατά γραμμή.

Υπό την ίδια έννοια, εφόσον το τμήμα πυρήνα του επόμενου εκπαιδευτή $\boldsymbol{\Omega}_2$ είναι (1,1), το αντίστοιχο εξελικτικό σύνολο δεδομένων που θα προκύψει θα αποτελείται από το μέσο όρο κάθε σχηματιζόμενου τμήματος στα πρωτογενή δεδομένα:

$$\boldsymbol{\Xi}^2 = \mathbf{X} \circ \boldsymbol{\Omega}_2 = \left[\mathbf{X}_m^1\, \mathbf{X}_m^2\, \mathbf{X}_m^4\, \mathbf{X}_m^3\right]_{m=1(1)8} \circ \left[\boldsymbol{\Omega}_2^1\, \boldsymbol{\Omega}_2^2\, \boldsymbol{\Omega}_2^4\, \boldsymbol{\Omega}_2^3\right] =$$
$$= \left[\mathbf{X}_m^1 \circ \boldsymbol{\Omega}_2^1\;\; \mathbf{X}_m^2 \circ \boldsymbol{\Omega}_2^2\;\; \mathbf{X}_m^4 \circ \boldsymbol{\Omega}_2^4\;\; \mathbf{X}_m^3 \circ \boldsymbol{\Omega}_2^3\right]_{m=1(1)8}$$

και συνεπώς

$$\mathbf{\Xi}^2 = \begin{bmatrix} 3 & 7\frac{1}{2} & 3\frac{3}{4} & 6\frac{2}{3} \\ 2 & 6\frac{1}{2} & 4 & 6 \\ 7 & 4\frac{1}{2} & 4\frac{1}{4} & 6 \\ 5 & 4\frac{1}{2} & 3\frac{3}{4} & 5\frac{1}{3} \\ 6 & 2\frac{1}{2} & 6 & 4\frac{2}{3} \\ 4 & 6\frac{1}{2} & 4\frac{1}{4} & 5\frac{1}{3} \\ 9 & 6\frac{1}{2} & 4\frac{2}{4} & 4\frac{2}{3} \\ 8 & 3 & 4\frac{1}{4} & 3 \end{bmatrix} \quad \Rightarrow \quad EDS_2 = \begin{bmatrix} 3 & 7\frac{1}{2} & 3\frac{3}{4} & 6\frac{2}{3} & 0 \\ 2 & 6\frac{1}{2} & 4 & 6 & 0 \\ 7 & 4\frac{1}{2} & 4\frac{1}{4} & 6 & 1 \\ 5 & 4\frac{1}{2} & 3\frac{3}{4} & 5\frac{1}{3} & 0 \\ 6 & 2\frac{1}{2} & 6 & 4\frac{2}{3} & 0 \\ 4 & 6\frac{1}{2} & 4\frac{1}{4} & 5\frac{1}{3} & 1 \\ 9 & 6\frac{1}{2} & 4\frac{2}{4} & 4\frac{2}{3} & 1 \\ 8 & 3 & 4\frac{1}{4} & 3 & 1 \end{bmatrix}$$

όπου το δεύτερο εξελικτικό σύνολο δεδομένων EDS_2 προκύπτει εφόσον ο πίνακας εξόδου Y προσαρτηθεί στο $\mathbf{\Xi}^2$ κατά γραμμή.

Δεδομένου ότι ο πίνακας των εκπαιδευτών αποτελείται από έξι μέλη, αντιστοίχως θα μορφοποιηθούν έξι εξελικτικά σύνολα δεδομένων, ένα από κάθε εκπαιδευτή, ο μηχανισμός παραγωγής των οποίων είναι παρόμοιος. Προς χάριν ολοκλήρωσης του παραδείγματος με όλους τους πιθανούς συνδυασμούς των γονιδίων του τμήματος πυρήνα, πιο κάτω παρατίθενται τα εξελικτικά σύνολα δεδομένων EDS_3 και EDS_4, των οποίων η συμπεριφορά υπαγορεύεται από τα εκπαιδευτικά τμήματα πυρήνα (0,1) και (1,0) αντίστοιχα. Συνεπώς:

$$\begin{aligned} &\mathbf{\Xi}^3 = \mathbf{X} \circ \mathbf{\Omega}_3 = \left[\mathbf{X}_m^2\, \mathbf{X}_m^3 \mathbf{X}_m^3\, \mathbf{X}_m^1\, \mathbf{X}_m^1\right]_{m=1(1)8} \circ \left[\mathbf{\Omega}_3^2 \mathbf{\Omega}_3^3 \mathbf{\Omega}_3^3 \mathbf{\Omega}_3^1 \mathbf{\Omega}_3^1\right] = \\ &= \left[\mathbf{X}_m^2 \circ \mathbf{\Omega}_3^2 \;\; \mathbf{X}_m^3 \circ \mathbf{\Omega}_3^3 \;\; \mathbf{X}_m^3 \circ \mathbf{\Omega}_3^3 \;\; \mathbf{X}_m^1 \circ \mathbf{\Omega}_3^1 \;\; \mathbf{X}_m^1 \circ \mathbf{\Omega}_3^1\right]_{m=1(1)8} \end{aligned}$$

και έτσι

$$\mathbf{\Xi}^3 = \begin{bmatrix} 5\frac{1}{2} & 7 & 4 & 9 & 6 \\ 5\frac{1}{2} & 4 & 3 & 8 & 7 \\ 6\frac{1}{2} & 3 & 4 & 6 & 8 \\ 4\frac{1}{2} & 5 & 2 & 7 & 8 \\ 4\frac{1}{2} & 4 & 8 & 1 & 4 \\ 5\frac{1}{2} & 5 & 3 & 9 & 5 \\ 7 & 4 & 7 & 4 & 2 \\ 5 & 7 & 5 & 3 & 1 \end{bmatrix} \quad \Rightarrow \quad EDS_3 = \begin{bmatrix} 5\frac{1}{2} & 7 & 4 & 9 & 6 & 0 \\ 5\frac{1}{2} & 4 & 3 & 8 & 7 & 0 \\ 6\frac{1}{2} & 3 & 4 & 6 & 8 & 1 \\ 4\frac{1}{2} & 5 & 2 & 7 & 8 & 0 \\ 4\frac{1}{2} & 4 & 8 & 1 & 4 & 0 \\ 5\frac{1}{2} & 5 & 3 & 9 & 5 & 1 \\ 7 & 4 & 7 & 4 & 2 & 1 \\ 5 & 7 & 5 & 3 & 1 & 1 \end{bmatrix}$$

Τέλος, το τέταρτο εξελικτικό σύνολο δεδομένων θα παραχθεί ως εξής:

$$\mathbf{\Xi}^4 = \mathbf{X} \circ \mathbf{\Omega}_4 = \left[\mathbf{X}^1_m \, \mathbf{X}^4_m \mathbf{X}^3_m \, \mathbf{X}^1_m \, \mathbf{X}^1_m\right]_{m=1(1)8} \circ \left[\mathbf{\Omega}^1_4 \mathbf{\Omega}^4_4 \mathbf{\Omega}^3_4 \mathbf{\Omega}^1_4 \mathbf{\Omega}^1_4\right] =$$
$$= \left[\mathbf{X}^1_m \circ \mathbf{\Omega}^1_4 \;\; \mathbf{X}^4_m \circ \mathbf{\Omega}^4_4 \;\; \mathbf{X}^3_m \circ \mathbf{\Omega}^3_4 \;\; \mathbf{X}^1_m \circ \mathbf{\Omega}^1_4 \;\; \mathbf{X}^1_m \circ \mathbf{\Omega}^1_4\right]_{m=1(1)8}$$

το οποίο θα δώσει

$$\mathbf{\Xi}^4 = \begin{bmatrix} 0 & 6 & 4 & 0 & 0 \\ 0 & 6 & 4 & 0 & 0 \\ 0 & 7 & 4 & 0 & 0 \\ 0 & 3 & 3 & 0 & 0 \\ 0 & 5 & 4 & 0 & 0 \\ 0 & 6 & 6 & 0 & 0 \\ 0 & 7 & 2 & 0 & 0 \\ 0 & 7 & 4 & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad EDS_4 = \begin{bmatrix} 0 & 6 & 4 & 0 & 0 & 0 \\ 0 & 6 & 4 & 0 & 0 & 0 \\ 0 & 7 & 4 & 0 & 0 & 1 \\ 0 & 3 & 3 & 0 & 0 & 0 \\ 0 & 5 & 4 & 0 & 0 & 0 \\ 0 & 6 & 6 & 0 & 0 & 1 \\ 0 & 7 & 2 & 0 & 0 & 1 \\ 0 & 7 & 4 & 0 & 0 & 1 \end{bmatrix}$$

Για περαιτέρω εξήγηση του τρόπου αποτύπωσης του χρωμοσώματος κάθε εκπαιδευτή επί των πρωτογενών δεδομένων, ας θεωρήσουμε μια περίπτωση όπως αυτή του Πίνακα 4.2

Στο εν λόγω παράδειγμα, εάν το τμήμα πυρήνα είναι 00, τότε ο εκπαιδευτής

Πίνακας 4.2: Αποτύπωση του χρωμοσώματος εκπαιδευτή επί πρωτογενούς χρονοσειράς και παραγωγή εξελικτικών δεδομένων σύμφωνα με το τμήμα πυρήνα

	Πρωτογενή Δεδομένα									
	44	32	17	8	8	12	1	18	30	48
	Χρωμόσωμα εκπαιδευτή									
	1	0	0	1	1	0	0	1	0	1
Πυρήνας	**Εξελικτικά Δεδομένα**									
00	44			8	8			18		48
11	31			8	7			24		48
01	32			8	8			24		48
10	27			0	11			12		0

εκτελεί τη λειτουργία «Απόρριψη-Ολων-Των-0». Έτσι, η πρωτογενής χρονοσειρά χάνει όλες τις τιμές της (32, 17, 12, 1, 30) για τις οποίες η αντίστοιχη θέση του εκπαιδευτή ισούται με 0, ώστε να προκύψει το εξελικτικό σχήμα 44, 8, 8, 18, 48. Η μορφολογία του τμήματος ενεργοποίησης του πυρήνα καθορίζει πέντε τμήματα για όλες τις υπόλοιπες περιπτώσεις. Όταν αυτό είναι 11, τότε ο εκπαιδευτής εκτελεί τη λειτουργία «Πρώτο-Ένα-Τελευταίο-0 μέσος», εξάγοντας το μέσο όρο των τμημάτων που ορίζονται από το πρώτο 1 και το τελευταίο 0 των γονιδίων του χρωμοσώματός του. Το πρώτο τμήμα καθορίζεται από την αλληλουχία (1, 0, 0) που αντιστοιχεί στις τιμές (44, 32, 17) της πρωτογενούς χρονοσειράς για την οποία εξάγεται ο μέσος όρος. Έτσι, ως πρώτη τιμή του προκύπτοντος εξελικτικού πακέτου δίδεται η τιμή 31. Το μοναδικό 1 στο τέταρτο γονίδιο του εκπαιδευτή, αποτελεί μονοδιάστατο τμήμα, διατηρώντας την αντίστοιχη τιμή 8 στο εξελικτικό σχήμα. Το τρίτο τμήμα καθορίζεται από τα γονίδια αριθμός 5, 6 και 7 του εκπαιδευτή (1, 0, 0). Με εξαγωγή του μέσου των αντίστοιχων τιμών της χρονοσειράς (8, 12, 1) λαμβάνουμε τον αριθμό 7 ως τρίτη τιμή του εξελικτικού πακέτου. Το τέταρτο τμήμα αποτελείται από τα γονίδια 8 και 9 (1, 0) τα οποία αντιστοιχούν στις τιμές 18 και 30 της πρωτογενούς χρονοσειράς. Η μέση τιμή τους (24) δίδει την τέταρτη τιμή της εξέλιξης. Το τελευταίο τμήμα που καθορίζει ο εκπαιδευτής είναι μοναδικής τιμής και διατηρεί την αντίστοιχη τιμή των πρωτογενών δεδομένων (48). Στην περίπτωση, λοιπόν, του δεύτερου εκπαιδευτή με τμήμα πυρήνα 11 προκύπτει η σειρά (31, 8, 7, 24, 48) ως μέλος του δευτερογενούς εξελικτικού πακέτου. Στα ίδια πρότυπα, τα δευτερογενή

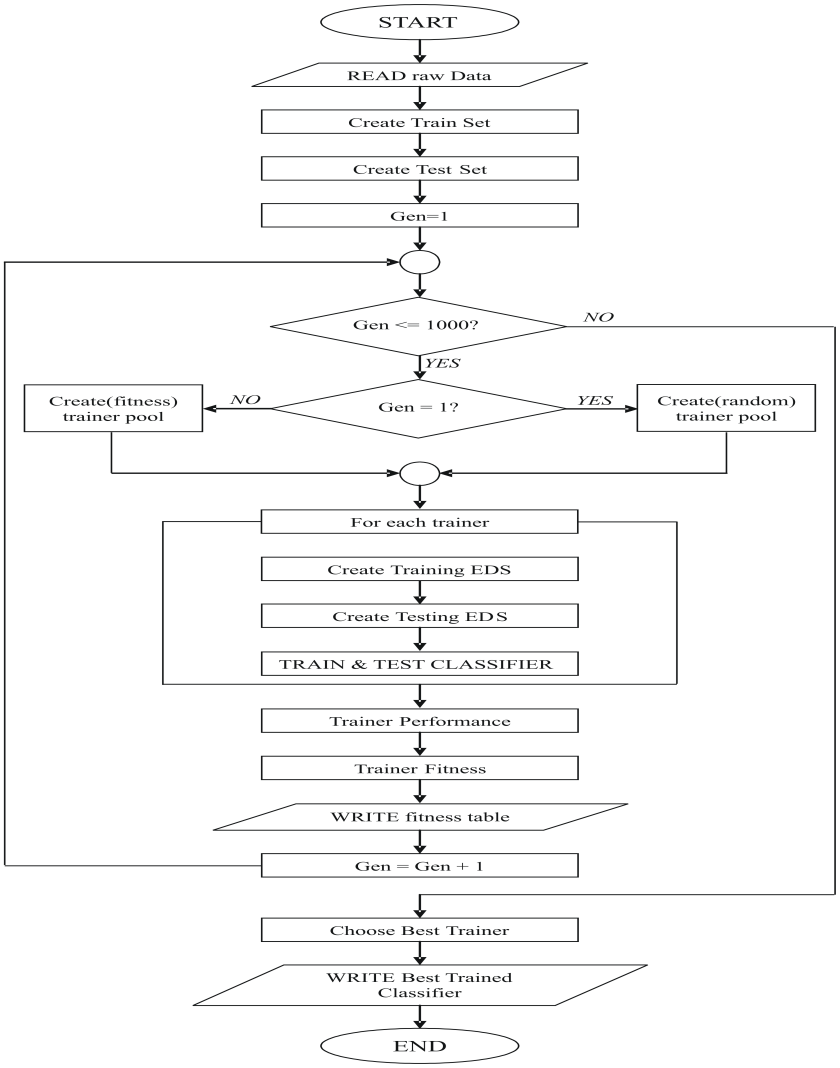
δεδομένα για τους συνδυασμούς πυρήνα 01 και 10 παράγονται με την υλοποίηση των μεθόδων του διάμεσου και της απόστασης μέγιστου - ελάχιστου αντίστοιχα. Με την αποτύπωση της δομής του χρωμοσώματος κάθε εκπαιδευτή στα πρωτογενή δεδομένα παράγεται ένας αριθμός εξελικτικών πακέτων ίσος με τον πληθυσμό των εκπαιδευτών κάθε γενεάς. Αυτά χρησιμοποιούνται στην εκπαίδευση του συστήματος και η προσαρμοστικότητα καθενός αποτιμάται από την απόδοση του εκάστοτε ταξινομητή. Στο κεφάλαιο που ακολουθεί θα παρουσιασθεί αυτός ο τρόπος ανταλλαγής της πληροφορίας μεταξύ των διαφόρων συνιστωσών του συστήματος.

Κεφάλαιο 5
ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ

Στη συνέχεια μελετώνται οι βασικές συνιστώσες του αλγορίθμου, οι οποίες αναλύονται και παρατίθενται ως ψευδοκώδικας, ενώ μέσω διαγράμματος ροής παρουσιάζονται οι σχέσεις και οι συνθήκες που διέπουν την ανταλλαγή πληροφορίας μεταξύ τους. Επίσης, μελετάται το περιβάλλον ανάπτυξης, το γραφικό περιβάλλον διεπαφής χρήστη και η παραμετροποίηση του εργαλείου λογισμικού, όπως αυτό αναπτύχθηκε ώστε να υλοποιήσει τη μέθοδο της Πρότυπης Εξελικτικής Τμηματοποίησης.

Κατά τη φάση του σχεδιασμού της προτεινόμενης μεθόδου, ως πρωταρχικός στόχος τέθηκε η αυξημένη δυνατότητα προσαρμογής του αλγορίθμου στα αρχικά δεδομένα. Η μεθοδολογία που αναπτύχθηκε ουσιαστικά αναζητεί το σχεδιασμό του καλύτερου δυνατού σχήματος τμηματοποίησης στα αρχικά δεδομένα χρονοσειράς, παρακολουθώντας τις επιδόσεις των εκπαιδευτών ενός συγκεκριμένου αριθμού γενεών, ενώ ταυτόχρονα επιλέγει τον καλύτερο ταξινομητή (ΤΝΔ ή ΜΔΥ) για το εκάστοτε πρόβλημα.

Σε αδρές γραμμές το σύστημα λειτουργεί ως εξής: Κατά την αρχικοποίηση, ανασύρει από τον αποθηκευτικό του χώρο την αρχική πρωτογενή χρονοσειρά, από την οποία δημιουργεί το εκπαιδευτικό και ελεγκτικό σύνολο δεδομένων, και στη συνέ-



Σχήμα 5.1: *Διάγραμμα ροής της λειτουργίας του συστήματος.*

χεια δημιουργεί την πρώτη και τις μετέπειτα γενεές. Μέσα σε κάθε γενεά, η πρώτη φάση λειτουργιών αφορά στη σταχυολόγηση μιας δεξαμενής εκπαιδευτών, καθένας από τους οποίους επεξεργάζεται το εκπαιδευτικό και ελεγκτικό σύνολο δεδομένων. Στην περίπτωση που πρόκειται για την πρώτη γενεά, οι εκπαιδευτές δημιουργούνται με τυχαίο τρόπο, ενώ από τη δεύτερη γενεά και στις επόμενες η δεξαμενή αποτελείται από τους εκπαιδευτές της προηγούμενης γενεάς οι οποίοι σημείωσαν τα καλύτερα αποτελέσματα και προκρίνονται βάσει της μεθόδου επιλογής της ρουλέτας. Στο σχήμα 5.1 δίνεται ένα πρωτόλειο διάγραμμα ροής του προτεινόμενου αλγορίθμου.

Listing 5.1: Ψευδοκώδικας Αλγορίθμου Πρότυπης Εξελικτικής Τμηματοποίησης

```
1  Algorithm_PES=PieceWise_Evolutionary_Segmentation:
2
3     READ initial time series data IN tsDat
4     READ structural data in strucDat
5     READ output data in outDat
6
7     crsvrProb := crossover_probability
8     mtnProb := mutation_probability
9     maxGens := maximum_generations
10
11    generation := 1
12
13    WHILE generation <= maxGens DO
14       begin
15          createEDS_Data(generation, tsDat, bestTrnrs, crsvrProb, mtnProb)
16          evaluateTrainer(generation, EDS_data)
17          selectBestTrainers(generation, trnrs)
18          generation := generation + 1
19       end
20
21    WRITE(best_trainer, best_trained_classifier)
```

Πιο συγκεκριμένα, ο αλγόριθμος διακρίνεται σε δύο διαδοχικές φάσεις (κώδικας 5.1), εκ των οποίων στην πρώτη προδιαγράφεται το περιβάλλον εργασίας και γίνεται ο χειρισμός των πρωτογενών δεδομένων (σειρές 3-5). Αρχικά καθορίζονται οι τρέχοντες κατάλογοι εργασίας και αποθήκευσης των αποτελεσμάτων με φυσικό τρόπο στο δίσκο του υπολογιστή και στη συνέχεια υπό αφηρημένη έννοια σε ειδικό αποθηκευτικό χώρο του αλγορίθμου.

Στη φάση αυτή επιτυγχάνεται ο διαχωρισμός των πρωτογενών δεδομένων σε δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση των ταξινομητών του συστήματος και σε δεδομένα μέσω των οποίων θα επιτευχθεί η αξιολόγηση των ταξινομητών και μέσω αυτών η αξιολόγηση των εξελικτικών δεδομένων εκπαίδευσης και κατ' επέκταση των αντίστοιχων εκπαιδευτών. Στη συνέχεια το εκπαιδευτικό και το ελεγκτικό σύνολο δεδομένων αναλύονται ώστε να καθορισθεί το τμήμα που αντιστοιχεί στη χρονοσειρά (σειρά 3), το τμήμα που σχετίζεται με πιθανούς δομι-

κούς παράγοντες εισόδου (σειρά 4) και τέλος το τμήμα που αποτελεί την έξοδο του συστήματος (σειρά 5). Κάθε ένα από αυτά τα τμήματα των αρχικών δεδομένων διαχωρίζεται από τα υπόλοιπα και αποθηκεύεται σε τοπικά δημιουργημένο εμφωλευμένο κατάλογο του αλγορίθμου. Το τμήμα που αντιστοιχεί στην αρχική χρονοσειρά χρησιμοποιείται ως πρότυπο για τη δημιουργία της πρώτης τυχαιοποιημένης γενεάς χρωμοσωμάτων του αλγορίθμου, όσον αφορά στον αριθμό των γονιδίων ανά χρωμόσωμα. Τα λοιπά δεδομένα που αντιστοιχούν σε τυχόντα δομικά ή δυναμικά δεδομένα εισόδου και τα οποία δεν σχετίζονται με χρονοσειρές τίθενται σε εφεδρεία για μελλοντική χρήση. Επίσης καθορίζονται οι πιθανότητες με τις οποίες θα συμβεί ανα-συνδυασμός (σειρά 7) και μετάλλαξη (σειρά 8), όπως και ο μέγιστος αριθμός γενεών (σειρά 9).

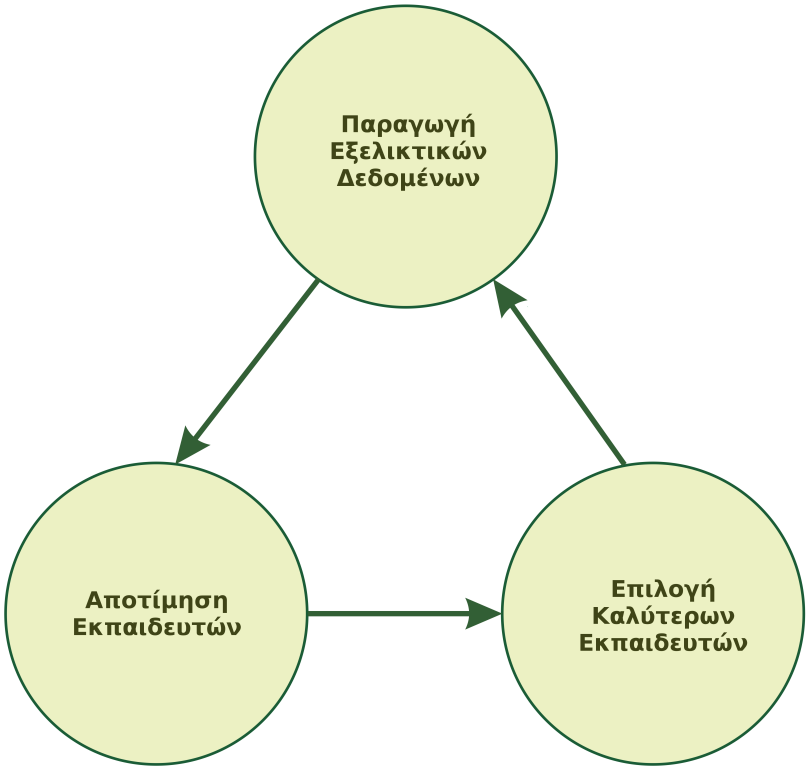
Η δεύτερη φάση του αλγορίθμου χαρακτηρίζεται βασικά από τρείς διακριτές και συνεργαζόμενες μεταξύ τους μεθόδους, μέσω των οποίων δεδομένα μετατρέπονται σε πληροφορία και τελικά αποθηκεύονται ως γνώση του συστήματος (κώδικας 5.1, σειρές 13-19), συνδεόμενες μεταξύ τους από πληροφορία εξελικτικού επιπέδου. Οι μέθοδοι αυτές εκτελούνται διαδοχικά σε σειρά, κάτω από συγκεκριμένο αριθμό επαναλήψεων, αποτελούν δε στην πραγματικότητα το γενετικό αλγόριθμο του συστήματος. Οι μέθοδοι αυτές είναι οι εξής:

α. **Δημιουργία εξελικτικών δεδομένων.** Πρόκειται ουσιαστικά για το σχηματισμό της βασικής γενεάς του αλγορίθμου με πρώτιστο στόχο τη συνάθροιση εξελικτικών εκπαιδευτών (σειρά 15). Στη φάση αυτή επίσης ενσωματώνεται η αποτύπωση του σχήματος τμηματοποίησης των εκπαιδευτών στα πρωτογενή δεδομένα και η παραγωγή εξελικτικών δευτερογενών δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση και τον έλεγχο των ταξινομητών του συστήματος.

β. **Αξιολόγηση εκπαιδευτών** κατά την οποία τα σχηματισθέντα εξελικτικά δεδομένα χρησιμοποιούνται για την εκπαίδευση των ταξινομητών (σειρά 16). Ο βαθμός προσαρμοστικότητας που αποδίδεται στον αντίστοιχο εκπαιδευτή εξαρτάται από την απόδοση κάθε ταξινομητή στη φάση ελέγχου.

γ. **Ενδιάμεση γενεά.** Πρόκειται για μια φάση μετάβασης από τη μια γενεά στην επόμενη με βασικό της σκοπό τη σταχυολόγηση των αποτελεσματικότερων

εκπαιδευτών και τη διενέργεια διαδικασιών αναπαραγωγής μέσω επιλογής, διασταύρωσης και μετάλλαξης (σειρά 17).



Σχήμα 5.2: *Ροή πληροφορίας μεταξύ των βασικών μεθόδων του αλγορίθμου.*

Τα βασικά συστατικά του εξελικτικού αλγορίθμου και οι μεταξύ τους διεργασίες απεικονίζονται στο σχήμα 5.2. Πρόκειται ουσιαστικά για μια επαναληπτική διαδικασία, η οποία συνίσταται στην ανταλλαγή πληροφορίας μεταξύ των βασικών μεθόδων και στην τελική αξιολόγηση των αποτελεσμάτων τους. Αρχικά πυροδοτείται η μέθοδος δημιουργίας εξελικτικών δεδομένων (EDS). Κατά τη μέθοδο αυτή, κάθε μέλος του πληθυσμού των εκπαιδευτών αποτυπώνει το χρωμόσωμά του στην αρχική χρονοσειρά παράγοντας δευτερογενή δεδομένα. Το εξελικτικό αυτό εκπαιδευτικό υλικό τροφοδοτεί την επόμενη μέθοδο όπου συντελείται η αξιολόγηση των εκπαιδευτών. Οι ταξινομητές του συστήματος εκπαιδεύονται με κάθε δευτερογενές σύνολο εκπαίδευσης και η απόδοσή τους ποσοτικοποιείται βάσει του αντίστοιχου δευτερογενούς εξελικτικού ελεγκτικού συνόλου δεδομένων. Η ακρίβεια ταξινόμησης που προκύπτει αποτελεί το χαρακτηριστικό βαθμό προσαρμοστικότητας

ο οποίος αποδίδεται σε κάθε εκπαιδευτή. Στη συνέχεια το σύνολο των εκπαιδευτών, εμπλουτισμένο με τους βαθμούς προσαρμοστικότητας καθενός από αυτούς, οδηγείται προς την τρίτη μέθοδο. Σε αυτή διενεργείται η επιλογή των καλύτερων εκπαιδευτών, οι οποίοι θα αποτελέσουν τον πληθυσμό της επόμενης γενεάς. Κατά τον τρόπο αυτό νέα εξελικτικά δεδομένα θα προκύψουν, για να ξεκινήσει η όλη διαδικασία από την αρχή.

Στη συνέχεια θα εξετασθούν λεπτομερέστερα οι τρεις αυτές βασικές μέθοδοι του αλγορίθμου.

5.1 Δημιουργία Εξελικτικών Δεδομένων

Κατά την εκκίνηση του βρόγχου της εξελικτικής διαδικασίας (κώδικας 5.1, σειρές 13-19), ο αλγόριθμος εκτελεί την πρώτη μέθοδο *CreateEDS_Data*, η οποία σχετίζεται με τη δημιουργία των εξελικτικών δεδομένων.

Listing 5.2: Ψευδοκώδικας Μεθόδου Δημιουργίας Εξελικτικών Δεδομένων

```
1  METHOD createEDS_Data(generation, tsDat, bestTrnrs, crsvrProb, mtnProb):
2
3      numTrnrs := number_of_Trainers_per_Generation
4      lenTrnr := length_of_Trainer_chromosome
5
6      CREATE_ARRAY newTrnrPop
7      CREATE_ARRAY EDS_data
8
9      newTrnrPop := Null
10     EDS_data := Null
11
12     IF generation = 1 THEN
13        begin
14     #a[k][m] denotes the element of array a in k-th line, m-th column
15           FOR i := 1 TO numTrnrs DO
16              FOR j := 1 TO lenTrnr DO
17                 newTrnrPop[i][j] = RANDOM_INT(0,1)
18           FOR i := 1 TO numTrnrs DO
19              begin
20     #a[k] denotes the k-th element of a
21                 newTrnrPop[i].insert(RANDOM_INT(0,1) at position 0)
```

```
22                  newTrnrPop[i].insert(RANDOM_INT(0,1) at position 0)
23              end
24          end
25      ELSE
26          begin
27              WHILE i <= numTrnrs DO
28                  begin
29                      trnr1 := RANDOM_CHOICE(bestTrnrs)
30                      trnr2 := RANDOM_CHOICE(bestTrnrs)
31
32                      crsvrPosition := RANDOM_INT(1, lenTrnr)
33                      crsvrActivate := RANDOM_INT(0,100)
34
35                      IF crsvrActivate < crsvrProb THEN
36                          begin
37      #a[:k] denotes the part of array a from element 0 till element k
38      #a[k:] denotes the part of array a from element k till the end
39                              newTrnrPop.append(trnr1[:crsvrPosition]+
40                                                trnr2[crsvrPosition:])
41                              newTrnrPop.append(trnr2[:crsvrPosition]+
42                                                trnr1[crsvrPosition:])
43                          end
44                      ELSE
45                          COPY_TRAINERS((trnr1,trnr2), newTrnrPop)
46
47                      mtnActivate := RANDOM_INT(0,1000)
48
49                      FOR each individual Trainer DO
50                          FOR each individual gene DO
51                              IF mtnActivate < mtnProb THEN
52                                  IF gene = 0 THEN
53                                      gene := 1
54                                  ELSE
55                                      gene := 0
56                      i := i + 1
57              end
58          end
59
60      FOR trainer := 1 TO numTrnrs DO
```

```
61          begin
62              IF (trainer[0] = 0 AND trainer[1] = 0) THEN
63                  EDS_data := Discard_Zeros(trainer[2:], tsDat)
64              IF (trainer[0] = 1 AND trainer[1] = 1) THEN
65                  EDS_data := FOLZ_average(trainer[2:], tsDat)
66              IF (trainer[0] = 0 AND trainer[1] = 1) THEN
67                  EDS_data := FOLZ_median(trainer[2:], tsDat)
68              IF (trainer[0] = 1 AND trainer[1] = 0) THEN
69                  EDS_data := FOLZ_minmax(trainer[2:], tsDat)
70          end
71
72      RETURN EDS_data
```

Αμέσως μετά την έναρξη της μεθόδου αυτής (κώδικας 5.2) ο αλγόριθμος εισέρχεται σε κλάδο απόφασης (γραμμές 12-58), όπου αξιολογείται αν πρόκειται για την πρώτη ή μια από τις επόμενες γενεές. Η απάντηση στο ερώτημα αυτό είναι κρίσιμη, καθώς η πρώτη γενεά διαφέρει από τις υπόλοιπες κατά το ότι ο πληθυσμός των εκπαιδευτών της συναθροίζεται με εντελώς τυχαίο τρόπο (γραμμές 12-24), ενώ στις υπόλοιπες γενεές προκύπτει κατόπιν διαδικασίας επιλογής (25-58). Συγκεκριμένα, η πρώτη γενεά αποτελείται από αριθμό εκπαιδευτών που καθορίζεται από το χρήστη, καθένας από τους οποίους περιλαμβάνει τυχαιοποιημένο δυαδικό τμήμα δραστηριοποίησης (γραμμή 17) και πυρήνα (γραμμές 21-22). Στην περίπτωση που δεν πρόκειται για την πρώτη γενεά, ο αλγόριθμος επιλέγει τυχαία δύο από τους καλύτερους εκπαιδευτές της προηγούμενης (γραμμές 29-30). Στη συνέχεια, είτε τους ανα-συνδυάζει σε τυχαία θέση του χρωμοσώματος (γραμμές 32-43), είτε τους επιλέγει αυτούσιους για την επόμενη γενεά (γραμμή 45) με συγκεκριμένη πιθανότητα. Επίσης τα γονίδια των θυγατρικών εκπαιδευτών υπόκεινται σε πιθανότητα μετάλλαξης (γραμμές 47-55). Στη συνέχεια ο πίνακας εκπαιδευτών ενημερώνεται με τον πληθυσμό της καινούριας γενεάς και ακολουθεί η διαδικασία αποτύπωσης του σχήματος τμηματοποίησης των εκπαιδευτών επί των πρωτογενών δεδομένων (γραμμές 60-70) για την παραγωγή των εξελικτικών δεδομένων.

