



ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Μελέτη της Μεθόδου Αναπαράστασης χρονοσειρών
SAX στην κατηγοριοποίηση χρονοσειρών»

«Εικόνα»

Του φοιτητή
Φωτίου Σγουρού
175060

Επιβλέπων
Λεωνίδας Καραμητόπουλος
Βαθμίδα

Σεπτέμβριος 2026

Τίτλος Δ.Ε.: Μελέτη της Μεθόδου αναπαράστασης χρονοσειρών SAX στην κατηγοριοποίηση
χρονοσειρών

Κωδικός Δ.Ε.25160

Ονοματεπώνυμο φοιτητή: Φώτιος Σγουρός

Ονοματεπώνυμο εισηγητή: Λεωνίδα Καραμητόπουλος

Ημερομηνία ανάληψης Δ.Ε 10-03-2025

Ημερομηνία περάτωσης Δ.Ε. : 06-09-2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Φωτίου Σγουρού που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητα και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

«Αφιέρωση»

Πρόλογος

Η ραγδαία αύξηση των δεδομένων χρονοσειρών σε σύγχρονες εφαρμογές και οι προκλήσεις που θέτει η κατάρτα των πολλών διαστάσεων αποτέλεσαν το βασικό κίνητρο για την επιλογή του συγκεκριμένου θέματος. Η ανάγκη αποδοτικής διαχείρισης σημάτων μεγάλου όγκου με ταυτόχρονη διατήρηση της δομικής πληροφορίας μου κίνησε έντονα το ενδιαφέρον, οδηγώντας με στη συγκριτική μελέτη προηγμένων μεθόδων μείωσης διαστάσεων, όπως η συμβολική προσέγγιση SAX και οι αριθμητικές αναπαραστάσεις PAA, DFT και PLA.

Μέσα από την εκπόνηση της εργασίας είχα την ευκαιρία να εμβαθύνω ουσιαστικά στις αρχές της εξόρυξης δεδομένων και της μηχανικής μάθησης, υλοποιώντας ένα πλήρες πειραματικό pipeline σε περιβάλλον RStudio πάνω σε σύνολα δεδομένων του αποθετηρίου UCR. Παράλληλα, εξοικειώθηκα με τη χρήση μοντέλων συλλογικής μάθησης (Random Forests, CatBoost) και την εφαρμογή αυστηρών μη παραμετρικών στατιστικών ελέγχων (Wilcoxon, Friedman, Conover).

Η διαδικασία αυτή συνέβαλε καθοριστικά στην ανάπτυξη της αναλυτικής μου σκέψης και στην απόκτηση πολύτιμων ερευνητικών και πρακτικών δεξιοτήτων για τη μετέπειτα επιστημονική και επαγγελματική μου πορεία.

Περίληψη

Η παρούσα πτυχιακή εργασία εξετάζει τη διαχείριση δεδομένων χρονοσειρών υψηλής διάστασης στο πλαίσιο της Εξόρυξης Δεδομένων και της Μηχανικής Μάθησης. Κεντρικός άξονας της έρευνας είναι η αξιολόγηση της μεθόδου Συμβολικής Αθροιστικής Προσέγγισης (SAX), η οποία μετατρέπει συνεχή σήματα σε διακριτές συμβολοσειρές εξασφαλίζοντας την ιδιότητα του κάτω ορίου (Lower Bounding) μέσω της συνάρτησης MINDIST. Η αποτελεσματικότητα της SAX συγκρίνεται τόσο με τα πλήρη δεδομένα (Raw) όσο και με τις καθιερωμένες αριθμητικές μεθόδους μείωσης διαστάσεων: την Τμηματική Αθροιστική Προσέγγιση (PAA), τον Διακριτό Μετασχηματισμό Φουριέ (DFT) και την Τμηματική Γραμμική Προσέγγιση (PLA).

Η πειραματική μεθοδολογία υλοποιήθηκε σε περιβάλλον R/Rstudio χρησιμοποιώντας δέκα αντιπροσωπευτικά σύνολα δεδομένων από το αποθετήριο UCR Times Series Classification Archive. Εξετάστηκαν τέσσερις οικογένειες αλγορίθμων ταξινόμησης: ο κατηγοριοποιητής k-Εγγύτερων Γειτόνων (k-NN), τα Μεμονομένα Δέντρα Απόφασης (CART), τα Τυχαία Δάση (Random Forests) και ο αλγόριθμος CatBoost. Η αξιολόγηση βασίστηκε στην ακρίβεια κατηγοριοποίησης και σε αυστηρούς μη παραμετρικούς στατιστικούς ελέγχους (Wilcoxon, Friedman και Conover post-hoc).

Τα ευρήματα αποδεικνύουν ότι η PAA επιτυγχάνει την κορυφαία και πιο σταθερή επίδοση (82,05% στον 1-NN και 83.06% στα Random Forests), υπερτερώντας συχνά των αρχικών δεδομένων χάρη στην εξομάλυνση του θορύβου. Ο DFT επιδεικνύει στατιστική ισοδυναμία με τα πλήρη σήματα (80.08% στον 1-NN), ενώ η PLA υστερεί συστηματικά (68.19%). Η SAX παρουσιάζει διττή συμπεριφορά: υπερέχει αισθητά σε συνθετικά και βιομηχανικά δεδομένα με θόρυβο (π.χ synthetic control, CBF, Wafer), αλλά εμφανίζει απώλειες λόγω σφάλματος κβάντισης σε πολυταξικά προβλήματα περιγραμμάτων. Παράλληλα, προσφέρει απaráμιλλο πλεονέκτημα συμπίεσης και ταχύτητας μέσω πινάκων αναζήτησης (Lookup Tables), διατηρώντας υψηλή ανταγωνιστικότητα σε μοντέλα συνόλου. Συμπερασματικά, η εργασία τεκμηριώνει ότι η κατάλληλη μείωση διαστάσεων βελτιστοποιεί ουσιαστικά την υπολογιστική αποδοτικότητα χωρίς απαραίτητα να θυσιάζει την ταξινομική ισχύ.

«Study of SAX representation method for time series classification»

Fotios Sgouros

Abstract

This thesis investigates the management of high-dimensional time series data within the context of Data Mining and Machine Learning. The primary objective is the evaluation of the Symbolic Aggregate Approximation (SAX) method, which transforms continuous signals into discrete symbolic strings while guaranteeing the lower bounding property via the MINDIST distance function. The performance of SAX is comparatively benchmarked against raw, uncompressed times series and established numeric dimensionality reduction techniques: Piecewise Aggregate Approximation (PAA), Discrete Fourier Transform (DFT), and Piecewise Linear Approximation (PLA).

The experimental pipeline was implemented in R/RStudio across ten benchmark datasets from the UCR Times Series Classification Archive. Four distinct classification algorithms were assessed: K-Nearest Neighbors (k-NN), Decision Trees (CART), Random Forests, and CatBoost. Model performance was evaluated using classification accuracy and rigorous non-parametric statistical hypothesis testing, including Wilcoxon signed-rank, Friedman, and post-hoc Conover tests with multiple testing corrections.

The experimental results demonstrate that PAA achieves the highest and most consistent performance (82.05% with the 1-NN and 83.06% with Random Forests), often surpassing the raw baseline by effectively filtering out stochastic noise. DFT exhibits statistical equivalence to full data representations (80.08% with 1-NN), whereas PLA systematically underperforms across all models (68.19%). SAX displays a dual nature: it achieves superior accuracy on noisy synthetic and industrial sensor signals (e.g. Synthetic Control, CFB, Wafer), but suffers from quantization error on complex image outline shapes with high class cardinality. Nevertheless, SAX provides substantial data compression and computational acceleration via precomputed lookup tables while maintaining solid competitiveness within ensemble models. In conclusion, the study validates that dimensionality reduction serves as a vital optimization step, successfully reducing computational overhead without inevitably compromising classification power.

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή, κ. Λεωνίδα Καραμητόπουλο, για την ανάθεση του θέματος, τη συνεχή επιστημονική υποστήριξη και τις καίριες συμβουλές του σε όλα τα στάδια εκπόνησης της παρούσας διπλωματικής εργασίας. Η καθοδήγηση και η επικοινωνιακή του ανατροφοδότηση αποτέλεσαν θεμέλιο λίθο για την ολοκλήρωση της έρευνας.

Περιεχόμενα

Πρόλογος.....	5
Περίληψη.....	6
Abstract.....	7
Ευχαριστίες.....	8
Περιεχόμενα.....	9
Κατάλογος Σχημάτων.....	12
Κατάλογος Πινάκων.....	12
Συντομογραφίες.....	13
Κεφάλαιο 1ο: Εισαγωγή	14
1.1 Εξόρυξη Δεδομένων.....	14
1.2 Knowledge Discovery in Databases.....	14
1.3 Προκλήσεις της Εξόρυξης Δεδομένων.....	15
1.4 Χρονοσειρές.....	16
1.5 Εξόρυξη πληροφορίας από χρονοσειρές.....	17
1.6 Διάσταση.....	17
Κεφάλαιο 2ο: Κατηγοριοποίηση	18
2.1 Εισαγωγή.....	18
2.2 Μετρα Ομοιότητας.....	19
2.2.1 Ευκλείδεια Απόσταση.....	19
2.2.2 Απόσταση Manhattan.....	20
2.2.3 Απόσταση Minkowski.....	20
2.3 Μέθοδοι Κατηγοριοποίησης.....	20
2.3.1 K-Εγγύτεροι Γείτονες – K-Nearest Neighbors.....	20
2.4 Δέντρο Απόφασης.....	22
2.5 Τυχαία Δάση – Random Forests.....	24
Κεφάλαιο 3ο: Μέθοδοι Αναπαράστασης Χρονοσειρών	32
3.1 Εισαγωγή.....	32
3.2 Διακριτός Μετασχηματισμός Φουριέ (Discrete Fourier Transform - DFT).....	32
3.3 Piecewise Linear Approximation.....	34
3.4 Piecewise Aggregate Approximation.....	36
3.5 Επίλογος.....	38
Κεφάλαιο 4ο: Συμβολική Αθροιστική Προσέγγιση	39

4.1.1	Μειονεκτήματα Προγενέστερων Συμβολικών Αναπαραστάσεων	39
4.1.2	Η Καινοτομία της SAX.....	39
4.2	Αρχιτεκτονική και Στάδια Μετασχηματισμού SAX.....	40
4.2.1	Προ-επεξεργασία: Κανονικοποίηση Τυπικής Μορφής (Z-score Normalization)	40
4.2.2	Ενδιάμεσο Στάδιο: Μείωση διαστάσεων μέσω PAA	40
4.2.3	Breakpoints και Γκαουσιανή Κατανομή.....	41
4.2.4	Συμβολική Αντιστοίχιση και Κατασκευή Λέξης (Word Construction)	41
4.3	Μέτρα Ομοιότητας και Θεωρητικές Ιδιότητες.....	42
4.3.1	Η συνάρτηση Απόστασης MINDIST	42
4.3.2	Κατασκευή Πίνακα Αναζήτησης (Lookup Table).....	43
4.3.3	Η ιδιότητα του Κάτω Ορίου (Lower Bounding Property).....	43
4.3.4	Αυστηρότητα Κάτω Ορίου (Tightness of Lower Bound-TLB) και Επιλογή Παραμέτρων(w, a).....	44
4.4	Μείωση Πληθικότητας (Numerosity Reduction).....	44
4.4.1	Τεχνική Ολισθαίνοντος Παραθύρου και το Πρόβλημα Όγκου	44
4.4.2	Συμπίεση Τύπου RLE (Run-Length Encoding).....	45
4.4.3	Διαχείριση Ειδικών Περιπτώσεων: Σχεδόν Σταθερές Υποακολουθίες (Near-Constant Subsequences).....	45
4.5	Τομείς Εφαρμογής και Αλγόριθμοι Εξόρυξης.....	46
4.5.1	Κατηγοριοποίηση (Classification).....	46
4.5.2	Ομαδοποίηση.....	46
4.5.3	Αναζήτηση Βάσει Περιεχομένου	46
4.5.4	Ανίχνευση Ανωμαλιών (Anomaly Detection).....	47
4.5.5	Ανακάλυψη Μοτίβων (Motif Discovery).....	47
4.6	Παραλλαγές και Επεκτάσεις της SAX στη Βιβλιογραφία	47
4.6.1	Προσεγγίσεις Προσαρμοσμένες και Βασισμένες στην Εντροπία	47
4.6.2	Ενσωμάτωση Χαρακτηριστικών Τάσης	48
4.7	Επίλογος.....	48
Κεφάλαιο 5ο: Πειραματική Μεθοδολογία		49
5.1	επιγραφή Συνόλων Δεδομένων (UCR Time Series Classification Archive).....	49
5.2	Ροή Εκτέλεσης Πειραμάτων (Experimental Pipeline).....	50
5.2.1	Προσέγγιση 1: Κατηγοριοποίηση k-NN με την Μετρική MINDIST.SAX.....	51
5.2.2	Προσέγγιση 2: Μοντέλα Δέντρων και Συλλογικής Μάθησης (Decision Trees, Random Forest, CatBoost).....	51
5.3	Περιβάλλον Υλοποίησης και Πακέτα Λογισμικού (R/RStudio).....	52

5.4	Μετρικές Αξιολόγησης (Evaluation Metrics).....	52
5.5	Στατιστική Μεθοδολογία και Έλεγχος Υποθέσεων.....	53
5.5.1	Συγκρίσεις Ζευγών: Έλεγχος Wilcoxon Signed-Rank.....	53
5.5.2	Συγκριτική Αξιολόγηση Πολλαπλών Αναπαραστάσεων: Έλεγχος Friedman.....	54
5.5.3	Post-Hoc Ανάλυση Conover και Διόρθωση Πολλαπλών Ελέγχων.....	54
5.6	Επίλογος.....	54
Κεφάλαιο 6ο: Παρουσίαση Αποτελεσμάτων		55
6.1	Εισαγωγή και Μεθοδολογικό Πλαίσιο Αξιολόγησης.....	55
6.2	Συγκεντρωτικά Πειραματικά Αποτελέσματα Ακρίβειας.....	56
6.3	Αναλυτική Συγκριτική Μελέτη ανά Μέθοδο Αναπαράστασης.....	58
6.3.1	Συμβολική Αθροιστική Προσέγγιση (SAX).....	58
6.3.2	Τμηματική Αθροιστική Προσέγγιση (PAA).....	58
6.3.3	Διακριτός Μετασχηματισμός Fourier (DFT).....	59
6.4	Στατιστικός Έλεγχος Υποθέσεων.....	59
6.4.1	Έλεγχος Ζευγών με Paired Wilcoxon Signed-Rank Test.....	59
6.4.2	Συνολικός Έλεγχος Διατάξεων Friedman και Post-Hoc Conover Test ανά Κατηγοριοποιητή.....	61
6.5	Συγκριτική Ανάλυση Υπολογιστικής Πολυπλοκότητας και Συμπίεσης.....	62
6.6	Σύνοψη Κεφαλαίου και Κύρια Ευρήματα.....	62
Κεφάλαιο 7ο: Συζήτηση και Συμπεράσματα		64
7.1	Επισκόπηση και Σκοπός της Έρευνας.....	64
7.2	Ερμηνεία των Κύριων Ευρημάτων.....	64
7.2.1	Η Διττή Φύση της Μεθόδου SAX: Εξομάλυνση Θορύβου έναντι Σφάλματος Κβάντισης	64
7.2.2	Σταθερότητα του DFT και Ανεπάρκεια της PLA.....	64
7.3	Συγκριτική Αξιολόγηση Αλγορίθμων Ταξινόμησης.....	65
7.3.1	k-Nearest Neighbors(1-NN).....	65
7.3.2	Τυχαία Δάση (Random Forests και Catboost.....	65
7.4	Στατιστική Εγκυρότητα και Μεθοδολογικές Παρατηρήσεις.....	65
7.5	Περιορισμοί της Παρούσας Εργασίας.....	66
7.6	Προτάσεις για Μελλοντική Έρευνα.....	66
7.7	Τελικό Συμπέρασμα.....	66
ΒΙΒΛΙΟΓΡΑΦΙΑ.....		67
ΠΑΡΑΡΤΗΜΑ Α : ΤΙΤΛΟΣ ΠΑΡΑΡΤΗΜΑΤΟΣ.....		69

Κατάλογος Σχημάτων

Σχήμα 1.1: Knowledge Discovery in Databases.....	16
Σχήμα 1.2: Time series visualization.....	18
Σχήμα 1.3: Dimensions in a dataset.....	20
Σχήμα 4: Classification Process.....	22
Σχήμα 5: KNN.....	24
Σχήμα 6: Decision Tree.....	27
Σχήμα 7: Representation Methods.....	34
Σχήμα 8: DFT Method.....	36
Σχήμα 9: Breakpoints.....	43
Σχήμα 10:Lookup table.....	45

Κατάλογος Πινάκων

Πίνακας 5.1: UCR Datasets.....	51
Πίνακας 6.1: KNN Accuracy.....	59
Πίνακας 6.2: Decision Tree Accuracy.....	59
Πίνακας 6.3: Random Forest Accuracy.....	60
Πίνακας 6.4: CatBoost Accuracy.....	60
Πίνακας 6.5: Wilcoxon Test.....	63

Συντομογραφίες

KDD	Knowledge Discovery in DataBases
KNN	K-Nearest Neighbors
DT	Decision Tree
CART	Classification and Regression Trees
RF	Random Forest
CB	CatBoost
SAX	Symbolic Aggregate Approximation
A-SAX	Adaptive Symbolix Aggregate Approximation
E-SAX	Entropy-based Symbolic Aggregate Approximation
PAA	Piecewise Aggregate Approximation
PLA	Piecewise Linear Approximation
DFT	Discrete Fourier Transform
DWT	Discrete Wavelet Transform
SVD	Singular Value Decomposition
TLB	Tightness of Lower Bound
ANN	Artificial Neural Networks
FN	False Negatives
FP	False Positives
GDBT	Gradient Boosted Decision Trees
XGBoost	extreme Gradient Boosting
TS:	Target Statistics
OOB	Out-of-Bag
UCR	University of California, Riverside
RLE	Run-Length Encoding

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εξόρυξη Δεδομένων

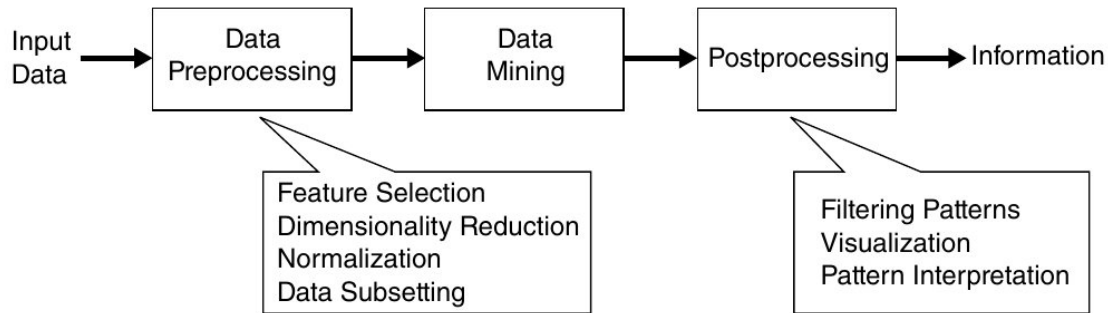
Η εξόρυξη δεδομένων είναι η διαδικασία η οποία μας επιτρέπει να εξάγουμε χρήσιμες, πληροφορίες από μεγάλες αποθήκες δεδομένων. Αλγόριθμοι εξόρυξης δεδομένων χρησιμοποιούνται για να εξερευνήσουν μεγάλες βάσεις δεδομένων ώστε να ανακαλυφθούν πιθανώς κρυμμένες πληροφορίες που δεν μπορούν να ανιχνευθούν με το ανθρώπινο μάτι. Μελετώντας ένα τεράστιο σύνολο δεδομένων, οι τεχνικές εξόρυξης δεδομένων μας βοηθούν στην επίτευξη εργασιών όπως η διατύπωση κανόνων συσχέτισης (Association Rules), η ομαδοποίηση όμοιων δεδομένων (Clustering), η παρατήρηση απόκλισης από μια αναμενόμενη συμπεριφορά (Anomaly Detection) καθώς επίσης και η πρόβλεψη μια μελλοντικής τιμής (Prediction).