5.2 Αξιολόγηση Εκπαιδευτών

Τα χρωμοσώματα-εκπαιδευτές που παράγονται κατά την έναρξη κάθε γενεάς συμπεριφέρονται με προκαθορισμένο τρόπο και τροποποιούν κατάλληλα την αρχική χρονοσειρά δημιουργώντας εξελικτικά παράγωγα, τα οποία βαθμονομούνται με συγκεκριμένη βαθμολογία προσαρμοστικότητας (fitness score). Η βαθμολογία αυτή εξάγεται από συνάρτηση που ποσοτικοποιεί τη συνάφεια καθενός προς την άριστη λύση του προβλήματος. Η καταλληλότητα κάθε εξελικτικού δευτερογενούς συνόλου δεδομένων, άρα και του αντίστοιχου εκπαιδευτή, εξάγεται με διαφορετικό τρόπο για κάθε τύπο προβλήματος. Ως μέτρο ποσοτικοποίησης της καταλληλότητας για τα προβλήματα ταξινόμησης λαμβάνεται η Ακρίβεια (Acc) του εκάστοτε ταξινομητή, ενώ για τα προβλήματα πρόβλεψης το σύστημα εξάγει το βαθμό προσαρμοστικότητας βάσει του μέσου τυπικού τετραγωνικού σφάλματος (RMSE: Root Mean Square Error) που αποδίδει κάθε ταξινομητής στον πυρήνα του συστήματος.

Κατά τη σύγκριση δύο συνόλων δεδομένων, εκ των οποίων το ένα προέρχεται από θεωρητική πρόβλεψη και το άλλο από πραγματικές μετρήσεις μιας μεταβλητής, το RMSE των διαφορών τους κατά ζεύγη αποτελεί ένα μέτρο αξιολόγησης του σφάλματος της θεωρητικής πρόβλεψης. Συνεπώς, εάν ο ταξινομητής εκτελεί n προβλέψεις $\hat{y}_i$ από ένα σύνολο δεδομένων, οι πραγματικές μετρήσεις του οποίου είναι y_i, όπου $i = 1, 2, \ldots, n$, τότε το μέσο τυπικό τετραγωνικό σφάλμα της πρόβλεψης δίνεται από τη σχέση:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{5.2.1}$$

Ως Ακρίβεια του ταξινομητή ορίζεται ο λόγος του αθροίσματος των Αληθώς Θετικά (TP: True Positive) και Αληθώς Αρνητικά (TN: True Negative) ταξινομημένων δειγμάτων, προς τα συνολικά δείγματα που περιλαμβάνονται στα δεδομένα ελέγχου. Συνεπώς, ο παρονομαστής του λόγου της ακρίβειας προκύπτει εάν στο άθροισμα των δύο παραπάνω όρων προστεθούν οι τιμές των Ψευδώς Θετικά (FP: False Positive) και Ψευδώς Αρνητικά (FN: False Negative) ταξινομημένων δειγμάτων. Έτσι:

$$\text{Acc} = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.2.2}$$

Αφού καθορισθεί ο πληθυσμός των εκπαιδευτών της τρέχουσας γενεάς του αλγορίθμου και παραχθούν τα αντίστοιχα εξελικτικά δευτερογενή δεδομένα, ξεκινά η αξιολόγηση της απόδοσης κάθε εκπαιδευτή, η οποία υλοποιείται μέσω της μεθόδου *evaluateTrainer* (κώδικας 5.1, σειρά 14). Οι διαδικασίες της μεθόδου εμφανίζονται στον κώδικα 5.3.

Listing 5.3: Ψευδοκώδικας Μεθόδου Αξιολόγησης των εκπαιδευτών

```
1  METHOD evaluateTrainer(generation,EDS_data):
2
3      CREATE_ARRAY trnrs
4      CREATE_ARRAY evaluator_type
5      CREATE_ARRAY EDS_trn
6      CREATE_ARRAY EDS_tst
7      CREATE_ARRAY classifiers
8      CREATE_ARRAY parameters
9
10     trnrs := Null
11     evaluator_type := ("Acc","RMSE") #"Acc" Accuracy
12                                      #"RMSE" Root Mean Square Error
13     EDS_trn := EDS_data(training_part)
14     EDS_tst := EDS_data(testing_part)
15     classifiers := (ANN,SVM) #ANN Artificial Neural Network object
16                              #SVM Support Vector Machine object
17     parameters := (ANN(numHidNeurs,ActivFuncs)
18                    SVM(C,Gamma,Degree))
19     iters := iterations_of_training
20     trnrEval := 0
21     trnrFit := 0
22     maxAcc := maximum_accuracy
23     minRMSE := minimum_root_mean_square_error
24
25     FOR clsfr IN classifiers DO
26         begin
27             clsfr.ASSEMBLE(parameters)
28             clsfr.TRAIN(EDS_trn,iters)
29             clsfr.SAVE_TRAINED
30             IF evaluator_type = "Acc" THEN
31                 begin
```

```
32                tp,fp,fn,tn := 0
33                IF clsfr.RUN(EDS_tst) = 1 AND EDS_tst.out = 1 THEN
34                   tp := tp+1
35                IF clsfr.RUN(EDS_tst) = 1 AND EDS_tst.out = 0 THEN
36                   fp := fp+1
37                IF clsfr.RUN(EDS_tst) = 0 AND EDS_tst.out = 1 THEN
38                   fn := fn+1
39                IF clsfr.RUN(EDS_tst) = 0 AND EDS_tst.out = 0 THEN
40                   tn := tn+1
41                trnrEval := (tp+tn)/(tp+fp+fn+tn)
42                trnrFit := 1/ABS(trnrEval-maxAcc)
43             end
44          IF evaluator_type = "RMSE" THEN
45             begin
46                trnrEval := RMSE(clsfr.RUN(EDS_tst), EDS_tst.out)
47                trnrFit := 1/ABS(trnrEval-minRMSE)
48             end
49          trnrs.UPDATE(generation,trainer,clsfr,trnrEval,trnrFit)
50       end
51
52    RETURN trnrs
```

Αρχικά καθορίζεται ο τύπος αξιολόγησης που θα εκτελεσθεί: υπολογισμός της ακρίβειας ή του μέσου τυπικού τετραγωνικού σφάλματος (γραμμή 11). Κάθε τύπος αξιολόγησης αντιστοιχεί και σε διαφορετικό τύπο προβλήματος: η ακρίβεια χρησιμοποιείται για τα προβλήματα ταξινόμησης, ενώ το μέσο τυπικό τετραγωνικό σφάλμα για τα προβλήματα πρόβλεψης. Τα εξελικτικά δεδομένα που παρήχθησαν από την προηγούμενη μέθοδο κανονικοποιούνται με τύπο κανονικοποίησης που ορίζει ο χρήστης, καθορίζεται ο αριθμός των εκπαιδευτικών επαναλήψεων και διαχωρίζονται σε εκπαιδευτικό και ελεγκτικό σύνολο (σειρές 13-14), ενώ ταυτόχρονα καθορίζονται οι ταξινομητές του συστήματος και οι παράμετροί τους (σειρές 15-18). Στη συνέχεια, ο αλγόριθμος προχωρεί στην εκπαίδευση του αντίστοιχου ταξινομητή (σειρές 25-29) και την αξιολόγηση κάθε εκπαιδευτή με βάση το ελεγκτικό σύνολο εξελικτικών δεδομένων (σειρές 30-43 και 44-48), εξάγοντας κατά περίπτωση την ακρίβεια ή το RMSE και υπολογίζοντας το βαθμό προσαρμοστικότητας του εκπαιδευτή. Τέλος, ο πίνακας των εκπαιδευτών ενημερώνεται ανάλογα (σειρά 49).

Το σύστημα επίσης είναι κατά τέτοιο τρόπο δομημένο ώστε να έχει τη δυνατότητα να παρακολουθεί την αποτελεσματικότητα κάθε εκπαιδευτή, τουτέστιν τη διακριτική του ικανότητα, επίσης και με την εξαγωγή συγκεκριμένων δεικτών. Συγκεκριμένα, στην απόδοση κάθε εκπαιδευτή, και με βάση το εκάστοτε πρόβλημα αλλά και τις ανάγκες του ερευνητή, σημαντικά μπορεί να αποδειχθούν τα μέτρα της Ευαισθησίας (Sens: Sensitivity), της Ειδικότητας (Spec: Specificity), καθώς επίσης και της Θετικής και Αρνητικής Προγνωστικής Αξίας (PPV: Positive Prediction Value και NPV: Negative Prediction Value αντίστοιχα). Τα αντίστοιχα μέτρα ορίζονται ως εξής:

- Ευαισθησία (Sens):
$$\text{Sens} = \frac{TP}{TP + FN} \tag{5.2.3}$$
- Ειδικότητα (Spec):
$$\text{Spec} = \frac{TN}{FP + TN} \tag{5.2.4}$$
- Θετική Προγνωστική Αξία (PPV):
$$\text{PPV} = \frac{TP}{TP + FP} \tag{5.2.5}$$
- Αρνητική Προγνωστική Αξία (NPV):
$$\text{NPV} = \frac{TN}{TN + FN} \tag{5.2.6}$$

Από τη στιγμή που εξάγεται το μέτρο ποσοτικοποίησης της προσαρμοστικότητας κάθε εξελικτικού συνόλου δεδομένων —Ακρίβεια ή RMSE —μετά τη φάση της εκπαίδευσης και του ελέγχου των ταξινομητών, το σύστημα προχωρεί στον υπολογισμό του βαθμού προσαρμοστικότητας που αποδίδεται στον αντίστοιχο εκπαιδευτή. Ουσιαστικά, πρόκειται για αξιολόγηση της διακριτικής δυνατότητας του εκάστοτε εκπαιδευτή η οποία ποσοτικοποιείται μέσω της εγγύτητάς της προς μια άριστη λύση που ορίζεται κατά την αρχικοποίηση του αλγορίθμου.

Εστω r_{ij} η ακρίβεια ενός ταξινομητή ο οποίος εκπαιδεύθηκε και ελέγχθηκε μέσω εξελικτικών δεδομένων που προέκυψαν από τον i-οστό εκπαιδευτή της j-οστής

γενεάς πληθυσμών του αλγορίθμου. Τότε, ο βαθμός προσαρμοστικότητας f_{ij} που αποδίδεται στον εκπαιδευτή αυτόν θα πρέπει να είναι

$$f_{ij} = \frac{1}{|r_{ij} - h|} \tag{5.2.7}$$

όπου h είναι το κατώφλι ακρίβειας μέσω του οποίου μεγιστοποιείται η συνάρτηση f_{ij}. Είναι φανερό ότι καθώς η τιμή της f_{ij} αυξάνει, η τιμή της r_{ij} πλησιάζει εγγύτερα στο h, με αποτέλεσμα το προτεινόμενο σύστημα να τείνει στην ιδεατή λύση h που τέθηκε αυτόματα κατά την αρχικοποίηση του αλγορίθμου. Η μεγιστοποίηση του μεγέθους f_{ij} έχει ως αποτέλεσμα την παραγωγή όλο και περισσότερο ισχυρών πληθυσμών εκπαιδευτών αυξημένης μέσης προσαρμοστικότητας έναντι των αρχικών δεδομένων του προβλήματος. Είναι πολύ σημαντικό στο σημείο αυτό να υπογραμμισθεί το γεγονός ότι η μέθοδος υπολογισμού της απόδοσης κάθε ταξινομητή εκτελείται κατά τη διάρκεια του ελέγχου της εκπαίδευσης και όχι κατά την εκπαίδευση αυτή καθ' εαυτήν. Το ελεγκτικό σύνολο δεδομένων έχει επίσης αναπαρασταθεί με βάση το σχήμα που προτείνει ο αντίστοιχος εκπαιδευτής και κατά τη διάρκεια του ελέγχου θεωρείται άγνωστο από τον ταξινομητή. Η στρατηγική αυτή εξασφαλίζει υψηλότερο βαθμό δι-επικύρωσης (cross validation) με άγνωστα υποδείγματα χρονοσειρών και προωθεί την παραγωγή όλο και ισχυρότερων ταξινομητών με αυξημένες ικανότητες γενίκευσης, αφού εμμέσως στοχεύει στην καταλληλότερη ρύθμιση του πλέγματος των συναπτικών βαρών τους. Στη συνέχεια ο αλγόριθμος ενημερώνει κάθε εκπαιδευτή, ενσωματώνοντας σε αυτόν το βαθμό προσαρμοστικότητας που του αντιστοιχεί.

5.3 Ενδιάμεση γενεά

Από το σημείο αυτό και μετά, οι επερχόμενες γενεές του αλγορίθμου μορφοποιούνται βάσει μιας πολιτικής επιλογής σύμφωνα με την οποία προκρίνονται οι καταλληλότεροι εκπαιδευτές για να επιβιώσουν στην επόμενη γενεά. Το συναρτησιακό αντικείμενο επιλογής ρουλέτας, όπως υλοποιήθηκε στον κώδικα του προτεινόμενου συστήματος, αποτελεί τον παράγοντα κατάρτισης της γενετικής δεξαμενής κάθε επόμενης γενεάς. Ως γενετική δεξαμενή ορίζουμε μια εδιάμεση φάση, στο μεσοδιάστημα μεταξύ δυο διαδοχικών γενεών, η οποία αποτελεί τον κύριο χώρο ανα-

παραγωγής νέων χρωμοσωμάτων/εκπαιδευτών για τη σχηματοποίηση της επόμενης γενεάς. Αφού αντιστοιχισθεί ο βαθμός προσαρμοστικότητας σε κάθε εκπαιδευτή, το σύστημα εκκινεί τη διαδικασία επιλογής και πλήρωσης της γενετικής δεξαμενής σύμφωνα με τη μέθοδο της ρουλέτας η οποία υλοποιείται από τον προτεινόμενο αλγόριθμο. Η μέθοδος αυτή καθορίζει ότι η πιθανότητα επιλογής p_{ij} του i-οστού εκπαιδευτή $(i = 1, 2, \ldots, V)$ της j-οστής γενεάς πληθυσμών $(j = 1, 2, \ldots, M)$ του αλγορίθμου, ισούται με το λόγο του βαθμού προσαρμοστικότητας του συγκεκριμένου εκπαιδευτή, προς το σύνολο της προσαρμοστικότητας όλων των εκπαιδευτών της συγκεκριμένης γενεάς:

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{V} f_{ij}}, \quad \forall\, j = 1, 2, ..., M \tag{5.3.1}$$

Η μέθοδος της ρουλέτας ενσωματώνει ένα χειριστή, ο οποίος επιλέγει να διαιωνίσει τα πλέον προσαρμοσμένα χρωμοσώματα κατ' αναλογία της προσαρμοστικότητάς τους. Η διαδικασία επιλογής των πλέον προσαρμοσμένων εκπαιδευτών υλοποιείται μέσω της μεθόδου *selectBestTrainers* (κώδικας 5.1, σειρά 17). Οι διαδικασίες της μεθόδου εμφανίζονται στον κώδικα 5.4. Η μέθοδος της ρουλέτας επιλογής των πλέον προσαρμοσμένων εκπαιδευτών ξεκινά δημιουργώντας τρία διαφορετικά αποθετήρια για το βαθμό προσαρμοστικότητας (*fitnessList*, σειρά 3), το σχετικό βαθμό προσαρμοστικότητας (*relProbList*, σειρά 4) και τη σωρευτική πιθανότητα επιλογής (*cumulList*, σειρά 5). Επίσης δημιουργεί το αποθετήριο των καλύτερων εκπαιδευτών (σειρά 6). Στο πρώτο αποθηκεύονται οι βαθμοί προσαρμοστικότητας των εκπαιδευτών της απερχόμενης γενεάς (σειρά 13), όπως υπολογίσθηκαν κατά την προηγούμενη μέθοδο και εξάγεται το άθροισμά τους (σειρά 15).

Listing 5.4: Ψευδοκώδικας μεθόδου επιλογής προσαρμοσμένων εκπαιδευτών

```
1 METHOD selectBestTrainers(generation, trnrs):
2
3     CREATE_ARRAY fitnessList
4     CREATE_ARRAY relProbList
5     CREATE_ARRAY cumulList
6     CREATE_ARRAY bestTrnrs
7
8     fitnessList := Null
```

```
 9     relProbList := Null
10     cumulList := Null
11     bestTrnrs := Null
12
13     fitnessList.APPEND(generation,trnrs.trnrFit)
14
15     fit_sum := SUM(fitnessList.trnrFit)
16
17     FOR i IN fitnessList DO
18         relProbList.APPEND(i/fit_sum)
19
20     cumulValue := 0
21
22     FOR prob IN relProbList DO
23         begin
24             cumulValue := cumulValue + prob
25             cumulList.APPEND(cumulValue)
26         end
27
28     WHILE length(bestTrnrs) <= 100 DO
29         begin
30             rndm := RANDOM(0,1)
31             FOR (index,cumulProb) IN ENUM(cumulList):
32                 IF rndm <= cumulProb THEN
33                     begin
34                         bestTrnrs.append(generation,trnrs[index])
35                         BREAK OPERATION
36                     end
37         end
38
39     RETURN bestTrnrs
```

Κάθε ατομικός βαθμός προσαρμοστικότητας διαιρείται με το συνολικό αυτό άθροισμα και τα αποτελέσματα αποθηκεύονται στον πίνακα σχετικής προσαρμοστικότητας (γραμμές 17-18). Οι σχετικοί βαθμοί προσαρμοστικότητας αθροίζονται ανά δύο και τα αποτελέσματα αποθηκεύονται στον πίνακα σωρευτικής πιθανότητας (*cumulList*) (σειρές 20-26). Στην επόμενη φάση ο αλγόριθμος ενημερώνει το αποθετήριο των καλύτερων εκπαιδευτών που έχει ήδη δημιουργηθεί κατά τη διάρ-

κεια αρχικοποίησης του αλγορίθμου με τους πιο προσαρμοσμένους εκπαιδευτές (*bestTrnrs*), εκείνους δηλαδή που παρουσιάζουν το μεγαλύτερο βαθμό προσαρμοστικότητας (σειρές 28-37). Το αποθετήριο αυτό έχει σταθερή χωρητικότητα εκατό εκπαιδευτών, και αποτελείται από μέλη που έχουν επιλεγεί περισσότερες της μιας φορές. Εξαιτίας του τρόπου λειτουργίας της μεθόδου της ρουλέττας, η συχνότητα των καλύτερων εκπαιδευτών στον πληθυσμό του αποθετηρίου θα είναι τόσο μεγαλύτερη, όσο μεγαλύτερος είναι ο βαθμός προσαρμοστικότητάς τους. Για κάθε επανάληψη, ο αλγόριθμος παράγει έναν τυχαίο αριθμό στο διάστημα $(0, 1)$ (σειρά 30) τον οποίο συγκρίνει με κάθε μέλος του αποθετηρίου σωρευτικής πιθανότητας (σειρά 32), σημειώνοντας ταυτόχρονα το δείκτη (index) του μέλους στο αποθετήριο. Στην περίπτωση που ο τυχαίος αριθμός είναι μικρότερος ή ίσος της σωρευτικής πιθανότητας, ο αλγόριθμος ενημερώνει τον πίνακα επιλογής με ένα αντίγραφο του αντίστοιχου εκπαιδευτή με βάση το δείκτη του (σειρά 34) και ταυτόχρονα τερματίζει τη διαδικασία για την τρέχουσα επανάληψη (σειρά 35).

Είναι επόμενο ότι αυτού του είδους η διαδικασία προωθεί την επιλογή εκπαιδευτών με υψηλότερους βαθμούς προσαρμοστικότητας ως καταλληλότερες λύσεις για το πρόβλημα. Αυτό επιτυγχάνεται εφόσον τέτοιου είδους χρωματοσώματα θα έχουν μεγαλύτερη συχνότητα εμφάνισης στο τελικό αποθετήριο επιλογής. Η γενετική αυτή δεξαμενή είναι δομημένη με τέτοιο τρόπο ώστε, καθώς η συχνότητα εμφάνισης των πλέον προσαρμοσμένων χρωμοσωμάτων αυξάνει, τόσο η πιθανότητα επιλογής τους για την επόμενη γενεά να αυξάνει επίσης, χωρίς παρόλα αυτά τα λιγότερο ικανά χρωμοσώματα να αποκλείονται παντελώς από την όλη διαδικασία. Έτσι, ακολουθείται μια πολιτική προστασίας των πλέον προσαρμοσμένων εκπαιδευτών, δηλαδή αυτών που παρουσιάζουν τους υψηλότερους βαθμούς προσαρμοστικότητας. Στο πλαίσιο αυτό, εκπαιδευτές με σχετικά υψηλό βαθμό προσαρμοστικότητας είναι λιγότερο πιθανό να εξαλειφθούν από τη γενετική δεξαμενή. Στον αντίποδα, λιγότερο προσαρμοσμένοι εκπαιδευτές, παρουσιάζουν μια μικρή πιθανότητα να εμφανισθούν στην τελική γενετική δεξαμενή επιλογής. Η πολιτική αυτή βασίζεται στην υπόθεση ότι λιγότερο ισχυρές λύσεις, αν και σαν ολοκληρωμένο χρωμόσωμα χαρακτηρίζονται από χαμηλές αποδόσεις, εντούτοις είναι πιθανόν να περιλαμβάνουν κατάλληλα δομικά συστατικά[1] για μελλοντικές διασταυρώσεις. Η

[1] *σχήματα* κατά την υπόθεση των δομικών στοιχείων (ΒΒΗ)

πολιτική που ακολουθεί ο προτεινόμενος αλγόριθμος έχει ως αποτέλεσμα κάποιες από τις πιο αδύναμες λύσεις του προβλήματος να επιβιώνουν της «από κάτω» σάρωσης του αλγόριθμου μεταφέροντας τα δυνητικά ισχυρά τους γονίδια ή γονιδιακά σχήματα στους απογόνους της νέας γενεάς.

Από τη στιγμή που ολοκληρώνεται ο καθορισμένος πληθυσμός της ενδιάμεσης γενεάς, οι γενετικοί μηχανισμοί του αλγορίθμου επανεκκινούν την πρώτη μέθοδο δημιουργίας εξελικτικών δεδομένων (κώδικας 5.1, σειρά 13). Εφόσον πρόκειται για γενεά διάφορη της πρώτης, η διαδικασία συνίσταται στη συγκέντρωση των εκπαιδευτών και το ζευγάρωμα μεταξύ τους ώστε να ξεκινήσει η αναπαραγωγή των απογόνων. Όπως έχει ήδη αναφερθεί, το σύστημα επιλέγει κάθε φορά ένα τυχαίο ζεύγος χρωμοσωμάτων από τη γενετική δεξαμενή για ζευγάρωμα. Για κάθε τέτοιο τυχαίο ζεύγος, υπάρχει μια προκαθορισμένη πιθανότητα επιλογής αυτούσιου του ενός από τα δύο χρωμοσώματα ή ανασυνδυασμού και μετάλλαξης αυτών. Ανασυνδυασμός ονομάζεται η διαδικασία κατά την οποία τα δύο πατρικά χρωμοσώματα συνεισφέρουν με διαφορετικά συμπληρωματικά τμήματα του γονιώματός τους στην παραγωγή του απογόνου. Στον προτεινόμενο αλγόριθμο το «σημείο τομής» κάθε πατρικού χρωμοσώματος είναι τυχαίο, εντός βεβαίως των ορίων του γονιώματός του. Η μετάλλαξη αναφέρεται στην αλλαγή ενός – ή περισσότερων – δυαδικού γονιδίου του θυγατρικού χρωματοσώματος, από τη μια κατάσταση (1 ή 0) στην άλλη (0 ή 1). Η βέλτιστη ρύθμιση των πιθανοτήτων ανασυνδυασμού και μετάλλαξης θεωρείται ως ύψιστης σημασίας παραμετροποίηση του αλγορίθμου [163]. Υπερβολικά μικροί ρυθμοί μετάλλαξης πιθανόν να οδηγήσουν σε γενετική παρέκκλιση (genetic drift), δηλαδή το στατιστικό αποτέλεσμα που προκαλείται από την πιθανότητα επιβίωσης των αλληλομόρφων. Στην προκείμενη περίπτωση πρόκειται για τις παραλλαγές 0 ή 1 ενός γονιδίου και την επίδραση που αυτή επιφέρει στο χρωμόσωμα. Θετική γενετική παρέκκλιση καθιστά τον αλληλόμορφο κυρίαρχο στη γενετική δεξαμενή, ενώ αντίθετα αρνητική τον αφανίζει. Αμφότερα τα δύο άκρα της παρέκκλισης ασκούν ιδιαίτερα αρνητικές πιέσεις στη γενετική δεξαμενή, αφού μειώνουν την παραλλακτικότητα σε απαράδεκτα επίπεδα. Το ίδιο ισχύει επίσης και για την πιθανότητα ανασυνδυασμού. Η διαδικασία ζευγαρώματος συνεχίζεται έως ότου συμπληρωθεί ο προκαθορισμένος αριθμός εκπαιδευτών της επόμενης γενεάς. Η διαδικασία αυτή εξασφαλίζει τη γενετική αρχή της προσαρμοστικότητας, σύμφωνα με την οποία η συχνότητα εμφάνισης κάθε εκπαιδευτή είναι ανάλογη της δυνατότητάς του να πα-

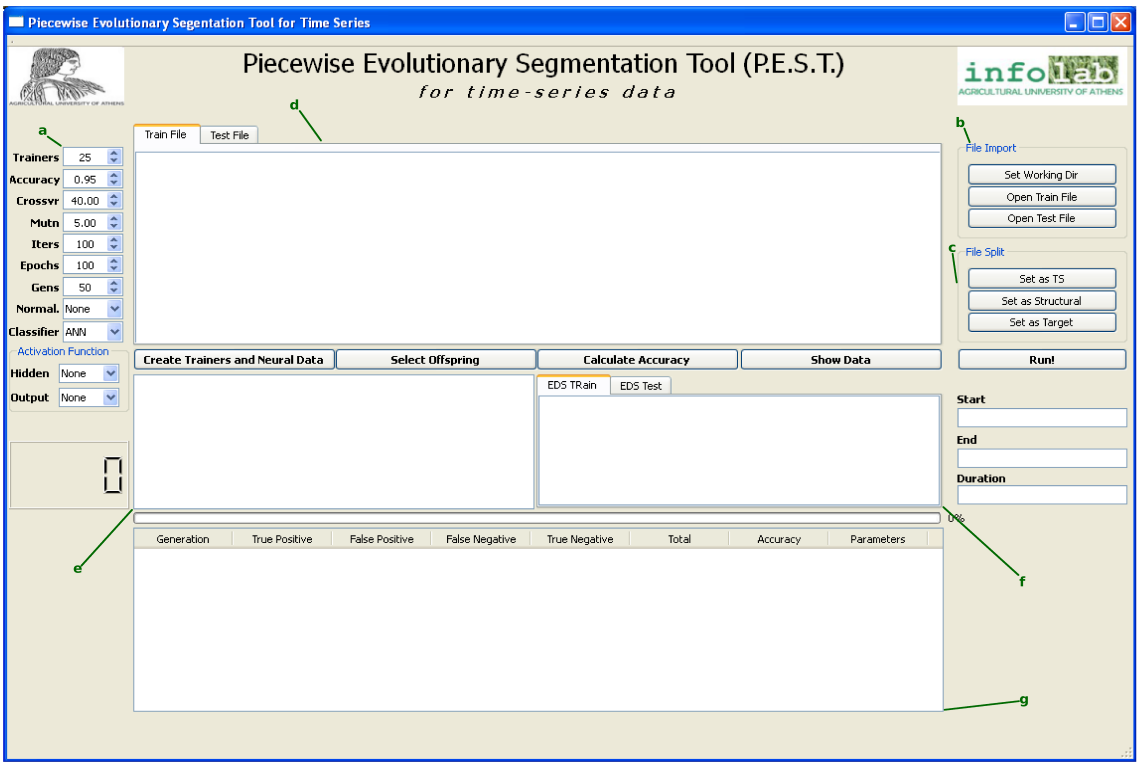
ράξει εξελικτικά δεδομένα υψηλής αποδοτικότητας. Εφόσον συμπληρωθεί ο πληθυσμός της επόμενης γενεάς, το σύστημα είναι πλέον έτοιμο να αρχίσει εκ νέου τη διαδικασία από την αρχή με παραγωγή νέων εξελικτικών δεδομένων.

Τελικά, μετά το πέρας του προκαθορισμένου αριθμού γενεών, το σύστημα επιλέγει τον καλύτερο εκπαιδευτή, ως το καλύτερα προσαρμοσμένο σχήμα τμηματοποίησης της αρχικής χρονοσειράς, καθώς επίσης και τον καλύτερο ταξινομητή, εκπαιδευμένο και έτοιμο για χρήση. Ένα ιδιαίτερα σημαντικό γνώρισμα της μεθόδου αποτελεί η παραγωγή μεγάλου αριθμού εκπαιδευτικών και ελεγκτικών συνόλων δευτερογενών δεδομένων, το καταλληλότερο από τα οποία τελικά καθιστά το δίκτυο συμπαγές και ακριβές στην πρόβλεψη ή ταξινόμηση άγνωστων δεδομένων. Η εξελικτική τμηματοποίηση της αρχικής χρονοσειράς είναι επωφελής έναντι των κλασσικών μεθόδων για έναν αριθμό λόγων. Κατ' αρχήν, με το προτεινόμενο σύστημα γίνεται αποτελεσματικότερος έλεγχος της διαστατικότητας των αρχικών δεδομένων βελτιστοποιώντας την ισορροπία μεταξύ του αριθμού των μελών του διανύσματος εισόδου και της ποσότητας της πληροφορίας που αυτό μεταφέρει εντός του δικτύου. Στη συνέχεια, με τη μέθοδο αυτή γίνεται ευκολότερη και οικονομικότερη η παραγωγή ενός μεγάλου αριθμού εκπαιδευτικών προτύπων για εκπαίδευση και έλεγχο με αποτέλεσμα το δίκτυο να επιδεικνύει αυξημένες δυνατότητες γενίκευσης, ενώ παράλληλα, τα λειτουργικά έξοδα της όλης διαδικασίας παραμένουν σε λογικά επίπεδα.