Ως παράδειγμα μπορούμε να χρησιμοποιήσουμε τις αγορές σε ένα σούπερ μάρκετ. Μελετώντας τις αγοραστικές συνήθειες των καταναλωτών μπορούμε να εξάγουμε κανόνες συσχέτισης μεταξύ προϊόντων που συνήθως αγοράζονται μαζί. Αντίστοιχα θα μπορούσαμε να προβλέψουμε με ένα υψηλό ποσοστό ακρίβειας αν ένας πελάτης σε μια αντιπροσωπεία αυτοκινήτων θα προχωρήσει σε αγορά με βάση το εισόδημα του.

Είναι σημαντικό να αναφέρουμε πως δεν υπάγονται όλες οι εργασίες εξεύρεσης πληροφοριών στον κλάδο της εξόρυξης των δεδομένων. Για παράδειγμα η αναζήτηση μιας εγγραφής σε μια βάσης δεδομένων με τη χρήση ενός Συστήματος Διαχείρισης Βάσεων Δεδομένων (DBMS), ή η αναζήτηση μιας ιστοσελίδας στον Παγκόσμιο Ιστό, ανήκουν στον κλάδο της Ανάκτησης Πληροφοριών (Information Retrieval).

1.2 Knowledge Discovery in Databases

Η εξόρυξη δεδομένων είναι ένα από τα βήματα της διαδικασίας που είναι γνωστή ως Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD). Η KDD περιγράφει όλα τα βήματα μετατροπής ακατέργαστων πληροφοριών σε χρήσιμα δεδομένα. Η διαδικασία της KDD περιγράφεται με το σχήμα 1[1]. Για την επίτευξη αυτής της διαδικασίας ακολουθείται μια συγκεκριμένη δομή. Αρχικά έχουμε ένα σύνολο δεδομένων που θέλουμε να μελετήσουμε. Πάνω σε αυτό το σύνολο δεδομένων γίνεται η επιλογή των χαρακτηριστικών που μας ενδιαφέρουν άμεσα. Γίνεται έτσι μια αρχική επεξεργασία. Στην συνέχεια γίνεται ο “καθαρισμός” των δεδομένων. Σε αυτό το βήμα αφαιρείται ο θόρυβος από το σύνολο των δεδομένων μας. Επίσης, εδώ αντιμετωπίζουμε τις ελλιπείς τιμές που μπορεί να έχουν τα δεδομένα. Τέλος αφαιρούμε ακραίες τιμές που μπορεί να επηρεάσουν αρνητικά τον αλγόριθμο μας και να μην έχουμε τα επιθυμητά αποτελέσματα. Το επόμενο βήμα είναι να μετασχηματίσουμε τα δεδομένα. Αυτή η διαδικασία γίνεται ώστε τα δεδομένα να έχουν μορφή που υποστηρίζεται από τους αλγόριθμους εξόρυξης δεδομένων που θα εφαρμόσουμε. Επίσης σε αυτό το βήμα μπορούμε να εφαρμόσουμε τεχνικές μείωσης διαστάσεων. Έπειτα, πρέπει να ληφθεί μια απόφαση σχετικά με το μοντέλο που θα χρησιμοποιήσουμε. Αυτό έχει να κάνει με το πρόβλημα που θέλουμε να επιλύσουμε και με το είδος των δεδομένων που χρησιμοποιούμε. Μόλις γίνει η επιλογή του μοντέλου, προχωράμε στην εφαρμογή αυτού πάνω στα δεδομένα μας, δηλαδή στην εκπαίδευση και την αξιολόγηση του μοντέλου. Το τελευταίο στάδιο στην KDD είναι η μελέτη και διερμηνεία των αποτελεσμάτων που προκύπτουν από την εφαρμογή των μοντέλων. Εδώ είναι σημαντικό να οπτικοποιήσουμε τα αποτελέσματα με κατάλληλο τρόπο, ώστε να γίνονται εύκολα κατανοητά από το κοινό στο οποίο θα τα παρουσιάσουμε.



Σχήμα 1: Η διαδικασία της Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων

1.3 Προκλήσεις της Εξόρυξης Δεδομένων

Οι παραδοσιακές τεχνικές της Εξόρυξης Δεδομένων αντιμετωπίζουν ορισμένες προκλήσεις που σχετίζονται με τα καινούρια σύνολα δεδομένων που εισάγονται στον κλάδο. Παρακάτω θα αναφερθούν αυτές οι προκλήσεις.

Ποιότητα των δεδομένων. Στον πραγματικό κόσμο τα σύνολα δεδομένων που χρησιμοποιούνται δεν είναι ποτέ τέλεια. Τα σύνολα περιέχουν ελλιπείς τιμές, διπλότυπες εγγραφές, ακραίες τιμές και γενικά θόρυβο. Όλα αυτά δυσκολεύουν τους αλγόριθμους εξόρυξης δεδομένων να αποδώσουν με ακρίβεια.

Δυνατότητα κλιμάκωσης. Με την πρόοδο και την συνεχή ανάπτυξη στον τομέα των της δημιουργίας των δεδομένων, είναι πλέον σύνηθες φαινόμενο τα σύνολα των δεδομένων να αθροίζονται στα μεγέθη των petabyte ή ακόμα και exabyte. Οι αλγόριθμοι εξόρυξης δεδομένων, είναι αναγκαίο, να μπορούν να διαχειριστούν αυτή την κλίμακα χωρίς να δαπανούν σημαντική ποσότητα πόρων. Δημιουργείται, λοιπόν, η ανάγκη για τους αλγόριθμους να μπορούν να κλιμακώνονται ανάλογα.

Υψηλός αριθμός διαστάσεων. Μία εξίσου σημαντική πρόκληση που αντιμετωπίζει ο κλάδος της ανάλυσης δεδομένων είναι ο υψηλός αριθμός διαστάσεων στα σύνολα δεδομένων. Όπως είδαμε, με τον όρο διάσταση εννοούμε τον αριθμό των χαρακτηριστικών που περιγράφουν κάθε εγγραφή σε ένα σύνολο δεδομένων. Όταν αυτός ο αριθμός είναι υψηλός, οι αλγόριθμοι εξόρυξης δεδομένων, τείνουν να υπολειτουργούν και τα αποτελέσματα που δίνονται, να μην είναι τα ιδανικά. Επιπλέον, όσο αυξάνονται οι διαστάσεις ενός συνόλου δεδομένων, τόσο αυξάνεται και το υπολογιστικό κόστος αρκετών αλγορίθμων εξόρυξης δεδομένων. Στην μοντέρνα εποχή είναι σύνηθες τα σύνολα δεδομένων που μελετώνται, μπορεί να περιέχουν εκατοντάδες, σε κάποιες περιπτώσεις ακόμα και χιλιάδες χαρακτηριστικά. Το πρόβλημα αυτό παρατηρείται ιδιαίτερα στα σύνολα δεδομένων που περιέχουν χρονοσειρές λόγω της φύσης της συγκεκριμένης δομής δεδομένων. Στον παρακάτω πίνακα[] παρουσιάζεται ένα σύνολο δεδομένων που περιλαμβάνει χρονοσειρές. Όπως έχει αναλυθεί, οι χρονοσειρές είναι δεδομένα τα οποία καταγράφονται σε μια μονάδα χρόνου. Αυτό σημαίνει ότι κάθε χαρακτηριστικό μιας χρονοσειράς είναι και μία καταγραφή των δεδομένων που μελετώνται. Για παράδειγμα, έχουμε ένα σύνολο δεδομένων το οποίο περιέχει χρονοσειρές που καταγράφουν την μέση θερμοκρασία μιας των πόλεων μιας χώρας κάθε μέρα, για ένα χρόνο. Αυτό σημαίνει ότι το σύνολο δεδομένων που προκύπτει θα είναι εκ φύσεως πολύ μεγάλο. Όπως αναφέρθηκε, όταν τα σύνολα δεδομένων είναι τόσο μεγάλα οι αλγόριθμοι δεν είναι όσο παραγωγικοί όσο θα θέλαμε, καθώς αναπτύχθηκαν για να λειτουργούν καλύτερα σε σύνολα δεδομένων με λίγες διαστάσεις. Το πρόβλημα των πολλών διαστάσεων στην Εξόρυξη των δεδομένων είναι γνωστό ως “η κατάρα των πολλών διαστάσεων (Curse of Dimensionality)”, και είναι από τα βασικά κίνητρα που ώθησαν τους ερευνητές

να δημιουργήσουν μεθόδους μείωσης των διαστάσεων όπως είναι και η Symbolic Aggregate Approximation (SAX), την οποία θα μελετήσουμε.

1.4 Χρονοσειρές

Οι χρονοσειρές είναι μια ακολουθία δεδομένων τα οποία καταγράφονται σε μια μονάδα χρόνου, συνήθως με σταθερή συχνότητα. Στον πραγματικό κόσμο παρατηρούνται χρονοσειρές σε πολλούς κλάδους. Στα οικονομικά, η ημερήσια τιμή μιας μετοχής για ένα τρίμηνο. Στην μετεωρολογία, η θερμοκρασία μιας πόλης κάθε ώρα της μέρας. Στον τομέα της υγείας, παρατηρούνται χρονοσειρές ως ηλεκτροκαρδιογραφήματα που μετρούν την ηλεκτρική δραστηριότητα της καρδιάς κάθε δευτερόλεπτο.

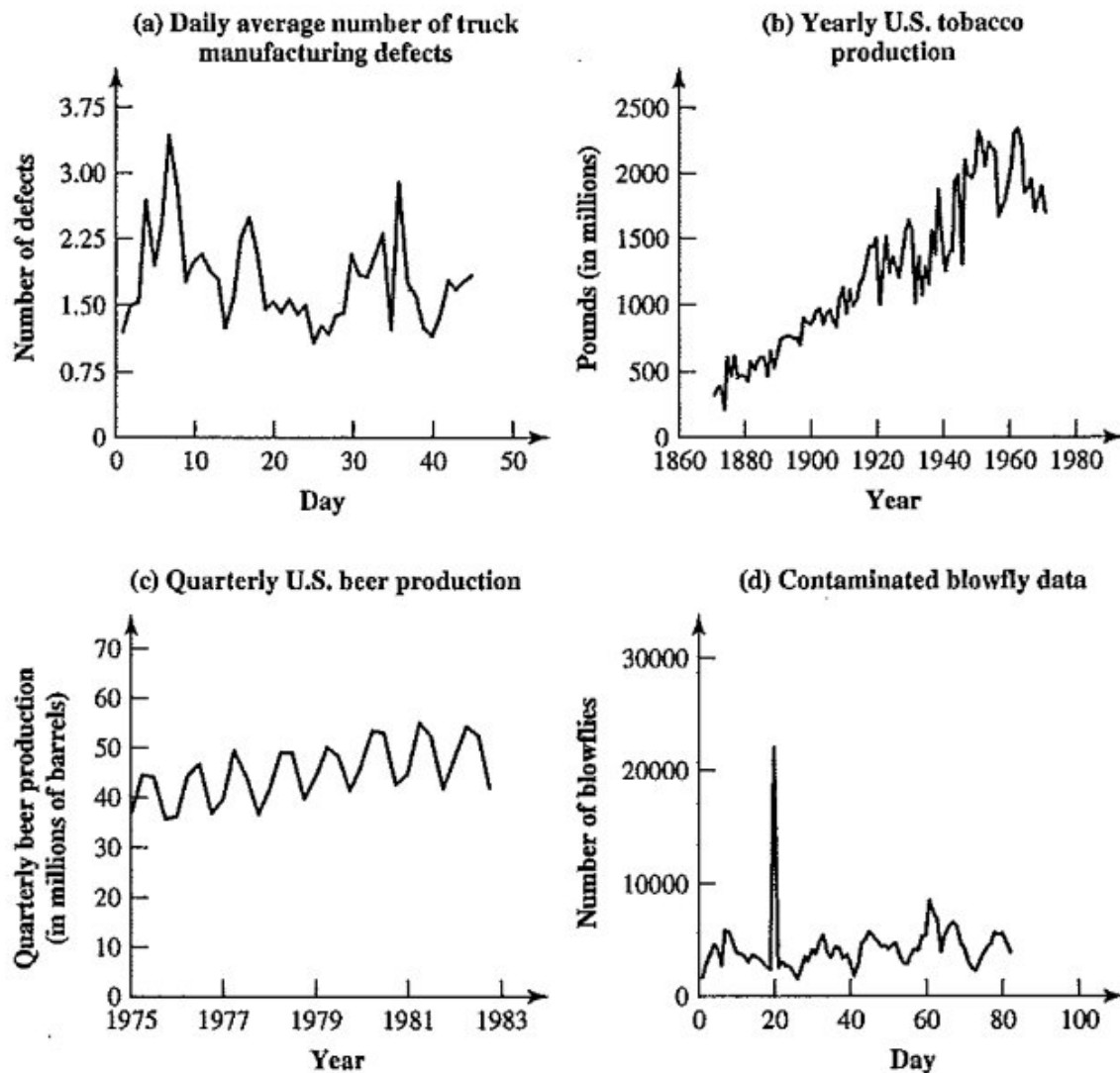
Οι χρονοσειρές συνήθως συμβολίζονται ως εξής:

$$T = (t_1, t_2, t_3, \dots, t_n) \quad (1.1)$$

Όπου t_i είναι η τιμή της μεταβλητής που μελετάται, την χρονική στιγμή i . Όπως φαίνεται και στο σχήμα 2[2], οι χρονοσειρές μπορούν να αναπαρασταθούν με διαγράμματα που στον άξονα οριζόντιο άξονα x εμφανίζεται η μονάδα του χρόνου κατά την οποία γίνεται η μέτρηση, ενώ στον κατακόρυφο άξονα y καταγράφεται η τιμή του χαρακτηριστικού που μελετάται.

Οι χρονοσειρές μπορούν να χωριστούν σε δύο κατηγορίες: τις συνεχείς και τις διακριτές. Ως συνεχείς χρονοσειρές, ορίζονται εκείνες που μπορούν να καταμετρηθούν συνεχόμενα στον χρόνο. Παράδειγμα μιας συνεχούς χρονοσειράς είναι το ηλεκτροκαρδιογράφημα. Η καταγραφή του ρεύματος είναι συνεχόμενη και δεν υπάρχει καμία διακοπή σε κανένα σημείο του γραφήματος. Διακριτές χρονοσειρές είναι εκείνες των οποίων η καταγραφή γίνεται με σταθερά κενά. Για παράδειγμα η καταγραφή της τιμής κλεισίματος μια μετοχής είναι μία διακριτή χρονοσειρά. Ο διαχωρισμός αυτός γίνεται γιατί η καταγραφή ορισμένων δεδομένων δεν χρειάζεται να είναι συνεχής. Μπορούμε να μετρήσουμε την θερμοκρασία μιας πόλης ανά millisecond, αλλά με αυτό το τρόπο θα έχουμε εκατομμύρια καταγραφές δεδομένων σχεδόν πανομοιότυπες. Επίσης το σύνολο δεδομένων θα περιλαμβάνει και αρκετό “θόρυβο”. Τέλος το σύνολο δεδομένων που θα προκύπτει θα είναι τεράστιο, καθιστώντας την αποθήκευση και την εφαρμογή αλγορίθμων εξόρυξης δεδομένων πολύ ακριβές διαδικασίες.

Οι χρονοσειρές έχουν ως κύριο χαρακτηριστικό το μεγάλο μέγεθός τους. Μια χρονοσειρά μπορεί να έχει καταγραφές σε χιλιάδες έως εκατοντάδες μονάδες χρόνου. Αυτό σημαίνει αυτομάτως ότι η συγκεκριμένη δομή δεδομένων έχει εξ ορισμού πολλές διστάσεις. Ένα σύνολο δεδομένων που αποτελείται από χρονοσειρές θα είναι, λοιπόν, τεράστιο σε διαστάσεις και σε όγκο κάτι που δυσκολεύει τους παραδοσιακούς αλγορίθμους εξόρυξης δεδομένων. Βλέπουμε, επομένως ότι υπάρχει ανάγκη για εφαρμογή τεχνικών και αλγορίθμων που θα μειώσουν τον αριθμό των διατάσεων για να βελτιωθεί η απόδοση αυτών των αλγορίθμων, χωρίς να χαθεί σημαντική πληροφορία που καταγράφεται σε αυτά τα σύνολα δεδομένων. Παρακάτω, αναλύονται μέθοδοι που χρησιμοποιούνται για να επιτευχθεί αυτός ο σκοπός και πως συμβάλουν στην βελτίωση της απόδοσης των αλγορίθμων εξόρυξης δεδομένων.



Σχήμα 2: Παραδείγματα οπτικοποίησης χρονοσειρών στους άξονες x και y

Multivariate και Univariate Χρονοσειρές

Για τις χρονοσειρές μπορούμε, επίσης, να διακρίνουμε άλλες δυο κατηγορίες, οι οποίες όμως δεν αφορούν το είδος των δεδομένων αλλά τον αριθμό των χαρακτηριστικών που καταγράφονται. Διαχωρίζουμε, άρα τις χρονοσειρές σε μονομεταβλητές (univariate) και πολυμεταβλητές (multivariate).

Univariate Time series

Οι χρονοσειρές που χαρακτηρίζονται ως μονομεταβλητές (univariate) είναι αυτές που αποτελούνται από μία μόνο μεταβλητή. Γίνεται καταγραφή μιας τιμής, ανά μια ορισμένη μονάδα χρόνου. Όπως είδαμε στο παράδειγμα με την μέτρηση της θερμοκρασίας. Έχουμε μόνο μια μεταβλητή που μας ενδιαφέρει και αυτή είναι η θερμοκρασία μας πόλης, την οποία καταγράφουμε κάθε μέρα.

Multivariate Time series

Πολυμεταβλητές (multivariate) χρονοσειρές, είναι αυτές που αποτελούνται από περισσότερες από μία τιμές. Σε αντίθεση με τις μονομεταβλητές χρονοσειρές, μια δεδομένη χρονική στιγμή t γίνεται καταγραφή πολλαπλών τιμών. Ένα παράδειγμα μιας πολυμεταβλητής χρονοσειράς στον τομέα της υγείας είναι η ταυτόχρονη καταγραφή καρδιακού ρυθμού, της πίεσης του αίματος και των επιπέδων οξυγόνου.

1.5 Εξόρυξη πληροφορίας από χρονοσειρές

Όπως και με κάθε άλλο είδος δεδομένων, έτσι και με τις χρονοσειρές υπάρχουν κάποια συγκεκριμένα βήματα που ακολουθούνται ώστε να επιτευχθεί η διαδικασία της εξόρυξης γνώσης. Αρχικά πρέπει να είναι ξεκάθαρος ο στόχος που θέτουμε (Κατηγοριοποίηση, ομαδοποίηση, εξαγωγή κανόνων ή πρόβλεψη. Εφόσον έχουν γίνει ξεκάθαροι οι στόχοι, το επόμενο βήμα είναι η επιλογή των δεδομένων που θα χρησιμοποιηθούν, ποια χαρακτηριστικά θα ληφθούν υπόψιν, ποιες παράμετροι θα οριστούν και τέλος ποιος αλγόριθμος θα εφαρμοστεί. Έπειτα, σειρά έχει η προ-επεξεργασία των δεδομένων που θα χρησιμοποιήσουμε για την διαδικασία της εξόρυξης. Τα σύνολα δεδομένων συχνά περιέχουν θόρυβο και αν δεν τον αφαιρέσουμε, τα αποτελέσματα που θα λάβουμε δεν θα είναι αξιόπιστα. Ένα συχνό φαινόμενο που παρατηρείται επίσης στα σύνολα δεδομένων είναι οι ελλιπείς τιμές. Κατά την καταγραφή των δεδομένων, συμβαίνει να μην έχουμε δεδομένα λόγω εξωτερικών παραγόντων. Στην μέτρηση της θερμοκρασίας για παράδειγμα, ενδέχεται μια συγκεκριμένη χρονική στιγμή να δυσλειτουργήσει ένας αισθητήρας. Αυτό θα έχει ως αποτέλεσμα μέσα στο σύνολο δεδομένων να μην υπάρχουν κάποιες μετρήσεις. Σε άλλες περιπτώσεις, μπορεί πολύ απλά να υπάρχει έλλειψη δεδομένων του αντικειμένου που μελετάμε. Οι περισσότεροι αλγόριθμοι εξόρυξης δεδομένων δεν αποδίδουν όταν τα σύνολα δεδομένων έχουν ελλιπείς τιμές. Ευτυχώς υπάρχουν τεχνικές που μπορούμε να χρησιμοποιήσουμε, να διαμορφώσουμε τα δεδομένα ώστε να αντισταθμιστούν αυτά τα προβλήματα και να εξασφαλίσουμε ότι οι αλγόριθμοι που θα εφαρμόσουμε στα δεδομένα μας, θα μπορέσουν να αποδώσουν με αξιοπιστία.

Οι χρονοσειρές είναι μια μορφή δεδομένων που παρατηρείται σε όλους τους τομείς, πλέον, επομένως είναι σημαντικό να μπορούμε να αξιοποιήσουμε τεχνικές και αλγορίθμους εξόρυξης δεδομένων πάνω σε αυτές.

1.6 Διάσταση

Διάσταση ενός συνόλου δεδομένων ονομάζεται το πλήθος των χαρακτηριστικών των αντικειμένων που βρίσκονται μέσα στο σύνολο. Τα δεδομένα που είναι μικρά σε διαστάσεις συνήθως είναι ποσοτικά διαφορετικά από αυτά με υψηλό πλήθος διαστάσεων. Στο σχήμα 3[1] μπορούμε να παρατηρήσουμε ένα σύνολο δεδομένων. Κάθε στήλη του πίνακα είναι και μία διάσταση. Σχεδόν όλοι οι αλγόριθμοι εξόρυξης δεδομένων και μηχανικής μάθησης υπολειτουργούν λόγω των πολλών διαστάσεων. Για αυτό τον λόγο η εφαρμογή μεθόδων μείωσης διαστάσεων, ως μέρος την προ-επεξεργασίας, σε σύνολα δεδομένων παρέχει σημαντικό όφελος. Πιο συγκεκριμένα, όταν διαχειριζόμαστε δεδομένα χρονοσειρών, εκ φύσεως οι διαστάσεις του συνόλου θα είναι πολλές. Είναι λοιπόν, καίριας σημασίας να εφαρμόσουμε σωστές τεχνικές μείωσης διαστάσεων ώστε να μειώσουμε όσο γίνεται περισσότερο τον όγκο των δεδομένων και να διευκολύνουμε τους αλγορίθμους να αποδώσουν καλύτερα, αλλά ταυτόχρονα να μη χάσουμε σημαντικό κομμάτι της πληροφορίας που μελετάμε.

Κεφάλαιο 1

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard	cold-blooded	scales	yes	yes	no	no	no	fish
shark								
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Σχήμα 3: Παράδειγμα ενός συνόλου δεδομένων όπου κάθε χαρακτηριστικό(στήλες) των προτύπων είναι και μία διάσταση

1.7 Επίλογος

Συνοψίζοντας, στο πρώτο αυτό κεφάλαιο παρουσιάστηκαν οι θεμελιώδεις έννοιες της Εξόρυξης Δεδομένων και τα διαδοχικά στάδια της διαδικασίας KDD. Ιδιαίτερη έμφαση δόθηκε στη δομή και τις ιδιαιτερότητες των χρονοσειρών, ανακεικνύοντας τις προκλήσεις που θέτει η “κατάρα των πολλών διαστάσεων” στους παραδοσιακούς αλγορίθμους. Καθίσταται σαφές ότι η αποτελεσματική επεξεργασία τέτοιων δεδομένων προϋποθέτει την εφαρμογή κατάλληλων τεχνικών προεπεξεργασίας και μάθησης. Για τον λόγο αυτό, στο Κεφάλαιο 2 που ακολουθεί, αναλύεται εκτενώς η διαδικασία της κατηγοριοποίησης, εξετάζοντας τα βασικά μέτρα ομοιότητας και τις αντιπροσωπευτικές οικογένειες κατηγοριοποιητών που θα αξιοποιηθούν στη μελέτη.

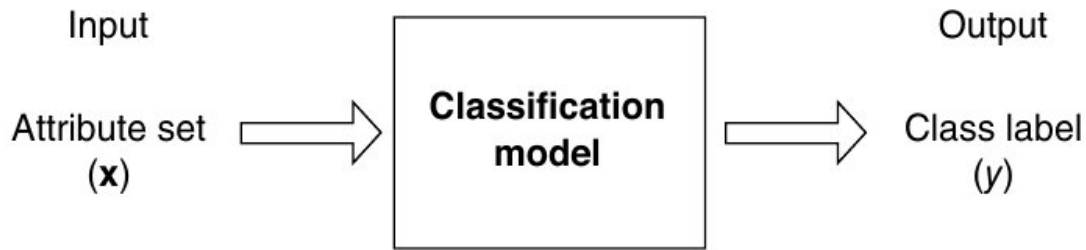
Κεφάλαιο 2ο: Κατηγοριοποίηση

2.1 Εισαγωγή

Κατηγοριοποίηση (classification) είναι η διαδικασία ανάθεσης ενός αντικειμένου σε κάποιες προκαθορισμένες κλάσεις. Παράδειγμα κατηγοριοποίησης είναι η διάκριση ενός μηνύματος ηλεκτρονικού ταχυδρομείου ως ανεπιθύμητο (spam) με βάση τα περιεχόμενα του. Η κατηγοριοποίηση χρησιμοποιείται για διακριτές μεταβλητές. Ως μαθηματική έκφραση ορίζεται η συνάρτηση στόχου f , που δέχεται σαν είσοδο το σύνολο χαρακτηριστικών x και παράγει μια έξοδο που αντιστοιχεί στην προβλεφθείσα κλάση. Για να θεωρηθεί σωστή η κατηγοριοποίηση ενός αντικειμένου, θα πρέπει να ισχύει:

$$f(x) = y \quad (2.1)$$

Η κατηγοριοποίηση είναι μια από τις διαδικασίες εξόρυξης δεδομένων. Σκοπός της κατηγοριοποίησης είναι να δημιουργηθεί ένα μοντέλο το οποίο θα δίνει σε άγνωστα αντικείμενα, προκαθορισμένες διακριτές ετικέτες. Αν ο αριθμός των κλάσεων δεν είναι διακριτός αλλά συνεχής, τότε η διαδικασία λέγεται παλινδρόμηση. Στο παρακάτω σχήμα[] μπορούμε να διακρίνουμε τη διαδικασία της κατηγοριοποίησης. Έχουμε ένα σύνολο χαρακτηριστικών x στα δεδομένα εισόδου. Πάνω σε αυτά εφαρμόζεται το μοντέλο κατηγοριοποίησης και η έξοδος του μοντέλου είναι οι προβλεπόμενες κλάσεις για αυτά τα χαρακτηριστικά. Για την επίτευξη της κατηγοριοποίησης χρειαζόμαστε ένα σύνολο δεδομένων το οποίο θα περιέχει εγγραφές που έχουν ήδη γνωστές κλάσεις. Το ονομάζουμε σύνολο δεδομένων εκπαίδευσης (training data set). Πάνω σε αυτό το σύνολο δεδομένων εκπαιδεύεται ο αλγόριθμος κατηγοριοποίησης, και έπειτα μπορεί να δοκιμαστεί σε άγνωστα δεδομένα. Στο σχήμα 3[1] βλέπουμε ένα παράδειγμα ενός συνόλου δεδομένων. Το σύνολο αυτό αποτελείται από κάποια είδη ζώων. Με τη βοήθεια της κατηγοριοποίησης, μπορούμε να δημιουργήσουμε ένα μοντέλο που θα εκπαιδευτεί πάνω σε αυτό το σύνολο δεδομένων, ώστε να μπορεί με βάση τα χαρακτηριστικά να κατηγοριοποιήσει τα ζώα σε θηλαστικά, πτηνά, αμφίβια και ψάρια. Οι αλγόριθμοι κατηγοριοποίησης χωρίζονται σε δύο κατηγορίες. Τους οκνηρούς (lazy) και τους πρόθυμους (eager) κατηγοριοποιητές. Οι οκνηροί κατηγοριοποιητές ονομάζονται με αυτό τον τρόπο γιατί δεν δημιουργούν κάποιο μοντέλο κατηγοριοποίησης. Η διαδικασία της ανάθεσης κατηγορίας στις εγγραφές του συνόλου γίνεται κατά την εξέταση των δεδομένων. Στην περίπτωση των πρόθυμων κατηγοριοποιητών, γίνεται το αντίθετο. Ο αλγόριθμος εκπαιδεύεται πάνω στο σύνολο των δεδομένων και εξάγεται ένα μοντέλο το οποίο μετά μπορεί να εφαρμοστεί σε καινούρια (άγνωστα) δεδομένα. Παράδειγμα οκνηρού κατηγοριοποιητή είναι ο αλγόριθμος των Εγγύτερων Γειτόνων, ενώ πρόθυμος κατηγοριοποιητής είναι ο αλγόριθμος του Δέντρου Απόφασης. Η κατηγοριοποίηση στον τομέα της Μηχανικής Μάθησης ονομάζεται Μάθηση με επίβλεψη. Η ονομασία προκύπτει από την διαδικασία της εκπαίδευσης του εκάστοτε αλγορίθμου, καθώς για να επιτευχθεί, είναι απαραίτητη η χρήση ενός συνόλου δεδομένων εκπαίδευσης, στο οποίο είναι ήδη γνωστές οι κλάσεις των εγγραφών. Η διαφορά με την ομαδοποίηση ή μάθηση χωρίς επίβλεψη είναι ότι εκεί δεν είναι γνωστές οι κλάσεις, και με τη χρήση των μέτρων ομοιότητας μπορούμε να δημιουργήσουμε συστάδες μεταξύ των εγγραφών που βρίσκονται σε κοντινές αποστάσεις, είναι δηλαδή όμοιες. Στο παρακάτω διάγραμμα ροής περιγράφεται η διαδικασία της κατηγοριοποίησης. Ξεκινάμε με το σύνολο δεδομένων το οποίο περιλαμβάνει τις ετικέτες. Χωρίζουμε το σύνολο σε δύο υποσύνολα, το υποσύνολο εκπαίδευσης και το υποσύνολο ελέγχου. Συνήθως για το σύνολο δεδομένων εκπαίδευσης χρησιμοποιούμε το 70-80% το αρχικού συνόλου και το υπόλοιπο 20-30% πάει στο σύνολο δεδομένων ελέγχου. Έπειτα, επιλέγουμε το μοντέλο κατηγοριοποίησης και το εφαρμόζουμε στα δεδομένα μας. Αφού εκπαιδευτεί το μοντέλο, περνάμε στο στάδιο του ελέγχου όπου δίνουμε το σύνολο δεδομένων ελέγχου χωρίς τις ετικέτες, ώστε να γίνει η πρόβλεψη. Τέλος συγκρίνουμε τις ετικέτες που προβλέπει το μοντέλο με τις πραγματικές και υπολογίζουμε τις μετρικές αξιολόγησης.



Σχήμα 4: Εφαρμογή μοντέλου κατηγοριοποίησης σε ένα σύνολο δεδομένων

Μετά την εκπαίδευση του μοντέλου, μπορούμε να κατασκευάσουμε έναν πίνακα που θα περιέχει τις σωστές και τις λάθος προβλέψεις του μοντέλου καθώς και τις πραγματικές κατηγορίες των εγγραφών. Αυτός ο πίνακας ονομάζεται πίνακας σύγχυσης και με την βοήθεια του μπορούμε να υπολογίσουμε τις μετρικές αξιολόγησης του μοντέλου κατηγοριοποίησης. Σε επόμενο κεφάλαιο θα αναλύσουμε σε βάθος τον πίνακα σύγχυσης και τις μετρικές αξιολόγησης.

2.2 Μέτρα Ομοιότητας

Τόσο στην κατηγοριοποίηση όσο και στην ομαδοποίηση δεδομένων πρέπει να είμαστε ικανοί να κρίνουμε κατά πόσο δύο εγγραφές δεδομένων μοιάζουν μεταξύ τους. Για να επιτύχουμε κάτι τέτοιο, χρησιμοποιούμε κάποια μέτρα ομοιότητας που μπορούμε να ορίσουμε.

2.2.1 Ευκλείδεια Απόσταση

Το πιο γνωστό και ευρέως διαδεδομένο από αυτά είναι η Ευκλείδεια απόσταση. Η ευκλείδεια απόσταση υπολογίζει την απόσταση μεταξύ δύο σημείων στον Ευκλείδειο χώρο και είναι αποτέλεσμα του Πυθαγόρειου θεωρήματος. Ο υπολογισμός της ευκλείδειας απόστασης στον χώρο δύο διαστάσεων γίνεται με τον μαθηματικό τύπο:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2.2)$$

Με αυτόν τον τρόπο, έχουμε την δυνατότητα να μετρήσουμε την απόσταση μεταξύ δύο εγγραφών στο σύνολο δεδομένων μας, και κατ' επέκταση, να συγκρίνουμε την ομοιότητα τους. Η ευκλείδεια απόσταση αποτελεί το μέτρο σύγκρισης για έναν από τους πιο γνωστούς αλγορίθμους κατηγοριοποίησης, τον αλγόριθμο K-Εγγύτερων Γειτόνων (K-Nearest Neighbors - KNN). Παρακάτω θα αναλύσουμε πως λειτουργεί ο συγκεκριμένος αλγόριθμος.

2.2.2 Απόσταση Manhattan

Ένα άλλο μέτρο ομοιότητας που χρησιμοποιείται συχνά είναι η απόσταση Manhattan. Με την απόσταση Manhattan μπορούμε επίσης να ορίσουμε την απόσταση μεταξύ δύο σημείων με βάση τον ακόλουθο τύπο. Έστω δύο σημεία $P = (x_1, y_1)$ και $Q = (x_2, y_2)$:

$$D(P, Q) = |x_1 - x_2| + |y_1 - y_2| \quad (2.3)$$

Όπως βλέπουμε στην σχέση, η απόσταση Manhattan για δύο σημεία υπολογίζεται αθροίζοντας τις απόλυτες τιμές των διαφορών των σημείων σε καρτεσιανές συντεταγμένες. Και αυτό το μέτρο σύγκρισης χρησιμοποιείται σε αλγορίθμους κατηγοριοποίησης όπως είναι ο K-NN.

2.2.3 Απόσταση Minkowski

Η απόσταση Minkowski αποτελεί μια γενίκευση τόσο της Ευκλείδειας απόστασης όσο και της απόστασης Manhattan στον διανυσματικό χώρο. Ορίζεται μέσω μιας παραμέτρου p ($p \geq 1$), η οποία καθορίζει τη μορφή της μετρικής. Έστω δύο σημεία $P = (x_1, y_2)$ και $Q = (x_2, y_2)$, η απόσταση Minkowski υπολογίζεται από τον παρακάτω μαθηματικό τύπο:

$$Di(P, Q) = (|x_1 - x_2|^p + |y_1 - y_2|^p)^{\frac{1}{p}} \quad (2.4)$$

Ανάλογα με την τιμή της παραμέτρου p , προκύπτουν διαφορετικά μέτρα ομοιότητας. Για $p=1$ η σχέση ταυτίζεται με την απόσταση Manhattan, ενώ για $p=2$ ανάγεται στην κλασσική Ευκλείδεια απόσταση. Λόγω της ευελιξίας που προσφέρει η ρύθμιση της παραμέτρου p , χρησιμοποιείται συχνά σε αλγόριθμους όπως ο K-NN όταν απαιτείται προσαρμογή του μέτρου σύγκρισης στη γεωμετρία των δεδομένων.