5.4 Πλατφόρμα ανάπτυξης λογισμικού

Το προτεινόμενο εργαλείο λογισμικού αναπτύχθηκε κάτω από σύστημα Ubuntu Linux και αποτελεί μια εφαρμογή λειτουργική ανεξαρτήτως πλατφόρμας, καθώς ως γλώσσα προγραμματισμού χρησιμοποιήθηκε η Python v.2.7, η οποία υποστηρίζει όλα τα γνωστά σήμερα λειτουργικά συστήματα. Στο σύνολό τους οι προκαταρκτικοί έλεγχοι και οι ρυθμίσεις παραμέτρων του εργαλείου διενεργήθηκαν με βάση τη βιβλιοθήκη FANN (Fast Artificial Neural Network libary) v.2.1.0, ενώ η φάση παραγωγής του εργαλείου υποστηρίχθηκε από τις βιβλιοθήκες PyBrain v.0.3 και Scikits.Learn v.0.8.1 για την ανάπτυξη των αντικειμένων ΤΝΔ και ΜΔΥ αντίστοιχα. Ο γενετικός αλγόριθμος μέσω του οποίου βελτιστοποιείται η εκπαίδευση των ταξινομητών με αποτελεσματική εξαγωγή των κυριότερων χαρακτηριστικών της εκάστοτε χρο-

νοσειράς αναπτύχθηκε ως βασικό αντικείμενο του συστήματος χωρίς την υποστήριξη οποιασδήποτε βιβλιοθήκης. Ανεκτίμητη βοήθεια και σημαντική καθοδήγηση για την υλοποίηση του γενετικού αλγορίθμου και την προσαρμογή του στη γλώσσα προγραμματισμού Python προσέφερε η εργασία των Wonjae Lee και Hak-Young Kim [109]. Το ολοκληρωμένο εργαλείο εφαρμόσθηκε σε σύστημα Η/Υ με κεντρικό επεξεργαστή Intel Core i7 στα 3.07 GHz και 6 GB μνήμης RAM.



Εικ. 5.1: *Γραφικό Περιβάλλον Διεπαφής Χρήστη του εργαλείου λογισμικού.*

Ως πλατφόρμα ανάπτυξης του γραφικού περιβάλλοντος διεπαφής χρησιμοποιήθηκε η γλώσσα Qt4 και πιο συγκεκριμένα η προβολή της σε Python PyQt4-8.1. Στο σχήμα 5.1 απεικονίζεται το γραφικό περιβάλλον του συστήματος. Στο περιβάλλον διακρίνονται περιοχές καθορισμού παραμέτρων και περιοχές επίδειξης δεδομένων και αποτελεσμάτων. Κατά την αρχικοποίηση του συστήματος δίνεται η δυνατότητα παραμετροποίησης της εξελικτικής διαδικασίας (περιοχή a), με επιλογές καθορισμού:

α. του αριθμού εκπαιδευτών ανά γενεά

β. της τελικής ακρίβειας ταξινόμησης (προεπιλογή 95%)

γ. της πιθανότητας ανα-συνδυασμού (προεπιλογή 40%)

δ. της πιθανότητας μετάλλαξης (προεπιλογή 4‰)

ε. του αριθμού των εκπαιδευτικών επαναλήψεων (προεπιλογή 10.000)

ζ. του τύπου κανονικοποίησης των πρωτογενών δεδομένων. Το σύστημα, όταν απαιτείται κανονικοποίηση, παρέχει τη δυνατότητα μετασχηματισμού των πρωτογενών δεδομένων στα διαστήματα [0, 1] ή [-1, 1].

η. του τύπου του ταξινομητή μεταξύ των ΤΝΔ και ΜΔΥ (RBF ή Polynomial)

θ. του τύπου της συνάρτησης ενεργοποίησης στο ενδιάμεσο επίπεδο και στο επίπεδο εξόδου, όταν ως ταξινομητής έχει επιλεγεί το ΤΝΔ

Επίσης, παρέχονται διαδικασίες καθορισμού του τρέχοντος καταλόγου εργασίας και φόρτωσης των πρωτογενών αρχείων δεδομένων (περιοχή b). Από τη στιγμή που καθορίζεται ο κατάλογος, το εργαλείο λογισμικού δημιουργεί κάτω από αυτόν συγκεκριμένο υπο-κατάλογο στον οποίο αποθηκεύονται τα αποτελέσματα της εξελικτικής διαδικασίας μετά από την εκπαίδευση και τον έλεγχο κάθε ταξινομητή. Μετά τη φόρτωση του αρχείου, τα χρονοσειριακά ή/και δομικά περιεχόμενά του εμφανίζονται στην περιοχή (d).

Listing 5.5: Ψευδοκώδικας μεθόδων κανονικοποίησης πρωτογενών δεδομένων

```
1  METHOD normalization(Lst, type:=["01","-11"]):
2
3      sumOfSquares := 0
4      FOR i in Lst DO
5          sumOfSquares := sumOfSquares + SUM(i*i)
6
7      rootOfSoS := SQUARE_ROOT(sumOfSquares)
8
9      avg := AVERAGE(Lst)
10
11     CREATE_ARRAY normResult
12     normResult := Null
```

```
13
14      IF type = "01" THEN
15          FOR i IN Lst DO
16              IF i < 0 THEN
17                  normResult.APPEND((i-MINIMUM(Lst))/rootOfSoS)
18              ELSE
19                  normResult.APPEND(i/rootOfSoS)
20
21      IF type = "-11" THEN
22          FOR i IN Lst DO
23              normResult.APPEND((i-avg)/rootOfSoS)
24
25      RETURN normResult
```

Ανάλογα με τον τύπο κανονικοποίησης που έχει επιλεγεί στην περιοχή (a), τα πρωτογενή δεδομένα μετασχηματίζονται, σύμφωνα με τον κώδικα 5.5. Κατά την κανονικοποίηση των πρωτογενών δεδομένων το σύστημα εξάγει τη ρίζα του αθροίσματος τετραγώνων κάθε εγγραφής (σειρές 3-7), καθώς επίσης και τη μέση τιμή αυτής (σειρά 9). Στη συνέχεια:

- εάν ζητείται κανονικοποίηση [0, 1] (σειρά 14), τότε εάν η τιμή είναι αρνητική (σειρά 16), αφαιρείται από αυτήν η ελάχιστη τιμή της εγγραφής και το αποτέλεσμα διαιρείται με τη ρίζα του αθροίσματος τετραγώνων (σειρά 17). Σε αντίθετη περίπτωση η τιμή διαιρείται κατευθείαν με τη ρίζα του αθροίσματος τετραγώνων (γραμμή 19).
- εάν ζητείται κανονικοποίηση [-1, 1] (σειρά 21), τότε από κάθε τιμή αφαιρείται η μέση τιμή της εγγραφής και το αποτέλεσμα διαιρείται με τη ρίζα του αθροίσματος τετραγώνων (γραμμή 23).

Το εργαλείο λογισμικού δίνει επίσης τη δυνατότητα να μην εκτελεστεί κανονικοποίηση στα πρωτογενή δεδομένα του προβλήματος.

Αφού φορτωθούν τα πρωτογενή δεδομένα, κανονικοποιημένα ή όχι, καθίσταται δυνατός ο διαχωρισμός τους σε δεδομένα δομικά, χρονοσειράς και εξόδου με επιλογή των αντίστοιχων στηλών και επικαιροποίηση της επιλογής μέσω των πλήκτρων της περιοχής (c). Τα αποτελέσματα της εξελικτικής διαδικασίας εμφανίζο-

νται μετά το πέρας αυτής στην περιοχή (g). Οι περιοχές (e) και (f) είναι βοηθητικές και μπορούν να περιλαμβάνουν τους εκπαιδευτές κάθε γενεάς (περιοχή e), και τα εξελικτικά δεδομένα που προκύπτουν από κάθε εκπαιδευτή (περιοχή f).

5.5 Ανάπτυξη ταξινομητών και παραμετροποίηση

Το εργαλείο λογισμικού περιλαμβάνει στον πυρήνα του δύο διαφορετικούς ταξινομητές - ένα ΤΝΔ και μια ΜΔΥ - οι οποίοι τροφοδοτούνται με εξελικτικά δεδομένα. Πρόκειται ουσιαστικά για εργαλεία που προέρχονται από τον τομέα της υπολογιστικής νοημοσύνης και η έξοδός τους χρησιμοποιείται με διττό τρόπο. Αφενός αποτελεί το βασικό μέτρο ποσοτικοποίησης της απόδοσης των εκπαιδευτών που παράγονται από την εξελικτική διαδικασία. Εκπαιδευτές που παρουσιάζουν μεγαλύτερη ακρίβεια ταξινόμησης - ή μικρότερο μέσο τυπικό τετραγωνικό σφάλμα - θεωρούνται αποδοτικότεροι και παρουσιάζουν αυξημένη πιθανότητα επιβίωσης. Αφετέρου, το εργαλείο λογισμικού τελικά καθορίζει τον πλέον αποδοτικό ταξινομητή και χρησιμοποιεί το αποτελεσματικότερο σχήμα κατάτμησης της χρονοσειράς σε όλες τις μελλοντικές του χρήσεις. Η αρχιτεκτονική των ταξινομητών αποτελεί βασική παράμετρο του εργαλείου λογισμικού και ο κρισιμότερος παράγοντας της επιτυχίας του. Κατά συνέπεια, κρίθηκε απαραίτητο να διενεργηθεί μια σειρά προκαταρκτικών ελέγχων και δοκιμών με στόχο τη ρύθμιση ποικίλων παραμέτρων της αρχιτεκτονικής των ταξινομητών πριν από την είσοδο του εργαλείου σε κατάσταση παραγωγής.

Σε γενικές γραμμές τα αντικείμενα που υπέστησαν τέτοιου είδους δοκιμαστικούς ελέγχους είτε αποτελούν δομικά στοιχεία οριζόντιας εφαρμογής, είτε σχετίζονται εξειδικευμένα με το γενετικό αλγόριθμο ή κάθε ταξινομητή ξεχωριστά. Συγκεκριμένα, στο πλαίσιο αυτό και σε σχέση με τις μελέτες περιπτώσεων στις οποίες εφαρμόσθηκε το προτεινόμενο σύστημα, επιχειρείται καθορισμός της πολιτικής για τη βέλτιστη επιλογή:

- της αρχιτεκτονικής του ΤΝΔ, σχετικά με τον αριθμό των επιπέδων και τον τύπο διασύνδεσης των νευρώνων μεταξύ των επιπέδων
- του αριθμού των νευρώνων για τα διάφορα επίπεδα του νευρωνικού ταξινομητή

- των συναρτήσεων δραστηριοποίησης των νευρώνων του ενδιάμεσου επιπέδου και του επιπέδου εξόδου του ΤΝΔ ανά περίπτωση
- των τιμών για τις βασικές παραμέτρους C και γ της ΜΔΥ
- της συνάρτησης πυρήνα της ΜΔΥ
- του βέλτιστου τρόπου περαίωσης του γενετικού αλγορίθμου
- του αριθμού των εκπαιδευτών ανά γενεά
- της πιθανότητας ανα –συνδυασμού και μετάλλαξης των εκπαιδευτών κατά τη μετάβαση από γενεά σε γενεά

5.5.1 Νευρωνικός ταξινομητής

Ιδιαίτερα σημαντική για την τελική ποιότητα της ταξινόμησης ή της πρόβλεψης του νευρωνικού ταξινομητή θεωρείται η γενικότερη τοπολογία[2] του. Παρότι τα ΤΝΔ έχουν κάνει την εμφάνισή τους εδώ και δεκαετίες, δεν έχει αναπτυχθεί θεωρία η οποία να διέπει την άριστη τοπολογία τους και να είναι γενικά και καθολικά παραδεκτή. Παραδοσιακά, ο καθορισμός του αριθμού των επιπέδων του νευρωνικού δικτύου, όπως επίσης και του αριθμού των νευρώνων τους ορίζεται μετά από διαδικασίες δοκιμής και λάθους, ή ακόμη και διαισθητικά. Επίσης έχουν εμφανισθεί εργασίες που εμπλέκουν γενετικούς αλγορίθμους [172, 191], ή άλλες εμπειρικές μεθόδους [103, 195] στον καθορισμό της βέλτιστης τοπολογίας, καμιά όμως από αυτές δεν τυγχάνει καθολικής αποδοχής. Σε κάποιες περιπτώσεις εμπλέκεται ο συνολικός αριθμός δειγμάτων του διανύσματος εισόδου, ενώ σε άλλες ο αριθμός των εξόδων του δικτύου. Σύμφωνα με το Σταθάκη [172], τοπολογίες με δύο ενδιάμεσα επίπεδα θα έπρεπε να περιλαμβάνουν κατ' ελάχιστο

$$2\sqrt{(m+2)N} \tag{5.5.1}$$

συνολικό αριθμό νευρώνων στα δύο επίπεδα, όπου m ο αριθμός των νευρώνων εξόδου και N ο συνολικός αριθμός των δειγμάτων του διανύσματος εισόδου [80]. Σύμ-

[2]Με τον όρο τοπολογία ορίζεται ο αριθμός των ενδιάμεσων επιπέδων, καθώς επίσης και ο αριθμός των νευρώνων καθενός από αυτά

φωνα με άλλες ακαδημαϊκές πηγές [25], ο βέλτιστος αριθμός νευρώνων προκύπτει από τη σχέση:

$$\frac{n+m}{2} + (0.1N) \tag{5.5.2}$$

ή

$$\log(n)\log(N) \tag{5.5.3}$$

όπου n και m ο αριθμός των δεδομένων εισόδου και εξόδου αντίστοιχα και N ο συνολικός αριθμός των δειγμάτων του διανύσματος εισόδου.

Κατά την αρχική φάση σχεδιασμού της αρχιτεκτονικής του νευρωνικού δικτύου, αντιμετωπίστηκε το δίλημμα της δημιουργίας μιας εξ ολοκλήρου νέας δομής μέσω της διαδικασίας δοκιμής και σφάλματος. Η μόνη τροποποίηση που δεν ήταν δυνατό να αποφευχθεί έγκειται στη δομή του διανύσματος εισόδου τουν ταξινομητή, εφόσον κάθε εκπαιδευτής του ΓΑ παράγει δεδομένα διαφορετικής διάστασης, ενώ η διάσταση της εξόδου παραμένει αμετάβλητη. Ο κώδικας που αναπτύχθηκε για τη φάση της δοκιμής εμπλουτίσθηκε με ένα αντικείμενο ΤΝΔ προερχόμενο από τη βιβλιοθήκη FANN το οποίο και χρησιμοποιήθηκε ως ο αξιολογητής για κάθε εκπαιδευτή, σχεδιάσθηκε δε και εκπαιδεύθηκε ώστε να ακολουθεί την αρχιτεκτονική του κλασσικού πολυ-επίπεδου perceptron (MLP: Multi Layered Perceptron) για αμφότερες τις περιπτώσεις. Στις αρχικές δοκιμές, η επιλογή για αρχιτεκτονική τριών επιπέδων για το νευρωνικό ταξινομητή υποδείχθηκε μετά από μια σειρά ελέγχων οι οποίοι υλοποιήθηκαν από τη συνάρτηση *cascadetrain_on_data* της βιβλιοθήκης FANN, μέσω της οποίας επιτρέπεται στο δίκτυο να ξεκινά με ένα μόνο νευρώνα στο ενδιάμεσο επίπεδο και στη συνέχεια, καθώς η εκπαιδευτική διαδικασία προχωρεί, να προσθέτει όλο και περισσότερους νευρώνες – ή ακόμη και επίπεδα εάν κάποιο συγκεκριμένο κατώφλι ξεπεραστεί - έως ότου βρεθεί η βέλτιστη αρχιτεκτονική του ΤΝΔ.

Στην παραγωγική του φάση όμως, το προτεινόμενο σύστημα καθορίζει την τοπολογία κάθε νευρωνικού που εκπαιδεύει κατόπιν δοκιμών. Το επίπεδο εισόδου του συστήματος προσαρμόζεται στη διάσταση του εκάστοτε εκπαιδευτικού πακέτου με την έννοια ότι είναι σε θέση να τροποποιεί τον αριθμό των νευρώνων της εισόδου, ανάλογα με το εξελικτικό σύνολο δεδομένων που εισάγεται κάθε φορά στο δίκτυο.

Θεωρώντας, λοιπόν, ότι ο αριθμός των δεδομένων εισόδου σχετίζεται άμεσα με το σχήμα τμηματοποίησης που ορίζει ο εκάστοτε εκπαιδευτής, το αντικείμενο του νευρωνικού ταξινομητή καθορίσθηκε να ξεκινάει με ένα ενδιάμεσο επίπεδο με αριθμό νευρώνων που ορίζεται από τη σχέση 5.5.2 για ευρείες χρονοσειρές. Σε περίπτωση που ο συνολικός αριθμός νευρώνων υπολείπεται εκείνου που δίδεται από τη σχέση 5.5.1, ο απαραίτητος αριθμός συμπληρώνεται σε δεύτερο ενδιάμεσο επίπεδο. Για την περίπτωση των μικρότερων χρονοσειρών το ΤΝΔ περιλαμβάνει ένα ενδιάμεσο επίπεδο με αριθμό νευρώνων που ορίζεται από τη σχέση 5.5.3.

Στο επίπεδο εξόδου περιλαμβάνεται ένας νευρώνας, ο οποίος είναι δυαδικός για την περίπτωση της ταξινόμησης και αντιστοιχίζει κάθε δείγμα του διανύσματος εισόδου σε μια από δύο προκαθορισμένες κλάσεις. Το σύστημα με την προαναφερόμενη παραμετροποίηση εφαρμόσθηκε στη μελέτη περίπτωσης της αναγνώρισης των φυτικών ιών, όπου η δυαδική έξοδος αντιστοιχίζει κάθε δείγμα χρονοσειράς σε έναν από τους δύο προκαθορισμένους ιούς του προβλήματος. Για την περίπτωση των προβλημάτων πρόβλεψης, η έξοδος είναι τύπου κινητής υποδιαστολής. Εφαρμόσθηκε με την αρχιτεκτονική αυτή στην περίπτωση της πρόβλεψης της χειμαρρικής επικινδυνότητας, ώστε να αποδίδει την πρόβλεψη για τη μέση ετήσια παροχή ύδατος του εκάστοτε ρεύματος.

Τέλος, για κάθε επανάληψη των πειραμάτων δοκιμάσθηκε μια πλειάδα συναρτήσεων ενεργοποίησης τόσο για το ενδιάμεσο επίπεδο, όσο και για το επίπεδο εξόδου. Οι συναρτήσεις αυτές, που δίνονται στον Πίνακα 5.1, προσφέρονται από το εργαλείο λογισμικού και για τα δύο επίπεδα του νευρωνικού ταξινομητή.

5.5.2 Ταξινομητής Διανυσμάτων Υποστήριξης

Ο ταξινομητής ΜΔΥ αναπτύχθηκε παράλληλα και ως βασικός του μηχανισμός επιλέχθηκε η χρήση των υπορουτίνων C-SVC και C-SVR για τις περιπτώσεις προβλημάτων ταξινόμησης και πρόβλεψης αντίστοιχα. Στις βασικές παραμέτρους του εφαρμόσθηκε μια πολιτική προσαρμογής στο σχήμα τμηματοποίησης του αντίστοιχου εκπαιδευτή. Έτσι, για κάθε παραγόμενο εξελικτικό σύνολο δεδομένων το σύστημα υπολογίζει τις καταλληλότερες τιμές για τις παραμέτρους C και γ του ΜΔΥ, υλοποίηση που επιτυγχάνεται μέσω μιας ειδικά για το σκοπό αυτό ανεπτυγμένης

Πίνακας 5.1: Έλεγχος συναρτήσεων ενεργοποίησης για τους νευρώνες του ενδιάμεσου επιπέδου και του επιπέδου εξόδου του ΤΝΔ

Συνάρτηση	**Όρια εξαρτημένης μεταβλητής**	**Περιγραφή**
Γραμμική	$-\infty < y < \infty$	$y = sx$ $d = 1s$
Βηματική		$y = \begin{cases} 0 & \text{εάν} \quad x < 0 \\ 1 & \text{εάν} \quad x \geq 0 \end{cases}$
Σιγμοειδής	$0 < y < 1$	$y = \frac{1}{1+e^{-2sx}}$ $d = 2sy(1-y)$
Σιγμοειδής Συμμετρική	$-1 < y < 1$	$y = \tanh(sx) = \frac{2}{1+e^{-2sx}} - 1$ $d = s(1-y^2)$
Gaussian	$0 < y < 1$	$y = e^{-xsxs}$ $d = -2xsys$
Gaussian Symmetric	$-1 < y < 1$	$y = e^{-xsxs}2 - 1$ $d = -2xs(y+1)s$
Softmax	$0 \leq y \leq 1$ $\sum_{i=1}^{n} y_i = 1$	$y_i = \frac{e^{q_i}}{\sum_{j=1}^{n} e^{q_j}}$

όπου x η είσοδος στη συνάρτηση, y η έξοδος, s η κλίση, d η παράγωγος, q_i το σήμα εισόδου στον i-οστό νευρώνα του δικτύου και n το σύνολο των νευρώνων του επιπέδου εξόδου

διαδικασίας αναζήτησης πλέγματος (grid-search procedure) σε σχήμα πενταπλής διεπικύρωσης (5-fold cross validation scheme) (κώδικας 5.6).

Listing 5.6: Ψευδοκώδικας καθορισμού παραμέτρων της ΜΔΥ μέσω αναζήτησης πλέγματος

```
METHOD SVMparams(Original_Data,Evolutionary_Data):

        orgnTiBit := Time_Bits_Of_Original_Timeseries
        evoTiBit := Time_Bits_Of_Evolutionary_Data

        CREATE_ARRAY folds
        folds := Stratified_K_Fold(Evolutionary_Data, 5)

        timeBits := orgnTiBit - evoTiBit

        CREATE_ARRAY lstGamma
        CREATE_ARRAY listC

        IF timeBits = 0
            begin
                lstGamma := RANGE(0-log(evoTiBit),log(orgnTiBit))
                FOR i := 100 TO 1000 STEP=100 DO
                    listForC.APPEND(log(evoTiBit)*i)
            end
        ELSE
            begin
                lstGamma := RANGE(0-log(evoTiBit),
                                  log(orgnTiBit, STEP=timeBits))
                FOR i := 100 TO 1000 STEP=100 DO
                    listForC.APPEND(log(timeBits)*i)
            end

        tuned_parameters := ('kernel' := RBF,
                             'gamma': ((2^j) FOR j in lstGamma),
                             'C': listForC)

        CREATE_ARRAY results_Gamma
        CREATE_ARRAY results_C
        results_Gamma := Null
```

```
35          results_C := Null
36
37          FOR fold in folds DO
38              begin
39                  classifier.ASSEMBLE(SVM, tuned_parameters)
40                  classifier.TRAIN(training_data := EDS-fold)
41                  classifier.RUN(testing_data := fold)
42                  results_Gamma.APPEND(classifier.best_Gamma)
43                  results_C.APPEND(classifier.best_C)
44              end
45
46          RETURN AVERAGE(results_Gamma), AVERAGE(results_C)
```

Η υπο-ρουτίνα αυτή προηγείται της πραγματικής εκπαιδευτικής φάσης για κάθε εξελικτικό πακέτο εκπαιδευτικών δεδομένων. Αρχικά κάθε τέτοιο πακέτο διαχωρίζεται ισομερώς σε πέντε υποσύνολα (σειρά 7), καθένα από τα οποία δοκιμάζεται με ένα συνδυασμό τιμών για τις παραμέτρους C και γ (σειρές 37-44). Οι τιμές αμφοτέρων των παραμέτρων που υπόκεινται στη δοκιμή λαμβάνονται από πίνακες που κατασκευάζονται με βάση την εξελικτική τμηματοποίηση που ορίζει ο αντίστοιχος εκπαιδευτής (σειρές 14-30). Ειδικά οι τιμές της παραμέτρου γ, προέρχονται από πίνακα δυνάμεων του δύο, στον οποίο ο εκθέτης εξαρτάται από τη σχηματοποίηση του εκπαιδευτή [32, 146]. Η υπο-ρουτίνα επιστρέφει το μέσο των καλύτερων συνδυασμών των δύο παραμέτρων για κάθε εξελικτικό σύνολο δεδομένων (σειρά 46). Με τον τρόπο αυτό, ο επιλεγμένος συνδυασμός των παραμέτρων C και γ της ΜΔΥ χρησιμοποιείται επίσης και κατά τη φάση ελέγχου.

5.6 Προεπισκόπιση προβλημάτων εφαρμογής

Η προτεινόμενη μεθοδολογία σχεδιάσθηκε με τρόπο ώστε το σύστημα να είναι σε θέση να αντιμετωπίζει προβλήματα χρονοσειρών που εμπίπτουν είτε στο πεδίο της ταξινόμησης, είτε σε αυτό της πρόβλεψης. Συνεπώς ο αλγόριθμος αναπτύχθηκε με σκοπό την αντιμετώπιση και των δύο βασικών τομέων προβλημάτων που απαντώνται στο πεδίο της εξόρυξης δεδομένων, στο πλαίσιο δε αυτό, το εργαλείο λογισμικού που αναπτύχθηκε δοκιμάστηκε στην επίλυση δύο διαφορετικών προβλημάτων.

Στην πρώτη περίπτωση [66] αντιμετωπίζεται το θέμα της αποτελεσματικής δια-

χείρισης υδατικών αποθεμάτων ορεινών όγκων και αντιμετώπισης πλημμυρικών κινδύνων. Ως περιοχή μελέτης επιλέχθηκε η Κύπρος, με κυριότερο κίνητρο - και ταυτόχρονα σημαντική πρόκληση - το γεγονός ότι παρά τις χρονοβόρες μελέτες που προηγήθηκαν κατά τα παρελθόντα έτη, μικρή πρόοδος είχε σημειωθεί προς την κατεύθυνση της ανάπτυξης μιας βιώσιμης λύσης στο πρόβλημα της διαχείρισης των υδατικών αποθεμάτων του νησιού. Ως είσοδοι επιλέχθηκαν τόσο δομικά, όσο και δυναμικά στοιχεία, ανάμεσα στα οποία η ετήσια και μηνιαία μέση τιμή των κατακρημνισμάτων αποδεικνύεται ως εξέχουσας σημασίας. Στην περίπτωση αυτή, δεδομένα χρονοσειρών προήλθαν από την ιστορική παρακολούθηση της μηνιαίας βροχόπτωσης σε σταθμούς μέτρησης τοποθετημένους σε συγκεκριμένες λεκάνες απορροής, ενώ τα δεδομένα κάλυψαν μια σχετικά ευρεία χρονική περίοδο. Το ζητούμενο στο πρόβλημα αυτό ήταν η ανάπτυξη μιας μεθοδολογίας εξαγωγής χαρακτηριστικών για την παραγωγή εξελικτικών δεδομένων εκπαίδευσης των ταξινομητών, στοχεύοντας στην αποτελεσματική πρόβλεψη της Μέσης Ετήσιας Παροχής Ύδατος (AAWS: Average Annual Water Supply) σε ετήσια βάση για κάθε ορεινή λεκάνη απορροής της Κύπρου. Ο υπολογισμός του παράγοντα αυτού είναι κρίσιμος για τη διαχείριση των ορεινών υδατικών διαθεσίμων, καθώς σχετίζεται στενά με τις διαδικασίες προσχώσεως, καθώς επίσης και με πιθανές περιβαλλοντικές πιέσεις, λόγω αυξημένου κινδύνου εκδήλωσης πλημμυρικών καταστάσεων στις εμπλεκόμενες περιοχές.

Η δεύτερη περίπτωση στην οποία εφαρμόσθηκε το σύστημα είναι ουσιαστικά ένα πρόβλημα ταξινόμησης. Στην έρευνα αυτή [65], η μέθοδος της Βιοηλεκτρικής Αναγνώρισης (BERA: Bioelectric Recognition Assay) [97] χρησιμοποιήθηκε για την παραγωγή δεδομένων εισόδου προς την κατεύθυνση της ανίχνευσης και ταυτοποίησης φυτικών ιών. Συγκεκριμένα, στόχους της έρευνας αποτέλεσαν οι ιοί του κροταλίσματος του καπνού (TRV: *Tobacco Rattle Virus*) και της πράσινης ποικιλοχλώρωσης με μωσαϊκό της αγγουριάς (CGMMV: *Cucumber Green Mottle Mosaic Virus*), η ανίχνευση των οποίων υλοποιείται από τη μέθοδο BERA με τη χρήση κατάλληλα προεπεξεργασμένων αντιδραστηρίων που λειτουργούν ως βιο-αισθητήρες. Κατά τη διάρκεια της συνεπαγόμενης αντίδρασης με τους εν λόγω βιο-αισθητήρες, κάθε ανιχνευόμενος ιός επιδεικνύει διακριτά πρότυπα βιο-αποκρίσεων επί συγκεκριμένου εύρους συγκεντρώσεων, καθιστώντας τα πρότυπα αυτά ως αποκλειστικά χαρακτηριστικά ενός εκάστου ιού, μια πραγματικά ψηφιακή υπογραφή αναγνώρισης. Πρό-

Πίνακας 5.2: Χρόνοι εκπαίδευσης του συστήματος για τις δύο μελέτες περιπτώσεων

	Ταυτοποίηση Ιών		**Χειμαρρική Επικινδυνότητα**	
Μήκος ΧΣ	331		12	
Εκπαιδευτές	15		15	
Γενεές	1.000		1.000	
Πλήθος EDS	15.000		15.000	
Χρόνοι	**ΤΝΔ**	**ΜΔΥ**	**ΤΝΔ**	**ΜΔΥ**
Ημέρες	8	3	6	0
Ώρες	15	7	23	16
Λεπτά	41	27	28	48
Δευτερόλεπτα	44	11	6	26

κειται, ουσιαστικά, για μια γραφική παράσταση βιοηλεκτρικών αποκρίσεων στη μονάδα του χρόνου, μια χρονοσειρά χαρακτηριστική για κάθε ιό η οποία οδηγεί στην ταυτοποίησή του.

Η ανάλυση χρονοσειρών αποδεικνύεται αποφασιστικής σημασίας για την επίλυση αμφότερων των προβλημάτων. Η προτεινόμενη μέθοδος καλείται να εξαλείψει τους ανασταλτικούς παράγοντες που ανακύπτουν λόγω του υψηλού βαθμού διάστασης και του θορύβου που ενέχονται στα πρωτογενή χρονοσειριακά δεδομένα. Η υλοποίηση εστιάσθηκε στο σχεδιασμό κατάλληλου εξελικτικού σχήματος τμηματοποίησης και στην αποτελεσματική αναπαράσταση της αρχικής χρονοσειράς με εξαγωγή των βασικότερων χαρακτηριστικών της, ούτως ώστε να παραχθούν εκπαιδευτικά δεδομένα που βελτιστοποιούν τη διακριτική ικανότητα και τις δυνατότητες πρόβλεψης των ενσωματωμένων ταξινομητών. Υπό αυτό το πρίσμα το πρόγραμμα δοκιμάσθηκε σε σύστημα Η/Υ με κεντρικό επεξεργαστή Intel Core i7 στα 3.07GHz με 6GB RAM και οι χρόνοι εκπαίδευσης δίδονται στον Πίνακα 5.2. Επίσης, το ολοκληρωμένο εργαλείο συμπεριλαμβάνεται σε CD, ενώ στο Παράρτημα δίνονται οδηγίες εγκατάστασης και χρήσης.

Στη συνέχεια θα παρουσιασθούν αναλυτικότερα τα δύο προβλήματα, καθώς επίσης και τα αποτελέσματα που σημειώθηκαν μετά την εφαρμογή της ΠΕΤ.