2.3 Μέθοδοι Κατηγοριοποίησης

2.3.1 K-Εγγύτεροι Γείτονες – K-Nearest Neighbors

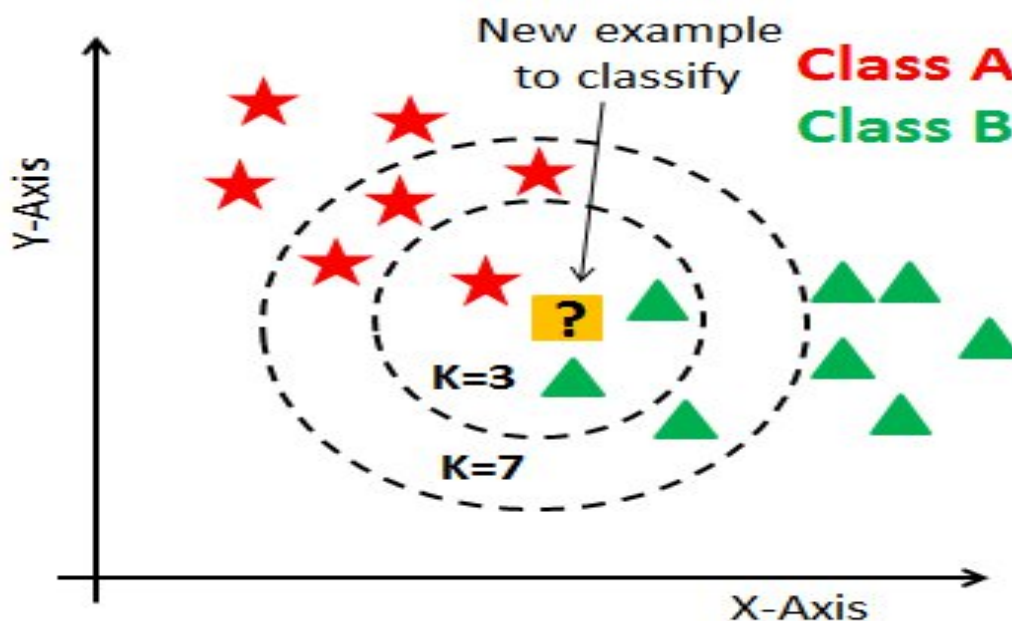
Ο αλγόριθμος των K-Εγγύτερων Γειτόνων (KNN) είναι ένας από τους πιο διαδεδομένους αλγόριθμους κατηγοριοποίησης καθώς επίσης και από τους πιο απλούς. Ο KNN χαρακτηρίζεται ως σκληρός κατηγοριοποιητής. Αυτό σημαίνει ότι δεν δημιουργείται κάποιο εκπαιδευμένο μοντέλο, απλώς αποθηκεύονται τα δεδομένα, και γίνονται προβλέψεις όταν ζητηθούν. Ο αλγόριθμος λειτουργεί, όπως οι περισσότεροι αλγόριθμοι εποπτευμένης μάθησης, με τα δεδομένα να χωρίζονται σε σύνολο δεδομένων εκπαίδευσης και ελέγχου. Εξετάζεται κάθε εγγραφή στο σύνολο δεδομένων ελέγχου, μετρώντας την απόσταση της από όλες τις εγγραφές στο σύνολο δεδομένων εκπαίδευσης. Τότε, επιλέγονται τα k πιο κοντινά σημεία και η εγγραφή του συνόλου εκπαίδευσης παίρνει την κλάση η οποία επικρατεί στα σημεία αυτά. Στο σχήμα 5[3] μπορούμε να διακρίνουμε την εφαρμογή του αλγορίθμου πάνω σε ένα καινούριο πρότυπο. Αν θέσουμε $k=3$ η κλάση που ανατίθεται είναι η κλάση B, ενώ για $k=5$ το πρότυπο κατηγοριοποιείται ως κλάση A. Για την μέτρηση της απόστασης χρησιμοποιούνται μετρικές όπως η Ευκλείδεια απόσταση, η απόσταση Manhattan, η απόσταση Minkowski. Η επιλογή της μετρικής της απόστασης γίνεται ανάλογα με τον τύπο των δεδομένων. Ο αλγόριθμος αναπτύχθηκε από τους στατιστικολόγους Evelyn Fix και Joseph Hodges ως μέρος μιας τεχνικής αναφοράς για την Πολεμική Αεροπορία των Ηνωμένων Πολιτειών το 1951[4]. Ο KNN είναι εύκολος στην υλοποίηση και μπορεί να εφαρμοστεί σε μεγάλο αριθμό προβλημάτων. Η θεωρητική θεμελίωση και η ισχύς του αλγορίθμου αποδείχθηκαν από τους Cover και Hart (1967)[5]. Οι ερευνητές απέδειξαν ότι, καθώς το μέγεθος του συνόλου εκπαίδευσης τείνει στο άπειρο, η πιθανότητα σφάλματος R του αλγορίθμου Nearest Neighbor (για $k=1$) περιορίζεται πάνω από το διπλάσιο του ελαχίστου δυνατού σφάλματος (Bayes error rate, R^*), δηλαδή ισχύει:

$$R^* \leq R \leq 2R^* \quad (2.5)$$

Το εύρημα αυτό είναι ιδιαίτερα σημαντικό, καθώς αποδεικνύει ότι τουλάχιστον η μισή πληροφορία κατηγοριοποίησης που εμπεριέχεται σε ένα άπειρο σύνολο δεδομένων εμπεριέχεται στον πρώτο πλησιέστερο γείτονα, καθιστώντας τον αλγόριθμο εξαιρετικά αποτελεσματικό παρά την απλότητα του. Μπορεί να εφαρμοστεί τόσο σε προβλήματα κατηγοριοποίησης όσο και σε προβλήματα παλινδρόμησης. Στο παρακάτω διάγραμμα μπορούμε να δούμε την λειτουργία του KNN στην πράξη. Έχουμε ένα σύνολο δεδομένων με εγγραφές που ανήκουν σε δύο κλάσεις. Με αστέρια αναπαριστούμε την κλάση A, ενώ με τρίγωνα αναπαριστούμε την κλάση B. Έστω ότι έχουμε μια καινούρια εγγραφή

(τετράγωνο) και θέλουμε να της αναθέσουμε μια κατηγορία. Μετράμε την απόσταση της εγγραφής από τα k κοντινότερα σημεία, και αναθέτουμε την κλάση η οποία υπερισχύει στους γείτονες. Για να αποδώσει καλά ο αλγόριθμος, τα δεδομένα μας χρειάζονται προ-επεξεργασία. Είναι απαραίτητο να εφαρμοστεί κανονικοποίηση σε όλες τις αριθμητικές τιμές του συνόλου δεδομένων. Αυτό συμβαίνει λόγω των μέτρων ομοιότητας που χρησιμοποιεί ο αλγόριθμος. Όταν ένα σύνολο δεδομένων έχει χαρακτηριστικά σε διαφορετικές κλίμακες, τα χαρακτηριστικά με το μεγαλύτερο εύρος τιμών θα υπερτερούν σε σχέση με τα υπόλοιπα πιο χαμηλής κλίμακας, με αποτέλεσμα τα τελευταία να μην έχουν βαρύτητα για τον αλγόριθμο.

Τα αρνητικά του αλγορίθμου KNN είναι τρία. Έχει υψηλό υπολογιστικό κόστος. Ο κατηγοριοποιητής πρέπει να υπολογίσει τις αποστάσεις για κάθε εγγραφή των συνόλων δεδομένων ελέγχου και εκπαίδευσης. Επιπλέον, υπάρχει ανάγκη για αποθήκευση των τεράστιων συνόλων δεδομένων σε βάσεις προσβάσιμες ανά πάσα στιγμή. Τέλος αυτός ο αλγόριθμος είναι επιρρεπής στον θόρυβο. Αν το σύνολο των δεδομένων που χρησιμοποιούμε έχει ακραίες τιμές, θα επηρεαστεί αρνητικά η απόδοσή του. Ευτυχώς, αυτά τα προβλήματα μπορούν να λυθούν. Με την χρήση δεικτών (indexing) μπορούμε να μειώσουμε το χρόνο και το υπολογιστικό κόστος του αλγορίθμου. Για την διαχείριση των συνόλων δεδομένων γίνεται εφαρμογή μεθόδων μείωσης διαστάσεων που βοηθούν στην συμπίκνωση των δεδομένων χωρίς να χαθεί σημαντική πληροφορία που θα προκαλέσει ανακρίβειες στον αλγόριθμο. Μια τέτοια μέθοδος είναι και η Symbolic Aggregate Approximation (SAX). Η μείωση των διαστάσεων λύνει, επίσης, σε σημαντικό βαθμό το πρόβλημα του θορύβου και των ακραίων τιμών.



Σχήμα 5: Παράδειγμα εφαρμογής του αλγορίθμου k -NN σε ένα καινούριο πρότυπο

2.4 Δέντρο Απόφασης

Τα δέντρα απόφασης (Decision Trees) αποτελούν μία από τις πιο διαδεδομένες διαισθητικές τεχνικές στον τομέα της Εξόρυξης Δεδομένων και της Μηχανικής Μάθησης. Σε αντίθεση με τους οκνηρούς κατηγοριοποιητές όπως ο αλγόριθμος KNN, το δέντρο απόφασης κατατάσσεται στους πρόθυμους κατηγοριοποιητές. Αυτό σημαίνει ότι κατά τη φάση της εκπαίδευσης κατασκευάζεται ένα αυτόνομο μοντέλο, το οποίο στη συνέχεια μπορεί να εφαρμόζεται άμεσα για την ταξινόμηση νέων, άγνωστων δεδομένων χωρίς να απαιτείται η διατήρηση ολόκληρου του συνόλου εκπαίδευσης στη μνήμη.

Η δομή ενός δέντρου απόφασης αποτελείται από τα εξής βασικά στοιχεία:

Ρίζα (Root Node): Ο αρχικός κόμβος που περιέχει το σύνολο των δεδομένων εισόδου πριν από οποιοδήποτε διαχωρισμό.

Εσωτερικοί Κόμβοι (Internal Nodes): Αντιστοιχούν σε ερωτήματα ή ελέγχους πάνω στα χαρακτηριστικά του συνόλου δεδομένων.

Ακμές / Κλάδοι (Edges / Branches): Αναπαριστούν τις πιθανές τιμές ή συνθήκες που προκύπτουν από τον έλεγχο ενός χαρακτηριστικού.

Φύλλα / Τερματικοί Κόμβοι (Leaves/ Terminal Nodes): Αντιστοιχούν στις τελικές προβλεπόμενες κλάσεις(ετικέτες).

Στο σχήμα 6[1] μπορούμε να δούμε την δομή ενός δέντρου απόφασης που έχει μοντελοποιηθεί πάνω στο σύνολο δεδομένων του πίνακα 1.

Η θεμελίωση της σύγχρονης μεθοδολογίας των δέντρων απόφασης οφείλεται κυρίως στους Breiman, Friedman, Olshen και Stone[6] με την εισαγωγή του αλγορίθμου CART (Classification and Regression Trees). Τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν τόσο σε προβλήματα κατηγοριοποίησης (Classification Trees), όσο και σε προβλήματα παλινδρόμησης (Regression Trees), όταν η μεταβλητή-στόχος λαμβάνει συνεχείς τιμές[6].

Γεωμετρικά, ο αλγόριθμος λειτουργεί χωρίζοντας αναδρομικά τον χώρο των δεδομένων X σε μικρότερες, μη επικαλυπτόμενες περιοχές (hyper-rectangles), με στόχο σε κάθε περιοχή να συγκεντρώνονται πρότυπα που ανήκουν κατά το δυνατόν στην ίδια κλάση (ομογενείς περιοχές).

Έστω ένας κόμβος t . Η αναλογία των εγγραφών στον κόμβο t που ανήκουν στην κλάση j συμβολίζεται ως $p(j|t)$. Η επιλογή του βέλτιστου διαχωρισμού (split) σε κάθε κόμβο βασίζεται στη μείωση της ακαθαρσίας (impurity) του κόμβου. Η συνάρτηση ακαθαρσίας $i(t)$ λαμβάνει τη μέγιστη τιμή της όταν όλες οι κλάσεις είναι ισοπίθανα κατανεμημένες στον κόμβο και μηδενίζεται όταν ο κόμβος περιέχει στοιχεία από μία μόνο κλάση (πλήρης ομογένεια).

Δύο από τα πιο δημοφιλή μέτρα ακαθαρσίας που χρησιμοποιούνται είναι:

Δείκτης Gini (Gini Index):

$$i(t) = 1 - \sum_{j=1}^J p(j|t)^2 \quad (2.6)$$

Εντροπία (Entropy/Information Gain):

$$i(t) = -\sum_{j=1}^J p(j|t) * \log_2 p(j|t) \quad (2.7)$$

Όταν ένας υποψήφιος διαχωρισμός s χωρίζει τον κόμβο t σε δύο απογόνους, τον αριστερό t_l (με ποσοστό δεδομένων p_l) και τον δεξιό t_r (με ποσοστό p_r), η ποιότητα/βελτίωση του διαχωρισμού $\Delta i(s,t)$ υπολογίζεται ως η μείωση της ακαθαρσίας:

$$\Delta i(s,t) = i(t) - p_l * i(t_l) - p_r * i(t_r) \quad (2.8)$$

Ο αλγόριθμος επιλέγει τον διαχωρισμό s που μεγιστοποιεί τη μείωση της ακαθαρσίας $\Delta i(s,t)$.

2.4.1 Τερματισμός και Υπερεκπαίδευση (Overfitting)

Αν το δέντρο αφεθεί να αναπτυχθεί πλήρως, υπάρχει ο κίνδυνος υπερεκπαίδευσης (overfitting). Σε αυτή την περίπτωση, το δέντρο αποστηθίζει τον θόρυβο και τις ιδιαιτερότητες του συνόλου εκπαίδευσης, με αποτέλεσμα να μειώνεται η ικανότητα γενίκευσής του σε νέα δεδομένα. Για την αντιμετώπιση αυτού του φαινομένου εφαρμόζεται η τεχνική της κλάδευσης (pruning), κατά την οποία αφαιρούνται τμήματα του δέντρου που δεν προσφέρουν σημαντική βελτίωση στην ακρίβεια πρόβλεψης.

2.4.2 Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα:

Ερμηνεία: Τα δέντρα απόφασης παρέχουν κανόνες της μορφής “IF-THEN”, καθιστώντας τα εύκολα κατανοητά από τον άνθρωπο.

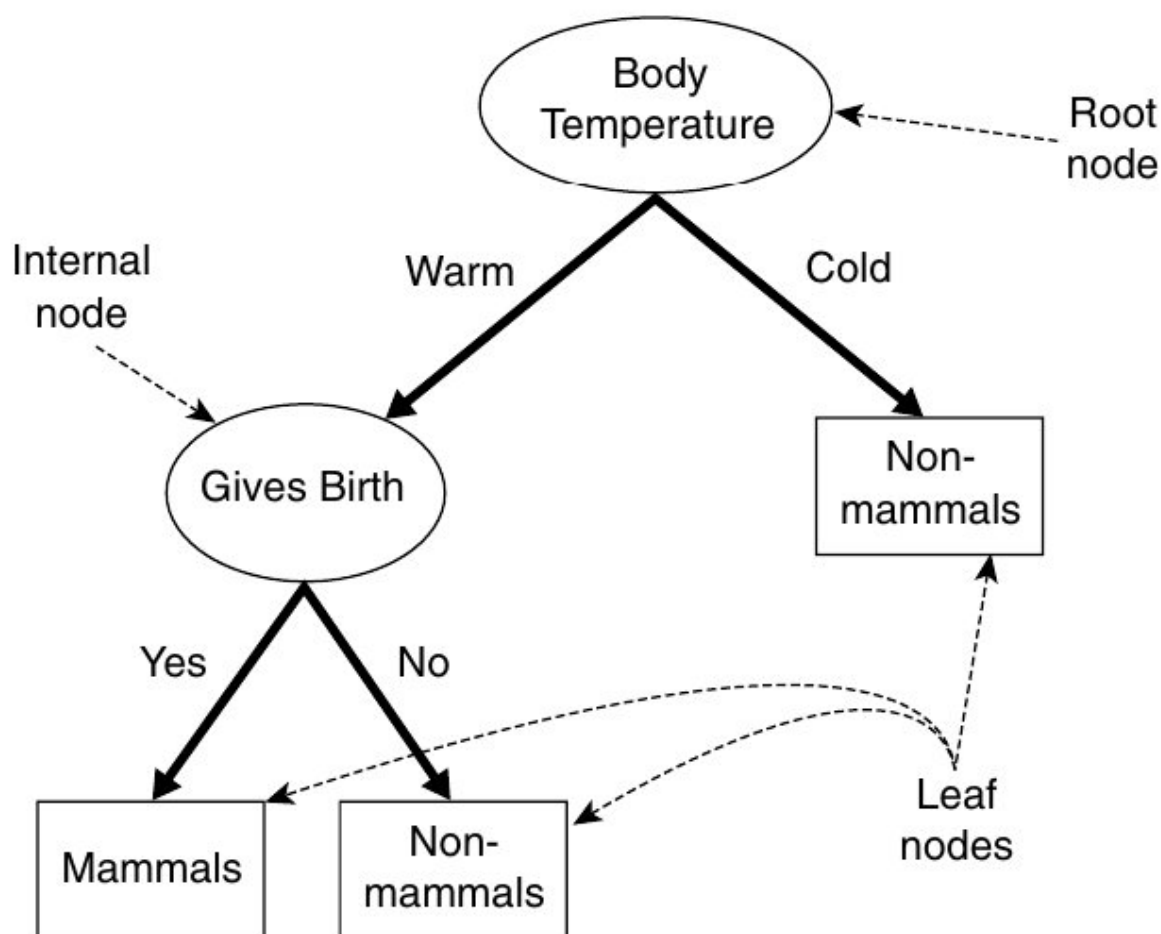
Διαχείριση Ετερογενών Δεδομένων: Μπορούν να επεξεργαστούν τόσο αριθμητικά όσο και κατηγορικά δεδομένα χωρίς να απαιτείται εκτεταμένη κανονικοποίηση.

Υπολογιστική Αποδοτικότητα κατά την πρόβλεψη: Η ταξινόμηση μιας νέας εγγραφής απαιτεί απλά περάσματα από τη ρίζα στα φύλλα, καθιστώντας την πολύ γρήγορη.

Μειονεκτήματα:

Αστάθεια (Instability): Μικρές αλλαγές στα δεδομένα εκπαίδευσης ενδέχεται να οδηγήσουν σε εντελώς διαφορετική δομή δέντρου.

Επιρρέπεια σε Overfitting: Χωρίς την κατάλληλη κλάδευση ή ρύθμιση παραμέτρων, τα δέντρα τείνουν να γίνονται πολύ σύνθετα.



Σχήμα 6: Παράδειγμα δομής ενός δέντρου απόφασης

2.5 Τυχαία Δάση – Random Forests

Τα Τυχαία Δάση (Random Forest) αποτελούν μία από τις πιο ισχυρές, δημοφιλείς και αποτελεσματικές τεχνικές Εξόρυξης Δεδομένων και Μηχανικής Μάθησης. Αναπτύχθηκαν από τον Leo Breiman το 2001 και ανήκουν στην κατηγορία της Συλλογικής / Ομαδικής Μάθησης (Ensemble Learning)[7]. Η βασική ιδέα της συλλογικής μάθησης είναι ο συνδυασμός πολλών ασθενών ή αυτόνομων μοντέλων (weak/individual base learners) με σκοπό τη δημιουργία ενός ισχυρού συνδυαστικού μοντέλου πρόβλεψης, το οποίο επιτυγχάνει υψηλότερη ακρίβεια και καλύτερη ικανότητα γενίκευσης σε σχέση με οποιοδήποτε μεμονωμένο μοντέλο[7].

Όπως αναλύθηκε στην προηγούμενη ενότητα, ένα μεμονωμένο Δέντρο Απόφασης (Decision Tree) είναι ένας πρόθυμος κατηγοριοποιητής με υψηλή δυνατότητα ερμηνείας, αλλά παρουσιάζει δύο σημαντικά μειονεκτήματα: υψηλή διακύμανση και μεγάλη ευαισθησία στα δεδομένα εκπαίδευσης. Μικρές αλλαγές στο σύνολο εκπαίδευσης μπορούν να οδηγήσουν σε εντελώς διαφορετική δομή δέντρου και σε φαινόμενα υπερεκπαίδευσης (overfitting). Ο αλγόριθμος των Τυχαίων Δασών επιλύει αυτά τα προβλήματα κατασκευάζοντας “δάση” από πολλά ανεξάρτητα δέντρα απόφασης και συνδυάζοντας τις προβλέψεις τους.

2.5.1 Μηχανισμός Λειτουργίας και Αρχιτεκτονική

Η επιτυχία του αλγορίθμου βασίζεται στην εισαγωγή τυχαιότητας κατά τη διαδικασία εκπαίδευσης κάθε δέντρου. Αυτή η τυχαιότητα επιτυγχάνεται μέσω δύο βασικών μηχανισμών:

Δειγματοληψία με Επανάθεση (Bootstrap Aggregating/Bagging): Για την κατασκευή κάθε δέντρου $b = 1, 2, \dots, B$ στο δάσος (όπου B το συνολικό πλήθος των δέντρων), δημιουργείται ένα νέο σύνολο εκπαίδευσης T_b μέσω τυχαίας δειγματοληψίας με επανάθεση (bootstrap sampling) από το αρχικό σύνολο δεδομένων εκπαίδευσης N εγγραφών. Το σύνολο T_b έχει το ίδιο μέγεθος N με το αρχικό, αλλά λόγω της επανάθεσης περιέχει διπλότυπες εγγραφές, ενώ ορισμένες εγγραφές παραλείπονται.

Τυχαία Επιλογή Χαρακτηριστικών (Random Subspace/Feature Selection): Κατά την ανάπτυξη κάθε κόμβου σε ένα δέντρο, ο διαχωρισμός (split) δεν αναζητείται στο σύνολο των M διαθέσιμων χαρακτηριστικών, αλλά σε ένα τυχαία επιλεγμένο υποσύνολο m χαρακτηριστικών ($m < M$). Η τυπική τιμή για προβλήματα κατηγοριοποίησης είναι:

$$m = \lfloor \sqrt{M} \rfloor \quad \text{ή} \quad m = \lfloor \log_2 M + 1 \rfloor \quad (2.9)$$

Η τυχαία επιλογή χαρακτηριστικών εξασφαλίζει ότι τα δέντρα του δάσους δεν είναι απολύτως συσχετισμένα μεταξύ τους (decoupling/de-correlation). Αν υπήρχε ένα κυρίαρχο χαρακτηριστικό στα δεδομένα, ένα απλό Bagging θα το επέλεγε στη ρίζα όλων των δέντρων, καθιστώντας τα παρόμοια. Με την τυχαία επιλογή m χαρακτηριστικών, εξαναγκάζονται τα δέντρα να εξερευνήσουν διαφορετικές δομές και εναλλακτικούς συνδυασμούς χαρακτηριστικών.

2.5.2 Διαδικασία Πρόβλεψης

Αφού εκπαιδευτούν B πλήρως ανεπτυγμένα δέντρα χωρίς κλάδευση (unpruned trees), η πρόβλεψη για μία νέα άγνωστη εγγραφή x πραγματοποιείται ως εξής:

Για Κατηγοριοποίηση (Classification): Κάθε δέντρο δίνει μία ψήφο για την επικρατέστερη κλάση. Η τελική πρόβλεψη του δάσους προκύπτει από την Πλειοψηφική Ψηφοφορία (Majority Voting):

$$\hat{y} = \arg \max_c \sum_{b=1}^B I(h_b(x) = c) \quad (2.10)$$

Όπου I είναι η συνάρτηση δείκτης που λαμβάνει την τιμή 1 αν η πρόβλεψη του b -οστού δέντρου ισούται με την κλάση c , και 0 διαφορετικά.

Για Παλινδρόμηση (Regression): Η τελική πρόβλεψη είναι ο μέσος όρος των προβλέψεων όλων των δέντρων:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B h_b(x) \quad (2.11)$$

2.5.3 Εκτίμηση Εκτός Δείγματος (Out-of-Bag OOB Error)

Ένα από τα σημαντικότερα πλεονεκτήματα των Τυχαίων Δασών είναι η ενσωματωμένη δυνατότητα αξιολόγησης της ακρίβειας τους χωρίς τη χρήση ξεχωριστού συνόλου επικύρωσης (validation set).

Λόγω της δειγματοληψίας bootstrap, η πιθανότητα μια εγγραφή να μην επιλεγεί σε ένα δείγμα μεγέθους N είναι:

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{N}\right)^N = \frac{1}{e} \approx 0.368$$

(2.12)

Αυτό σημαίνει ότι περίπου το 36.8% των εγγραφών παραμένει εκτός εκπαίδευσης για κάθε δέντρο b . Αυτές οι εγγραφές ονομάζονται Out-of-Bag (OOB) δεδομένα.

Για να υπολογιστεί το σφάλμα OOB του μοντέλου:

1. Για κάθε εγγραφή του αρχικού συνόλου, λαμβάνεται η πρόβλεψη χρησιμοποιώντας μόνο τα δέντρα για τα οποία η εγγραφή αυτή ήταν OOB.
2. Συγκρίνεται η πλειοψηφική πρόβλεψη των OOB δέντρων με την πραγματική ετικέτα y

Το συνολικό OOB σφάλμα αποτελεί αμερόληπτη εκτίμηση του σφάλματος γενίκευσης (generalization error), καθιστώντας περιττή την εφαρμογή διασταυρωμένης επικύρωσης (cross-validation) σε πολλές εφαρμογές.

2.5.4 Σημαντικότητα Μεταβλητών (Variable Importance)

Τα τυχαία Δάση προσφέρουν έναν μηχανισμό της σημαντικότητας των χαρακτηριστικών (Feature Importance) μέσω των OOB δεδομένων:

1. Υπολογίζεται η ακρίβεια πρόβλεψης στα OOB δεδομένα.
2. Τυχαία διαταράσσονται(ανακατεύονται) οι τιμές του χαρακτηριστικού X στα OOB δεδομένα, διατηρώντας τα υπόλοιπα χαρακτηριστικά σταθερά.
3. Υπολογίζεται εκ νέου η ακρίβεια πρόβλεψης

Αν η διαταραχή των τιμών του χαρακτηριστικού X προκαλέσει μεγάλη αύξηση στο σφάλμα ταξινόμησης, το χαρακτηριστικό αυτό θεωρείται ιδιαίτερα σημαντικό για το μοντέλο.

2.5.5 Θεωρητική Θεμελίωση και Σύγκλιση

Ο Breiman απέδειξε μέσω του Ισχυρού Νόμου των Μεγάλων Αριθμών ότι καθώς το πλήθος των δέντρων B αυξάνεται στο άπειρο, το σφάλμα γενίκευσης PE^* συγκλίνει σχεδόν βέβαια σε ένα όριο[7]:

$$\lim_{B \rightarrow \infty} PE^* = (P_{\theta}(h(X, \theta) = Y) - \max_{j \neq Y} P_{\theta}(h(X, \theta) = j) < 0) \quad (2.13)$$

Το αποτέλεσμα αυτό εξηγεί θεωρητικά γιατί τα Τυχαία Δάση δεν υποφέρουν από υπερεκπαίδευση (overfitting) όσο αυξάνεται ο αριθμός των δέντρων. Επιπλέον, το άνω όριο του σφάλματος γενίκευσης εξαρτάται από δύο παραμέτρους:

1. Τη μέση ισχύ(strength, s) των μεμονωμένων δέντρων.
2. Τη μέση συσχέτιση(correlation) μεταξύ των raw margin συναρτήσεων των δέντρων.

Το άνω όριο του σφάλματος δίνεται από τη σχέση:

$$PE^* \leq \frac{\bar{\rho}^*(1-s^2)}{s^2} \quad (2.14)$$

Στόχος του αλγορίθμου είναι η ελαχιστοποίηση του λόγου $\frac{\bar{\rho}}{s^2}$, δηλαδή η διατήρηση υψηλής ακρίβειας στα επιμέρους δέντρα (s) με ταυτόχρονη ελαχιστοποίηση της μεταξύ τους συσχέτισης ($\bar{\rho}$).

2.5.6 Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα

Υψηλή Ακρίβεια & Γενίκευση: Ανήκει στους πιο ακριβείς αλγορίθμους ταξινόμησης και παλινδρόμησης, συγκρίσιμος και με μεθόδους Boosting(π.χ. AdaBoost)

Ανθεκτικότητα στην Υπερεκπαίδευση και στον Θόρυβο: Λόγω του νόμου των μεγάλων αριθμών και της τυχαίας επιλογής δείγματος/ χαρακτηριστικών, παρουσιάζει μεγάλη αντοχή σε θορυβώδη δεδομένα και ακραίες τιμές(outliers).

Διαχείριση Μεγάλου Όγκου και διαστάσεων: Αποδίδει εξαιρετικά σε σύνολα δεδομένων με χιλιάδες μεταβλητές εισόδου

Παράλληλη εκπαίδευση: Η εκπαίδευση κάθε δέντρου είναι ανεξάρτητη, καθιστώντας τον αλγόριθμο εύκολα προσαρμόσιμο σε συστήματα πολλαπλών πυρήνων.

Μειονεκτήματα:

Απώλεια Ερμηνευσιμότητας (Black Box): Σε αντίθεση με ένα μοναδικό δέντρο απόφασης, η δομή εκατοντάδων δέντρων δεν οπτικοποιείται εύκολα, καθιστώντας το μοντέλο λιγότερο άμεσα ερμηνεύσιμο.

Υπολογιστικό Κόστος και Μνήμη: Η αποθήκευση και η πρόβλεψη μέσω εκατοντάδων δέντρων απαιτεί περισσότερους πόρους μνήμης και υπολογιστικό χρόνο σε σχέση με απλούστερα μοντέλα.

2.6 Categorical Boosting - CatBoost

Ο αλγόριθμος CatBoost(Categorical Boosting), που αναπτύχθηκε από τους Prokhorenkova et al., αποτελεί μία από τις πλέον εξελιγμένες και αποτελεσματικές υλοποιήσεις της μεθοδολογίας Gradient Boosted Decision Trees (GBDT)[8]. Ενώ οι κλασικές μέθοδοι συλλογικής μάθησης, όπως τα Τυχαία Δάση που αναλύθηκαν προηγουμένως, βασίζονται στην παράλληλη και ανεξάρτητη κατασκευή δέντρων μέσω δειγματοληψίας bootstrap, οι αλγόριθμοι ενίσχυσης βαθμίδας (gradient boosting) λειτουργούν σειριακά, προσθέτοντας διορθωτικά δέντρα που ελαχιστοποιούν το σφάλμα των προηγούμενων στάσεων. Παρά την υψηλή τους ακρίβεια, οι παραδοσιακοί αλγόριθμοι GBDT (όπως το XGBoost και το LightGBM) αντιμετωπίζουν δύο σημαντικά θεωρητικά και πρακτικά προβλήματα: κατά τη διαδικασία της ενίσχυσης, καθώς και την αναποτελεσματική διαχείριση κατηγορικών χαρακτηριστικών υψηλής πληθικότητας(high-cardinality categorical features). Το CatBoost σχεδιάστηκε ειδικά για την αντιμετώπιση αυτών των αδυναμιών μέσω της εισαγωγής της αρχής της διάταξης (ordering principle)[8].

2.6.1 Διαχείριση Κατηγορικών Χαρακτηριστικών και Στατιστικά Στόχου

Σε πολλά προβλήματα της πραγματικότητας, τα δεδομένα εισόδου περιέχουν κατηγορικές μεταβλητές με μεγάλο πλήθος διακριτών τιμών(π.χ. αναγνωριστικά χρηστών, ταχυδρομικοί κώδικες). Η κλασική τεχνική της μετατροπής τους σε δυαδικές μεταβλητές (One-Hot Encoding) οδηγεί σε υπερβολική αύξηση των διαστάσεων του χώρου χαρακτηριστικών, καθιστώντας την εκπαίδευση υπολογιστικά ασύμφορη. Μια δημοφιλής εναλλακτική προσέγγιση είναι η αντικατάσταση κάθε κατηγορίας από ένα αριθμητικό Στατιστικό Στόχου(Target Statistic – TS), το οποίο εκτιμά την αναμενόμενη τιμή της μεταβλητής y εξαρτημένη από την εκάστοτε κατηγορία.

Στην απλούστερη μορφή της (Greedy TS), η τιμή του στατιστικού για μια κατηγορία υπολογίζεται ως ο μέσος όρος των τιμών στόχου των εγγραφών εκπαίδευσης που ανήκουν στην ίδια κατηγορία, ενισχυμένος με έναν όρο ομαλοποίησης (smoothing parameter $\alpha > 0$) και μια εκ των προτέρων πιθανότητα p [8]:

$$\widehat{x}_k^i = \frac{\sum_{j=1}^n I\{x_j^i = x_k^i\} * y_i + a * p}{\sum_{j=1}^n I\{x_j^i = x_k^i\} + \alpha} \quad (2.15)$$

Όπου I είναι η συνάρτηση δείκτη κι x είναι η τιμή του i -οστού κατηγορικού χαρακτηριστικού για την k -οστή εγγραφή. Ωστόσο, η παραπάνω προσέγγιση εισάγει άμεση διαρροή στόχου (target leakage), καθώς η τιμή y της ίδιας της εγγραφής χρησιμοποιείται για τον υπολογισμό του χαρακτηριστικού της x . Αυτό προκαλεί υπό συνθήκη διαταραχή (conditional shift) μεταξύ των κατανομών των δεδομένων εκπαίδευσης και ελέγχου, οδηγώντας το μοντέλο σε σοβαρή υπερεκπαίδευση.

Για την επίλυση αυτού του φαινομένου, το CatBoost εισάγει τα Ordered Target Statistics. Η βασική ιδέα εμπνέεται από τους αλγόριθμους διαδικτυακής μάθησης (online learning). Παράγεται μια τυχαία μετάθεση σ των εγγραφών του συνόλου εκπαίδευσης, η οποία προσομοιώνει μια τεχνητή χρονική ακολουθία. Για τον υπολογισμό του στατιστικού στόχου μιας εγγραφής x_k , ορίζεται το υποσύνολο των εγγραφών που προηγούνται αυτής στη μετάθεση σ , δηλαδή $D_k = \{x_j : \sigma(j) < \sigma(k)\}$. Τότε ο τύπος υπολογισμού διαμορφώνεται ως εξής []:

$$(\widehat{x}_k)^t = \frac{\sum_{x_i \in D_k} I\{x_j^i = x_k^i\} * y_i + a * p}{\sum_{x_i \in D_k} I\{x_j^i = x_k^i\} + \alpha} \quad (2.16)$$

Με αυτόν τον τρόπο εξαλείφεται η διαρροή στόχου, ενώ παράλληλα αξιοποιείται το σύνολο των δεδομένων εκπαίδευσης χωρίς την απώλεια πληροφορίας που παρατηρείται σε άλλες τεχνικές διαχωρισμού (όπως το Holdout TS). Για τη μείωση της διακύμανσης που προκαλείται από μια μεμονωμένη διάταξη, το CatBoost χρησιμοποιεί πολλαπλές ανεξάρτητες τυχαίες μεταθέσεις κατά τη διάρκεια της εκπαίδευσης.

2.6.2 Διαταραχή Πρόβλεψης και Ordered Boosting

Πέρα από την διαρροή στόχου στα κατηγορικά χαρακτηριστικά, οι Prokhorenkova et al. απέδειξε ότι όλοι οι κλασικοί αλγόριθμοι Gradient Boosting υποφέρουν από ένα βαθύτερο θεωρητικό πρόβλημα: τη διαταραχή πρόβλεψης (prediction shift). Σε κάθε βήμα t της κλασικής διαδικασίας boosting, οι κλίσεις (gradients) $g^t(x_k, y_k)$ που χρησιμοποιούνται για την εκπαίδευση του νέου ασθενούς κατηγοριοποιητή h υπολογίζονται πάνω στις ίδιες εγγραφές εκπαίδευσης στις οποίες έχει βασιστεί η κατασκευή του ήδη υπάρχοντος μοντέλου F^{t-1} . Αυτό συνεπάγεται ότι η εξαρτημένη κατανομή των κλίσεων για τα δεδομένα εκπαίδευσης μετατοπίζεται σε σχέση με την αντίστοιχη κατανομή σε άγνωστα δεδομένα ελέγχου, εισάγοντας συστηματικό σφάλμα (bias) στο εκπαιδευμένο μοντέλο [].

Η λύση που προτείνει το CatBoost είναι ο αλγόριθμος Ordered Boosting. Σύμφωνα με αυτή τη μεθοδολογία, διατηρούνται n διαφορετικά υποστηρικτικά μοντέλα $M_1, M_2, \dots, M_n$. Το μοντέλο M_i εκπαιδεύεται χρησιμοποιώντας αποκλειστικά τις πρώτες i εγγραφές σύμφωνα με μια τυχαία μετάθεση σ . Σε κάθε επανάληψη, προκειμένου να υπολογιστεί το υπόλοιπο (residual) ή η κλίση για την j -οστή εγγραφή, χρησιμοποιείται το μοντέλο M_{j-1} , το οποίο έχει εκπαιδευτεί αυστηρά χωρίς τη συμπερίληψη της εγγραφής αυτής. Επειδή η διατήρηση n ξεχωριστών μοντέλων αυξάνει την υπολογιστική και μνημονική πολυπλοκότητα κατά ένα παράγοντα n , οι δημιουργοί του CatBoost ανέπτυξαν μια βελτιστοποιημένη υλοποίηση. Αντί για n μοντέλα, αποθηκεύονται και ενημερώνονται μόλις $O(\log_2 n)$

υποστηρικτικά μοντέλα ανά μετάθεση, μειώνοντας την ασυμπτωτική πολυπλοκότητα στο επίπεδο των κλασσικών αλγορίθμων GBDT[8].