Κεφάλαιο 6

ΤΑΞΙΝΟΜΗΣΗ ΦΥΤΙΚΩΝ ΙΩΝ

Στο κεφάλαιο αυτό παρουσιάζεται η εφαρμογή της προτεινόμενης μεθόδου στο πρόβλημα της αναγνώρισης / ταξινόμησης φυτικών ιών. Τα δεδομένα εισόδου προέρχονται από βιο-ηλεκτρικές αντιδράσεις των ιών με εξειδικευμένα αντιδραστήρια ακινητοποιημένα σε πήγμα στη μονάδα του χρόνου. Πρόκειται δηλαδή για δεδομένα χρονοσειράς που αποτελούν χαρακτηριστικό κάθε ιού. Στο παρόν κεφάλαιο αναφέρονται πληροφορίες για το αντικείμενο του προβλήματος, καθώς επίσης και για τον τρόπο συλλογής και χειρισμού των δεδομένων, ενώ στο τέλος παρατίθενται αναλυτικά τα αποτελέσματα της εφαρμογής και επιχειρείται κριτικός σχολιασμός και προσέγγιση αυτών.

Οι φυτικοί ιοί, αν και πολύ λιγότερο γνωστοί και κατανοητοί απ' ότι τα ανθρώπινα αντίστοιχά τους, είναι ιδιαίτερα μολυσματικά δια-κυτταρικά παράσιτα, τα οποία στερούνται της ικανότητας για αυτόνομη θρέψη και επιβίωση επί μακρόν ή της ικανότητας για αναπαραγωγή, χωρίς την υποστήριξη κάποιου ξενιστή. Μέχρι σήμερα αρκετές εκατοντάδες ιών έχουν περιγραφεί και, σταδιακά, ταυτοποιούνται όλο και περισσότεροι.

Εικ. 6.1: *Πρασινοκίτρινες ομόκεντρες κηλιδώσεις σε φύλλα παιωνίας (Paeonia Sara Bernhardt) που οφείλονται στον ιό του κροταλίσματος του καπνού (κατά Chastagner και Pappu, http://www.apsnet.org).*

6.1 Συμπτωματολογία των ιών TRV και CGMMV

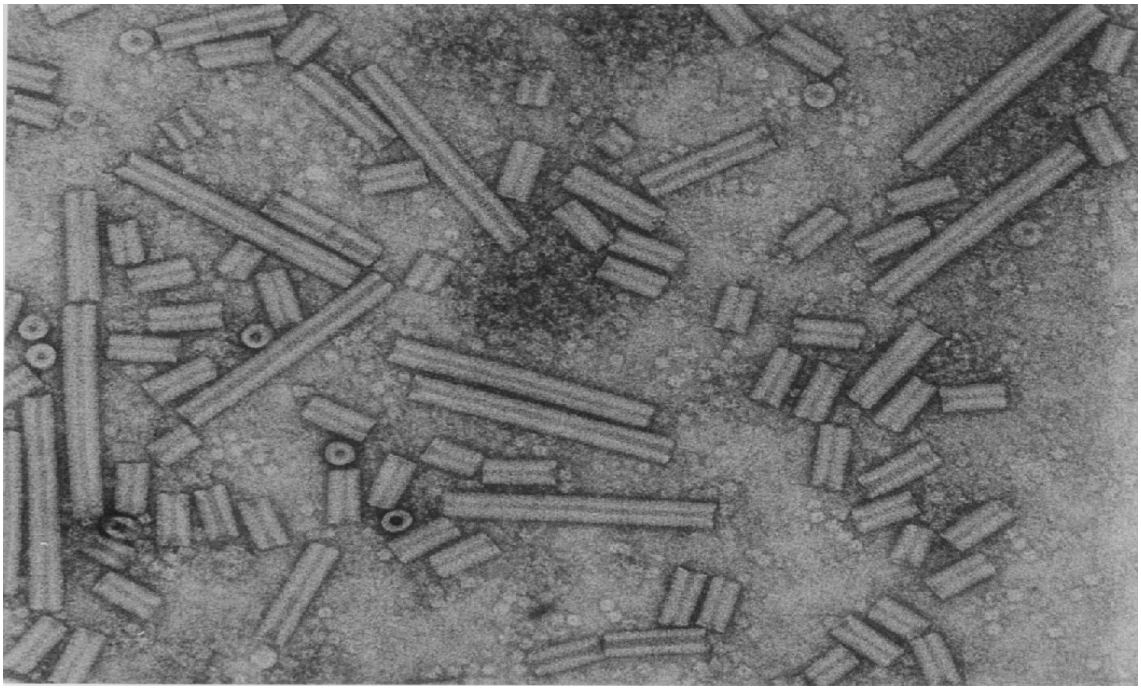
Δύο από τους γνωστότερους εδώ και αρκετά χρόνια φυτικούς ιούς, οι οποίοι προξενούν ιδιαίτερα σοβαρές καταστροφές ακόμη και σήμερα στις καλλιέργειες, είναι οι CGMMV και TRV. Αμφότεροι ανήκουν στην κατηγορία ιών, των οποίων το μολυσματικό υλικό εδράζεται κυρίως στα νημάτια RNA μέσω των οποίων προξενούν σοβαρές παραγωγικές πιέσεις στους ξενιστές. Στα κυριότερα συμπτώματα των μολυσμένων φυτών περιλαμβάνονται λεπιδόμορφες τομές στα φύλλα, κίτρινο, υποκίτρινο έως κιτρινοπράσινο φαινοτυπικά ομόκεντρο δακτυλιωτό μωσαϊκό στα φύλλα και τον καρπό, ποικιλόχρωση, τοπικές νεκρωτικές κακώσεις, στρεβλή ανάπτυξη και συστηματική νέκρωση.

Η ονομασία του ιού του κροταλίσματος του καπνού οφείλεται στο γεγονός ότι ο εν λόγω ιός ανιχνεύθηκε για πρώτη φορά σε καλλιέργεια καπνού (*Nicotiana tabacum*). Παρ' όλ' αυτά, ο TRV δεν περιορίζεται στο φυτό του καπνού, αλλά διαθέτει ένα μεγάλο εύρος ξενιστών. Σε αυτούς περιλαμβάνεται μια πλειάδα φυτικών ειδών, άνω των 350 περίπου, στα οποία συγκαταλέγονται λαχανικά όπως η τομάτα και η πατάτα (Εικ. 6.2) και καλλωπιστικά είδη όπως η τουλίπα, ο ασφόδελος, η γλαδιόλα, η παιωνία (Εικ. 6.1), η καλενδούλα και το ηλιοτρόπιο. Επίσης τα τεύτλα, αλλά και πολυάριθμα ζιζάνια αποτελούν δυνητικούς στόχους του ιού.

Εικ. 6.2: *Καστανοκίτρινες τοξοειδείς κηλιδώσεις στη σάρκα με προβολές στην επιφάνεια καρπού πατάτας (Solanum tuberosum) που οφείλονται στον ιό του κροταλίσματος του καπνού (Πηγή: United Nations Economic Commission for Europe, http://www.unece.org).*

Πρόκειται για ένα παθογόνο του οποίου η καταπολέμηση παρουσιάζει πολύ μεγάλες δυσκολίες, σε πολλές δε περιπτώσεις η προσβολή θεωρείται μη αναστρέψιμη. Συχνά, μοναδική λύση υπ' αυτές τις συνθήκες, αποτελεί το ολοκληρωτικό ξερίζωμα των προσβεβλημένων καλλιεργειών. Στην πράξη κυριότερος τρόπος μετάδοσης και έναρξης της ασθένειας είναι μέσω ήδη μολυσμένου πολλαπλασιαστικού υλικού, οπότε η εξάπλωση συμβαίνει ιδιαίτερα γρήγορα και η μόλυνση είναι απότομη με τα φυτά να επιδεικνύουν αμέσως συμπτώματα. Μια από τις βασικότερες οδούς διάδοσης του ιού είναι μέσω νηματωδών του γένους *Trichodorus* οι οποίοι διαβιούν στο έδαφος και τρέφονται από φυτικές ρίζες μεταδίδοντας δι' επαφής τον ιό από φυτό σε φυτό. Αρκετά συνηθισμένος επίσης τρόπος μετάδοσης είναι μέσω μηχανικής μεταφοράς κατά τη σπορά ή το ξεβοτάνισμα, στις περισσότερες περιπτώσεις εξαιτίας μολυσμένων γεωργικών εργαλείων, ενώ δεν έχει αναφερθεί σε καμιά περίπτωση μόλυνση φυτού μέσω επαφής με άλλο φυτό. Η χημική καταπολέμηση του εν λόγω νηματώδους καθίσταται δύσκολη λόγω του ότι τα σχετικά νηματοκτόνα είναι ιδιαίτερα τοξικά και όχι ευρέως διαθέσιμα.

Στον αντίποδα, οι ατραποί που ακολουθεί ο ιός της πράσινης ποικιλοχλώρωσης με μωσαϊκό της αγγουριάς είναι διαφορετικοί. Ο εν λόγω ιός ανήκει στο γένος *Tobamovirus* και οι ξενιστές του προέρχονται κυρίως από την οικογένεια *Cucurbitaceae*. Σημαντικότερες υποψηφιότητες για προσβολή θέτουν το αγγούρι *Cucumis sativus*, το καρπούζι *Citrullus vulgaris*, καθώς επίσης και η κολοκύθα *Lagenaria siceraria*, χωρίς



Εικ. 6.3: *Ο ιός του κροταλίσματος του καπνού (σάρωση από φωτογραφία σε ανάλυση 150 dpi στα 8 bits/pixel greyscale, πηγή: Rothamsted Research, http://www.rothamsted.ac.uk/ppi/links/pplinks/virusems/).*

να αποκλείεται η επέκταση της προσβολής επίσης και σε άλλα λαχανικά ή καλλωπιστικά φυτά.

Η προσβολή από τον CGMMV εστιάζεται κυρίως σε Ευρώπη και Ασία, ιδιαίτερα δε στην Ελλάδα, Μεγάλη Βρετανία, την Ινδία και την Ιαπωνία. Οι τρόποι μετάδοσης του CGMMV περιλαμβάνουν το πολλαπλασιαστικό υλικό και τις φερτές ύλες, όπως το νερό ή τα διάφορα εδαφικά σωματίδια.

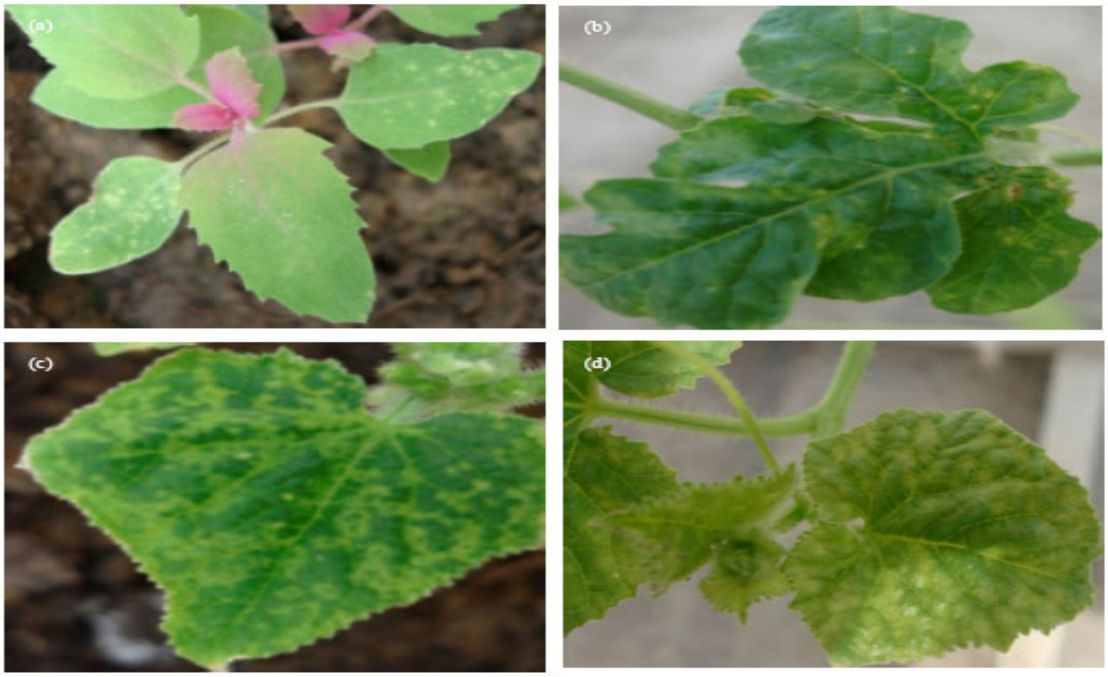
Συνοψίζοντας, πρόκειται για δύο ιούς μεγάλης οικονομικής σημασίας. Η έγκυρη αναγνώριση των επιδημιών που οφείλονται σε αυτούς, και κυρίως η έγκαιρη πρόληψη της προσβολής που προξενούν αποτελεί υψίστης σημασίας παράγοντα για τις καλλιέργειες και την παραγωγή. Το γενικό πλαίσιο της έρευνας που σχετίζεται με την περίπτωση αυτή είναι βασικά η συνεισφορά στο πρόβλημα της ταυτοποίησης της προσβολής που οφείλεται στους δύο προαναφερόμενους ιούς, ενώ πιθανή επέκταση του συστήματος με ενσωμάτωση της δυνατότητας αναγνώρισης και άλλων παρόμοιων ιών θα αποτελέσει σημαντικό εργαλείο για τη φυτοπαθολογία.

Εικ. 6.4: *Συμπτώματα του ιού της πράσινης ποικιλοχλώρωσης με μωσαϊκό σε φυτά Cucurbitaceae. (a) Σοβαρή προσβολή σε φύλλο Cucumis sativus θερμοκηπίου (b) ήπια προσβολή σε φύλλο Cucumis melo ([131]).*

6.2 Παραμετροποίηση του συστήματος

Η τοπολογία του νευρωνικού ταξινομητή εξαρτάται σε μεγάλο βαθμό από το σχήμα τμηματοποίησης που ορίζεται από τον εκάστοτε εκπαιδευτή, τόσο όσον αφορά στη διάσταση του ενδιάμεσου επιπέδου, όσο και σε αυτήν του επιπέδου εξόδου (Πίνακας 6.1). Ο νευρωνικός ταξινομητής είναι πλήρους συνδεσιμότητας μεταξύ όλων των επιπέδων του, εκ των οποίων αυτό της εξόδου περιλαμβάνει ένα δυαδικό νευρώνα (0 ή 1) για την ταξινόμηση των ιών.

Μετά από σειρά ελέγχων, ως συνάρτηση ενεργοποίησης για το ενδιάμεσο επίπεδο νευρώνων χρησιμοποιήθηκε η συμμετρική σιγμοειδής, ενώ ως συνάρτηση ενεργοποίησης των νευρώνων του επιπέδου εξόδου χρησιμοποιήθηκε η Softmax. Οι εκπαιδευτικοί αλγόριθμοι που δοκιμάσθηκαν περιλαμβάνουν τους Incremental Training, Batch Training και τον αλγόριθμο Rprop [156, 83]. Ο πρώτος αποτελεί το βασικότερο αλγόριθμο οπισθόδρομης μετάδοσης (backpropagation) του σφάλματος κατά τον οποίο ο πίνακας βαρών ενημερώνεται μετά από κάθε εγγραφή των εκπαιδευτικών δεδομένων, ενώ ο batch training ενημερώνει τον πίνακα βαρών του δικτύου αφού ολόκληρο το εκπαιδευτικό πακέτο έχει περάσει από τον ταξινομητή και έχει υπολογισθεί η έξοδος για όλα τα εκπαιδευτικά δείγματα. Στην κατηγορία αυτή



Εικ. 6.5: *Προσβολές CGMMV μέτριας έως σοβαρής έντασης σε φύλλα φυτών θερμοκηπίου (a) χλωρωτική κάκωση επί Chenopodium amaranticolor (b) μωσαϊκό και εξέλκωση σε Citrullus lanatus (c) και (d) έντονη προσβολή Cucumis sativus και Cucumis melo αντίστοιχα (κατά [131]).*

επίσης εμπίπτει και ο αλγόριθμος Rprop, μέσω του οποίου επιτεύχθηκαν τα καλύτερα αποτελέσματα, σε συνδυασμό με τη συμμετρική σιγμοειδή (Sigmoid Symmetric - Hyperbolic Tangent) συνάρτηση ενεργοποίησης. Η διαδικασία ελέγχου επιβεβαίωσε την αποδεκτή εφαρμογή της παραπάνω δομής για τις περισσότερες περιπτώσεις.

Η οριζόντια παραμετροποίηση του συστήματος αφορά στη ρύθμιση τεσσάρων βασικά παραμέτρων, καθεμιά από τις οποίες είναι δυνατόν να ρυθμισθούν και από τον τελικό χρήστη του εργαλείου λογισμικού κατά την αρχικοποίησή του. Για κάθε γενετικό αλγόριθμο, η εξελικτική διαδικασία συνεχίζεται έως ότου ικανοποιηθεί μια συγκεκριμένη συνθήκη περαίωσης, οι πλέον συνηθισμένες από τις οποίες περιλαμβάνουν την ικανοποίηση ενός κριτηρίου ελαχιστοποίησης (ή μεγιστοποίησης) σε κάποια γενεά, την παγίδευση μιας γενεάς σε τοπικό ελάχιστο και τη μη συνέχιση της βελτιστοποίησης μετά παρέλευση συγκεκριμένου αριθμού γενεών και, τέλος, με την ολοκλήρωση συγκεκριμένου αριθμού γενεών. Ο προτεινόμενος αλγόριθμος ΠΕΤ

Πίνακας 6.1: Παραμετροποίηση του συστήματος για τη μελέτη περίπτωσης της ταξινόμησης των φυτικών ιών

	ΤΝΔ	**ΜΔΥ**
Τύπος ταξινομητή	MLP	C-SVC
Αριθμός νευρώνων	((In+Out)/2)+0.1Rec	
Ενδιάμεσο Ενεργοποίηση	Σιγμοειδής συμμετρική	
Εξοδος Ενεργοποίηση	Softmax	
Κανονικοποίηση	[-1, 1]	[0, 1]
Πυρήνας		RBF
Παράμετρος C	Αναζήτηση πλέγματος	
Παράμετρος γ	Αναζήτηση πλέγματος	
Πληθυσμός εκπαιδευτών	15	
Πλήθος γενεών	1.000	
Μέθοδος επιλογής	Roulette wheel	
Πιθανότητα ανα-συνδυασμού	40%	
Πιθανότητα μετάλλαξης	5‰	

εμπίπτει στην τελευταία κατηγορία, ώστε να αποφευχθεί η παγίδευση του συστήματος σε αέναη λειτουργία ένεκα του εύρους εφαρμογής του. Για τις ανάγκες του συστήματος επιλέχθηκε ο τερματισμός του αλγορίθμου μετά την παρέλευση 1.000 γενεών για αμφότερα τα προβλήματα, καθώς επίσης και ο σχηματισμός 15 ταξινομητών ανά γενεά. Επίσης, ποικίλες πιθανότητες ανα –συνδυασμού και μετάλλαξης δοκιμάσθηκαν για τις δύο περιπτώσεις όπου εφαρμόσθηκε η προτεινόμενη μεθοδολογία, πριν τελικά καθορισθούν οι πιθανότητές τους στα επίπεδα του 40% και 5‰αντίστοιχα.

6.3 Χειρισμός των δεδομένων και αποτελέσματα

Το γεγονός ότι ένας ταξινομητής αποδίδει καλά αποτελέσματα έναντι του εκπαιδευτικού πακέτου δεδομένων δε σημαίνει απαραίτητα ότι είναι και αποτελεσματικός. Η μόνη θετική ένδειξη της ποιότητάς του προέρχεται από τη δυνατότητα γενίκευσης, δηλαδή την ικανότητά του να αποδίδει εξίσου καλά και σε νέα δεδομένα τα οποία δεν έχει προηγουμένως συναντήσει. Για το λόγο αυτό, στην περίπτωση εφαρμογής του προτεινόμενου εργαλείου ακολουθήθηκε η κλασσική και

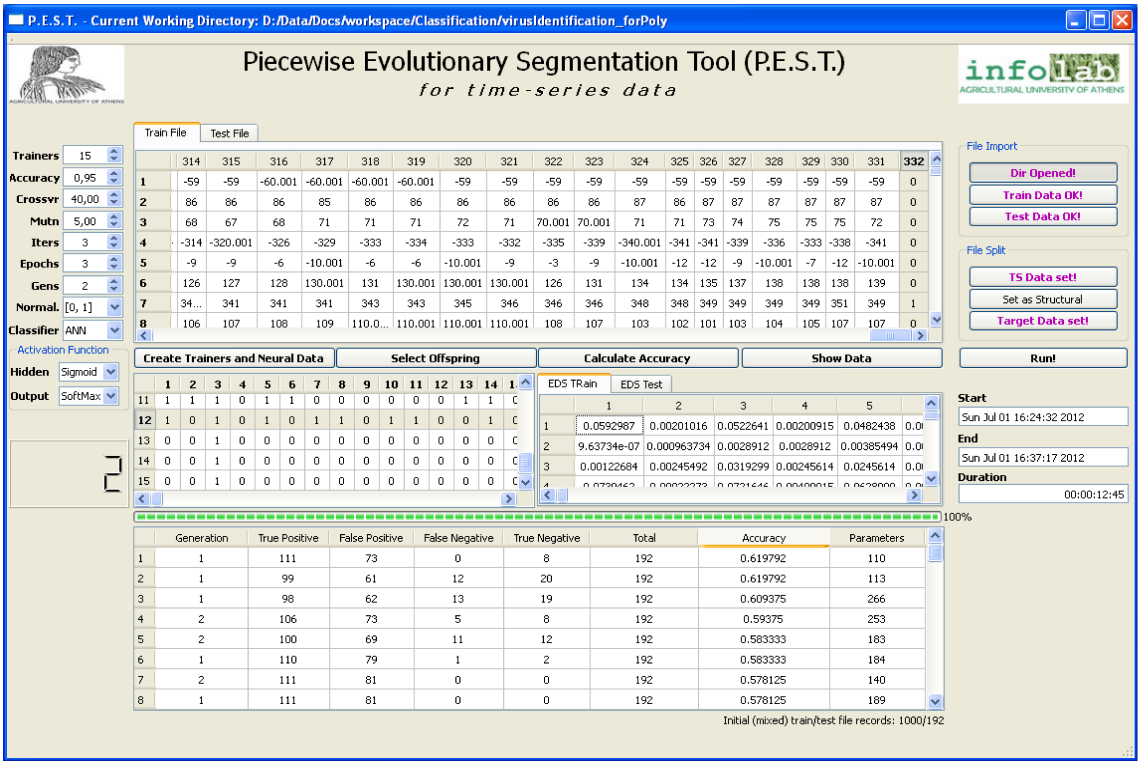
καλά τεκμηριωμένη [75] διαδικασία σύμφωνα με την οποία τμήμα των δεδομένων δεν ενεπλάκη στην εκπαίδευση, αλλά συναθροίστηκαν στο ελεγκτικό πακέτο με το οποίο αξιολογήθηκε η απόδοση του ταξινομητή.

Πίνακας 6.2: Δεδομένα εκπαίδευσης, ελέγχου και αξιολόγησης για το πρόβλημα αναγνώρισης των φυτικών ιών

	Σύνολα Δεδομένων			
	Εκπαιδευτικό	**Ελεγκτικό**	**Αξιολόγησης**	**Σύνολο**
CGMMV	500	91	49	640
TRV	500	101	30	631
Σύνολο	1000	192	79	1271

Όπως αναλύεται στον Πίνακα 6.2, τα πρωτογενή δεδομένα όπως προήλθαν από την αντίδραση των βιο-αισθητήρων, περιελάμβαναν 1.271 δείγματα χρονοσειρών, καθένα από τα οποία αποτελούσε ουσιαστικά μια χρονική ακολουθία 331 στιγμών. Χρησιμοποιώντας αναλογία 3/1, τα δεδομένα αυτά διαχωρίσθηκαν σε εκπαιδευτικό και ελεγκτικό σύνολο, αποτελούμενα από 1.000 και 192 δείγματα αντίστοιχα. Παρατηρείται ότι λήφθηκε μέριμνα για τη δημιουργία όσο το δυνατόν πιο ισορροπημένων συνόλων δεδομένων, ώστε οι δείκτες που θα εξαχθούν να μην είναι μεροληπτικοί. Τέλος, 79 δείγματα αφαιρέθηκαν τελείως από την εκπαίδευση για να συνθέσουν το σύνολο αξιολόγησης, το οποίο χρησιμοποιήθηκε στο τελικό στάδιο ελέγχου, μετά την υπόδειξη από το σύστημα για το βέλτιστο σχήμα αναπαράστασης της χρονοσειράς. Στην εικόνα 7.3 εμφανίζεται το εργαλείο λογισμικού με φορτωμένα δεδομένα του προβλήματος και αποτελέσματα της εξελικτικής διαδικασίας στη δεύτερη γενεά του αλγορίθμου.

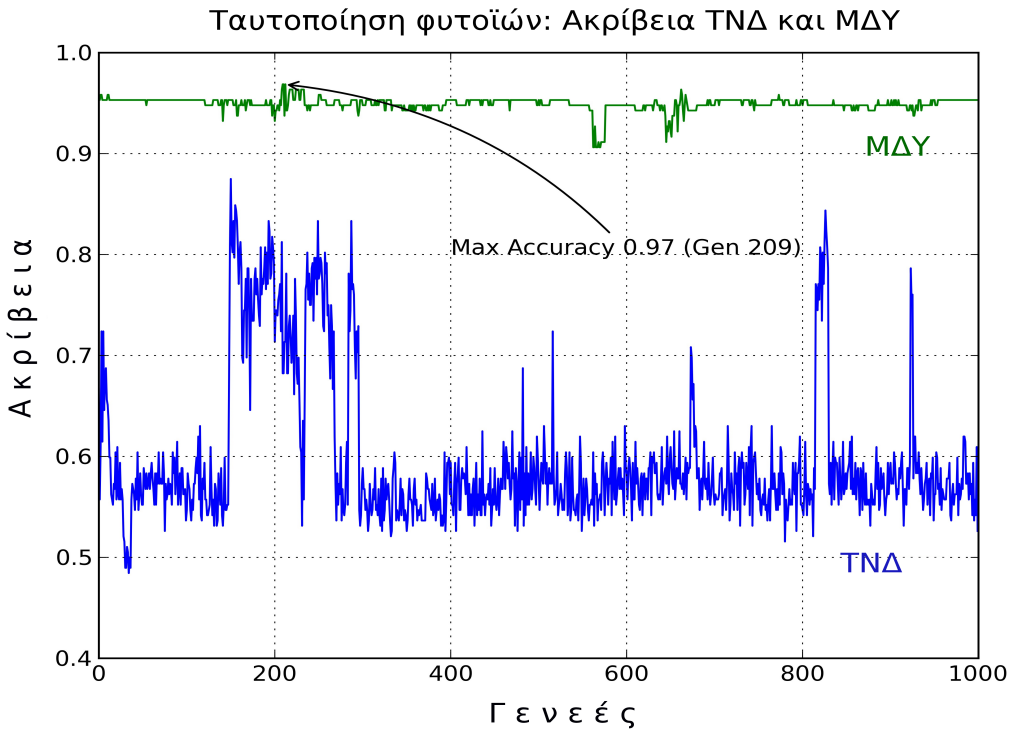
Η διαδικασία απόδοσης του βαθμού προσαρμοστικότητας, η οποία εφαρμόσθηκε μέσω του ΓΑ, λαμβάνει υπόψη το σφάλμα που καταγράφεται κατά τη διαδικασία ελέγχου. Υπό αυτό το πρίσμα, τα αρχικά δεδομένα διαχωρίσθηκαν σε τρία σύνολα, στα οποία εφαρμόζεται κάθε φορά το σχήμα τμηματοποίησης κάθε εκπαιδευτή. Το πρώτο πακέτο που σχηματοποιήθηκε για να λειτουργήσει ως εκπαιδευτικό σύνολο δεδομένων, χρησιμοποιήθηκε κατά τη φάση της εκπαίδευσης των ταξινομη-



Εικ. 6.6: *Πρωτογενή και εξελικτικά δεδομένα και αποτελέσματα δεύτερης γενεάς για την περίπτωση ταυτοποίησης ιών, στο εργαλείο λογισμικού.*

τών. Γνωρίζοντας την επιθυμητή έξοδο του εκπαιδευτικού πακέτου και ενημερώνοντας τον πίνακα βαρών τους, οι ταξινομητές ήταν σε θέση να βελτιώνουν τη διακριτική τους ικανότητα μέσω της επανάληψης της εκπαιδευτικής διαδικασίας. Το δεύτερο σύνολο δεδομένων αποτελεί το πακέτο ελέγχου, μέσω του οποίου επιτεύχθηκε η αξιολόγηση της δυνατότητας γενίκευσηςτης διακριτικής ικανότητας έκαστου ταξινομητή σε δεδομένα τα οποία δεν είχε συναντήσει κατά τη διάρκεια της εκπαίδευσης. Τέλος, το τρίτο πακέτο αξιολόγησης χρησιμοποιήθηκε για την τελική αξιολόγηση του ταξινομητή, αφού έχει πλέον αποκαλυφθεί το βέλτιστο σχήμα αναπαράστασης των αρχικών δεδομένων.

Το εξελικτικό υποσύστημα του προτεινόμενου εργαλείου λογισμικού παρήγαγε μια σειρά εκπαιδευτών ανά γενεά, καθένας από τους οποίους αποτύπωσε το σχήμα τμηματοποίησης που έφερε στο χρωμόσωμά του τόσο στο εκπαιδευτικό όσο και στο αντίστοιχο σύνολο ελέγχου. Το πρώτο χρησιμοποιήθηκε στην εκπαίδευση των ταξινομητών ώστε να ενημερωθεί κατάλληλα ο πίνακας των συναπτικών βαρών τους,



Σχήμα 6.1: *Καταγραφή της Ακρίβειας των ταξινομητών ΤΝΔ και ΜΔΥ καθ' όλες τις γενεές του προτεινόμενου αλγορίθμου για το πρόβλημα της αναγνώρισης των φυτικών ιών.*

ενώ το δεύτερο στην εξαγωγή των αποτελεσμάτων ταξινόμησης καθενός από αυτούς.

Στη συνέχεια τα αποτελέσματα ταξινόμησης ενσωματώθηκαν στον αντίστοιχο εκπαιδευτή, έτσι ώστε να εξαχθεί ο βαθμός προσαμοστικότητάς του, ο οποίος αποτέλεσε το μέτρο επιβίωσης προς τις επόμενες γενεές. Τα αποτελέσματα της διαδικασίας ταυτοποίησης των ιών, όπως υποδείχθηκαν από το προτεινόμενο σύστημα μετά παρέλευση 1.000 γενεών, επιδεικνύονται στον πίνακα 6.3 και στο σχήμα 6.1.

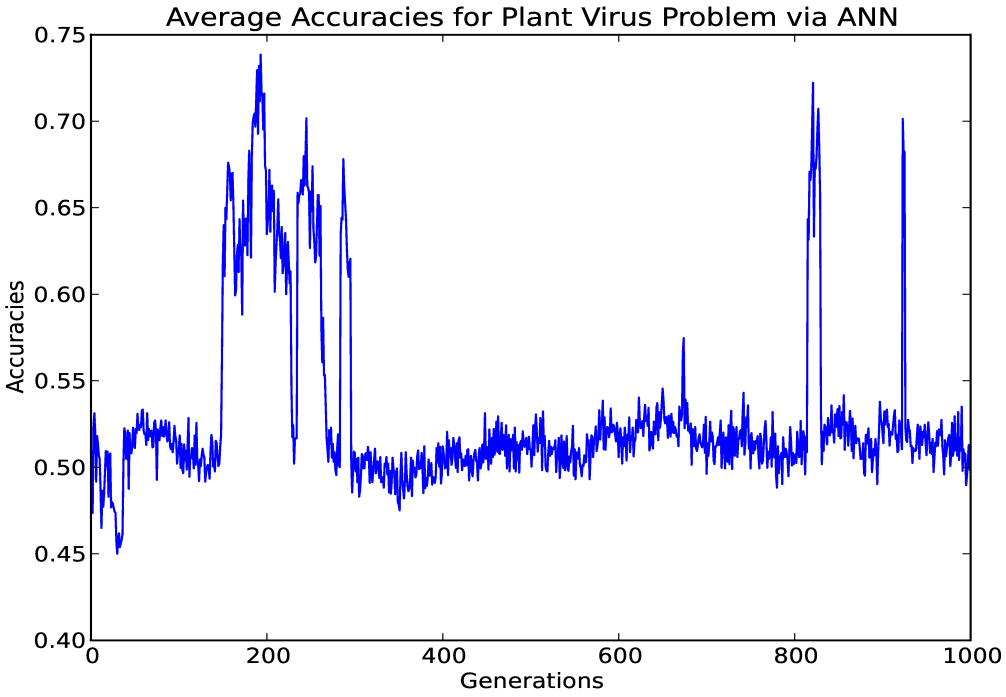
Για το νευρωνικό ταξινομητή, οι μέσες τιμές ακρίβειας ανά γενεά επιδεικνύονται στο γράφημα 6.2, ενώ οι αντίστοιχες μέγιστες τιμές στο γράφημα 6.3

Όσον αφορά στη ΜΔΥ, στα γραφήματα 6.4 και 6.3 απεικονίζονται οι μέσες και μέγιστες τιμές ακρίβειας αντίστοιχα για το πρόβλημα της αναγνώρισης των ιών.