2.6.3 Αρχιτεκτονική Λήψης Αποφάσεων και Συνδυασμοί Χαρακτηριστικών

Σε αντίθεση με άλλες υλοποιήσεις που χρησιμοποιούν ασύμμετρα δέντρα, το CatBoost χρησιμοποιεί ως βασικούς δομικούς λίθους Λησμονούντα Δέντρα Απόφασης (Oblivious Decision Trees). Στα δέντρα αυτά η ίδια συνθήκη διαχωρισμού (splitting criterion) εφαρμόζεται ομοιόμορφα σε όλα τα επίπεδα του δέντρου. Αυτή η ισορροπημένη δομή προσφέρει τρία σημαντικά πλεονεκτήματα:

1. Δρα ως φυσικός μηχανισμός τακτικοποίησης (regularization), αποτρέποντας την υπερεκπαίδευση.
2. Εξασφαλίζει εξαιρετικά γρήγορη εκτέλεση κατά τη φάση της πρόβλεψης, καθώς οι αποφάσεις μπορούν να υπολογιστούν παράλληλα μέσω δυαδικών πράξεων bitwise.
3. Προσφέρει σταθερότητα στην αξιολόγηση της σημαντικότητας των χαρακτηριστικών.

Επιπλέον, το CatBoost εισάγει έναν αυτόματο μηχανισμό κατασκευής συνδυασμών κατηγορικών χαρακτηριστικών (Feature Combinations). Καθώς η χειροκίνητη δημιουργία συνδυασμών όλων των κατηγορικών μεταβλητών οδηγεί σε εκθετική αύξηση της πολυπλοκότητας, το CatBoost δημιουργεί συνδυασμούς με άπληστο τρόπο (greedy strategy) κατά τη διάρκεια της κατασκευής του δέντρου. Συγκεκριμένα, συνδυάζει όλα τα κατηγορικά χαρακτηριστικά που έχουν ήδη χρησιμοποιηθεί σε προηγούμενους διαχωρισμούς του τρέχοντος δέντρου με όλα τα υπόλοιπα κατηγορικά χαρακτηριστικά του συνόλου δεδομένων, μετατρέποντάς τα άμεσα σε Ordered TS.

2.6.4 Σύγκριση και Πλεονεκτήματα

Η θεωρητική ανωτερότητα του CatBoost επιβεβαιώνεται από ευρέα πειραματικά δεδομένα σε ποικίλα σύνολα δεδομένων ταξινόμησης και παλινδρόμηση. Σε σύγκριση με καθιερωμένα εργαλεία όπως το XGBoost και το LightGBM, το CatBoost επιτυγχάνει σταθερά χαμηλότερο σφάλμα ταξινόμησης (zero-one loss) και βέλτιστες τιμές συνάρτησης απώλειας (logloss), ιδιαίτερα σε σύνολα δεδομένων που περιέχουν έντονο κατηγορικό στοιχείο. Επιπλέον, λόγω της εξάλειψης της διαταραχής πρόβλεψης, παρουσιάζει εξαιρετική απόδοση ακόμα και σε μικρού μεγέθους σύνολα δεδομένων εκπαίδευσης, στα οποία το συστηματικό σφάλμα των κλασσικών αλγορίθμων ενίσχυσης είναι εντονότερο.

Συνοψίζοντας, το CatBoost αντιπροσωπεύει την εξέλιξη των αλγορίθμων συλλογικής μάθησης βασισμένων σε δέντρα αποφάσεων. Συνδυάζοντας την αυστηρή θεωρητική θεμελίωση για την αποφυγή της διαρροής στόχου με υψηλά βελτιστοποιημένες υπολογιστικές τεχνικές καθίσταται ένα ισχυρό εργαλείο για την επίλυση σύνθετων προβλημάτων μηχανικής μάθησης σε ετερογενή δεδομένα[8].

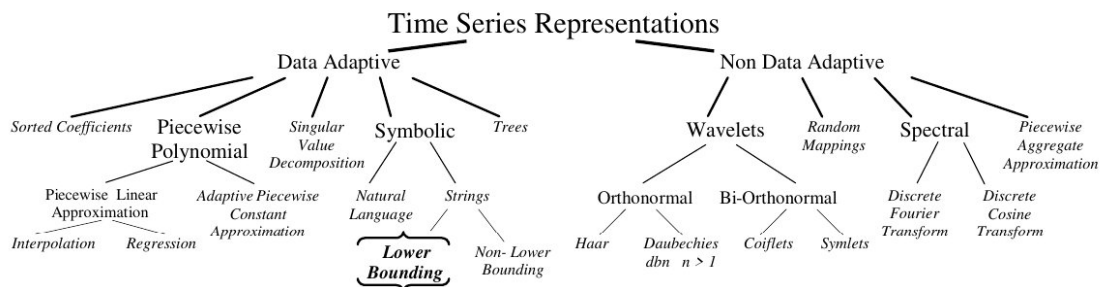
2.7 Επίλογος

Στο παρόν κεφάλαιο εξετάστηκε η κατηγοριοποίηση ως βασική εργασία μάθησης, αναλύοντας τόσο τα γεωμετρικά μέτρα ομοιότητας (Ευκλείδεια, Manhattan, Minkowski) όσο και τη λειτουργία διακριτών κατηγοριοποιητών: από τον οκνηρό αλγόριθμο K-NN και τα μεμονωμένα Δέντρα Απόφασης (CART), έως προηγμένους μεθόδους συλλογικής μάθησης όπως τα Τυχαία Δάση (Random Forests) και το CatBoost. Παρά τη θεωρητική ισχύ αυτών των μοντέλων, η άμεση εφαρμογή τους σε ανεπεξέργαστα σήματα χρονοσειρών συχνά προσκρούει στο υψηλό υπολογιστικό κόστος και στον θόρυβο. Έτσι, στο Κεφάλαιο 3 παρουσιάζονται οι σημαντικότερες τεχνικές αναπαράστασης και μείωσης διαστάσεων χρονοσειρών, οι οποίες στοχεύουν στη συμπίεση των δεδομένων χωρίς απώλεια κρίσιμης πληροφορίας.

Κεφάλαιο 3ο: Μέθοδοι Αναπαράστασης Χρονοσειρών

3.1 Εισαγωγή

Σε προηγούμενο κεφάλαιο αναλύθηκαν τα προβλήματα που προκαλούνται στην Εξόρυξη δεδομένων όταν ένα σύνολο δεδομένων που χρησιμοποιείται έχει πολύ μεγάλο αριθμό διαστάσεων. Όπως φάνηκε, οι χρονοσειρές είναι μια μορφή δεδομένων που εκ φύσεως πάσχει από το πρόβλημα των υπερβολικά πολλών διαστάσεων. Για αυτό τον λόγο έχουν αναπτυχθεί μέθοδοι που εφαρμόζονται στα δεδομένα με σκοπό τη μείωση των διαστάσεων τους. Στο διάγραμμα 7[9] βλέπουμε διάφορες μεθόδους αναπαράστασης χρονοσειρών που έχουν προταθεί στο πεδίο της Εξόρυξης Δεδομένων για να καταπολεμήσουν την “κατάρα των πολλών διαστάσεων”.



Σχήμα 7: Μέθοδοι Αναπαράστασης Χρονοσειρών

Όπως παρατηρούμε στον διάγραμμα, οι τεχνικές αναπαράστασης χρονοσειρών ταξινομούνται σε δύο θεμελιώδεις κατηγορίες, ανάλογα με το αν η διαδικασία μετασχηματισμού εξαρτάται από τα ίδια τα δεδομένα ή όχι:

1. Προσαρμοσμένες στα δεδομένα (Data Adaptive): Σε αυτή την κατηγορία οι παράμετροι της αναπαράστασης μεταβάλλονται ανάλογα με τα χαρακτηριστικά της εκάστοτε χρονοσειράς (π.χ. τοπικά μέγιστα/ ελάχιστα, κατανομή τιμών). Εδώ συναντάμε προσεγγίσεις όπως η Τμηματική Πολυωνμική (Piecewise Polynomial) στην οποία ανήκει η Τμηματική Γραμμική Προσέγγιση(PLA), καθώς και οι Συμβολικές αναπαραστάσεις(Symbolic).
2. Μη προσαρμοσμένες στα δεδομένα(Non Data Adaptive): Οι μέθοδοι αυτές εφαρμόζουν τον ίδιο καθολικό μετασχηματισμό σε όλες τις χρονοσειρές, ανεξάρτητα από τη μορφολογία τους. Χαρακτηριστικά παραδείγματα αποτελούν οι φασματικές αναλύσεις(Spectral) όπως ο Διακριτός Μετασχηματισμός Fourier(DFT), οι Κυματομορφές (Wavelets), αλλά και η Τμηματική Συσσωρευτική Προσέγγιση(Piecewise Aggregate Approximation – PAA).

Μια ιδανική μέθοδος αναπαράστασης οφείλει να μειώνει δραστικά τον όγκο των δεδομένων, διατηρώντας παράλληλα τα κρίσιμα χαρακτηριστικά τους (όπως η γενική τάση και το σχήμα) και επιτρέποντας τον υπολογισμό αποστάσεων που εγγυώνται θεωρητικά ότι δεν θα χαθούν πραγματικά όμοια ζεύγη (ιδιότητα του Lower Bounding, η οποία επισημαίνεται και στο σχήμα, θα αναλυθεί σε επόμενο κεφάλαιο[9][10][11]. Στις υπό-ενότητες που ακολουθούν, αναλύονται εκτενέστερα τρεις από τις πιο διαδεδομένες μεθόδους μείωσης διαστάσεων: ο μετασχηματισμός DFT, η προσέγγιση PLA και η τεχνική PAA.

3.2 Διακριτός Μετασχηματισμός Φουριέ (Discrete Fourier Transform - DFT)

Ο Διακριτός Μετασχηματισμός Φουριέ (Discrete Fourier Transform) αποτελεί μία από τις πιο διαδεδομένες και θεμελιώδεις τεχνικές αναπαράστασης χρονοσειρών[10]. Ο DFT είναι μια μέθοδος μη προσαρμοσμένη στα δεδομένα που εφαρμόζεται(non data adaptive). Η βασική του αρχή στηρίζεται στη

μεταφορά ενός σήματος από το πεδίο του χρόνου στο πεδίο των συχνοτήτων, αναλύοντας τη χρονοσειρά σε ένα άθροισμα ημιτονοειδών και συνημιτονοειδών κυμάτων διαφορετικών συχνοτήτων και πλατών[10][11].

3.2.1 Μαθηματικός Ορισμός

Για μια διακριτή χρονοσειρά $x = [x_0, x_1, \dots, x_{n-1}]$ μήκους n , ο DFT παράγει μια ακολουθία n μιγαδικών αριθμών $X = [X_0, X_1, \dots, X_{n-1}]$, όπου κάθε συντελεστής X_F υπολογίζεται από τον τύπο[10]:

$$X_F = \frac{1}{\sqrt{n}} \cdot \sum_{i=0}^{n-1} x_i \exp\left(-j \frac{2\pi Fi}{n}\right), F = 0, 1, \dots, n-1 \quad (3.1)$$

Όπου $j = \sqrt{-1}$ είναι η φανταστική μονάδα. Ο αρχικός σηματοφορέας x μπορεί να ανακτηθεί πλήρως χωρίς απώλεια πληροφορίας μέσω του Αντίστροφου Διακριτού Μετασχηματισμού Φουριέ(IDFT)[10]:

$$x_i = \frac{1}{\sqrt{n}} \cdot \sum_{F=0}^{n-1} X_F \exp\left(j \frac{2\pi Fi}{n}\right), i = 0, 1, \dots, n-1 \quad (3.2)$$

3.2.2 Μηχανισμός Μείωσης Διαστάσεων και Αποκοπή Συντελεστών

Η μείωση διαστάσεων μέσω του DFT επιτυγχάνεται με τη διαδικασία της αποκοπής (truncation). Αντί να διατηρηθούν και οι n συντελεστές Fourier, επιλέγονται μόνο οι πρώτοι f συντελεστές ($f \ll n$), οι οποίοι αντιστοιχούν στις χαμηλότερες συχνότητες του σήματος.

Η προσέγγιση αυτή βασίζεται στο γεγονός ότι οι περισσότερες πραγματικές χρονοσειρές(π.χ. Τιμές μετοχών, μετεωρολογικά δεδομένα, αισθητήρες) εμπίπτουν στην κατηγορία του “χρωματισμένου θορύβου(colored noise)”[10]. Σε αυτές τις χρονοσειρές, το φάσμα ενέργειας είναι συγκεντρωμένο κυρίως στις χαμηλές συχνότητες, ενώ οι υψηλές συχνότητες αντιπροσωπεύουν κυρίως τον τυχαίο θόρυβο. Συνεπώς, κρατώντας τους πρώτους f συντελεστές (που αντιστοιχούν σε $2f$ πραγματικούς αριθμούς λόγω του πραγματικού και φανταστικού μέρους των μιγαδικών), διατηρείται η πλειονότητα της ενέργειας και της συνολικής μορφολογίας της χρονοσειράς.

Διατήρηση Αποστάσεων (Parseval's Theorem and Lower Bounding)

Ένα από τα σημαντικότερα πλεονεκτήματα της χρήσης του DFT στην εξόρυξη δεδομένων χρονοσειρών είναι η διατήρηση της Ευκλείδειας απόστασης. Σύμφωνα με τον Θεώρημα του Parseval, ο DFT διατηρεί την συνολική ενέργεια του σήματος[10][11]:

$$\sum_{i=0}^{n-1} |x_i|^2 = \sum_{F=0}^{n-1} |X_F|^2 \quad (3.3)$$

Επειδή ο μετασχηματισμός Φουριέ είναι γραμμικός και ορθογώνιος, συνεπάγεται ότι η Ευκλείδεια απόσταση μεταξύ δύο χρονοσειρών x και y στο πεδίο του χρόνου είναι απολύτως ίση με την Ευκλείδεια απόσταση των συντελεστών τους στο πεδίο των συχνοτήτων[10]:

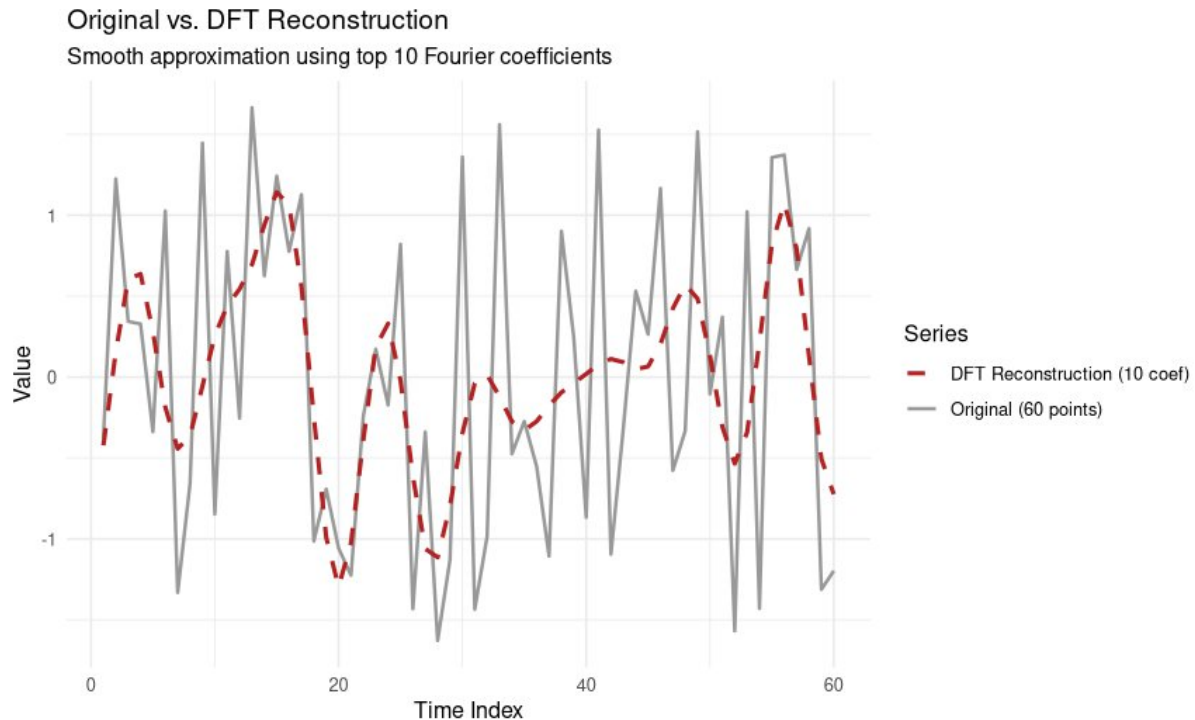
$$D_{object}(x, y) = D_{feature}(X, Y) \quad (3.4)$$

Όταν αποκόπτουμε το φάσμα κρατώντας μόνο f συντελεστές, η υπολογιζόμενη απόσταση στο χώρο των f χαρακτηριστικών υποεκτιμά την πραγματική απόσταση στο χώρο των n διαστάσεων[10]:

$$D_{feature}(F(x), F(y)) \leq D_{object}(x, y) \quad (3.4)$$

Η ιδιότητα αυτή (Lower Bounding Property) είναι καθοριστικής σημασίας για την ευρετηρίαση και την αναζήτηση ομοιότητας (similarity search), καθώς εγγυάται τη μη ύπαρξη ψευδώς αρνητικών αποτελεσμάτων (no false dismissals)[10][11]. Οποιοδήποτε ζεύγος χρονοσειρών είναι όμοιο στο αρχικό

πεδίο θα ανακτηθεί υποχρεωτικά από το ευρετήριο στο πεδίο των χαρακτηριστικών, ενώ τυχόν ψευδώς θετικά(false alarms) απορρίπτονται εύκολα σε δεύτερο στάδιο[10].



Σχήμα 8: Εφαρμογή μεθόδου DFT σε μια χρονοσειρά

3.3 Piecewise Linear Approximation

Η Τμηματική Γραμμική Προσέγγιση (Picewise Linear Approximation – PLA) αποτελεί μία από τις πιο διδεδομένες τεχνικές αναπαράστασης χρονοσειρών που ανήκουν στην κατηγορία των προσαρμοσμένων στα δεδομένα (data adaptive) μεθόδων. Σε αντίθεση με ολικές μετασχηματιστικές τεχνικές (όπως ο DFT), οι οποίες εφαρμόζουν μια καθολική συνάρτηση σε ολόκληρο το σχήμα, η PLA χωρίζει τη χρονοσειρά σε επιμέρους χρονικά τμήματα (segments) και προσεγγίζει τα δεδομένα σε κάθε τμήμα χρησιμοποιώντας ευθύγραμμα τμήματα.

Η βασική φιλοσοφία της PLA στηρίζεται στο ότι οι περισσότερες πραγματικές χρονοσειρές παρουσιάζουν τοπικές τάσεις (trends), ακρότατα και μεταβολές, τα οποία μπορούν να αποδοθούν με εξαιρετική ακρίβεια από μια σειρά συνδεδεμένων ή διακεκριμένων ευθύγραμμων τμημάτων.

3.3.1

Έστω μια χρονοσειρά $X=[x_1, x_2, \dots, x_n]$ μήκους n . Η μέθοδος PLA χωρίζει το διάστημα της χρονοσειράς σε K υποδιαστήματα ($K \ll n$). Για κάθε τμήμα $i \in [1, K]$, κατασκευάζεται μια γραμμική συνάρτηση της μορφής:

$$(t) \hat{x} = a_i t + b_i \quad (3.5)$$

Όπου a_i είναι η κλίση (slope) και b_i η σταθερά (intercept) της ευθείας στο i -οστό τμήμα. Η εκτίμηση των παραμέτρων a και b πραγματοποιείται συνήθως με δύο κύριες προσεγγίσεις:

1. Γραμμική Παρεμβολή (Linear Interpolation): Η ευθεία γραμμή ορίζεται συνδέοντας απευθείας το πρώτο και το τελευταίο σημείο του κάθε τμήματος. Είναι υπολογιστικά ταχεία, αλλά μπορεί να είναι ευαίσθητη σε τοπικό θόρυβο.
2. Γραμμική Παλινδρόμηση (Linear Regression/ Least Squares): Οι παράμετροι υπολογίζονται μέσω της μεθόδου των Ελαχίστων Τετραγώνων (Ordinary Least Squares – OLS), ελαχιστοποιώντας το τετραγωνικό σφάλμα μεταξύ των πραγματικών τιμών της χρονοσειράς και της προσεγγιστικής ευθείας στο εκάστοτε τμήμα[12].

Με αυτόν τον τρόπο, αντί για την αποθήκευση των n αρχικών σημείων, η χρονοσειρά αναπαρίσταται συμπυκνωμένα από τα σημειακά όρια των τμημάτων (slit points) και τις αντίστοιχες παραμέτρους της ευθείας, επιτυγχάνοντας σημαντική μείωση διαστάσεων.

3.3.2 Αλγόριθμοι Κατάτμησης

Η αποτελεσματικότητα της PLA εξαρτάται από τον άμεσο τρόπο με τον οποίο επιλέγονται τα όρια των τμημάτων. Στην βιβλιογραφία χρησιμοποιούνται τρεις κύριες κατηγορίες αλγορίθμων κατάτμησης:

1. Top-Down: Ο αλγόριθμος ξεκινά θεωρώντας ολόκληρη τη χρονοσειρά ως ένα ενιαίο τμήμα. Στη συνέχεια, διαιρεί επαναληπτικά το τμήμα στο σημείο που παρουσιάζει το μεγαλύτερο σφάλμα προσέγγισης, μέχρι να καταλήξει σε έναν προκαθορισμένο αριθμό τμημάτων K ή μέχρι το συνολικό σφάλμα να πέσει κάτω από ένα κατώφλι ϵ .
2. Bottom-Up: Ξεκινά διαιρώντας τη χρονοσειρά σε $n/2$ μικρά τμήματα. Σε κάθε βήμα, συγχωνεύει γειτονικά τμήματα που προκαλούν την ελάχιστη αύξηση στο σφάλμα προσέγγισης, συνεχίζοντας έως ότου ικανοποιηθεί το κριτήριο τερματισμού.
3. Sliding Window: Ένα παράθυρο μετακινείται κατά μήκος της χρονοσειράς αυξάνοντας το μέγεθος του όσο το σφάλμα της γραμμικής προσαρμογής παραμένει κάτω από το όριο ϵ . Μόλις ξεπεραστεί το όριο, δημιουργείται ένα νέο τμήμα.

Παράλληλα, υπάρχουν και επαναληπτικοί αλγόριθμοι βελτιστοποίησης (iterative optimization algorithms), όπως αυτός των Pitukhin και Shchegoleva, οι οποίοι προσαρμόζουν δυναμικά τα σημεία τομής (split points) υπολογίζοντας τις διασταυρώσεις των γραμμών παλινδρόμησης γειτονικών τμημάτων. Οι αλγόριθμοι αυτοί επιτυγχάνουν ταχεία σύγκλιση (π.χ. Εντός 30 επαναλήψεων) και επιτρέπουν την εξαιρετική διατήρηση των αιχμών (peaks) και των οξέων μεταβολών της χρονοσειράς, κάτι ιδιαίτερα σε ιατρικά ή φυσικά δεδομένα υψηλής μεταβλητότητας[12].

3.3.3 Πλεονεκτήματα, Μειονεκτήματα και Ιδιότητα Lower Bounding

Πλεονεκτήματα:

1. Προσαρμογή στα δεδομένα (Data adaptive): Προσαρμόζεται τοπικά στη μορφολογία του σήματος, διατηρώντας τα κρίσιμα μορφολογικά χαρακτηριστικά, όπως αιχμές (peaks), κοιλώματα και απότομες τάσεις, τα οποία συχνά εξομαλύνονται από μεθόδους όπως ο DFT[10].
2. Δυνατότητα Ερμηνείας (Interpretability): Η αναπαράσταση αποτελείται από φυσικά ευθύγραμμα τμήματα, επιτρέποντας την εύκολη οπτικοποίηση και την εξαγωγή κανόνων (π.χ. αυξητική τάση στο διάστημα t_1 έως t_2).
3. Ευελιξία: Επιτρέπει τη χρήση τμημάτων μεταβλητού μήκους, εστιάζοντας περισσότερες διαστάσεις (περισσότερα τμήματα) στις περιοχές με υψηλή μεταβλητότητα.

Μειονεκτήματα και Lower Bounding:

1. Υπολογιστικό κόστος κατάτμησης: Η διαδικασία εύρεσης των βέλτιστων σημείων κατάτμησης είναι υπολογιστικά πιο απαιτητική σε σύγκριση με μη προσαρμοσμένους μετασχηματισμούς.
2. Υπολογισμός αποστάσεων: Σε αντίθεση με τον μετασχηματισμό Φουριέ (όπου το θεώρημα του Parseval εγγυάται άμεσα την ισότητα των Ευκλείδειων αποστάσεων), η κατασκευή μιας αυστηρής συνάρτησης κάτω ορίου (Lower Bounding distance function) στην PLA απαιτεί

ειδικούς μαθηματικούς ορισμούς (όπως η μέτρηση αποστάσεων μεταξύ των bounding envelopes των τμημάτων) για να διασφαλιστεί η απουσία ψευδώς αρνητικών αποτελεσμάτων (no false dismissals) κατά την αναζήτηση ομοιότητας.

3.4 Piecewise Aggregate Approximation

Η Τμηματική Αθροιστική Προσέγγιση (Piecewise Aggregate Approximation – PAA) είναι άλλη μια από τις πιο γνωστές μεθόδους αναπαράστασης και μείωσης διαστάσεων χρονοσειρών. Θεωρείται μη προσαρμοσμένη στα δεδομένα (non data adaptive). Αναπτύχθηκε από τους Keogh, Chakrabarti, Pazzani και Mehrotra. Ο βασικός στόχος της μεθόδου όπως και όλων των μεθόδων μείωσης διαστάσεων είναι η καταπολέμηση της κατάρας των πολλών διαστάσεων σε μεγάλες βάσεις δεδομένων που περιλαμβάνουν χρονοσειρές[11].

Η βασική ιδέα της PAA στηρίζεται στην παρατήρηση ότι στις περισσότερες πραγματικές χρονοσειρές εμφανίζεται ισχυρή συσχέτιση (auto-correlation) μεταξύ γειτονικών χρονικών σημείων. Συνεπώς, μια τοπική περιοχή (περίοδος/παράθυρο) δεδομένων μπορεί να προσεγγιστεί με υψηλή ακρίβεια απλά και μόνο από τη μέση τιμή των σημείων που την απαρτίζουν[11].

3.4.1 Μαθηματικός Ορισμός και Μηχανισμός

Έστω μια αρχική χρονοσειρά $X = [x_1, x_2, \dots, x_n]$ μήκος n διαστάσεων. Στόχος της PAA είναι η υποβίβαση των διαστάσεων από n σε N , όπου $N \ll n$ [11]. Για λόγους απλότητας των μαθηματικών εκφράσεων, υποτίθεται συνήθως ότι το N είναι διαιρέτης του n , ορίζοντας έτσι ένα σταθερό μέγεθος παραθύρου/πλαισίου (frame) ίσο με $w = \frac{n}{N}$ [11].

Η αναπαράσταση της χρονοσειράς στον μειωμένο χώρο N διαστάσεων συμβολίζεται ως $\overline{X} = [x_1, x_2, \dots, x_N]$, όπου το i -οστό στοιχείο x_i , υπολογίζεται ως ο αριθμητικός μέσος όρος των τιμών της χρονοσειράς που εμπίπτουν στο i -οστό πλαίσιο[11]:

$$\overline{x}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j \quad (3.6)$$

Γεωμετρικά, ο μετασχηματισμός αυτός ισοδυναμεί με την προσέγγιση της αρχικής καμπύλης από μια σταθερή τμηματική συνάρτηση (piecewise constant approximation), δηλαδή μια κλιμακωτή γραμμή που αποτελείται από N οριζόντια ευθύγραμμα τμήματα ίσου μήκους, όπου το ύψος κάθε τμήματος αντιστοιχεί στη μέση τιμή των σημείων του παραθύρου[11].

3.4.2 Υπολογιστική Πολυπλοκότητα και Κατασκευή Ευρετηρίου

Ένα από τα σημαντικότερα πλεονεκτήματα της PAA έναντι ανταγωνιστικών μεθόδων είναι η εξαιρετικά χαμηλή υπολογιστική της πολυπλοκότητα:

1. Γραμμική Κλιμάκωση $O(n)$: Για μεμονωμένη χρονοσειρά μήκους n , ο υπολογισμός των N μέσων όρων απαιτεί χρόνο $O(n)$, γεγονός που καθιστά την PAA σημαντικά ταχύτερη από τον μετασχηματισμό DFT ($O(n \log_2 n)$) και την SVD (Singular value Decomposition ($O(n^3)$) ή $O(nN)$ με προσεγγιστικές μεθόδους[11].
2. Ταχεία Ενημέρωση Ολισθαίνοντος Παραθύρου (Sliding Window): Στο πρόβλημα της αναζήτησης υποακολουθιών (sub-sequence matching), όπου ένα παράθυρο μήκους n ολισθαίνει κατά μήκος μιας μεγάλης ακολουθίας m σημείων, η PAA δεν απαιτεί τον επαναυπολογισμό του τύπου του μέσου όρου από την αρχή σε κάθε βήμα. Αντίθετα, η μέση τιμή του κάθε πλαισίου ενημερώνεται σε σταθερό χρόνο $O(1)$, αφαιρώντας την επίδραση του νέου σημείου που εισέρχεται[11].

3.4.3 Συνάρτηση Αποστάσεων και Ιδιότητα Lower Bounding

Για να διασφαλιστεί η ορθότητα της αναζήτησης σε δομές ευρετηρίασης (Spatial Access Methods – SAMs, όπως R-trees, K-d trees, ή Hybrid trees) μέσω του πλαισίου GEMINI, η συνάρτηση απόστασης στον χώρο των μειωμένων διαστάσεων οφείλει να ικανοποιεί αυστηρά την ιδιότητα του κάτω ορίου (Lower Bounding)[9][10][11]:

$$DR(\bar{X}, \bar{Y}) \leq D(X, Y) \quad (3.7)$$

Για την PAA, η υπολογιζόμενη τροποποιημένη Ευκλείδεια απόσταση $DR(\bar{X}, \bar{Y})$ μεταξύ δύο αναπαραστάσεων X και Y υπολογίζεται ως[9][11]:

$$DR(\bar{X}, \bar{Y}) = \sqrt{\frac{n}{N}} * \sqrt{\sum_{i=1}^N (\bar{x}_i, \bar{y}_i)^2} \quad (3.8)$$

Οι Keogh et al, απέδειξαν μαθηματικά ότι η συνάρτηση $DR(\bar{X}, \bar{Y})$ υποεκτιμά πάντα (ή στην καλύτερη περίπτωση είναι ίση με) την πραγματική Ευκλείδεια απόσταση $D(X, Y)$ των αρχικών χρονοσειρών. Η τήρηση αυτής της ανισότητας εγγυάται απόλυτη ακρίβεια χωρίς απώλεια σχετικών δεδομένων (zero false dismissals): κανένα ζεύγος χρονοσειρών που είναι πραγματικά όμοιο στο αρχικό πεδίο δεν πρόκειται να απορριφθεί από το ευρετήριο[9][11].

3.4.4 Ισοδυναμία με τα Κύματα Haar (Haar Wavelets)

Ένα εξαιρετικά σημαντικό θεωρητικό εύρημα που αποδείχθηκε στην εργασία των Keogh et al είναι η πλήρης ισοδυναμία της PAA με τον διακριτό Μετασχηματισμό Κυματιδίων Haar (Discrete Wavelet Transform – DWT). Συγκεκριμένα, όταν ο αριθμός των διαστάσεων N επιλέγεται να είναι δύναμη του 2 (δηλαδή $N = 2^k$)[11]:

1. Η ανακατασκευή του σήματος που προκύπτει από τους συντελεστές PAA είναι απολύτως ταυτόσημη με την ανακατασκευή που προκύπτει από τους αντίστοιχους συντελεστές Haar.
2. Οι εκτιμώμενες αποστάσεις κάτω ορίου (Lower Bound distances) στον χώρο των χαρακτηριστικών είναι αριθμητικά ίσες και στις δύο μεθόδους.

Συνεπώς, η PAA προσφέρει όλη την αποδεδειγμένη ισχύ απόρριψης μη ομοίων δεδομένων (pruning power) του μετασχηματισμού Haar, απαλλαγμένη όμως από τους αυστηρούς περιορισμούς του.

3.4.5 Πλεονεκτήματα της PAA

Η PAA παρουσιάζει μια σειρά από καθοριστικά πλεονεκτήματα που την καθιστούν μία από τις πιο ευέλικτες τεχνικές στην εξόρυξη δεδομένων[11]:

1. Ευελιξία Μήκους Ερωτημάτων(Queries of Various Lengths): Οι μετασχηματισμοί DFT και DWT απαιτούν αυστηρά προκαθορισμένο μήκος χρονοσειρών (με τον DWT να περιορίζεται αποκλειστικά σε δυνάμεις του 2). Αντίθετα, ένα ευρετήριο δομημένο με PAA μπορεί να εξυπηρετήσει ερωτήματα μικρότερου ή μεγαλύτερου μήκους από το μήκος n για το οποίο κατασκευάστηκε το ευρετήριο, χωρίς να απαιτείται ανακατασκευή της δομής.
2. Υποστήριξη Σταθμισμένης Ευκλείδειας Απόστασης (Weighted Euclidean Distance): Επειδή κάθε συντελεστής της PAA αντιπροσωπεύει μια συγκεκριμένη, τοπική χρονική περιοχή (σε αντίθεση με τους συντελεστές DFT/SVD που είναι ολικοί και επεκτείνονται σε ολόκληρο το μήκος του σήματος), η PAA επιτρέπει την εισαγωγή βαρών w_i σε κάθε τμήμα. Αυτό καθιστά εφικτή τη χρήση σταθμισμένων μέτρων απόστασης (DW), δίνοντας τη δυνατότητα στον

χρήστη να εστιάσει σε συγκεκριμένα μορφολογικά χαρακτηριστικά(π.χ. Αιχμές ή κοιλώματα) μέσω τεχνικών όπως η ανάδραση σχετικότητας (relevance feedback).

3. Δυναμικές Ενημερώσεις σε Σταθερό Χρόνο: Η εισαγωγή ή διαγραφή νέων χρονοσειρών στο ευρετήριο πραγματοποιείται τοπικά και ακαριαία, σε αντίθεση με μεθόδους όπως η SVD, η οποία είναι ολική και απαιτεί πλήρη εκ νέου υπολογισμό, ολόκληρου του ευρετηρίου σε κάθε αλλαγή των δεδομένων.
4. Απλότητα Υλοποίησης: Η μέθοδος είναι εξαιρετικά απλή στην κατανόηση, τον προγραμματισμό και την ενσωμάτωσή της σε πολύπλοκα συστήματα εξόρυξης δεδομένων.

Όλες οι παραπάνω μέθοδοι αναπαράστασης έχουν ένα κοινό. Τα χρονοσειρές που προκύπτουν διατηρούν τον αρχικό τύπο των δεδομένων τους. Δηλαδή οι μειωμένες χρονοσειρές εξακολουθούν να περιλαμβάνουν αριθμητικά δεδομένα. Στο επόμενο κεφάλαιο θα αναλύσουμε την μέθοδο Συμβολικής Αθροιστικής Προσέγγισης (Symbolic Aggregate Approximation – SAX) η οποία έχει σαν αποτέλεσμα, η χρονοσειρά που προκύπτει από την μετατροπή να μην έχει πλέον αριθμητικά δεδομένα. Γίνεται μια συμβολική αναπαράσταση των αρχικών δεδομένων.

3.5 Επίλογος

Στο τρίτο κεφάλαιο παρουσιάστηκε η ταξινόμηση των τεχνικών αναπαράστασης χρονοσειρών σε προσαρμοσμένες και μη προσαρμοσμένες στα δεδομένα, με αναλυτική παρουσίαση του Διακριτού Μετασχηματισμού Φουριέ(DFT), της Τμηματικής Γραμμικής Προσέγγισης (PLA) και της Τμηματικής Αθροιστικής Προσέγγισης (PAA). Κοινό γνώρισμα των παραπάνω μεθόδων αποτελεί η διατήρηση των δεδομένων σε συνεχή αριθμητική μορφή. Στο επόμενο κεφάλαιο (Κεφάλαιο 4) πραγματοποιείται το επόμενο λογικό βήμα: η διεξοδική ανάλυση της Συμβολικής Αθροιστικής Προσέγγισης (SAX), η οποία μετασχηματίζει τις χρονοσειρές σε διακριτές συμβολοσειρές εξασφαλίζοντας παράλληλα τη μαθηματική ιδιότητα του κάτω ορίου (lower bounding).

Κεφάλαιο 4ο: Συμβολική Αθροιστική Προσέγγιση

Σε αυτό το κεφάλαιο θα αναλυθεί εκτενώς η μέθοδος συμβολικής αναπαράστασης χρονοσειρών SAX.

4.1 Κίνητρο της SAX

Όπως αναλύθηκε σε προηγούμενα κεφάλαια, το κύριο εμπόδιο στην επεξεργασία και ανάλυση χρονοσειρών μεγάλου όγκου είναι η “κατάρρα των πολλών διαστάσεων”. Η πλειονότητα των πραγματικών χρονοσειρών αποτελείται από συνεχείς, πραγματικές τιμές (real-valued time series) που καταγράφονται σε βάθος χρόνου. Παράλληλα, στον κλάδο της επιστήμης δεδομένων παρατηρείται μια αλματώδης ανάπτυξη αλγορίθμων για τη διαχείριση συνεχών ροών δεδομένων (data streams).