Στην προκείμενη περίπτωση, και με βάση τις εξισώσεις 5.2.2, 5.2.3, 5.2.4, 5.2.5, 5.2.6, ως TP ορίζονται τα δείγματα που είναι στην πραγματικότητα CGMMV και

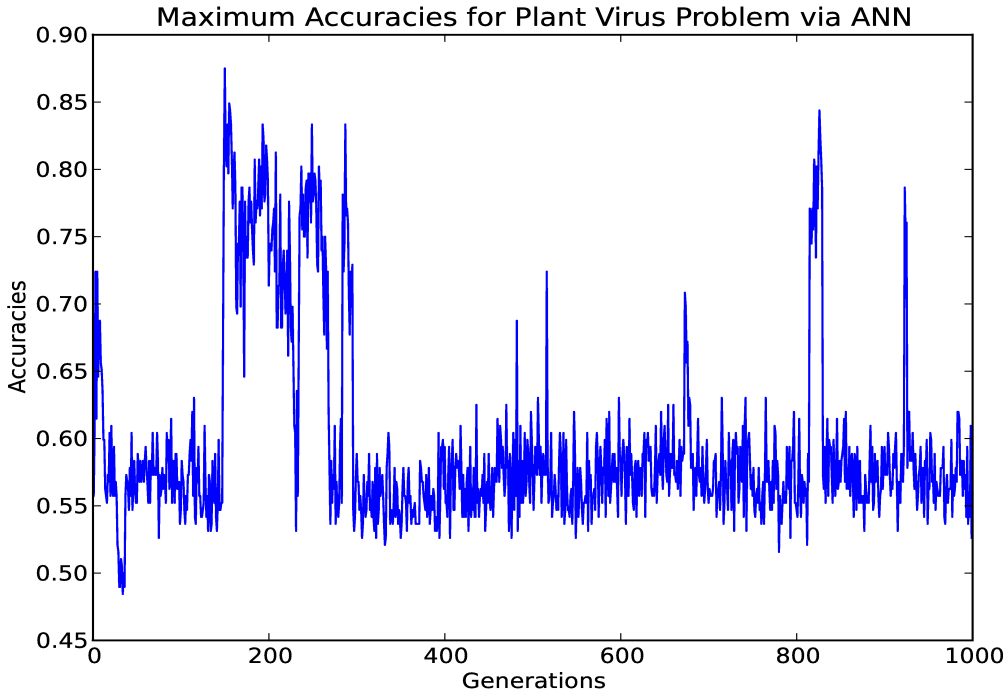
Πίνακας 6.3: Αποτελέσματα της εφαρμογής του προτεινόμενου εργαλείου στην ταυτοποίηση των φυτικών ιών CGMMV και TRV: Δεδομένα της πρωτογενούς χρονοσειράς έναντι εξελικτικών δεδομένων υπό ΤΝΔ και ΜΔΥ, με την καλύτερη ακρίβεια (Acc)

Τεχνητό Νευρωνικό Δίκτυο										
Γενεά	**TP**	**FP**	**FN**	**TN**	**Sens**	**Spec**	**PPV**	**NPV**	**Acc**	**Νευρώνες**
150	74	17	7	94	0,914	0,847	0,813	0,931	0,875	193
155	67	15	14	96	0,827	0,865	0,817	0,873	0,849	194
826	72	21	9	90	0,889	0,811	0,774	0,909	0,844	190
287	71	22	10	89	0,877	0,802	0,763	0,899	0,833	182
826	76	28	5	83	0,938	0,748	0,731	0,943	0,828	190
Πρωτογενή Δεδομένα	72	84	9	27	0,889	0,243	0,462	0,750	0,516	182
Μηχανή Διανυσμάτων Υποστήριξης											
Γενεά	**TP**	**FP**	**FN**	**TN**	**Sens**	**Spec**	**PPV**	**NPV**	**Acc**	**C**	**γ**
209	80	5	1	106	0,988	0,955	0,941	0,991	0,969	756,5	11,2
662	80	6	1	105	0,988	0,946	0,930	0,991	0,963	757,1	7,5
773	79	6	2	105	0,975	0,946	0,929	0,981	0,958	756,1	46,2
999	78	6	3	105	0,963	0,946	0,929	0,972	0,953	756,8	53,1
998	78	7	3	104	0,963	0,937	0,918	0,972	0,948	754,4	49,4
Πρωτογενή Δεδομένα	67	8	14	103	0,827	0,928	0,893	0,880	0,885	786,8	37,4



Σχήμα 6.2: *Μέσες τιμές ακρίβειας ανά γενεά ΤΝΔ (αναγνώριση φυτικών ιών).*

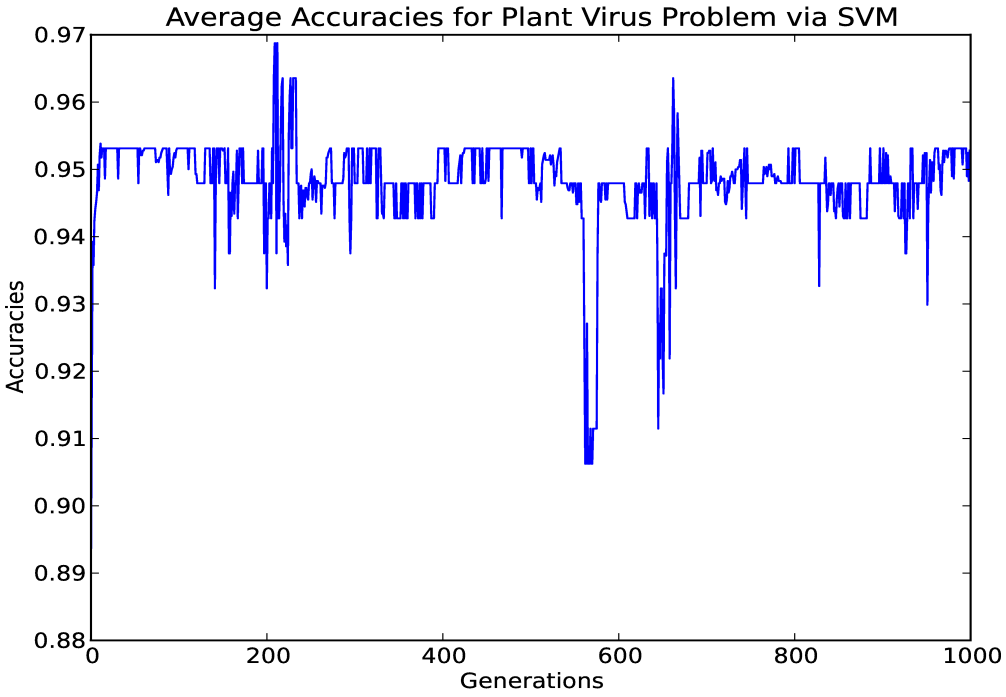
ταυτοποιούνται ως τέτοια, ενώ ως FP ορίζονται τα δείγματα που είναι CGMMV, αλλά ταυτοποιούνται λανθασμένα ως TRV. Ομοίως, ως TN ορίζονται τα δείγματα που είναι στην πραγματικότητα TRV και ταυτοποιούνται σωστά, ενώ ως FN ορίζονται εκείνα που αντιστοιχούν σε TRV, αλλά ταυτοποιούνται ως CGMMV. Στην περίπτωση του προβλήματος των φυτικών ιών, οι δείκτες Ευαισθησίας και Ειδικότητας ποσοτικοποιούν την πιθανότητα του εργαλείου για αποτελεσματική ταυτοποποίηση των δειγμάτων. Με την πρώτη υπολογίζεται η αναλογία των δειγμάτων CGMMV τα οποία το εργαλείο ταυτοποιεί σωστά, ενώ με τη δεύτερη αποδίδεται το αντίστοιχο ποσοστό για τον ιό TRV. Οι δείκτες PPV και NPV υπολογίζουν την πιθανότητα να αποδειχθεί σωστή η ταξινόμηση ενός συγκεκριμένου ιού που έχει ήδη ταυτοποιηθεί ως CGMMV ή TRV αντίστοιχα.



Σχήμα 6.3: *Μέγιστες τιμές ακρίβειας ανά γενεά ΤΝΔ (αναγνώριση φυτικών ιών).*

6.4 Σχολιασμός

Η αποτελεσματική αντιμετώπιση επιδημικών καταστάσεων, είτε όσον αφορά στην ανθρώπινη υγεία, είτε στην υγεία των φυτών, πριν από τη φάση της λήψης των κατάλληλων μέτρων, απαιτεί κυρίως τον προσδιορισμό της φύσης του παθογόνου, εάν αυτό ανήκει σε γνωστή και ορισμένη οικογένεια ή πρόκειται για ένα τελείως νέο είδος. Στο πλαίσιο αυτό, η ακριβής αναγνώριση παθογόνων ιών είναι ιδιαίτερα σημαντική τόσο για την ανθρώπινη υγεία, όσο και για την υγεία φυτών και καλλιεργειών. Το σύστημα που προτείνεται στη διατριβή τροφοδοτείται με δεδομένα που προέρχονται από τη μέθοδο BERA και χρησιμοποιούνται για την αναγνώριση δυο παθογόνων των φυτών εξέχουσας οικονομικής σημασίας, του ιού του κροταλίσματος του καπνού (TRV) και αυτού της πράσινης ποικιλοχλώρωσης με μωσαϊκό της αγγουριάς (CGMMV), οι οποίοι προξενούν σημαντικά προβλήματα στην παραγωγή. Η μέθοδος BERA χρησιμοποιεί εξειδικευμένους βιο-αισθητήρες που περιέχουν συγκεκριμένα αντιδραστήρια ακινητοποιημένα σε πήγμα τα οποία,

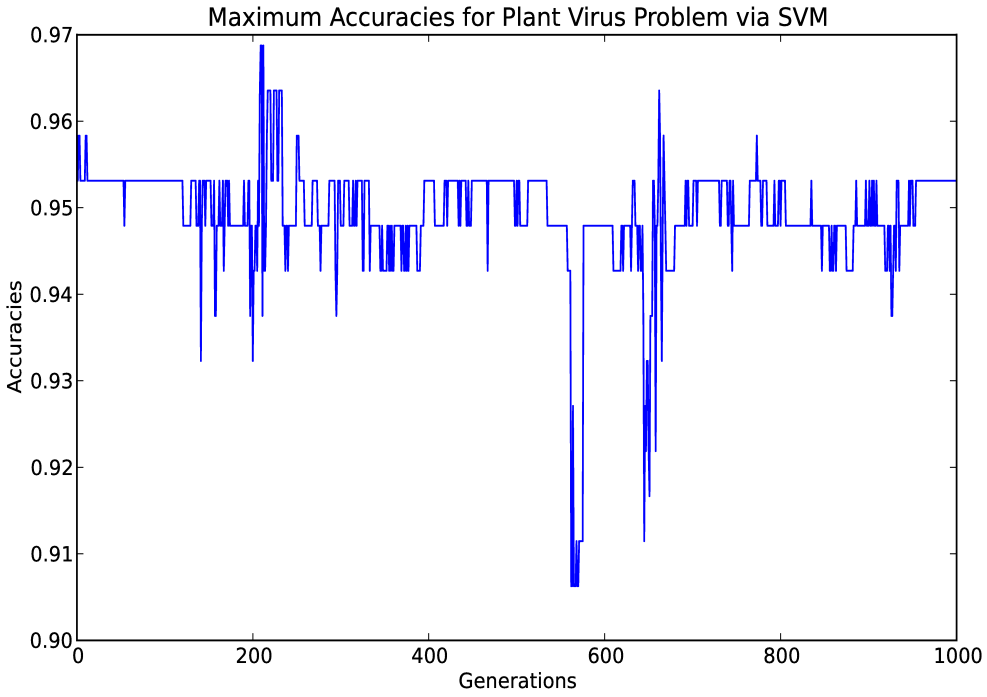


Σχήμα 6.4: *Μέσες τιμές ακρίβειας ανά γενεά ΜΔΥ (αναγνώριση φυτικών ιών).*

αντιδρώντας με τους εν λόγω ιούς, παράγουν σήματα ηλεκτρικού φορτίου. Η αντίδραση αυτή διαρκεί για συγκεκριμένο χρόνο, συνεπώς τα σήματα που παράγονται είναι ουσιαστικά χρονοσειρές – υπογραφές χαρακτηριστικές για κάθε ιό. Εξετάζοντας προσεκτικότερα τα αποτελέσματα που εξήχθησαν και συγκρίνοντας την απόδοση των ταξινομητών μετά την εκπαίδευσή τους με και χωρίς την εξελικτική προεπεξεργασία που προτείνεται από τη μέθοδο ΠΕΤ, θα μπορούσαν να σημειωθούν οι ακόλουθες παρατηρήσεις:

- Το προτεινόμενο σύστημα ανάλυσης χρονοσειρών αποτέλεσε σημαντικό παράγοντα για την εύρεση της άριστης λύσης στο πρόβλημα της ταξινόμησης των ιών. Ο αλγόριθμος ήταν σε θέση να διακρίνει τα κρισιμότερα χαρακτηριστικά κάθε χρονοσειράς, σημειώνοντας σημαντική βελτίωση στην ποιότητα της αναπαράστασης των αρχικών δεδομένων και στην εξομάλυνσή τους.

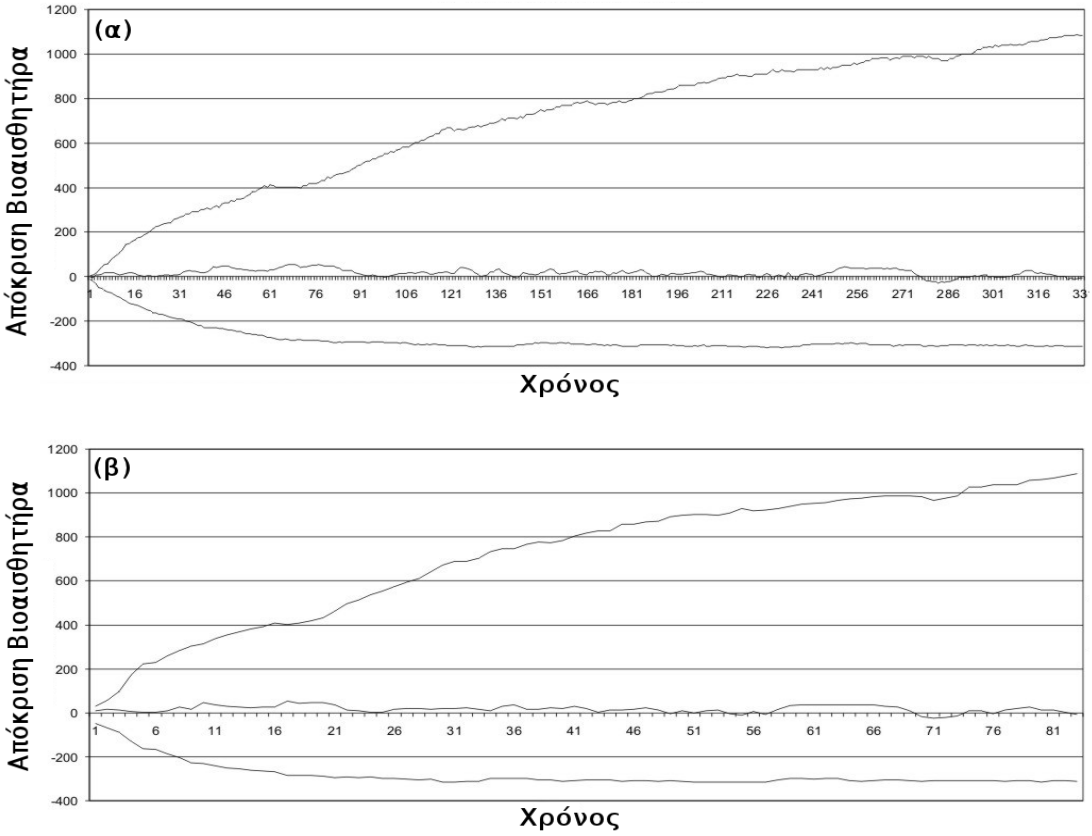
  Χαρακτηριστικό είναι το παράδειγμα εξομάλυνσης που επιτεύχθηκε μέσω της ΠΕΤ και δίνεται στο γράφημα του σχήματος 6.6. Πρόκειται για απεικόνιση



Σχήμα 6.5: *Μέγιστες τιμές ακρίβειας ανά γενεά ΜΔΥ (αναγνώριση φυτικών ιών).*

των αποτελεσμάτων της διαδικασίας που ακολουθήθηκε για το πρόβλημα αναγνώρισης ιών και συγκεκριμένα για ένα τυχαίο δείγμα που αντιστοιχεί στον ιό TRV. Η πρώτη εικόνα (α) περιλαμβάνει την αρχική χρονοσειρά, όπως αυτή συλλέχθηκε από τον βιο-αισθητήρα της μεθόδου BERA. Στην περίπτωση αυτή η πληροφορία αποτελείται από 331 μετρήσεις σε μια οδοντωτή γραμμή με πολλές πτυχώσεις. Στην επόμενη εικόνα (β) εμφανίζονται τα αποτελέσματα της εξελικτικής διαδικασίας που ακολουθήθηκε. Η προτεινόμενη μέθοδος πέτυχε να μειώσει το πλήθος των στοιχείων της εισόδου που απαιτούνται για την αναγνώριση κάθε ιού σε 84 μετρήσεις ανά υπογραφή, με αποτέλεσμα να παράγονται πιο λείες γραμμές.

- Η μέθοδος ΠΕΤ βελτίωσε σημαντικά τις δυνατότητες πρόβλεψης και ταξινόμησης και των δύο ταξινομητών (Πίνακας 6.3. Η βελτίωση αυτή γίνεται ιδιαίτερα εμφατική στην παρούσα περίπτωση της ταυτοποίησης ιών των φυτών, όπου η χρονοσειρά είναι αρκετά ευρεία και η αναπαράστασή της αποδεικνύ-



Σχήμα 6.6: *Αρχική (α) και εξελικτικά επεξεργασμένη μέσω της μεθόδου ΠΕΤ (β) χρονοσειρά αποκρίσεων του ιού TRV.*

εται πολύ αποτελεσματική.

- Ισχυρή βελτίωση της διακριτικής ικανότητας έχουμε στην περίπτωση του προβλήματος της ταυτοποίησης ιών των φυτών, όπου ένα άλμα της τάξης του 36% (από 51,6% σε 87,5%) σημειώνεται στην περίπτωση χρήσης ταξινομητή ΤΝΔ. Η μέθοδος ΠΕΤ είχε σημαντικά μικρότερη συμβολή στη βελτίωση της διακριτικής ικανότητας της ΜΔΥ, στα επίπεδα του 8,5% (από 88,5% σε 96,9%) (πίνακας 6.3). Παρ' όλ' αυτά, το σύστημα προκρίνει τη ΜΔΥ ως το βέλτιστο ταξινομητή για το συγκεκριμένο πρόβλημα, λόγω της αντικειμενικής του υπεροχής, καθότι φαίνεται ότι αποδίδει καλύτερα από την αρχή της εξελικτικής διαδικασίας.

Εκτός της σημαντικής επίδρασης της μεθόδου στη διακριτική ικανότητα των ταξινομητών, κυριότερη ωφέλεια που προκύπτει στο πλαίσιο της διατριβής είναι η δημιουργία ενός προτύπου για την ανάπτυξη ενός εξελισσόμενου συστήματος ικανού να ταξινομεί σωστά ποικίλους ιούς αξιολογώντας τις αντιδράσεις τους οι οποίες καταγράφονται από τη μέθοδο BERA. Εκτός της χρησιμότητάς του για τους ιολόγους, το σύστημα επίσης συγκρατεί τα λειτουργικά του έξοδα σε αρκετά ανεκτά επίπεδα, καθώς οι μοναδικές του απαιτήσεις εντοπίζονται στην απόκτηση των αρχικών σημάτων και στη λειτουργία του αλγορίθμου αναγνώρισης. Η πρώτη εναπόκειται σε μια καλά μελετημένη μέθοδο, ενώ η δεύτερη έχει απαιτήσεις μόνο σε υπολογιστική ισχύ. Εξαιτίας της μεγάλης παραγωγής δευτερογενών δεδομένων, αυξάνει η παραλλακτικότητα εντός της γενετικής δεξαμενής, αυξάνοντας παράλληλα την πιθανότητα επιλογής κατάλληλου γενετικού δομικού υλικού για την επόμενη γενεά. Αφού ολοκληρωθεί η εξελικτική διαδικασία με την αποκάλυψη του καλύτερου εκπαιδευτή ολοκληρώνεται και η εξαγωγή χαρακτηριστικών από την αρχική χρονοσειρά. Στην περίπτωση αυτή φαίνεται να επιτυγχάνεται δραστική μείωση του θορύβου και του βαθμού διάστασης της αρχικής πληροφορίας, με προφανείς ωφέλειες, τόσο όσον αφορά στο χρόνο εκπαίδευσης και στην απαιτούμενη υπολογιστική ισχύ, όσο και στη διακριτική ικανότητα του τελικού ταξινομητή. Φυσικά, από τη στιγμή που ολοκληρώνεται η εξελικτική διαδικασία για ένα συγκεκριμένο πρόβλημα, το προτεινόμενο σύστημα επιδεικνύει παρόμοιες απαιτήσεις σε χρόνο και υπολογιστική ισχύ όπως και οι υπόλοιποι ταξινομητές. Παράλληλα, είναι ιδιαίτερα ανταγωνιστικό έναντι του χρόνου και της ακρίβειας με την οποία επιτελεί την αναγνώριση ένας εμπειρογνώμονας.

Κεφάλαιο 7

ΠΡΟΒΛΕΨΗ ΧΕΙΜΑΡΡΙΚΗΣ ΕΠΙΚΙΝΔΥΝΟΤΗΤΑΣ

Στο κεφάλαιο αυτό παρουσιάζεται η εφαρμογή της ΠΕΤ στη διαχείριση υδατικών αποθεμάτων και ειδικά στην πρόβλεψη της χειμαρρικής επικινδυνότητας. Για την περίπτωση αυτή δίδονται πληροφορίες σχετικά με την περιοχή μελέτης του προβλήματος, το χειρισμό των δεδομένων και τα αποτελέσματα που προέκυψαν από την εφαρμογή της μεθόδου.

Στις ημέρες μας αυταπόδεικτο θεωρείται το γεγονός ότι τα υδάτινα διαθέσιμα μιας περιοχής παίζουν σημαντικό έως πρωτεύοντα ρόλο στην ευζωία των κατοίκων της, ενώ η ορθολογική τους διαχείριση ανάγεται σε μια από τις πλέον κρίσιμες συνιστώσες για την ολοκληρωμένη ανάπτυξή της. Η αύξηση των πιέσεων στο υδατικό περιβάλλον καθιστά αναγκαία την εφαρμογή ολοκληρωμένων πολιτικών ανάπτυξης και διαχείρισης των υδατικών διαθεσίμων, μέσω σχεδιασμού, υλοποίησης και βέλτιστης λειτουργίας έργων υποδομής και παρεμβάσεων διαχείρισης. Μια ορθολογική πολιτική ανάπτυξης οφείλει να λαμβάνει υπόψη, εκτός από την σε βάθος χρόνου προστασία των υδάτων, επίσης και την αποτελεσματική διαχείριση ακραίων φαινομένων και κρίσεων όπως τα προβλήματα λειψυδρίας και πλημμυρών. Ειδικότερα για την Ελλάδα αξίζει να σημειωθεί ότι, αν και παρατηρείται έντονη αναντιστοιχία μεταξύ της χρονικής και κυρίως της χωρικής κατανομής των βροχοπτώ-

σεων με τις χρονικές και χωρικές κατανομές της ζήτησης, η χώρα μας δεν παύει να είναι μια σχετικά ευνοημένη υδρολογικά χώρα της Μεσογείου. Παρά το γεγονός όμως αυτό, οι προαναφερόμενοι ανασταλτικοί παράγοντες σε συνδυασμό με την υψηλή παραλλακτικότητα του εδαφικού ανάγλυφου, έχουν δημιουργήσει προβλήματα έλλειψης νερού, όπως και σημαντικά προβλήματα πλημμυρικών καταστάσεων σε ποικίλες περιοχές.

7.1 Πλημμυρικά φαινόμενα

Ο συνηθέστερος λόγος πρόκλησης πλημμυρικών φαινομένων σε συγκεκριμένο τόπο και χρόνο είναι η κατάσταση στην οποία βρίσκονται αλλά και ο τρόπος με τον οποίο μεγάλα υδατικά αποθέματα υπερχειλίζουν, ή παλίρροιες σαρώνουν ηπειρωτικές περιοχές, είτε εξαιτίας σημαντικής αύξησης του ποσοστού βροχόπτωσης είτε λόγω απότομης και ευρείας κλίμακας τήξης παγωμένων όγκων ύδατος, τα οποία επιβαρύνουν απότομα και ανεξέλεγκτα την υδατοϊκανότητα παρακείμενων φυσικών ή τεχνητών υδατικών ταμιευτήρων. Η Ομοσπονδιακή Υπηρεσία Διαχείρισης Εκτάκτων Καταστάσεων των Η.Π.Α. (FEMA: Federal Emergency Management Agency, http://www.fema.gov), στο σχετικό εθνικό της πρόγραμμα (National Flood Insurance Program) χαρακτηρίζει πλημμυρική την κατάσταση ύπαρξης περίσσειας επιφανειακού ύδατος σε έδαφος που υπό κανονικές συνθήκες είναι ξηρότερο. Συγκεκριμένα, ως πλημμύρα ορίζεται μια γενική και προσωρινή κατάσταση μερικού ή ολικού κατακλυσμού δύο ή περισσότερων εκταρίων κανονικά ξηρών γαιών εξαιτίας:

- Υπερχείλισης ενδοχώριων ή παλιρροϊκών υδάτινων μαζών
- Ασυνήθους και ταχείας συσσώρευσης επιφανειακών νερών απορροής ή ιλύος από οποιαδήποτε πηγή
- Κατάρρευσης, καθίζησης ή/και ταπείνωσης της στάθμης γαιών ως αποτέλεσμα διάβρωσης, αποσάθρωσης ή τεχνητής ή φυσικής υπονόμευσης από παρακείμενες μεγάλες υδάτινες επιφάνειες ή ρεύματα τα οποία υπερβαίνουν τα όρια των συνηθισμένων επαναλαμβανόμενων κύκλων τους.

Η ύπαρξη παρακείμενων μεγάλων υδάτινων επιφανειών δεν αποτελεί αναγκαία συνθήκη για τη δημιουργία πλημμυρικών προϋποθέσεων. Ανεξαρτήτως υψομέτρου, γεωγραφικού μήκους και πλάτους, ακαριαίες πλημμύρες (flash floods) είναι δυνατόν να επισυμβούν οπουδήποτε, στις περιπτώσεις όπου υπερβολικά μεγάλοι όγκοι υδάτινων μετεωρολογικών κατακρημνισμάτων και υετού πέφτουν σε μικρό σχετικά χρονικό διάστημα στην ίδια περιοχή.

Είναι επίσης κοινός τόπος το γεγονός ότι χείμαρροι που υπερχειλίζουν και δρουν ανεξέλεγκτα είναι δυνατόν να προκαλέσουν σοβαρές πλημμύρες, η επικινδυνότητα των οποίων επηρεάζεται άμεσα από το καθεστώς των ατμοσφαιρικών κατακρημνισμάτων και τον τύπο του περιβάλλοντος τοπίου. Οι διάφορες πλημμυρικές καταστάσεις ευνοούνται όσο η εδαφική υδατοπερατότητα, απορροφητικότητα και αποστράγγιση μειώνονται και όσο αυξάνει το μέσο ύψος βροχής. Το γεγονός ότι έντονοι χείμαρροι συνήθως επισυμβαίνουν ανά τακτά χρονικά διαστήματα σε κάποιες περιοχές, έχει οδηγήσει τους μελετητές στην ανάλυση διαφόρων ειδών δεδομένων που προέρχονται από ποικίλες πηγές, με πλέον χαρακτηριστικό τα δεδομένα χρονοσειρών που σχετίζονται με το ύψος βροχής μιας περιοχής σε ιστορικό βάθος χρόνου. Μελετώνται επίσης και άλλοι παράγοντες που σχετίζονται με τη μηχανική συμπεριφορά του εδάφους, το υψόμετρο, την εδαφική δομή και κλίση, την εδαφοκάλυψη, καθώς επίσης και τη φύση της κοίτης και της λεκάνης απορροής του χειμάρρου, αλλά δεν θεωρούνται τόσο σημαντικοί όσο η ανάλυση χρονοσειρών η οποία σχετίζεται με τα πάσης φύσεως κατακρημνίσματα.

7.2 Η περιοχή μελέτης

Η Κύπρος, το τρίτο σε μέγεθος νησί της Μεσογείου μετά τη Σικελία και τη Σαρδηνία, τοποθετείται γεωγραφικά νοτίως της Ανατολικής χερσονήσου της Ασίας. Συνορεύει βορειοδυτικά με την Ελλάδα, προς ανατολάς με τη Συρία, το Λίβανο και το Ισραήλ και βόρεια με την Τουρκία. Συγκαταλέγεται ανάμεσα στα κράτη-μέλη της Ευρωπαϊκής Ενωσης και συνήθως αναφέρεται ως τμήμα της Μέσης Ανατολής.

Το ανάγλυφο του εδάφους, ως επί το πλείστον ορεινό, περιλαμβάνει την κεντρική πεδιάδα της Μεσαορίας η οποία περικλείεται από τα όρη της Κυρήνειας και του Πενταδάχτυλου προς βορρά και την οροσειρά Τρόοδος προς τα νοτιοδυτικά, ενώ υπάρχουν σκόρπιες μικρότερες ήσσονος σημασίας πεδιάδες κατά μήκος

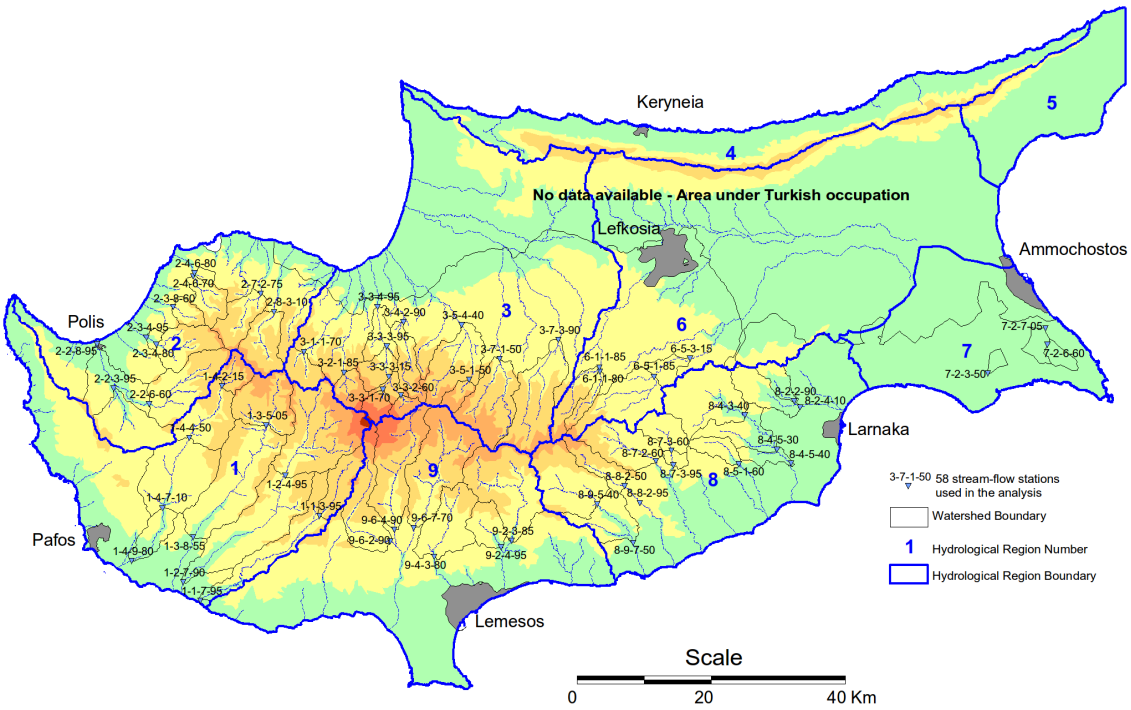
της νότιας ακτής του νησιού. Το κλίμα είναι ήπιο Μεσογειακό και χαρακτηρίζεται από ξηρά καλοκαίρια τα οποία ακολουθούνται από συνήθως βροχερούς χειμώνες.

Οι καλοκαιρινές θερμοκρασίες κυμαίνονται από μετρίως υψηλές στα χαμηλότερα υψόμετρα του όρους Τρόοδος σε υψηλότερες στις πεδιάδες. Οι χειμερινές θερμοκρασίες είναι ηπιότερες σε μικρό υψόμετρο, όπου σπάνια επικρατούν χιονοπτώσεις, αλλά σημειώνονται σημαντικά χαμηλότερες στην οροσειρά Τρόοδος. Κατά τη διάρκεια των τελευταίων ετών έχει σημειωθεί μεταβολή του κλίματος προς την επικράτηση σημαντικά μακρύτερων ξηρών περιόδων, οι οποίες εντείνουν το πρόβλημα έλλειψης πόσιμου νερού ως μια σημαντική πίεση έναντι του πληθυσμού. Υπ' αυτές τις συνθήκες, οι έντονες πλημμύρες, οι χαμηλές θερμοκρασίες και οι χειμαρρικές καταστάσεις που σημειώνονται κατά καιρούς στα ηπειρωτικά κατά τη διάρκεια της υγρής περιόδου συνιστούν μια ασυνήθιστη εικόνα για την εποχή αυτή και προκαλούν διάβρωση, αποσάθρωση και σημαντικές καταστροφές στις εγκαταστάσεις, τους οικισμούς και γενικότερα την υποδομή, ιδιαίτερα των γεωργικών εκμεταλλεύσεων. Δεν υπάρχει αμφιβολία ότι η αποτελεσματική διαχείριση των υδατικών διαθεσίμων αποτελεί παράγοντα κλειδί όχι μόνο για την ευζωία των κατοίκων και την ικανοποίηση των καθημερινών αναγκών τους, αλλά επίσης και για την επίτευξη της ολοκληρωμένης ανάπτυξης στο νησί.

7.3 Χειρισμός των δεδομένων

Η περιοχή έρευνας [84] καλύπτει όλες τις ορεινές λεκάνες απορροής που υπόκεινται στη Διοίκηση της Δημοκρατίας της Κύπρου. Συγκεκριμένα, το νησί διαιρείται σε εννέα διοικητικές υδατικές περιφέρειες οι οποίες διαχειρίζονται συνολικά εβδομήντα χειμαρρικά ρεύματα. Όπως έχει ήδη σημειωθεί, τα σημαντικότερα στοιχεία του εδαφικού αναγλύφου αποτελούν τα όρη Κυρήνεια και Πενταδάχτυλο προς το βορρά με υψόμετρο που αγγίζει τα 1.000 m και μήκος τα 160 km περίπου, ενώ το όρος Τρόοδος στα νοτιοδυτικά φθάνει σε υψόμετρο τα 1.951 m.

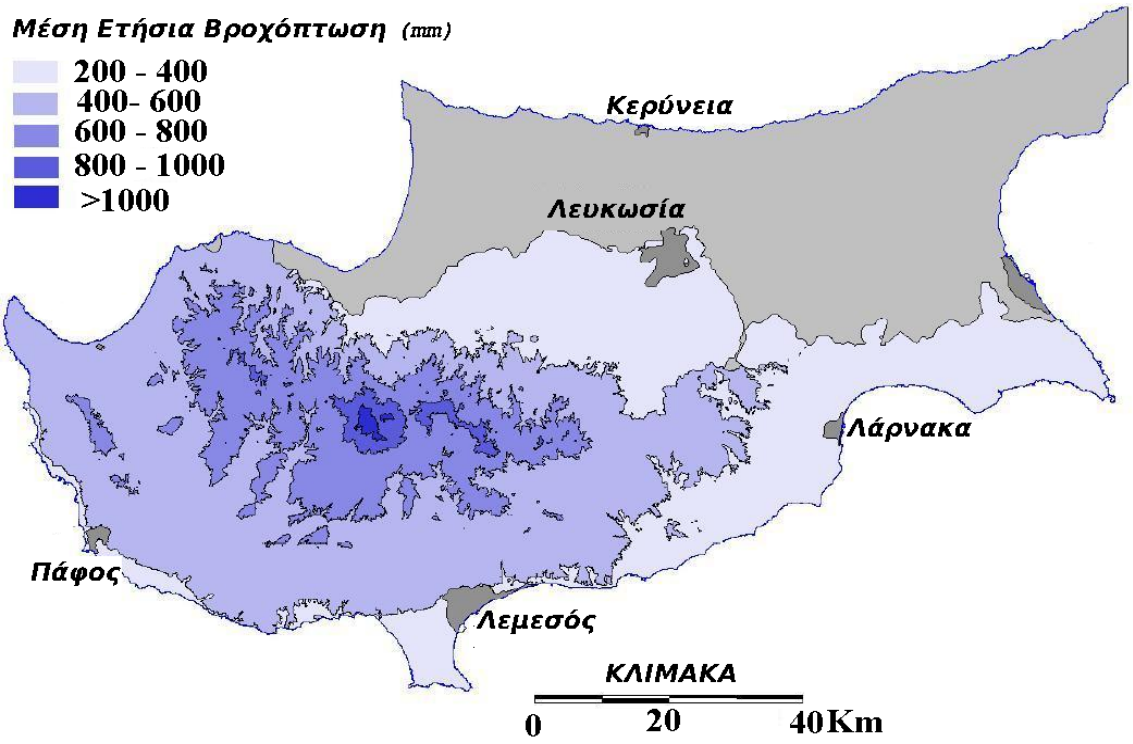
Η συλλογή του πρωτογενούς συνόλου δεδομένων επιτεύχθηκε από 78 σταθμούς τοποθετημένους κατά μήκος των 70 χειμαρρικών ρευμάτων όπως φαίνεται στην Εικ. 7.1. Η αρχική αυτή πληροφορία καλύπτει μια χρονική περίοδο 28 ετών, από το 1965 έως το 1993 για τις μετρήσεις των περισσότερων σταθμών. Στην έρευνα δε λαμβάνεται υπόψη το μέσο ετήσιο, αλλά το μέσο μηνιαίο ύψος βροχής για κάθε χρόνο.



Εικ. 7.1: *Γενική άποψη του υδρογραφικού δικτύου της Δημοκρατίας της Κύπρου.*

Τρεις δομικές και δύο δυναμικές παράμετροι εισόδου παίζουν σημαντικό ρόλο στην πρόβλεψη [84]. Συγκεκριμένα, στην πρώτη κατηγορία εμπίπτουν η επιφάνεια της λεκάνης απορροής, το υψόμετρο και η κλίση του εδάφους, ενώ στη δεύτερη δυο δομικές, η μέση ετήσια και η μέση μηνιαία βροχόπτωση. Για κάθε δυναμικό παράγοντα ο μέσος όρος υπολογίστηκε σε ετήσια βάση και το αποτέλεσμα αποτέλεσε μια τιμή εισόδου του συστήματος. Σημαντικό στοιχείο και παράγοντας κλειδί για την έρευνα αποτέλεσε το γεγονός ότι χρησιμοποιήθηκαν μόνο δύο δυναμικές παράμετροι και από αυτές μόνο το ύψος βροχής ήταν απαραίτητο να παρακολουθείται και να καταγράφεται σε μηνιαία και ετήσια βάση, καθιστώντας με τον τρόπο αυτό την απόκτηση των δεδομένων του συστήματος άκοπη και σχετικά χαμηλού κόστους, τόσο σε οικονομικό επίπεδο, όσο και σε επίπεδο ανθρώπινων πόρων. Στην προτεινόμενη μεθοδολογία, ενώ χρησιμοποιείται ως μια από τις βασικές συνιστώσες το ετήσιο μέσο ύψος βροχής, αυτό πλέον δεν παίζει τον σπουδαιότερο ρόλο στο διάνυσμα εισόδου. Αντίθετα, στην περίπτωση αυτή προτιμήθηκε ολόκληρη η χρονοσειρά του μέσου μηνιαίου ύψους βροχής από την οποία εξήχθησαν οι βέλτιστοι συνδυα-

σμοί ως είσοδοι στον εκάστοτε ταξινομητή. Ιδιαίτερο χαρακτηριστικό αυτής της προσέγγισης αποτελεί το ότι ενώ κρατάει σε χαμηλά επίπεδα τα κόστη παραγωγής και ανάκτησης δεδομένων, ταυτόχρονα μειώνει το συνολικό βαθμό διάστασης του διανύσματος εισόδου, καθιστώντας το ακριβέστερο. Από τη στιγμή που εκπαιδευθεί, ο ταξινομητής παραμένει ιδιαίτερα ευπροσάρμοστος σε ποικίλες περιφέρειες, υπό την προϋπόθεση ότι τηρείται η τμηματοποίηση του διανύσματος εισόδου που υπαγορεύει ο ΓΑ.



Εικ. 7.2: *Δημοκρατία της Κύπρου: Μέσο ετήσιο ύψος βροχής (σε mm) από το 1970 έως το 2000.*

Στην εικόνα 7.1 απεικονίζεται μια γενική άποψη του υδρογραφικού δικτύου του νησιού, ενώ στην εικόνα 7.2 παρουσιάζεται το μέσο ετήσιο ύψος βροχής (σε mm) από το 1970 έως το 2000. Έξάλλου, στον Πίνακα 7.1 παρουσιάζεται ένα ενδεικτικό δείγμα των δεδομένων όπως συγκεντρώθηκαν για τις ανάγκες της έρευνας. Οι 78 μετεωρολογικοί σταθμοί, καθώς επίσης και τα δεδομένα που συγκεντρώθηκαν για όλη τη χρονική διάρκεια της μελέτης, ανήκουν στο Υπουργείο Γεωργίας, Φυσικών Πόρων και Περιβάλλοντος της Κύπρου [84].

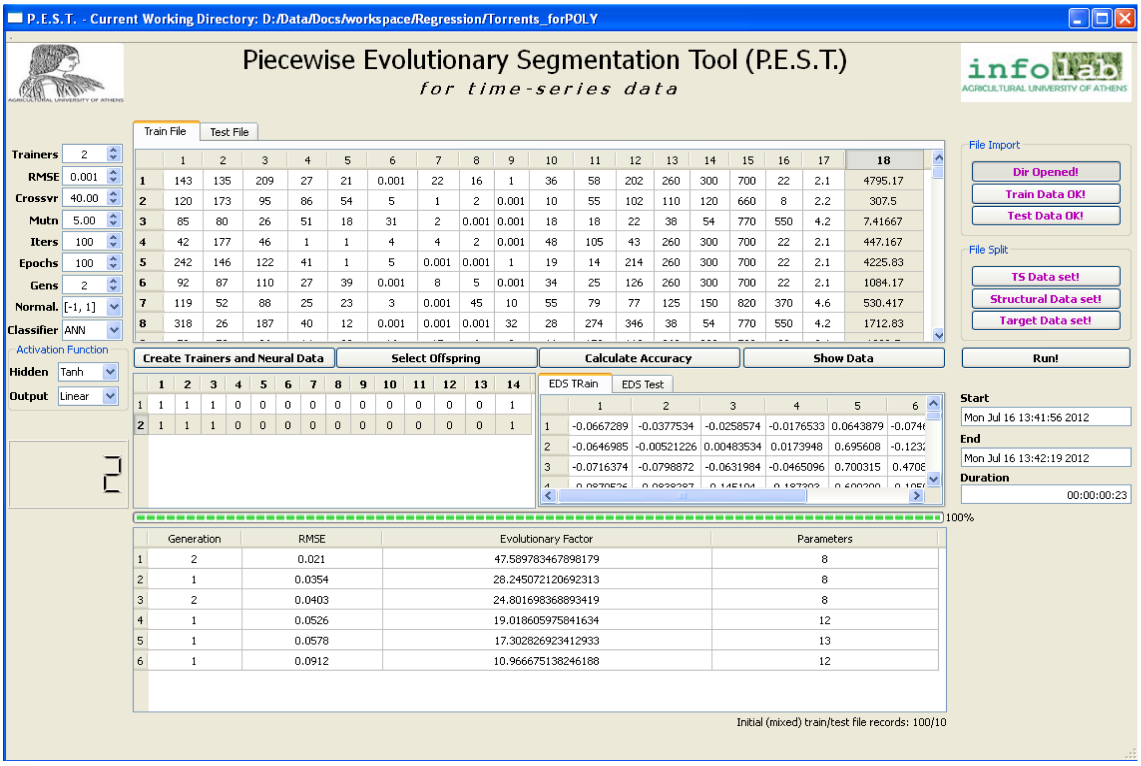
Πίνακας 7.1: Δείγμα δεδομένων για το πρόβλημα πρόβλεψης πλημμυρικών κινδύνων: M1-M12:Μήνες του έτους, Fn: επιφάνεια της λεκάνης απορροής, Q_{max}: μέγιστη παροχή ύδατος, P: μέση ετήσια βροχόπτωση, H: απόλυτο υψόμετρο, Jk: απόλυτη κλίση, Qmy: μέση ετήσια παροχή ύδατος.

	Σταθμός Α	**Σταθμός Β**	**Σταθμός Γ**
Έτος	**1965-66**	**1965-66**	**1979-80**
M1	229	201	195
M2	62	67	165
M3	89	77	135
M4	13	8	25
M5	4	4	9
M6	0	0	0
M7	0	0	0
M8	0	1	5
M9	70	54	1
M10	146	102	60
M11	23	19	133
M12	143	146	169
Fn (Km^2)	38	110	22
Q_{max}	54	120	8.6
P (mm)	770	660	810
H (m)	550	8	600
Jk (%)	4.2	2.2	8
Qmy (m^3/s)	420.62	611.95	330.6

Τα δεδομένα εισόδου αποτελούνται από παράγοντες οι οποίοι, ανάλογα με τον τύπο τους, είναι δυνατόν να ταξινομηθούν σε δυο κατηγορίες. Η πρώτη περιλαμβάνει τα δομικά δεδομένα με την έννοια ότι παραμένουν σταθερά καθ' όλη τη χρονική διάρκεια της μελέτης, σε αντίθεση με τη δεύτερη κατηγορία, η οποία περιλαμβάνει τα δυναμικά δεδομένα, με την έννοια ότι αναφέρεται σε στοιχεία τα οποία μεταβάλλονται κατά τη διάρκεια της καταγραφής δεδομένων.

Οσον αφορά στο συγκεκριμένο πρόβλημα, συνολικά συγκεντρώθηκαν 1.273 εγγραφές από τις ποικίλες λεκάνες απορροής της Κύπρου. Με βάση αυτό το σύνολο

δεδομένων σχηματίζονται τα πακέτα εκπαίδευσης, ελέγχου και αξιολόγησης, αποτελούμενα από 1.100, 100 και 73 δείγματα αντίστοιχα το καθένα.



Εικ. 7.3: *Πρωτογενή και εξελικτικά δεδομένα και αποτελέσματα για το πρόβλημα της χειμαρρικής επικινδυνότητας στο εργαλείο λογισμικού.*

Το διάνυσμα εισόδου περιλαμβάνει την επιφάνεια της λεκάνης απορροής (σε km^2), το απόλυτο υψόμετρο (σε μέτρα) και την απόλυτη κλίση (σε ποσοστό επί τοις εκατό) ως δομικά δεδομένα, ενώ στα δυναμικά δεδομένα περιλαμβάνονται η μέγιστη παροχή ύδατος, η μέση ετήσια, καθώς επίσης και η μέση μηνιαία βροχόπτωση. Ως μόνη έξοδο το δίκτυο καθορίζει τη Μέση Ετήσια Παροχή Ύδατος (σε m^3/s).

7.4 Αποτελέσματα

Τα αποτελέσματα που προέκυψαν από την έρευνα δίνονται στον Πίνακα 7.2 και στην Εικόνα 7.1, όπου απεικονίζεται η επίδραση που έχει η προτεινόμενη μέθοδος ΠΕΤ στην ικανότητα πρόβλεψης αμφότερων των ταξινομητών. Η παρούσα περίπτωση σχετίζεται με τη διαχείριση υδάτινων αποθεμάτων και ουσιαστικά αποτελεί

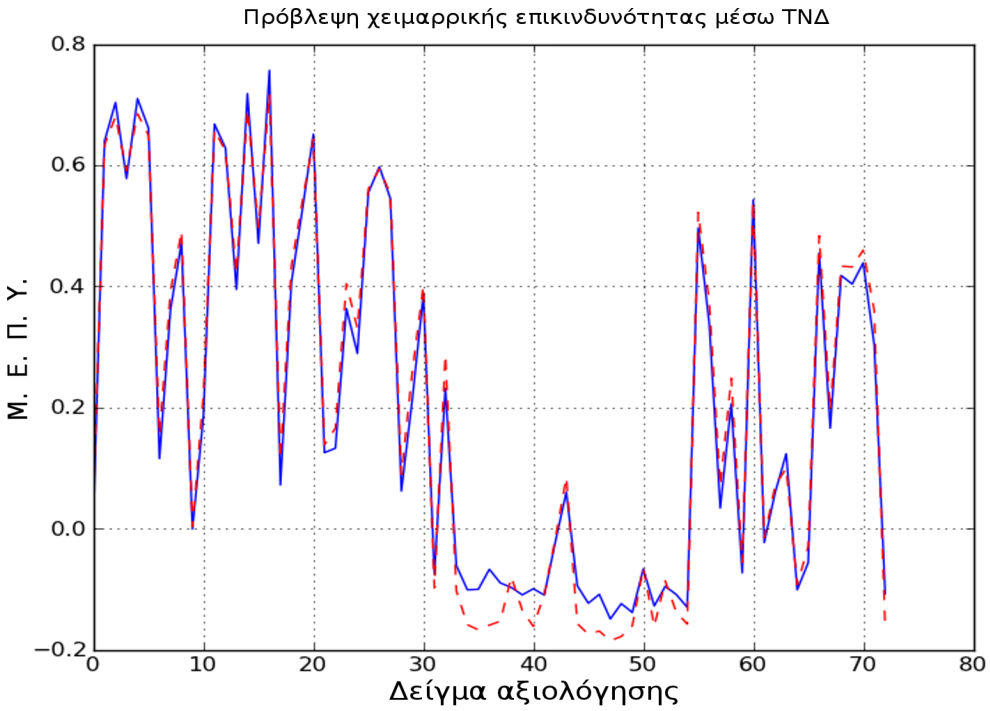
Πίνακας 7.2: Αποτελέσματα πρόβλεψης χειμαρρικής επικινδυνότητας: Αρχική χρονοσειρά έναντι εξελικτικών δεδομένων υπό ΤΝΔ και ΜΔΥ, ταξινόμηση με αύξον RMSE

ΤΝΔ			**ΜΔΥ**			
Γενεά	**RMSE**	**Νευρώνες**	**Γενεά**	**RMSE**	**C**	**γ**
28	59e-4	14	1	64e-3	1000	8
49	62e-4	15	4	65e-3	1000	8
90	63e-4	16	394	66e-3	1000	4
31	72e-4	13	329	67e-3	400	4
29	76e-4	14	1	68e-3	700	8
Πρωτογενή Δεδομένα	64e-3	19	Πρωτογενή Δεδομένα	68e-3	1000	4

ένα πρόβλημα πρόβλεψης. Επίσης και στην περίπτωση αυτή, όπως και στην προηγούμενη, φαίνεται ότι η εξαγωγή χαρακτηριστικών από τα δεδομένα της χρονοσειράς που επιτυγχάνεται μέσω της μεθόδου ΠΕΤ, έχει θετικά αποτελέσματα, καθώς βελτιώνει τη δυνατότητα πρόβλεψης και των δύο ταξινομητών (πίνακας 7.2).

Η απόδοση κάθε εκπαιδευτή υπολογίζεται σε συνάρτηση με το μέσο τυπικό τετραγωνικό σφάλμα (RMSE: Root Mean Square Error) κατά τη διάρκεια της φάσης ελέγχου του ταξινομητή. Μέσω του δείκτη RMSE ποσοτικοποιείται η ποιότητα της αναπαράστασης του προτύπου, με τις χαμηλότερες τιμές του δείκτη να αντιστοιχούν σε αποτελεσματικότερες αναπαραστάσεις.

Για το πρόβλημα της χειμαρρικής επικινδυνότητας, το σύστημα προκρίνει τη χρήση ταξινομητή ΤΝΔ ως λειτουργική μονάδα πρόβλεψης. Παρά το γεγονός ότι η βελτίωση του μέσου τυπικού τετραγωνικού σφάλματος σε αμφότερους τους ταξινομητές δεν είναι τόσο εντυπωσιακή όπως αυτή που παρατηρείται όσον αφορά στην ακρίβεια της ταξινόμησης του προηγούμενου προβλήματος, αναμφίβολα η μέθοδος επηρεάζει θετικά την έκβαση της πρόβλεψης. Εξάλλου, εφόσον ο επικρατέστερος ταξινομητής εκπαιδευθεί με το σχήμα τμηματοποίησης που προτείνεται από το βέλτιστο εκπαιδευτή, το σύστημα φαίνεται να προβλέπει σωστά τις περιπτώσεις που αποτελούν πραγματική απειλή για χειμαρρική δραστηριοποίηση, ενώ καθίσταται χαλαρότερο για μικρότερες τιμές της Μέσης Ετήσιας Παροχής Ύδατος (Εικ. 7.1).



Σχήμα 7.1: *Πρόβλεψη χειμαρρικής επικινδυνότητας με εφαρμογή ΤΝΔ εκπαιδευμένου με το καλύτερο σχήμα εξελικτικής τμηματοποίησης της 28ης γενεάς και εφαρμοσμένου στο σετ αξιολόγησης. Η γαλάζια συνεχής γραμμή αναπαριστά την πρόβλεψη του συστήματος, ενώ η κόκκινη διακεκομμένη αναπαριστά πραγματικές τιμές.*

7.5 Σχολιασμός

Πρόσφατες τάσεις και κλιματικές αλλαγές κατέστησαν φανερή την ανάγκη για αποτελεσματική διαχείριση των υδατικών αποθεμάτων, ειδικά για περιοχές όπως η Κύπρος στις οποίες η υφιστάμενη περιβαλλοντική πίεση έχει ασκήσει έντονη ανισορροπία στα υδατικά αποθέματα. Το πρόβλημα χαρακτηρίζεται ως ιδαίτερα κρίσιμο, ειδικά εάν ληφθεί υπόψη το γεγονός ότι μετά από επίπονες και χρονοβόρες μελέτες και ανάλυση δεδομένων, τα αρμόδια όργανα της Δημοκρατίας της Κύπρου δεν έχουν κατασταλάξει σε μια βιώσιμη λύση. Παράλληλα, η κλασσική στατιστική ανάλυση απέτυχε να αποδώσει ευοίωνα αποτελέσματα, παρότι η μελέτη του προβλήματος υπήρξε εντατική και επίμονη. Τα υδατικά διαθέσιμα του νησιού είναι στην παρούσα φάση πολύ χαμηλότερα από την αρχική θεώρηση, με επίπεδα απόκλισης ακόμη και 40% χαμηλότερα, γεγονός που επιτάσσει την ανάληψη πρωτο-

βουλιών για την προώθηση καινοτόμων και πιο αξιόπιστων λύσεων. Η πρόβλεψη της μέσης ετήσιας παροχής ύδατος θεωρείται ως ο κρισιμότερος παράγοντας, καθώς είναι στενά συνδεδεμένος με τη χειμαρρική επικινδυνότητα, ειδικά σε ορεινές περιοχές. Η ανάπτυξη εργαλείων για την πρόβλεψη τέτοιων φαινομένων σε βάθος χρόνου, θεωρείται μεγάλης σημασίας. Επίσης, θεωρείται σημαντικό να παραμένει το λειτουργικό τους κόστος, τόσο σε οικονομικό επίπεδο, όσο και σε επίπεδο ανθρώπινων διαθεσίμων, σε αποδεκτά όρια.

Βασικότερο επίτευγμα της διατριβής στη μελέτη αυτής της περίπτωσης είναι ο σχεδιασμός και η αποτελεσματική εφαρμογή της στην εξελικτική παραγωγή δευτερογενών δεδομένων τα οποία χρησιμοποιούνται για την πρόβλεψη ενός παράγοντα κλειδί στη διαχείριση των υδατικών διαθεσίμων και της χειμαρρικής επικινδυνότητας. Τα δεδομένα αυτά παράγονται μέσω εξελικτικής διαδικασίας στην προσπάθεια ελαχιστοποίησης του θορύβου και του υψηλού βαθμού διάστασης των πρωτογενών δεδομένων. Στο πλαίσιο της διατριβής σχεδιάσθηκε η ανάπτυξη ενός αυτό-προσαρμοζόμενου προτύπου βασισμένου στην εξελικτική υπολογιστική, το οποίο υλοποιεί ειδικά σχεδιασμένο γενετικό αλγόριθμο. Σκοπός του συστήματος είναι η εκτέλεση προσθιόδρομης αναζήτησης στο διάνυσμα εισόδου για την ανεύρεση κατάλληλων συνδυασμών γονιδίων που παράγουν δεδομένα ελαχιστοποίησης του σφάλματος των ταξινομητών. Μέσω του συστήματος επιχειρείται σε ετήσια βάση η πρόβλεψη της μέσης ετήσιας παροχής ύδατος για κάθε ορεινή λεκάνη απορροής της Κύπρου.

Σημαντικότερο συμπέρασμα σχετικά με την περίπτωση αυτή είναι το γεγονός ότι τελικά η μέθοδος που προτείνεται στο πλαίσιο αυτής της διατριβής, απαιτεί την καταβολή ελαχίστου κόστους, καθώς υπολογίζει μόνο δυο δυναμικούς παράγοντες εισόδου. Προφανώς οι δομικοί παράγοντες παραμένουν αμετάβλητοι για τη σχετικά μικρή χρονική περίοδο μελέτης, συνεπώς ο μόνος παράγοντας ο οποίος απαιτεί προσεκτική παρακολούθηση και μετρήσεις σε καθημερινή βάση είναι το ύψος βροχής. Επιπλέον, το σύστημα είναι δυνατό να επανεκπαιδευθεί εφόσον υπάρχουν διαθέσιμα νέα δεδομένα, υπό την προϋπόθεση ότι καινούριες μετρήσεις καθίστανται διαθέσιμες σε νέες περιφέρειες. Κρίσιμη συνεισφορά του προτεινόμενου αλγορίθμου είναι η εξελικτική τμηματοποίηση και δειγματοληψία που ασκείται επί της αρχικής χρονοσειράς δεδομένων, με χαρακτηριστικά πλεονεκτήματα. Αρχικά,

μετά την εφαρμογή του συστήματος ο μελετητής έχει στη διάθεσή του ένα μεγάλο αριθμό εκπαιδευτικών και ελεγκτικών συνόλων δεδομένων, διότι πλέον δεν είναι απαραίτητη η εξαγωγή του ετήσιου μέσου όρου βροχόπτωσης. Αντίθετα, ολόκληρος ο χρόνος αποτελεί πλέον ένα διάνυσμα εισόδου στο σύστημα, πολλαπλασιάζοντας σημαντικά τα διαθέσιμα δεδομένα, αλλά και διατηρώντας πιθανά χρήσιμη πληροφορία. Η εξελικτική διαδικασία μειώνει τη διάσταση και το θόρυβο της εισόδου με αποτέλεσμα δευτερογενή δεδομένα αυξημένης ικανότητας γενίκευσης.

Κεφάλαιο 8
ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτό το καταληκτικό κεφάλαιο της διατριβής επιχειρείται μια συνολική αποτίμηση της μεθόδου της Πρότυπης Εξελικτικής Τμηματοποίησης. Παρουσιάζονται τα καινοτόμα στοιχεία, τα κυριότερα πλεονεκτήματα, αλλά και οι περιορισμοί της προτεινόμενης μεθόδου, όπως προκύπτουν μετά την εφαρμογή του εργαλείου λογισμικού που την υλοποιεί στα επί μέρους προβλήματα ταυτοποίησης και πρόβλεψης. Μέσω της κριτικής σύνθεσης των συμπερασμάτων αξιολογείται η συμβολή της διατριβής στο γνωστικό διεπιστημονικό πεδίο της εξαγωγής χαρακτηριστικών από χρονοσειρές βιο- και γεω-δεδομένων. Η διατριβή κλείνει με την αναφορά σε πιθανές μελλοντικές επεκτάσεις της μεθόδου προς την κατεύθυνση της περαιτέρω βελτίωσης της αποτελεσματικότητάς της και τη διεύρυνση του κύκλου εφαρμογής της.

8.1 Περιβάλλον της έρευνας

Οι χρονοσειρές αποτελούν μια ειδική περίπτωση δεδομένων, η ανάλυση των οποίων διαδραματίζει ένα πολύ σημαντικό ρόλο σε πλήθος εφαρμογών ποικίλων επιστημονικών τομέων όπως η μηχανική, η οικονομετρία, η βιολογία, η κλινική φαρμακευτική, η μετεωρολογία, η υδραυλική, η δασολογία και η φυτική και ζωική παραγωγή στη γεωργία. Ιστορικά δεδομένα χρονοσειρών έχουν επίσης χρησι-

μοποιηθεί σε πολλές περιπτώσεις κατά την προτυποποίηση του περιβάλλοντος με σκοπό τη διευκόλυνση της περιβαλλοντικής πολιτικής και λήψης αποφάσεων. Στην τυπική της μορφή, μια χρονική ακολουθία αντιστοιχεί σε παρατηρήσεις που λαμβάνονται κατά τακτά χρονικά διαστήματα κατά τη διάρκεια ενός φαινομένου. Σε πολλές περιπτώσεις, τα δεδομένα χρονοσειρών έχουν ως άνω φράγμα τη χρονική στιγμή λήξης του φαινομένου, ενώ σε άλλες οι μετρήσεις συνεχίζονται ακατάπαυστα. Τέλος, υπάρχει και η περίπτωση κατά την οποία η μέτρηση τερματίζεται μετά την παρέλευση συγκεκριμένου χρονικού διαστήματος ώστε να ληφθεί ένα δείγμα δεδομένων που αποτελεί ένα είδος φυσικής υπογραφής του φαινομένου. Στις περισσότερες περιπτώσεις οι χρονοσειρές αντιστοιχούν σε προβλήματα ταξινόμησης ή πρόβλεψης. Η ανάλυση αυτού του είδους των δεδομένων περιλαμβάνει διαδικασίες που στοχεύουν στην προτυποποίηση του μηχανισμού παραγωγής των αρχικών πληροφοριών, είτε προς την κατεύθυνση της ταξινόμησης σε ένα πλήθος κατηγοριών, είτε για την πρόβλεψη κάποιας μελλοντικής κατάστασης του συστήματος, βάσει του υπολογισμού και της ποσοτικοποίησης ιστορικών στοιχείων του υπό παρατήρηση φαινομένου.

Μια από τις σημαντικότερες αναλυτικές συνιστώσες για τέτοιου είδους πληροφορίες αποτελεί η εξαγωγή χαρακτηριστικών, για την οποία έχουν προταθεί ποικίλοι αλγόριθμοι. Σε αυτούς συμπεριλαμβάνονται διαδικασίες δειγματοληψίας, μέσου όρου και εκθετικής εξομάλυνσης, συμβολικής αποτύπωσης και μετασχηματισμοί Fourier, ενώ, σχετικά πρόσφατα, διάφορα πρότυπα ARCH, γενικευμένα ή κανονικά, έχουν κάνει την εμφάνισή τους ως εναλλακτικές της παραδοσιακής στατιστικής ανάλυσης.

Ίσως μια από τις συχνότερα χρησιμοποιούμενες μεθόδους στην ανάλυση χρονοσειρών αποτελεί η Τμηματική Γραμμική Αναπαράσταση (PLR). Πρόκειται για μια κατηγορία προτύπων σύμφωνα με την οποία μια χρονοσειρά μήκους n αναπαρίσταται από K ευθείες, όπου το μέγεθος K είναι τυπικά πολύ μικρότερο από το n. Κάθε αλγόριθμος ο οποίος δέχεται στην είσοδο δεδομένα χρονοσειρών και επιστρέφει με οποιονδήποτε τρόπο μια τμηματική αναπαράστασή τους εμπίπτει στην κατηγορία των αλγορίθμων τμηματοποίησης χρονοσειρών, εφόσον κατά το σχηματισμό της καταλληλότερης αναπαράστασης μετέρχεται έναν αυθαίρετο αριθμό τμημάτων, ενώ το μέγιστο και μέσο σφάλμα αυτών να είναι μικρότερο ενός συγκε-

κριμένου κατωφλίου. Ως αποτέλεσμα των προϋποθέσεων αυτών συνάγεται ότι η αναπαράσταση των δεδομένων χρονοσειρών αποτελεί μια διαδικασία ιδιαίτερα δύσκολη, ένεκα των σημαντικών παρεκκλίσεων που διαπιστώνονται στην ταυτόχρονη τήρησή τους ως αναπόσπαστο τμήμα των προδιαγραφών ενός αλγορίθμου τμηματοποίησης. Παρ' όλ' αυτά, κατά τη διάρκεια της τελευταίας δεκαετίας παρατηρείται μια έκρηξη στη χρήση αλγορίθμων τμηματοποίησης για διάφορα προβλήματα στα οποία εμπλέκονται δεδομένα χρονοσειρών. Αυτό οφείλεται κυρίως στις δυσκολίες που παρουσιάζει η ανάλυση λόγω της μη-γραμμικότητας, του αυξημένου βαθμού διάστασης και του ποσοστού θορύβου στα πρωτογενή δεδομένα. Μια από τις πλέον συνήθεις μεθόδους τμηματοποίησης είναι ο αλγόριθμος του διολισθαίνοντος παραθύρου, μέσω του οποίου εξάγεται κάθε φορά ο μέσος του τμήματος ή άλλο στατιστικό περιγραφικό μέτρο. Το εύρος του παραθύρου είναι στην περίπτωση αυτή πολύ σημαντικό για την ποιότητα της τελικής αναπαράστασης και επιλέγεται αυθαίρετα ή μετά από διαδικασίες δοκιμής-λάθους. Η φύση του αλγορίθμου δεν επιτρέπει την πλήρη επισκόπηση ολόκληρης της σειράς των πρωτογενών δεδομένων. Επίσης, το εύρος του παραθύρου είναι σταθερό καθ' όλη τη διάρκεια της τμηματοποίησης, με αυξημένη πιθανότητα ενσωμάτωσης άχρηστης πληροφορίας ή κατάτμησης σε σημείο συσχετιζόμενης πληροφορίας. Τα χαρακτηριστικά αυτά έχουν συχνά ως αποτέλεσμα αναπαραστάσεις χαμηλής ποιότητας.

Στις εναλλακτικές μεθόδους περιλαμβάνονται ποικίλα εργαλεία ανεκτικά στο σφάλμα, όπως συστήματα ασαφούς λογικής και ταξινομητών υπολογιστικής νοημοσύνης υπό εκπαίδευση, ενώ μεγάλο ποσοστό έρευνας έχει αφιερωθεί στη μελέτη της αρχιτεκτονικής των ταξινομητών αυτών. Παράλληλα, αναπτύσσονται μέθοδοι προεπεξεργασίας των εκπαιδευτικών συνόλων με τη χρήση κανόνων μετα-εκμάθησης. Στο πεδίο αυτό, η διαστατικότητα του ανύσματος δεδομένων εισόδου έχει μελετηθεί επαρκώς και θεωρείται εξέχουσας σημασίας για την ανάλυση, καθώς στις περιπτώσεις κατά τις οποίες η διαστατικότητα αποδίδεται μικρότερη από την πραγματική, τότε ο ταξινομητής στερείται κρίσιμης πληροφορίας. Στην αντίθετη περίπτωση ενέχονται μεγάλοι κίνδυνοι υπερ-προσαρμογής της εκπαίδευσης στην οποία συμμετέχουν μεγάλα ποσοστά περίσσειας άχρηστης πληροφορίας που περνά στην επεξεργαστική διαδικασία του δικτύου.

8.2 Πρότυπη Εξελικτική Τμηματοποίηση

Στην παρούσα διατριβή υποστηρίζεται ο σχεδιασμός, η ανάπτυξη και η εφαρμογή ενός εναλλακτικού αλγορίθμου για την αναπαράσταση δεδομένων χρονοσειρών. Πρόκειται για μια καινοτόμο μεθοδολογία προς την κατεύθυνση αναβάθμισης της διακριτικής ικανότητας ταξινομητών υπολογιστικής νοημοσύνης, οι οποίοι εφαρμόζονται σε προβλήματα που περιλαμβάνουν χρονοσειρές. Ο προτεινόμενος αλγόριθμος, όπως αναπτύχθηκε στο πλαίσιο της διατριβής, εστιάζει στην αναπαράσταση χρονοσειρών μέσω εξελικτικής μεθόδου προ-επεξεργασίας. Τα παραγόμενα δευτερογενή εξελικτικά δεδομένα χρησιμοποιούνται στην εκπαίδευση κατάλληλα διαμορφωμένων λειτουργικών μονάδων ΤΝΔ και ΜΔΥ, οι οποίες είναι ενσωματωμένες στον πυρήνα του συστήματος με το ρόλο ταξινομητών.

Ως βασικότερος στόχος τίθεται ο καθορισμός ενός δυναμικού, αυτο-προσαρμοζόμενου σχήματος αναπαράστασης των πρωτογενών χρονοσειριακών δεδομένων και τελικά η βέλτιστη αναπαράσταση της χρονοσειράς σε χώρους μικρότερου βαθμού διάστασης . Στην υλοποίηση της μεθόδου περιλαμβάνεται μια λειτουργική μονάδα μέσω της οποίας εφαρμόζονται πολλαπλά σχήματα τμηματοποίησης στο διάνυσμα εισόδου των αρχικών δεδομένων. Καθε τέτοιο αντικείμενο αναπαράστασης μορφοποιείται από τον προτεινόμενο αλγόριθμο με εξελικτικό τρόπο και ελέγχεται όσον αφορά στην αποδοτικότητά του. Το πλέον προσαρμοσμένο δευτερογενές σύνολο εκπαίδευσης περιλαμβάνει τα ουσιωδέστερα χαρακτηριστικά της αρχικής πρωτόλειας πληροφορίας και αποτελεί ένα σχήμα εκπαίδευσης που εισάγει σημαντικές βελτιώσεις στην ικανότητα διάκρισης και πρόβλεψης καθενός από τους δύο ενσωματωμένους ταξινομητές. Παράλληλα, το σύστημα λαμβάνει αποφάσεις σχετικά με την επιλογή του καταλληλότερου ταξινομητή για κάθε περίπτωση.

8.3 Συμπεράσματα και κριτική αξιολόγηση

Σημαντικότερη συνεισφορά της διατριβής αποτελεί η ανάπτυξη μιας καινοτόμου μεθόδου αποτελεσματικής αναπαράστασης χρονοσειρών σε χώρους μικρότερης διάστασης. Συνδυάζοντας τα βασικά της ανάλυσης των χρονοσειρών με ταξινομητές υπολογιστικής νοημοσύνης και τεχνικές εξελικτικής υπολογιστικής, ο προτεινόμενος αλγόριθμος αποτελεί ουσιαστικά μια αυτοματοποιημένη εξελικτική μέθοδο

προ-επεξεργασίας των πρωτογενών δεδομένων δια μέσου ενός εργαλείου λογισμικού. Η διαδικασία αποδίδει γενετικά βελτιωμένα δευτερογενή σύνολα δεδομένων με τον αποτελεσματικό έλεγχο του θορύβου και της διαστατικότητας του αρχικού διανύσματος εισόδου. Με τον τρόπο αυτό ο αλγόριθμος αναζητά την αποκάλυψη του βέλτιστου σχήματος αναπαράστασης των πρωτογενών δεδομένων για το εκάστοτε πρόβλημα.

Η προτεινόμενη μεθοδολογία βασίζεται στην υπόθεση ότι με την πάροδο των γενεών του αλγορίθμου και την ολοένα και αυξανόμενη μέση απόδοση των εκπαιδευτών, θα βρεθεί εν τέλει ο πλέον προσαρμοσμένος, εκείνος που είναι σε θέση να προτυποποίησει αποτελεσματικότερα τα αρχικά δεδομένα, βελτιστοποιώντας την εκπαίδευση κάθε ταξινομητή. Σε γενικές γραμμές, η Πρότυπη Εξελικτική Τμηματοποίηση παρουσιάζει τα εξής χαρακτηριστικά:

- *Μηχανισμό εξόρυξης σημαντικών χαρακτηριστικών.* Ο εντοπισμός και καθορισμός των σημαντικών στοιχείων των πρωτογενών δεδομένων επιτυγχάνεται μέσω εξελικτικής μεθόδου.

- *Συνδυασμό ταξινομητών σε παράλληλη ή σε σειρά λειτουργία.* Στο σύστημα έχουν ενσωματωθεί δύο αντικείμενα ταξινομητών, ΤΝΔ και ΜΔΥ, τα οποία μπορούν να λειτουργούν είτε σε σειρά, είτε παράλληλα. Στην πρώτη περίπτωση κάθε ταξινομητής ελέγχει διαφορετικό σύνολο εκπαιδευτών, ενώ στη δεύτερη ο ίδιος εκπαιδευτής υποβάλλεται στον έλεγχο αμφότερων των ταξινομητών.

- *Μη γραμμικό μετασχηματισμό των αρχικών δεδομένων.* Η εξελικτική διαδικασία περιλαμβάνει περιγραφικά στατιστικά μέτρα τα οποία αποτυπώνονται στα πρωτογενή δεδομένα. Μέσω μιας σειράς μη γραμμικών διαδοχικών μετασχηματισμών κάθε εκπαιδευτής παράγει ένα δευτερογενές σύνολο δεδομένων που χρησιμοποιείται ως βάση της εκπαίδευσης

- *αυτο-οργανούμενη δομή*, με την έννοια της επιλογής του βέλτιστου ταξινομητή και του βέλτιστου σχήματος τμηματοποίησης για το εκάστοτε πρόβλημα

- *Αυτοματοποιημένη παραμετροποίηση.* Ο αλγόριθμος διαθέτει κατάλληλες λειτουργικές μονάδες για την επιλογή των βέλτιστων παραμέτρων των ταξινο-

μητών.

Η τελική αναπαράσταση που αποτυπώνεται από τον πιο προσαρμοσμένο εκπαιδευτή καθορίζει συγκεκριμένο αριθμό και τρόπο σχηματισμού τμημάτων, καθένα απο τα οποία αντιπροσωπεύει ένα σημαντικό χαρακτηριστικό των πρωτογενών δεδομένων. Η περίπτωση του αλγορίθμου ΠΕΤ είναι ουσιαστικά ένα εξελικτικό διολισθαίνον παράθυρο. Διαφέρει από τον αλγόριθμο αυτό κατ' αρχήν στο εύρος του παραθύρου, το οποίο στην ΠΕΤ καθορίζεται με ευέλικτο τρόπο, με αποτέλεσμα να μειώνονται οι πιθανότητες παράβλεψης σημαντικών χαρακτηριστικών. Επίσης, η φύση του αλγορίθμου ΠΕΤ επιτρέπει την πλήρη επισκόπηση της χρονοσειράς πριν την αποτύπωση του σχήματος τμηματοποίησης.

Αναμφίβολα ο εμπλουτισμός του συστήματος με δύο διαφορετικούς ταξινομητές, ΤΝΔ και ΜΔΥ δίνει τη δυνατότητα για εφαρμογή του σε ευρύτερο φάσμα περιπτώσεων. Επιπροσθέτως, για κάθε πρόβλημα που μελετάται, επιλέγεται ο καταλληλότερος ταξινομητής, σε συνδυασμό με το πλέον προσαρμοσμένο εξελικτικό σύνολο δευτερογενών δεδομένων. Με τον τρόπο αυτό συγκρίνονται ταυτόχρονα δύο εναλλακτικές λύσεις, ενώ το υπολογιστικό κόστος κρατείται σε αποδεκτά επίπεδα. Εξάλλου, η δυνατότητα λειτουργίας των ταξινομητών σε σειρά ή παράλληλα προσδίδει στο σύστημα ευελιξία όσον αφορά στη διάρκεια του χρόνου λειτουργίας, καθώς επίσης και στο εύρος του χώρου αναζήτησης. Στην πρώτη περίπτωση, όπου οι ταξινομητές λειτουργούν σε σειρά, ο χώρος αναζήτησης μειώνεται, αλλά η αξιολόγηση γίνεται εξαντλητικότερη λόγω του ότι ο ίδιος εκπαιδευτής υποβάλλεται στη δοκιμασία αμφότερων των ταξινομητών. Στη δεύτερη περίπτωση διευρύνεται ο χώρος αναζήτησης, αλλά ταυτόχρονα συρρικνώνονται οι δυνατότητες αξιολόγησης των εκπαιδευτών.

Η αποτελεσματικότητα της μεθόδου ελέγχθηκε με εφαρμογή του εργαλείου λογισμικού σε δύο διαφορετικά προβλήματα, τα οποία περιλαμβάνουν βιο- και γεωδεδομένα. Συγκεκριμένα, μετά την ολοκλήρωση των φάσεων σχεδιασμού, ανάπτυξης, παραμετροποίησης και ελέγχου, η προτεινόμενη μέθοδος ΠΕΤ εφαρμόσθηκε στα προβλήματα ταυτοποίησης των ιών των φυτών CGMMV και TRV, καθώς επίσης και στην πρόβλεψη της χειμαρρικής επικινδυνότητας σε συγκεκριμένες περιοχές της Κύπρου. Οι δύο περιπτώσεις, οι οποίες αντιστοιχούν σε ταξινόμηση αντικειμένων και πρόβλεψη της πορείας ενός φαινομένου σε βάθος χρόνου, χαρακτη-

ρίζονται από πρωτογενείς χρονοσειρές υψηλού βαθμού διάστασης και θορύβου, ο έλεγχος των οποίων τεθηκε ως πρώτιστος στόχος. Σε γενικές γραμμές ο αλγόριθμος απέδωσε ενθαρρυντικά αποτελέσματα και φαίνεται να δημιουργεί σημαντική εξομάλυνση και αποτελεσματική αναπαράσταση των χρονοσειρών σε αμφότερες τις περιπτώσεις, ενώ παράλληλα αναβαθμίστηκε η απόδοση και των δυο ταξινομητών συγκρινόμενη με την απόδοσή τους υπό καθεστώς εκπαίδευσης με τα αρχικά δεδομένα. Στα βασικά πλεονεκτήματα της προτεινόμενης προσέγγισης συγκαταλέγονται:

- *Δυνατότητες αποτελεσματικής αναπαράστασης σε ευρείες χρονοσειρές.* Το προτεινόμενο σύστημα δείχνει να ανταποκρίνεται αρκετά ικανοποιητικά έναντι του υψηλού βαθμού διάστασης και θορύβου που ενέχονται σε συγκεκριμένα δεδομένα. Πράγματι, παρατηρείται αποτελεσματικότερη αναπαράσταση σε ιδιαίτερα ευρείες χρονοσειρές, στις περιπτώσεις που αυτές αποτελούν ουσιαστικά μια χαρακτηριστική υπογραφή συγκεκριμένης τάξης. Τέτοια είναι η περίπτωση δεδομένων που προκύπτουν από ερευνητικές εργασίες είτε προς την κατεύθυνση ταυτοποίησης φυτικών ιών – όπως στην περίπτωση της παρούσας μελέτης – είτε ιών ζωικού κυττάρου, είτε στην αναγνώριση και ανίχνευση υπολειμμάτων φυτοφαρμάκων, τα αρχικά δεδομένα των οποίων συνήθως προέρχονται από πειράματα με αισθητήρες. Συνεπώς, το εργαλείο λογισμικού που αναπτύχθηκε στο πλαίσιο της παρούσας εργασίας είναι δυνατόν να έχει εφαρμογή σε γεωργική, βιολογική ή φαρμακευτική και κλινική έρευνα, οι οποίες συνήθως σχετίζονται με προβλήματα χρονοσειρών μεγάλου εύρους και διάστασης. Πολυέξοδες επαναλήψεις του ίδιου πειράματος επίσης είναι δυνατόν να περιοριστούν σημαντικά, εξαιτίας του μεγάλου διαθέσιμου όγκου πιθανών αναπαραστάσεων της αρχικής χρονοσειράς που προκύπτουν από την ΠΕΤ.

- *Βελτιστοποιημένη και αυτοματοποιημένη παραμετροποίηση.* Η επιτυχής εφαρμογή της προτεινόμενης μεθοδολογίας απαιτεί προσεκτική και πολυέξοδη σε χρόνο και υπολογιστική ισχύ ρύθμιση όσον αφορά στις παραμέτρους τόσο του γενετικού αλγορίθμου, όσο και των δύο ταξινομητών. Σε συνάρτηση με τη φύση της αρχικής χρονοσειράς, η ανάλυση πρέπει να καθορίζει τη βέλτιστη τιμή για τον αριθμό των νευρώνων στα ενδιάμεσα επίπεδα του ΤΝΔ, καθώς

επίσης και για τις παραμέτρους C και γ για την ΜΔΥ. Συνεπώς, κρίθηκε απαραίτητη η ανάπτυξη και ενσωμάτωση στον κώδικα υπο-ρουτίνας αναζήτησης πλέγματος για βέλτιστη παραμετροποίηση.

- *Σχεδιασμός συστήματος βαρέων καθηκόντων.* Λαμβάνοντας υπόψη το γεγονός ότι ο αλγόριθμος σχεδιάσθηκε έτσι ώστε να είναι σε θέση να αντιμετωπίσει ιδιαίτερα μεγάλο πλήθος πληροφοριών προερχόμενο από σημαντικές ερευνητικές βάσεις δεδομένων, σε συνδυασμό με τις ρουτίνες βέλτιστης παραμετροποίησης και εκπαίδευσης των ταξινομητών, είναι φυσικό να αναμένονται μεγαλύτεροι χρόνοι περάτωσης. Από τη στιγμή όμως που οριστικοποιείται το τελικό σχήμα τμηματοποίησης, οι υπολογιστικές απαιτήσεις μειώνονται καθώς η ΠΕΤ, εκτός της αποτελεσματικής εξομάλυνσης των αρχικών δεδομένων, επίσης επιτυγχάνει να μειώσει δραστικά το συνολικό αριθμό των στοιχείων του διανύσματος εισόδου, χωρίς να χάνει σημαντική πληροφορία σε καμιά από τις εξεταζόμενες περιπτώσεις. Η συμπεριφορά αυτή παίζει σημαντικό ρόλο στην οικονομία της εκπαίδευσης καθώς, από τη στιγμή που καθορίζεται ο πλέον προσαρμοσμένος εκπαιδευτής, το σύστημα μετατρέπει κάθε νέο άγνωστο δείγμα προς επεξεργασία σύμφωνα το σχήμα τμηματοποίησης που ορίζεται από αυτόν.

8.4 Μελλοντική επέκταση

Ο αλγόριθμος που παρουσιάστηκε στην παρούσα διατριβή είναι ανοιχτής αρχιτεκτονικής, επιτρέποντας την προσαρμογή του σε νέους ταξινομητές που πιθανόν προκύψουν στο μέλλον ή αλγορίθμων περαίωσης της εξελικτικής διαδικασίας. Σε γενικές γραμμές, η μελλοντική επέκταση του συστήματος περιλαμβάνει συνιστώσες προς την κατεύθυνση της βελτίωσης του συνολικού δυναμικού αριστοποίησης του συστήματος, της ενσωμάτωσης διαδικασιών για την ελάφρυνση των υπολογιστικών απαιτήσεων, καθώς επίσης και της αναδιάταξης (refactoring) συγκεκριμένων λειτουργικών μονάδων του κώδικα, ώστε να καταστεί το σύστημα φιλικότερο στη χρήση.

- Ενα βασικό πρόβλημα που παρατηρείται κατά τη χρήση του συστήματος, είναι η περίπτωση να μη δύναται ο αλγόριθμος να ανακαλύψει το βέλτιστο

σχήμα τμηματοποίησης, όταν ως συνθήκη περαίωσης επιλεγεί η υπέρβαση υψηλού κατωφλίου ακρίβειας εκ μέρους του ταξινομητή. Στις περιπτώσεις αυτές υπάρχει ο κίνδυνος ο αλγόριθμος να υποπέσει σε αέναο βρόγχο αναζήτησης, εφόσον καταπίπτει η συνθήκη περαίωσης. Σημαντική αναβάθμιση στο προτεινόμενο σύστημα θα ήταν η εισαγωγή διερευνητικής διαδικασίας των πρωτογενών δεδομένων και χρήση τεχνικών γενετικού προγραμματισμού για τον καθορισμό της συνθήκης περαίωσης.

- Η εμπειρία που αποκομίσθηκε μετά την εφαρμογή της προτεινόμενης μεθόδου στα δύο προβλήματα είναι ότι ο χρόνος που απαιτείται για την ολοκλήρωση της εξελικτικής διαδικασίας είναι σχετικά μεγάλος. Είναι χαρακτηριστικό ότι για το πρόβλημα των ιών των φυτών απαιτήθηκαν περίπου οκτώ ημέρες για το νευρωνικό ταξινομητή και περίπου τρεις για τη ΜΔΥ, έως ότου κλείσει ένας κύκλος εκπαίδευσης. Η σχετική συμπεριφορά μπορεί έως ένα σημείο να εξηγηθεί από τη φύση του μεταφραστή τον οποίο χρησιμοποιεί η γλώσσα προγραμματισμού. Εξάλλου, δεν πρέπει να παραβλεφθεί ο ιδιαίτερα μεγάλος όγκος των παραγόμενων εξελικτικών αναπαραστάσεων της αρχικής χρονοσειράς, σε συνδυασμό με τις υπο-ρουτίνες παραμετροποίησης του συστήματος, παράγοντες που ασκούν ιδιαίτερες πιέσεις στο υπολογιστικό κόστος. Πράγματι, κατά τη διάρκεια της εξελικτικής διαδικασίας, συνολικά παρήχθησαν 15.000 δευτερογενή σύνολα δεδομένων για κάθε πρόβλημα, για καθένα από τα οποία η παραμετροποίηση της ΜΔΥ και η ρύθμιση της αρχιτεκτονικής του ΤΝΔ προηγούντο κάθε φορά της εκπαιδευτικής διαδικασίας αυτής καθ' εαυτής. Η μετάβαση σε άλλο εργαλείο λογισμικού (C, C++) για να παρακαμφθούν οι περιορισμοί που τίθενται από τη γλώσσα προγραμματισμού που χρησιμοποιήθηκε κατά την ανάπτυξη της προτεινόμενης μεθοδολογίας, πιθανόν να αντιμετώπιζε έως ένα βαθμό το πρόβλημα.

- Μια πολύ ενδιαφέρουσα προσέγγιση επίσης θα ήταν η αναζήτηση άλλων τύπων απόδοσης της γενετικής αναπαράστασης των εκπαιδευτών. Τη θέση της δυαδικής αναπαράστασης κάθε γονιδίου είναι δυνατό να λάβουν αριθμητικές αναπαραστάσεις κινητής υποδιαστολής ή συμβολοσειρών ή ακόμη και εντελώς νέων τύπων. Μια τέτοια προσέγγιση είναι πιθανό να έχει θετικά αποτελέσματα, εκτός από την τελική ποιότητα αναπαράστασης της χρονοσειράς,

επίσης και στις υπολογιστικές απαιτήσεις του συστήματος.

- Ο βασικός πυρήνας του αλγορίθμου αποτελείται από δύο ταξινομητές. Το πεδίο εφαρμογής όμως του συστήματος είναι πιθανό να διευρυνθεί με τη χρήση επιπλέον στοιχειωδών ταξινομητών, όπως των εγγύτερων γειτόνων, του δένδρου αποφάσεων ή του ταξινομητή Bayes.

- Διεύρυνση του πεδίου εφαρμογής του συστήματος επίσης είναι πιθανό να επιτευχθεί με τη χρήση περισσότερων στατιστικών μέτρων στο τμήμα πυρήνα του χρωμοσώματος των εκπαιδευτών. Πολύ ενδιαφέρουσα εξέλιξη προς την κατεύθυνση αυτή θα ήταν η ενσωμάτωση υπορουτίνας εύρεσης Αντιληπτικά Σημαντικών Σημείων (PIP) ή του αλγόριθμου Ramer/Douglas-Peucker σε τμήματα το εύρος των οποίων ξεπερνά μια συγκεκριμένη τιμή κατωφλίου.

- Ενδιαφέρουσα επίσης κατεύθυνση για μελλοντική έρευνα θα ήταν η ανάπτυξη παρόμοιου συστήματος πελάτη-εξυπηρετητή ή εφαρμογής (δια)δικτύου, σε αποκλειστικό εξυπηρετητή. Πηγές άντλησης δεδομένων είναι δυνατό να αποτελέσουν η βάση δεδομένων του Ολοκληρωμένου Συστήματος Διαχείρισης και Ελέγχου (Ο.Σ.Δ.Ε.) και το σύνολο των θεματικών χαρτών του Υπουργείου Αγροτικής Ανάπτυξης και Τροφίμων. Οπωσδήποτε σημαντικούς ανασταλτικούς παράγοντες προς την κατεύθυνση αυτή αποτελούν οι υπολογιστικές απαιτήσεις του αλγορίθμου, οι οποίες θα αυξηθούν ακόμη περισσότερο με την ταυτόχρονη παράλληλη επεξεργασία πολλών προβλημάτων από πολλούς χρήστες.

- Στο άμεσο μέλλον το σύστημα θα εφαρμοσθεί στο πρόβλημα της ανίχνευσης υπολειμμάτων φυτοφαρμάκων σε τρόφιμα. Τα δεδομένα θα προέλθουν από βιο-αισθητήρες με μεγάλο εύρος εφαρμογής σε φυτοπροστατευτικές ουσίες. Κρίνεται συνεπώς απαραίτητο να τροποποιηθεί όσον αφορά στην έξοδο με δυνατότητες ταξινόμησης πέραν των δύο και μόνο κατηγοριών.

Βιβλιογραφία

[1] A. Abraham. Meta learning evolutionary artificial neural networks. *Neurocomputing*, 56(03):1--38, 2004.

[2] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. *Foundations of Data Organization and Algorithms*, 8958546:69--84, 1993.

[3] R. Agrawal, K-I. Lin, S. S. Harpreet, and K. Shim. Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. In *Data Base*, pages 490--501. Morgan Kaufmann Publishers Inc., 1995.

[4] R. Agrawal and J. K. Singh. Application of a Genetic Algorithm in the Development and Optimisation of a Non-linear Dynamic Runoff Model. *Biosystems Engineering*, 86(1):87--95, 2003.

[5] M. Aladjem. Recursive training of neural networks for classification. *IEEE Transactions on Neural Networks*, 11(2):496--503, 2000.

[6] E. Alpaydin. Techniques for Combining Multiple Learners. In E Alpaydin, editor, *Proceedings of Engineering of Intelligent Systems*, volume 2, pages 6--12. ICSC Press, 1998.

[7] F. Anctil, N. Lauzon, V. Andreassian, L. Oudin, and C. Perrin. Improvement of rainfall-runoff forecasts through mean areal rainfall optimization. *Journal of Hydrology*, 328(3-4):717--725, 2006.

[8] B. Andrews. Rank-Based Estimation for Autoregressive Moving Average Time Series Models. *Journal of Time Series Analysis*, 29(1):51--73, 2008.

[9] J. Arroyo and C. Maté. Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting*, 25(1):192--207, 2009.

[10] K. J. Astrom. On the choice of sampling rates in parametric identification of time series. *Information Sciences*, 1(3):273--278, 1969.

[11] L. Auret and C. Aldrich. Change point detection in time series data with random forests. *Control Engineering Practice*, 18(8):990--1002, 2010.

[12] S. Avramidis and L. Iliadis. Wood-water sorption isotherm prediction with artificial neural networks: A preliminary study. *Holzforschung*, 59(3):336--341, 2005.

[13] T. G. Bali, K. O. Demirtas, and H. Levy. Is There an Intertemporal Relation between Downside Risk and Expected Returns? *Journal of Financial and Quantitative Analysis*, 44(04):883--909, 2009.

[14] T. G. Bali and R. F. Engle. Investigating ICAPM with Dynamic Conditional Correlations. *New York*, (646), 2008.

[15] P. Banerjee and A. K. Bhunia. Mammalian cell-based biosensors for pathogens and toxins. *Trends in biotechnology*, 27(3):179--188, 2009.

[16] Z. Bankó, L. Dobos, and J. Abonyi. Dynamic Principal Component Analysis in Multivariate Time-Series Segmentation. *Evolution*, 1(1):11--24, 2011.

[17] M. Bertinelli, A. Catelli, C. Combi, and F. Pinciroli. Data compression applied to dynamic electrocardiography. *Medical & Biological Engineering & Computing*, 27(1):33--40, 1989.

[18] E. Boczko, W. Kalies, and K. Mischaikow. Polygonal approximation of flows. *Topology and its Applications*, 154(13):2501--2520, 2007.

[19] L. Bodri. Prediction of extreme precipitation using a neural network: application to summer flood occurrence in Moravia. *Advances in Engineering Software*, 31(5):311--321, 2000.

[20] T. Bollerslev and E. Ghysels. Periodic Autoregressive Conditional Heteroscedasticity. *Journal of Business & Economic Statistics*, 14(2):139--151, 1996.

[21] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory COLT 92*, 6(8):144--152, 1992.

[22] D. Bouchard. Automated Time Series Segmentation for Human Motion Analysis. *Time*, 2006.

[23] G. E. P. Box and G. M. Jenkins. Time series analysis forecasting and control (revised edition). *Holden Day Oakland*, 1976.

[24] X. Cai, D. C. McKinney, and L. S. Lasdon. Solving nonlinear water management models using a combined genetic algorithm and linear programming approach. *Advances in Water Resources*, 24(6):667--676, 2001.

[25] California Scientific Software CORPORATE. *BrainMaker application guide (3rd ed.)*. California Scientific Software, 1994.

[26] B. Cannas, A. Fanni, L. See, and G. Sias. Data preprocessing for river flow forecasting using neural networks: Wavelet transforms and data partitioning. *Physics and Chemistry of the Earth Parts ABC*, 31(18):1164--1171, 2006.

[27] H. Cao, F. R.ecknagel, G-J. Joo, and D-K. Kim. Discovery of predictive rule sets for chlorophyll-a dynamics in the Nakdong River (Korea) by means of the hybrid evolutionary algorithm HEA. *Ecological Informatics*, 1(1):43--53, 2006.

[28] S-G. Cao, Y-B. Liu, and Y-P. Wang. A forecasting and forewarning model for methane hazard in working face of coal mine based on LS-SVM. *Journal of China University of Mining and Technology*, 18(2):172--176, 2008.

[29] A. Carmona-Poyato, F. J. Madrid-Cuevas, R. Medina-Carnicer, and R. Muñoz Salinas. Polygonal approximation of digital planar curves through break point suppression. *Pattern Recognition*, 43(1):14--25, 2010.

[30] M. Ceccarelli and A. Maratea. Virtual genetic coding and time series analysis for alternative splicing prediction in C. elegans. *Artificial Intelligence in Medicine*, 45(2-3):109--115, 2009.

[31] F. K. P. Chan, A. W. C. Fu, and C. Yu. Haar wavelets for efficient similarity search of time-series: with and without time warping, 2003.

[32] C-C. Chang and C-J. Lin. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1--39, 2011.

[33] C. L. Chang, S. L. Lo, and S. L. Yu. Applying fuzzy theory and genetic algorithm to interpolate precipitation. *Journal of Hydrology*, 314(1-4):92--104, 2005.

[34] P-C. Chang, C-Y. Fan, and C-H. Liu. Integrating a Piecewise Linear Representation Method and a Neural Network Model for Stock Trading Points Prediction, 2009.

[35] K. W. Chau. A split-step particle swarm optimization algorithm in river stage forecasting. *Journal of Hydrology*, 346(3-4):131--135, 2007.

[36] Y. Chen and C. Lin. Dynamic parameter optimization of evolutionary computation for on-line prediction of time series with changing dynamics. *Applied Soft Computing*, 7(4):1170--1176, 2007.

[37] C. T. Cheng, C. P. Ou, and K. W. Chau. Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall–runoff model calibration. *Journal of Hydrology*, 268(1-4):72--86, 2002.

[38] C-T. Cheng, M-Y. Zhao, K. W. Chau, and X-Y. Wu. Using genetic algorithm and TOPSIS for Xinanjiang model calibration with a single procedure. *Journal of Hydrology*, 316(1-4):129--140, 2006.

[39] H-H. Cho, S-H. Kim, T-K. Cho, and M-R. Choi. Efficient image enhancement technique by decimation method, 2005.

[40] C-S. J. Chu. Time Series Segmentation: A Sliding Window Approach. *Information Sciences*, 85:147--173, 1995.

[41] H-H. Chu, Y-C. Shiau, and T-S. Kuo. The comparison of SSD algorithm with other ECG sampling algorithms. *Biomedical Engineering - Applications, Basis & Communications*, 18(3):124--127, 2006.

[42] F. L. Chung, T. C. Fu, and R. Luk. Flexible time series pattern matching based on perceptually important points. *Joint Conference on Artificial Intelligence Workshop*, pages 1--7, 2001.

[43] F-L. Chung, T-C. Fu, V. Ng, and R. W. P. Luk. An evolutionary approach to pattern-based time series segmentation, 2004.

[44] J. Coelho, P. Demouraoliveira, and J. Cunha. Greenhouse air temperature predictive control using the particle swarm optimisation algorithm. *Computers and Electronics in Agriculture*, 49(3):330--344, 2005.

[45] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273--297, 1995.

[46] P. Coulibaly and N. Evora. Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology*, 341(1-2):27--41, 2007.

[47] C. Damle and A. Yalcin. Flood prediction using Time Series Data Mining. *Journal of Hydrology*, 333(2-4):305--316, 2007.

[48] G. Das, K-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. *Knowledge Discovery and Data Mining*, pages 16--22, 1998.

[49] M. Datar. The sliding-window Computation model and results. *Minos*, pages 149--167, 2007.

[50] N. Davey, S. Hunt, and R. Frank. Time series prediction and neural networks. *Journal of Intelligent and Robotic Systems*, 31(1):91--103, 2001.

[51] E. D. Dawson, C. L. Moore, J. A. Smagala, D. M. Dankbar, M. Mehlmann, M. B. Townsend, C. B. Smith, N. J. Cox, R. D. Kuchta, and K. L. Rowlen. MChip: a tool for influenza surveillance. *Analytical Chemistry*, 78(22):7610--7615, 2006.

[52] L. A. Díaz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, and J. A. Moncada-Herrera. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment*, 42(35):8331--8340, 2008.

[53] D. H. Douglas and T. K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica The International Journal for Geographic Information and Geovisualization*, 10(2):112--122, 1973.

[54] H. Du and N. Zhang. Time series prediction using evolving radial basis function networks with new encoding scheme. *Neurocomputing*, 71(7-9):1388--1400, 2008.

[55] D. A. Elizondo, R. Birkenhead, M. Góngora, E. Taillard, and P. Luyima. Analysis and test of efficient methods for building recursive deterministic perceptron neural networks. *Neural Networks*, 20(10):1095--1108, 2007.

[56] M. H. Fazel Zarandi, B. Rezaee, I. B. Turksen, and E. Neshat. A type-2 fuzzy rule-based expert system model for stock price analysis. *Expert Systems with Applications*, 36(1):139--154, 2009.

[57] F. J. Fernandez, A. Seco, J. Ferrer, and M. A. Rodrigo. Use of neurofuzzy networks to improve wastewater flow-rate forecasting. *Environmental Modelling and Software*, 24(6):686--693, 2009.

[58] A. J. P. Filho and C. C. dos Santos. Modeling a densely urbanized watershed with an artificial neural network, weather radar and telemetric data. *Journal of Hydrology*, 317(1-2):31--48, 2006.

[59] E. Fink. Compression of time series by extracting major extrema. *Journal of Experimental Theoretical Artificial*, 23(2):1--22, 2011.

[60] S. Forrest and M. Mitchell. Relative Building-Block Fitness and the Building Block Hypothesis. In L Darrell Whitley, editor, *Foundations of Genetic Algorithms 2*, pages 109--126. Morgan Kaufmann, 1993.

[61] T-C. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164--181, 2011.

[62] T-C. Fu, F-L. Chung, R. Luk, and C-M. Ng. Preventing Meaningless Stock Time Series Pattern Discovery by Changing Perceptually Important Point Detection. *Learning*, pages 1171 -- 1174, 2005.

[63] S. Gaur and M. Deo. Real-time wave forecasting using genetic programming. *Ocean Engineering*, 35(11-12):1166--1172, 2008.

[64] F. Gilfeather, V. Hamine, P. Helman, J. Hutt, T. Loring, C. R. Lyons, and R. Veroff. Learning and modeling biosignatures from tissue images. *Computers in Biology and Medicine*, 37(11):1539--1552, 2007.

[65] T. J. Glezakos, G. Moschopoulou, T. A. Tsiligiridis, S. Kintzios, and C. Yialouris. Plant virus identification based on neural networks with evolutionary preprocessing. *Computers and Electronics in Agriculture*, 70(2):263--275, 2010.

[66] T. J. Glezakos, T. A. Tsiligiridis, L. S. Iliadis, C. Yialouris, F. P. Maris, and K. P. Ferentinos. Feature extraction for time-series data: An artificial neural network evolutionary training model for the management of mountainous watersheds. *Neurocomputing*, 73(1-3):49--59, 2009.

[67] L. Gonzalez-Abril, F. Velasco, J. A. Ortega, and F. J. Cuberos. A new approach to qualitative learning in time series. *Expert Systems with Applications*, 36(6):9924--9927, 2009.

[68] W. Guanghuo, A. Dobermann, C. Witt, S. Quingzhu, and F. Rongxing. Performance of site-specific nutrient management for irrigated rice in southeast China. *Agronomy journal*, 93(4):869--878, 2001.

[69] J. Guerrero, A. Berlanga, J. García, and J. Molina. Piecewise Linear Representation Segmentation as a Multiobjective Optimization Problem. *Distributed Computing and Artificial Intelligence*, pages 267--274, 2010.

[70] C. Guo, H. Li, and D. Pan. An Improved Piecewise Aggregate Approximation Based on Statistical Features for Time Series Mining. *Knowledge Science Engineering and Management*, pages 234--244, 2010.

[71] C. Hamzacebi. Improving artificial neural networks' performance in seasonal time series forecasting. *Information Sciences*, 178(23):4550--4559, 2008.

[72] J. V. Hansen, J. B. McDonald, and R. D. Nelson. Time Series Prediction With Genetic-Algorithm Designed Neural Networks: An Empirical Comparison With Modern Statistical Models. *Computational Intelligence*, 15(3):171--184, 1999.

[73] C. Harpham and C. W. Dawson. The effect of different basis functions on a radial basis function network for time series prediction: A comparative study. *Neurocomputing*, 69(16-18):2161--2170, October 2006.

[74] G. H. Haydon, R. Jalan, M. Ala-Korpela, Y. Hiltunen, J. Hanley, L. M. Jarvis, C. A. Ludlum, and P. C. Hayes. Prediction of cirrhosis in patients with chronic hepatitis C infection by artificial neural network analysis of virus and clinical factors. *Journal of Viral Hepatitis*, 5(1):9--17, 2002.

[75] S. Haykin. *Neural Networks and Learning Machines*. Number v. 10 in Neural networks and learning machines. Prentice Hall, 2008.

[76] J. H. Holland. *Adaptation in Natural and Artificial Systems*, volume Ann Arbor. University of Michigan Press, 1975.

[77] W. Hong. Electric load forecasting by support vector model. *Applied Mathematical Modelling*, 33(5):2444--2454, 2009.

[78] J-H. Horng. Improving fitting quality of polygonal approximation by using the dynamic programming technique. *Pattern Recognition Letters*, 23(14):1657--1673, 2002.

[79] C-W. Hsu, C-C. Chang, and C-J. Lin. A Practical Guide to Support Vector Classification. *Bioinformatics*, 1(1):1--16, 2010.

[80] G-B. Huang. Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Transactions on Neural Networks*, 14(2):274--281, 2003.

[81] S-C. Huang and T-K. Wu. Integrating GA-based time-scale feature extractions with SVMs for stock index forecasting. *Expert Systems with Applications*, 35(4):2080--2088, 2008.

[82] J. Hung. A genetic algorithm approach to the spectral estimation of time series with noise and missed observations. *Information Sciences*, 178(24):4632--4643, 2008.

[83] C. Igel and M. Hüsken. Improving the Rprop learning algorithm. *Symposium A Quarterly Journal In Modern Foreign Literatures*, pages 115--121, 2000.

[84] L. S. Iliadis and F. Maris. An Artificial Neural Network model for mountainous water-resources management: The case of Cyprus mountainous watersheds. *Environmental Modelling Software*, 22(7):1066--1072, 2007.

[85] A. Jain and A. M. Kumar. Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing*, 7(2):585--592, 2007.

[86] S. M. Jalaleddine, C. G. Hutchens, R. D. Strattan, and W. A. Coberly. ECG data compression techniques--a unified approach. *IEEE Transactions on Biomedical Engineering*, 37(4):329--343, 1990.

[87] E. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, 3(3):263--286, 2001.

[88] E. Keogh, S. Chu, D. Hart, and M. J. Pazzani. An Online Algorithm for Segmenting Time Series. In Nick Cercone, Tsau Young Lin, and Xindong Wu, editors, *Proceedings IEEE International Conference on Data Mining*, pages 289--296. Citeseer, IEEE Computer Society, 2001.

[89] E. Keogh, S. Chu, D. Hart, and M. J. Pazzani. Segmenting time series: A survey and novel approach. *Work*, 57:1--21, 2003.

[90] E. Keogh and M. J. Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *Proceedings of the 4 th International Conference of*, 1998.

[91] E. Keogh and M. J. Pazzani. An indexing scheme for fast similarity search in large time series databases. *In Proc of 11th Int Conf on SSDBMs*, pages 56--67, 1999.

[92] E. Keogh and P. Smyth. A probabilistic approach to fast pattern matching in time series databases. *Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining*, M(1994):52--57, 1997.

[93] R. Kerachian and M. Karamouz. A stochastic conflict resolution model for water quality management in reservoir–river systems. *Advances in Water Resources*, 30(4):866--882, 2007.

[94] T. Kerh and C. S. Lee. Neural networks forecasting of flood discharge at an unmeasured station using river upstream information. *Advances in Engineering Software*, 37(8):533--543, 2006.

[95] K. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2):307--319, 2003.

[96] A. Kimura, K. Kashino, T. Kurozumi, and H. Murase. A Quick Search Method for Audio Signals Based on a Piecewise Linear Representation of Feature Trajectories. *Ieee Transactions On Audio Speech And Language Processing*, 16(2):20, 2007.

[97] S. Kintzios, F. Bem, O. Mangana, K. Nomikou, P. Markoulatos, N. Alexandropoulos, C. Fasseas, V. Arakelyan, A. L. Petrou, K. Soukouli, G. Moschopoulou, C. P. Yialouris, and A. Simonian. Study on the mechanism of Bioelectric Recognition Assay: evidence for immobilized cell membrane interactions with viral fragments. *Biosensors and Bioelectronics*, 20(4):907--916, 2004.

[98] S. Kintzios, E. Pistola, J. Konstas, F. Bem, T. Matakiadis, N. Alexandropoulos, I. Biselis, and R. Levin. The application of the bioelectric recognition assay for the detection of human and plant viruses: definition of operational parameters. *Biosensors and Bioelectronics*, 16(7-8):467--480, 2001.

[99] S. Kintzios, E. Pistola, P. Panagiotopoulos, M. Bomsel, N. Alexandropoulos, F. Bem, G. Ekonomou, J. Biselis, and R. Levin. Bioelectric recognition assay (BERA). *Biosensors and Bioelectronics*, 16(4-5):325--336, 2001.

[100] A. Kolesnikov and P. Fränti. Polygonal approximation of closed discrete curves. *Pattern Recognition*, 40(4):1282--1293, 2007.

[101] V. Kumar, S. C. Saxena, V. K. Giri, and D. Singh. Improved modified AZTEC technique for ECG data compression: Effect of length of parabolic filter on reconstructed signal. *Computers & Electrical Engineering*, 31(4-5):334--344, June 2005.

[102] L. Kuncheva, J. C. Bezdek, and M. A. Sutton. On combining multiple classifiers by fuzzy templates. *1998 Conference of the North American Fuzzy Information Processing Society NAFIPS Cat No98TH8353*, pages 193--197, 1998.

[103] V. Kurkova. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5(3):501--506, 1992.

[104] R. Kurzweil. The Age of Intelligent Machines "Chronology", 1996.

[105] R. K. Lai, C-Y. Fan, W-H. Huang, and P-C. Chang. Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, 36(2):3761--3773, 2009.

[106] N. A. Laskaris, S. P. Zafeiriou, and L. Garefa. Use of random time-intervals (RTIs) generation for biometric verification. *Pattern Recognition*, 42(11):2787 -- 2796, 2009.

[107] C-M. Lee and C-N. Ko. Time series prediction using RBF neural networks with a nonlinear time-varying evolution PSO algorithm. *Neurocomputing*, 73(1-3):449--460, 2009.

[108] J-H. Lee, M. Marzelli, F. A. Jolesz, and S-S. Yoo. Automated classification of fMRI data employing trial-based imagery tasks. *Medical Image Analysis*, 13(3):392--404, 2009.

[109] W. Lee and H-Y. Kim. Genetic algorithm implementation in Python. *Fourth Annual ACIS International Conference on Computer and Information Science ICIS05*, pages 8--11, 2005.

[110] M. Lei and G. Meng. Symplectic Principal Component Analysis: A New Method for Time Series Analysis. *Mathematical Problems in Engineering*, 2011:1--14, 2011.

[111] D. Lemire. A Better Alternative to Piecewise Linear Time Series Segmentation. *Arxiv preprint cs0605103*, 2007:545--550, 2006.

[112] S. Lhermitte, J. Verbesselt, W. W. Verstraeten, and P. Coppin. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment*, 115(12):3129--3152, 2011.

[113] C-S. Li, P. S. Yu, and V. Castelli. MALM: a framework for mining sequence database at multiple abstraction levels. *CIKM*, 1998.

[114] F. Liu, G. Ng, and C. Quek. RLDDE: A novel reinforcement learning-based dimension and delay estimator for neural networks in time series prediction. *Neurocomputing*, 70(7-9):1331--1341, 2007.

[115] X. Liu, Z. Lin, and H. Wang. Novel Online Methods for Time Series Segmentation, 2008.

[116] B. Lkhagva, Y. Suzuki, and K. Kawagoe. Extended SAX : Extension of Symbolic Aggregate Approximation for Financial Time Series Data Representation. *Analysis*, 7, 2006.

[117] H. Locteau, R. Raveaux, S. Adam, Y. Lecourtier, P. Heroux, and E. Trupin. Polygonal Approximation of Digital Curves Using a Multi-objective Genetic Algorithm. In Liu Wenyin and Josep Llados, editors, *Graphics Recognition Ten Years Review and Future Perspectives*, volume 3926 of *Lecture Notes in Computer Science*, pages 300--311. Springer Berlin Heidelberg, 2006.

[118] W-Z. Lu and W-J. Wang. Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. *Chemosphere*, 59(5):693--701, 2005.

[119] P. Man and M-H. Wong. Efficient and robust feature extraction and pattern matching of time series by a lattice structure. *Proceedings of the tenth international conference on Information and knowledge management CIKM01*, pages 271--278, 2001.

[120] M. Marwan, M. Fuad, and P. F. Marteau. Enhancing the Symbolic Aggregate Approximation Method Using Updated Lookup Tables. *Time*, pages 420--431, 2010.

[121] J. Mccarthy. Programs with common sense. *Science*, pages 1--15, 1959.

[122] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4):115--133, 1943.

[123] V. Megalooikonomou, Q. Wang, G. Li, and C. Faloutsos. A Multiresolution Symbolic Representation of Time Series. *21st International Conference on Data Engineering ICDE05*, (Icde):668--679, 2005.

[124] M. Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8--30, 1961.

[125] M. Minsky. A framework for representing knowledge. In P Winston, editor, *The Psychology of Computer Vision*, The Psychology of Computer Vision, pages 211--277. McGraw-Hill, 1975.

[126] M. Mitchell and S. Forrest. Genetic algorithms and artificial life. *Artificial Life*, 1(3):267--289, 1994.

[127] Y. Molkov, D. Mukhin, E. Loskutov, A. Feigin, and G. Fidelin. Using the minimum description length principle for global reconstruction of dynamic systems from noisy time series. *Physical Review E*, 80(4):046207--1--6, 2009.

[128] R. Monetti, W. Bunk, and F. Jamitzky. Characterizing Synchronization in Time Series using Information Measures Extracted from Symbolic Representations. *Physical Review E*, 79(4):9, 2008.

[129] M. Monfared, H. Rastegar, and H. Kojabadi. A new strategy for wind speed forecasting using artificial intelligent methods. *Renewable Energy*, 34(3):845--848, 2009.

[130] L. Mora-Lopez, J. Mora, R. Morales-Bueno, and M. Sidrach-de Cardona. Modeling time series of climatic parameters with probabilistic finite automata. *Environmental Modelling Software*, 20(6):753--760, 2005.

[131] Z. Moradi and B. Jafarpour. First report of coat protein sequence of Cucumber Green Mottle Mosaic Virus in cucumber isolated from Khorasan in Iran. *International Journal of Virology*, 7(1):1--12, 2011.

[132] F. Mörchen and A. Ultsch. Efficient mining of understandable patterns from multivariate interval time series. *Data Mining and Knowledge Discovery*, 15(2):181--215, 2007.

[133] T. Morimoto, J. De Baerdemaeker, and Y. Hashimoto. An intelligent approach for optimal control of fruit-storage process using neural networks and genetic algorithms. *Computers and Electronics in Agriculture*, 18:205--224, 1997.

[134] J. R. Ni and A. Xue. Application of artificial neural network to the rapid feedback of potential ecological risk in flood diversion zone. *Engineering Applications of Artificial Intelligence*, 16(2):105--119, 2003.

[135] H. Nielsen, S. Brunak, and G. Von Heijne. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, 12(1):3--9, 1999.

[136] H. Niska, T. Hiltunen, A. Karppinen, J. Ruuskanen, and M. Kolehmainen. Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence*, 17(2):159--167, 2004.

[137] G. Niu and B-S. Yang. Dempster–Shafer regression for multi-step-ahead time-series prediction towards data-driven machinery prognosis. *Mechanical Systems and Signal Processing*, 23(3):740--751, 2009.

[138] R. Nygaard, G. Melnikov, and A. K. Katsaggelos. A rate distortion optimal ECG coding algorithm. *IEEE Transactions on Biomedical Engineering*, 48(1):28--40, 2001.

[139] S. Ovchinnikov. Discrete piecewise linear functions. *Arxiv preprint arXiv08073364*, page 17, 2008.

[140] A. Palmer, J. Josemontano, and A. Sese. Designing an artificial neural network for forecasting tourism time series. *Tourism Management*, 27(5):781--790, 2006.

[141] S. Papadimitriou, J. Sun, and P. Yu. Local Correlation Tracking in Time Series. *Sixth International Conference on Data Mining ICDM06*, pages 456--465, 2006.

[142] S. Park, S-W. Kim, and W. W. Chu. SBASS : Segment based approach for subsequence searches in sequence databases. *Computer Systems Science and Engineering*, pages 37--46, 2007.

[143] S. Park, D. Lee, and W. W. Chu. Fast Retrieval of Similar Subsequences in Long Sequence Databases. In *KDEX 99 Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*, pages 60--67. IEEE Comput. Soc, 1999.

[144] M. T. Parvez and S. A. Mahmoud. Polygonal approximation of digital planar curves through adaptive optimizations. *Pattern Recognition Letters*, 31(13):1997--2005, 2010.

[145] M. P. Paulraj, S. B. Yaacob, A. N. Abdullah, and S. K. Natraj. Segmentation of Voiced Portion for Voice Pathology Classification Using Fuzzy Logic. *Energy*, pages 560--566, 2010.

[146] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825--2830, 2011.

[147] J. C. Perez and E. Vidal. Optimum polygonal approximation of digitized curves. *Pattern Recognition Letters*, 15(8):743--750, 1994.

[148] C-S. Perng, H. Wang, S. R. Zhang, and D. S. Parker. Landmarks: A New Model for Similarity-Based Pattern Querying in Time Series Databases. *Data Engineering International Conference on*, 0:33, 2000.

[149] D. Preston, P. Protopapas, and C. Brodley. Event Discovery in Time Series. *Computer*, page 12, 2009.

[150] R. B. C. Prudêncio and T. B. Ludermir. Meta-learning approaches to selecting time series models. *Neurocomputing*, 61:121--137, 2004.

[151] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani. Data compression and learning in time sequences analysis. *Physica D: Nonlinear Phenomena*, 180(1-2):15, 2002.

[152] I. Pulido-Calvo and M. M. Portela. Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds. *Journal of Hydrology*, 332(1-2):1--15, 2007.

[153] U. Ramer. An Iterative Procedure for the Polygonal Approximation of Plane Curves. *Computer Graphics and Image Processing*, 1(3):244--256, 1972.

[154] F. Recknagel. Applications of machine learning to ecological modelling. *Ecological Modelling*, 146(1-3):303--310, 2001.

[155] C. Rhodes and M. Morari. False-nearest-neighbors algorithm and noise-corrupted time series. *Physical Review E*, 55(5):6162--6170, 1997.

[156] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In H Ruspini, editor, *IEEE International Conference on Neural Networks*, volume 1 of *Proceedings of the IEEE International Conference on Neural Networks*, pages 586--591. IEEE, Ieee, 1993.

[157] F. Rossi, N. Delannay, B. Conan-Guez, and M. Verleysen. Representation of Functional Data in Neural Networks. *Neurocomputing*, 64(October):183--210, 2007.

[158] J. Rydén. Statistical analysis of temperature extremes in long-time series from Uppsala. *Theoretical and Applied Climatology*, 105(1-2):193--197, 2010.

[159] G. B. Sahoo, C. Ray, and E. H. De Carlo. Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. *Journal of Hydrology*, 327(3-4):525--538, 2006.

[160] M. Salotti. An efficient algorithm for the optimal polygonal approximation of digitized curves. *Pattern Recognition Letters*, 22(2):215--221, 2001.

[161] M. Salotti. Optimal polygonal approximation of digitized curves using the sum of square deviations criterion. *Pattern Recognition*, 35:435--443, 2002.

[162] S. C. Saxena, A. Sharma, and S. C. Chaudhary. Data compression and feature extraction of ECG signals Data compression and feature extraction of ECG signals. *International Journal of Systems Science*, 28(5):483--498, 2007.

[163] L. M. Schmitt. Theory of genetic algorithms II: models for genetic operators over the string-tensor representation of populations and convergence to global optima for arbitrary fitness function under scaling. *Theoretical Computer Science*, 310(1-3):181--231, 2004.

[164] H. I. Shahein and H. M. Abbas. ECG data compression via cubic-splines and scan-along polygonal approximation. *Signal Processing*, 35(3):269--283, 1994.

[165] C. E. Shannon. Programming a Computer for Playing Chess. *Philosophical Magazine*, 41(4):256--275, 1950.

[166] H. Shatkay and S. B. Zdonik. Approximate queries and representations for large data sequences. In *Proceedings of the Twelfth International Conference on Data Engineering*, number CS-95-03, pages 536--545. Department of Computer Science, Brown University, IEEE Comput. Soc. Press, 1996.

[167] J. Shen, S-D. Bao, L-C. Yang, and Y. Li. The PLR-DTW method for ECG based biometric identification, 2011.

[168] G. Simon, J. A. Lee, and M. Verleysen. Unfolding preprocessing for meaningful time series clustering. *Neural Networks*, 19(6-7):877--888, 2006.

[169] R. K. Sivagaminathan and S. Ramakrishnan. A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert Systems with Applications*, 33(1):49--60, 2007.

[170] P. Smyth and E. Keogh. Clustering and Mode Classification of Engineering Time Series Data. *Work*, pages 1--11, 1997.

[171] S. Srinivasulu and A. Jain. A comparative analysis of training methods for artificial neural network rainfall–runoff models. *Applied Soft Computing*, 6(3):295--306, 2006.

[172] D. Stathakis. How many hidden layers and nodes? *International Journal of Remote Sensing*, 30(8):2133--2147, 2009.

[173] J. Sun, C. Zheng, Y. Zhou, Y. Bai, and J. Luo. Nonlinear noise reduction of chaotic time series based on multidimensional recurrent LS-SVM. *Neurocomputing*, 71(16-18):3675--3679, 2008.

[174] J. Takeuchi and K. Yamanishi. A unifying framework for detecting outliers and change points from time series, 2006.

[175] F. E. H. Tay and L. Cao. Application of support vector machines in financial time series forecasting. *Omega*, 29(4):309--317, 2001.

[176] C. Thornton. The Building Block Fallacy. *Complexity International*, 4:1--7, 1997.

[177] E. Toth, A. Brath, and A. Montanari. Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology*, 239(1-4):132--147, 2000.

[178] A. M. Turing. Computing Machinery and Intelligence. *Mind*, 59(236):433--460, 1950.

[179] P. F. Velleman and D. C. Hoaglin. *Applications, basics, and computing of exploratory data analysis*, volume 142. Duxbury Press, 1981.

[180] C. Wang and X. S. Wang. Supporting Content-based Searches on Time Series via Approximation. In *Proc Int Conf on Scientific and Statistical Database Management*, pages 69--81, 2000.

[181] Q. Wang and V. Megalooikonomou. A Dimensionality Reduction Technique for Efficient Time Series Similarity Analysis. *Information Systems Journal*, 33(1):115--132, 2008.

[182] Y. Wei, W. Xu, F. Ying, and H. Tasi. Artificial neural network based predictive method for flood disaster. *Computers & Industrial Engineering*, 42(2-4):383--390, 2002.

[183] A. S. Weigend, M. Mangeas, and A. N. Srivastava. Nonlinear gated experts for time series: discovering regimes and avoiding overfitting. *International Journal of Neural Systems*, 6(4):373--399, 1995.

[184] X. Weng and J. Shen. Detecting outlier samples in multivariate time series dataset. *Knowledge-Based Systems*, 21(8):807--812, 2008.

[185] C. Wu, G. Tzeng, and R. Lin. A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems with Applications*, 36(3):4725--4735, 2009.

[186] C. L. Wu, K. W. Chau, and Y. S. Li. Methods to improve neural network performance in daily flows prediction. *Journal of Hydrology*, 372(1-4):80--93, 2009.

[187] S-T. Wu. A non-self-intersection Douglas-Peucker Algorithm. *Computer Graphics and Image Processing 2003 SIBGRAPI 2003 XVI Brazilian Symposium on*, pages 60--66, 2003.

[188] Z. Xiao, S-J. Ye, B. Zhong, and C-X. Sun. BP neural network with rough set for short term load forecasting. *Expert Systems with Applications*, 36(1):273--279, 2009.

[189] K. Yamaguchi. Reexamination of stock price reaction to environmental performance: A GARCH application. *Ecological Economics*, 68(1-2):345--352, 2008.

[190] J. Yang, W. Wang, and P. S. Yu. Mining asynchronous periodic patterns in time series data, 2000.

[191] X. Yao. A review of evolutionary artificial neural networks. *International Journal of Intelligent Systems*, 3(1):585--601, 1993.

[192] P. Yin. A discrete particle swarm algorithm for optimal polygonal approximation of digital curves. *Journal of Visual Communication and Image Representation*, 15(2):241--260, 2004.

[193] I. W. H. Yip and M. K. P. So. Simplified specifications of a multivariate generalized autoregressive conditional heteroscedasticity model. *Mathematics and Computers in Simulation*, 80(2):327--340, 2009.

[194] G. U. Yule. Why do we sometimes get nonsense-correlations between Time-Series?–a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 89(1):1--63, 1926.

[195] G. P. Zhang. Neural networks for classification: a survey, 2000.

[196] G. P. Zhang and M. Qi. Neural network forecasting for seasonal and trend time series. *European Journal Of Operational Research*, 160(2):501--514, 2005.

[197] H. Zhang and J. Guo. Optimal polygonal approximation of digital planar curves using meta heuristics. *Pattern Recognition*, 34(7):1429--1436, 2001.

[198] H. Zhang and S. Wang. Linearly constrained global optimization via piecewise-linear approximation. *Journal of Computational and Applied Mathematics*, 214(1):111--120, 2008.

ΠΑΡΑΡΤΗΜΑ

Για την εγκατάσταση του προγράμματος σε περιβάλλον Windows, απαιτείται πρώτα η εγκατάσταση των παρακάτω εργαλείων απο τον υποκατάλογο Setup του CD, με την εξής σειρά και με χρήση των προεπιλογών τους:

- python-2.7.msi
- scipy-0.10.1-win32-superpack-python2.7.exe
- numpy-1.6.1-win32-superpack-python2.7.exe
- setuptools-0.6c11.win32-py2.7.exe
- scikits.learn-0.5.win32-py2.7.exe
- matplotlib-1.0.0.win32-py2.7.exe
- ενημέρωση της μεταβλητής περιβάλλοντος path του λειτουργικού σας με τις τιμές:

  C:\Python27;C:\Python27\Lib;C: \Python27\Lib\site-packages;C: \Python27\Lib\site-packages\setuptools;C:\MinGW;
- αντιγραφή του υποκαταλόγου pyBrain σε κάποιο προσωρινό κατάλογο στο δίσκο του υπολογιστή σας, είσοδος σε αυτόν και:

  python setup.py install <ENTER>
- αντιγραφή του υποκαταλόγου MinGW στον κατάλογο ρίζας του δίσκου C:
- qt-win-opensource-4.8.1-mingw.exe
- PyQt-Py2.7-x86-gpl-4.9.1-1.exe

Αφού εγκατασταθούν όλα τα επί μέρους εργαλεία, αντιγράφετε τον υποκατάλογο GUI που βρίσκεται στο CD σε οποιοδήποτε κατάλογο του υπολογιστή σας (μπορείτε να τον μετονομάσετε ελεύθερα), εισέρχεστε στον κατάλογο, δίνετε την εντολή

python main.py

και επιλέγετε το είδος του προβλήματος που θα αντιμετωπιστεί.

Το εργαλείο λογισμικού είναι ανοικτού κώδικα και έχει λάβει άδεια Creative Commons.