Ωστόσο, δημιουργείται ένα παράδοξο: οι περισσότεροι αλγόριθμοι που αναπτύχθηκαν για τη διαχείριση ροών δεδομένων προϋποθέτουν εξ ορισμού ότι τα εισαγόμενα δεδομένα είναι διακριτά (discrete), όπως συμβαίνει στα αλφαριθμητικά κείμενα ή στα δεδομένα αλληλουχιών DNA. Αντίθετα, οι χρονοσειρές είναι συνεχείς, πραγματικές συναρτήσεις. Η μετατροπή των πραγματικών τιμών μιας χρονοσειράς σε μια διακριτή, συμβολική αναπαράσταση (symbolic representation) αποτελεί τη γέφυρα για την εφαρμογή αυτών των προηγμένων αλγορίθμων.

4.1.1 Μειονεκτήματα Προγενέστερων Συμβολικών Αναπαραστάσεων

Η ιδέα της συμβολικής αναπαράστασης δεν είναι καινούρια, καθώς στη βιβλιογραφία έχουν προταθεί δεκάδες τεχνικές διακριτοποίησης[13][14]. Παρ’όλα αυτά, η ευρεία υιοθέτησή τους στην εξόρυξη δεδομένων εμποδιζόταν από τρία δομικά μειονεκτήματα:

1. Δεν επιτύγχαναν μείωση των διαστάσεων. Ο αριθμός των διαστάσεων από το αρχικό σήμα δεν άλλαζε. Αν και οι παλαιότερες μέθοδοι μείωναν τον απαιτούμενο χώρο αποθήκευσης σε επίπεδο bits (λόγω της χρήσης ακεραίων ή χαρακτήρων αντί για αριθμούς κινητής υποδιαστολής), διατηρούσαν τον ίδιο ακριβώς αριθμό διαστάσεων με το αρχικό σήμα. Καθώς οι αλγόριθμοι εξόρυξης κλιμακώνονται εκθετικά με βάση τον αριθμό των διαστάσεων, οι τεχνικές αυτές παρέμεναν ανεπαρκείς για υψηλές διαστάσεις.
2. Οι συναρτήσεις απόστασης στο συμβολικό πεδίο δεν είχαν συσχέτιση με την Ευκλείδεια απόσταση των αρχικών χρονοσειρών. Στις υπάρχουσες συμβολικές προσεγγίσεις, οι οριζόμενες συναρτήσεις απόστασης μεταξύ των συμβόλων δεν είχαν ουσιαστική συσχέτιση με την πραγματική Ευκλείδεια απόσταση των αρχικών χρονοσειρών. Δεν υπήρχε καμία εγγύηση ότι η συμβολική υποεκτιμά την πραγματική (Lower Bounding guarantee). Αυτό καθιστούσε αδύνατη τη διασφάλιση της ακρίβειας σε εργασίες αναζήτησης ομοιότητας, προκαλώντας υψηλά ποσοστά της ακρίβειας σε εργασίες αναζήτησης ομοιότητας, προκαλώντας υψηλά ποσοστά ψευδώς αρνητικών αποτελεσμάτων (false dismissals).
3. Απαιτούσαν πρόσβαση σε όλα τα δεδομένα εκ των προτέρων (batch process), εμποδίζοντας τη χρήση σε streaming αλγόριθμους. Οι περισσότερες προγενέστερες τεχνικές απαιτούσαν πρόσβαση στο σύνολο των δεδομένων δεδομένων εκ των προτέρων προκειμένου να καθορίσουν τα όρια διακριτοποίησης. Αυτή η ιδιότητα ακύρωνε πλήρως τη δυνατότητα χρήσης τους σε πραγματικό χρόνο ή σε αλγορίθμους ροών(streaming algorithms).

4.1.2 Η Καινοτομία της SAX

Για την αντιμετώπιση των παραπάνω περιορισμών, οι Lin et al. εισήγαγαν τη μέθοδο Symbolic Aggregate Approximation (SAX)[9]. Η SAX αποτελεί την πρώτη συμβολική αναπαράσταση χρονοσειρών στη βιβλιογραφία που καταφέρνει να συνδυάσει ταυτόχρονα:

- Δραστική Μείωση Διαστάσεων: Μέσω της ενδιάμεσης χρήσης της μεθόδου PAA
- Εγγύηση κάτω Ορίου(Lower Bounding Property): Εισαγωγή μιας συνάρτησης ελάχιστης απόστασης (MINDIST) στο συμβολικό πεδίο, η οποία αποδεδιγμένα υποεκτιμά την Ευκλείδεια απόσταση των αρχικών δεδομένων.
- Δυνατότητα Επεξεργασίας Ροών: Μετατροπές χαμηλής υπολογιστικής πολυπλοκότητας σε πραγματικό χρόνο.

Χάρη στην ιδιότητα του κάτω ορίου, η SAX επιτρέπει την εκτέλεση αλγορίθμων εξόρυξης δεδομένων(όπως η κατηγοριοποίηση, η ομαδοποίηση, η ανίχνευση ανωμαλιών και η αναζήτηση μοτίβων) πάνω σε συμπίεσμένες συμβολικές λέξεις, παράγοντας αποτελέσματα εγγυημένα ισοδύναμα με εκείνα που θα προέκυπταν από την επεξεργασία των αρχικών, ανεπεξέργαστων δεδομένων.

4.2 Αρχιτεκτονική και Στάδια Μετασχηματισμού SAX

Η SAX επιτρέπει τη μετατροπή μιας χρονοσειράς $C = c_1, c_2, \dots, c_n$ μήκους n σε μια συμβολική ακολουθία (λέξη) $\hat{C} = \hat{c}_1, \hat{c}_2, \dots, \hat{c}_w$ μήκους w ($w \ll n$). Το μέγεθος του χρησιμοποιούμενου αλφαβήτου συμβολίζεται με το a , όπου a είναι ένας ακέραιος αριθμός με $a > 2$.

Για την επίτευξη του μετασχηματισμού, η SAX εισάγει μια πρωτότυπη διεργασία δύο σταδίων: χρησιμοποιεί την PAA ως ενδιάμεση αναπαράσταση και στη συνέχεια εφαρμόζει διακριτοποίηση πάνω στους συντελεστές PAA για την παραγωγή των συμβόλων.

4.2.1 Προ-επεξεργασία: Κανονικοποίηση Τυπικής Μορφής (Z-score Normalization)

Πριν από την εφαρμογή οποιουδήποτε μετασχηματισμού, είναι απαραίτητη η κανονικοποίηση κάθε χρονοσειράς του συνόλου δεδομένων. Στην εξόρυξη δεδομένων χρονοσειρών είναι ευρέως αποδεκτό ότι η σύγκριση δύο ακολουθιών χωρίς προηγούμενη εξάλειψη των μετατοπίσεων (offsets) και των διαφορών πλάτους (amplitudes) οδηγεί σε λανθασμένα συμπεράσματα[15].

Για τον σκοπό αυτό, κάθε χρονοσειρά C κανονικοποιείται ώστε να έχει μέση τιμή $\mu = 0$ και τυπική απόκλιση $\sigma = 1$:

$$\mu = \frac{1}{n} \sum_{i=1}^n c_i, \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (c_i - \mu)^2} \quad (4.1)$$

Κάθε στοιχείο c_i της αρχικής χρονοσειράς μετασχηματίζεται σύμφωνα με τον τύπο:

$$\hat{c}_i = \frac{c_i - \mu}{\sigma} \quad (4.2)$$

Εφαρμόζοντας τη Z-normalization, εξασφαλίζεται ότι οι μεταβλητές της χρονοσειράς ακολουθούν μια τυποποιημένη κατανομή, η οποία είναι απαραίτητη για τη φάση της συμβολικής αναπαράστασης.

4.2.2 Ενδιάμεσο Στάδιο: Μείωση διαστάσεων μέσω PAA

Αφού κανονικοποιηθεί η χρονοσειρά C μήκους n , μετασχηματίζεται στον w -διάστατο χώρο PAA, παράγοντας το διάνυσμα $\overline{c} = \overline{c}_1, \overline{c}_2, \dots, \overline{c}_w$

Για τη διαίρεση της χρονοσειράς σε w ισομεγέθη χρονικά πλαίσια (frames), το i -οστό στοιχείο $\overline{c}_i$ του διανύσματος PAA υπολογίζεται ως αριθμητικός μέσος όρος των τιμών που εμπίπτουν στο αντίστοιχο παράθυρο:

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \quad (4.3)$$

Γεωμετρικά, η διαδικασία αυτή αντιστοιχεί στην προσέγγιση του αρχικού συστήματος μέσω μιας γραμμικής συνδυαστικής σειράς ορθογώνιων συναρτήσεων βάσης (box basis functions). Η PAA προσφέρει δύο θεμελιώδη πλεονεκτήματα στη SAX:

1. Δραστική Μείωση Διαστάσεων: Αξιοποιείται η υπολογιστική απλότητα και η αποδεδειγμένη ισχύς συμπίεσης της PAA
2. Θεμελίωση του Lower Bounding: Η απόδειξη ότι η συμβολική απόσταση υποεκτιμά την πραγματική Ευκλείδεια απόσταση βασίζεται στη μεταβατική ιδιότητα, αποδεικνύοντας πρώτα ότι η συμβολική απόσταση υποεκτιμά την απόσταση PAA[16][17].

4.2.3 Breakpoints και Γκαουσιανή Κατανομή

Μετά τον υπολογισμό των συντελεστών PAA, απαιτείται ο μετασχηματισμός των συνεχών αυτών τιμών σε διακριτά σύμβολα. Για να είναι η διακριτοποίηση αποτελεσματική, είναι επιθυμητό τα παραγόμενα σύμβολα να εμφανίζονται με ισοπιθανότητα (equiprobability)[18].

Επειδή οι χρονοσειρές έχουν κανονικοποιηθεί (Z-score), η αθροιστική κατανομή των τιμών τους ακολουθεί με μεγάλη ακρίβεια την Κανονική (Γκαουσιανή) Κατανομή $N(0,1)$. Με βάση την ιδιότητα αυτή, μπορούν να οριστούν κατάλληλα Σημεία Θραύσης (Breakpoints) που χωρίζουν την περιοχή κάτω από τη Γκαουσιανή καμπύλη σε α περιοχές ίσου εμβαδού, όπου α είναι το μέγεθος του αλφαβήτου.

Τα σημεία θραύσης υπολογίζονται εύκολα από στατιστικούς πίνακες της πρότυπης κανονικής κατανομής. Το σχήμα 9[12] παρουσιάζει τα σημεία θραύσης για μεγέθη αλφαβήτου α από 3 έως 10:

α	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Figure 9 Breakpoints για αλφάβητο α από 3 έως 10

4.2.4 Συμβολική Αντιστοίχιση και Κατασκευή Λέξης (Word Construction)

Αφού οριστούν τα σημεία θραύσης για το επιλεγμένο μέγεθος αλφαβήτου α , η διαδικασία της διακριτοποίησης ολοκληρώνεται με την αντιστοίχιση κάθε συντελεστή PAA σε ένα σύμβολο.

Όλοι οι συντελεστές PAA που είναι μικρότεροι από το μικρότερο σημείο θραύσης β_1 αντιστοιχίζονται στο πρώτο σύμβολο του αλφαβήτου (π.χ. “a”). Οι συντελεστές που είναι μεγαλύτεροι ή ίσοι με το β_1 και μικρότεροι από το β_2 αντιστοιχίζονται στο δεύτερο σύμβολο (π.χ. “b”), και ούτω καθεξής.

Ορισμός 2: Συμβολική Λέξη(SAX Word)

Μια υποακολουθία C μήκους n αναπαρίσταται ως συμβολική λέξη $\hat{C} = \hat{c}_1, \hat{c}_2, \dots, \hat{c}_w$. Έστω α_i το i -οστό στοιχείο του αλφαβήτου(δηλαδή $\alpha_1 = "a", \alpha_2 = "b", \dots$, κ.ο.κ.). Η απεικόνιση από το διάνυσμα PAA $\overline{C}$ στη λέξη $\hat{C}$ ορίζεται ως:

$$\hat{c}_i = \alpha_j \iff \beta_{j-1} \leq \overline{c}_i < \beta_j \quad (4.4)$$

Για παράδειγμα, αν μια κανονικοποιημένη χρονοσειρά μήκους $n = 128$ συμπιεστεί σε $w = 8$ συντελεστές PAA και χρησιμοποιηθεί αλφάβητο μεγέθους $\alpha = 3$ (με σύμβολα “a”, “b”, “c”), οι συντελεστές PAA συγκρίνονται με τα αντίστοιχα breakpoints ($\beta_1 = -0.43, \beta_2 = 0.43$). Αν προκύψει η αλληλουχία συμβόλων “baabcbe”, αυτή η πρόταση συμπυκνώνει ολόκληρη τη μορφολογία της αρχικής χρονοσειράς.

Συνοψίζοντας, η SAX μετασχηματίζει ένα συνεχές σήμα n διαστάσεων σε μια διακριτή συμβολική αλληλουχία w χαρακτήρων, επιτυγχάνοντας ταυτόχρονα μείωση διαστάσεων και συμπίεση πληροφορίας.

4.3 Μέτρα Ομοιότητας και Θεωρητικές Ιδιότητες

Ένα από τα σημαντικότερα προβλήματα των προγενέστερων συμβολικών μεθόδων ήταν η αδυναμία ορισμού μια συνάρτησης απόστασης στο συμβολικό χώρο που να διατηρεί ουσιαστική συσχέτιση με τις πραγματικές αποστάσεις των αρχικών δεδομένων. Η μέθοδος SAX επιλύει το πρόβλημα αυτό εισάγοντας τη συνάρτηση ελάχιστης απόστασης MINDIST, η οποία εγγυάται τη διατήρηση της ιδιότητας του κάτω ορίου(Lower Bounding property).

4.3.1 Η συνάρτηση Απόστασης MINDIST

Το πιο διαδεδομένο μέτρο ομοιότητας για τη σύγκριση δύο χρονοσειρών Q και C ίδιου μήκους n είναι η Ευκλείδεια Απόσταση:

$$D(Q, C) = \sqrt{\sum_{i=1}^n (\overline{q}_i - \overline{c}_i)^2} \quad (4.5)$$

Αν μετασχηματίσουμε τις χρονοσειρές Q και C στις αντίστοιχες αναπαραστάσεις PAA $\overline{Q}$ και $\overline{C}$ μήκους w , η υπολογιζόμενη Ευκλείδεια απόσταση στο χώρο PAA (DR) παρέχει ένα κάτω όριο της πραγματικής απόστασης:

$$D(\overline{Q}, \overline{C}) = \sqrt{\frac{n}{w} \sum_{i=1}^w (\overline{q}_i - \overline{c}_i)^2} \quad (4.6)$$

Στη συνέχεια όταν οι συντελεστές PAA μετατραπούν σε συμβολικές λέξεις SAX $\hat{Q}$ και $\hat{C}$ μήκους w , ορίζεται η συνάρτηση MINDIST, η οποία επιστρέφει την ελάχιστη δυνατή απόσταση μεταξύ των αρχικών χρονοσειρών που αντιπροσωπεύονται από τις δύο λέξεις:

$$MINDIST(\hat{Q}, \hat{C}) = \sqrt{\frac{n}{w} \sum_{i=1}^w (dist(\hat{q}_i, \hat{c}_i))^2} \quad (4.7).$$

Ο παράγοντας $\sqrt{\frac{n}{w}}$ λειτουργεί ως σταθμιστής κλίμακας (compression ratio), προσαρμόζοντας την απόσταση του συμπιεσμένου χώρου w στο μέγεθος του αρχικού χώρου n διαστάσεων.

4.3.2 Κατασκευή Πίνακα Αναζήτησης (Lookup Table)

Η υπό-συνάρτηση $dist(\hat{q}_i, \hat{c}_i)$ υπολογίζει την ελάχιστη απόσταση μεταξύ δύο μεμονωμένων συμβόλων $\hat{q}_i$ και $\hat{c}_i$. Για την αποφυγή περιττών υπολογισμών κατά την εκτέλεση αλγορίθμων, οι τιμές της $dist()$ προϋπολογίζονται και αποθηκεύονται σε έναν πίνακα Αναζήτησης (Lookup Table).

Το στοιχείο στη γραμμή r και στήλη c του πίνακα αναζήτησης υπολογίζεται από τη σχέση:

$$cell_{r,c} = \begin{cases} 0 & \text{if } |r - c| \leq 1 \\ \beta_{\max(r,c)-1} - \beta_{\min(r,c)}, & \text{otherwise} \end{cases} \quad (4.7)$$

Όπως φαίνεται στον τύπο (4.7):

1. Αν δύο σύμβολα είναι ταυτόσημα ($r = c$) ή γειτονικά ($|r - c| = 1$), η ελάχιστη μεταξύ τους απόσταση είναι 0, καθώς οι πραγματικές τιμές που αντιπροσωπεύουν θα μπορούσαν να βρίσκονται αυστηρά εκατέρωθεν του κοινού σημείου θραύσης.
2. Αν τα σύμβολα απέχουν περισσότερο από μια θέση ($|r - c| > 1$), η ελάχιστη απόσταση ισούται με την απόσταση μεταξύ των εσωτερικών σημείων θραύσης των δύο περιοχών.

Το σχήμα 10[12] παρουσιάζει παραδείγματα αναζήτησης για μέγεθος αλφαβήτου $\alpha = 4$

	a	b	c	d
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

Σχήμα 10: Lookup table για αλφάβητο $\alpha=4$

Για παράδειγμα, $dist(a,b) = 0$, ενώ $dist(a,c) = \beta_2 - \beta_1 = 0.00 - (= 0.67) = 0.67$.

4.3.3 Η ιδιότητα του Κάτω Ορίου (Lower Bounding Property)

Η σπουδαιότητα της συνάρτησης MINDIST έγκειται στην αυστηρή μαθηματική εγγύηση ότι υποεκτιμά πάντα (ή στην καλύτερη περίπτωση είναι ίση με) την πραγματική Ευκλείδεια απόσταση των αρχικών χρονοσειρών.

Θεώρημα: Lower Bounding Guarantee της SAX. Για οποιεσδήποτε δύο χρονοσειρές Q και C μήκους n και τις αντίστοιχε συμβολικές αναπαραστάσεις SAX $\hat{Q}$, και $\hat{C}$ μήκους w , ισχύει η ανισότητα:

$$MINDIST(\hat{Q}, \hat{C}) \leq DR(\overline{Q}, \overline{C}) \leq D(Q, C) \quad (4.8)$$

Σημασία στην Εξόρυξη Δεδομένων:

Στο πλαίσιο της αναζήτησης ομοιότητας (Similarity Search) και της ευρετηρίασης (indexing), η τήρηση της ανισότητας (4.8) διασφαλίζει ότι δεν υπάρχουν ψευδώς αρνητικά αποτελέσματα (zero false dismissals)[19]. Οποιοδήποτε ζεύγος χρονοσειρών είναι όμοιο στο αρχικό πεδίο $D(Q, C)$, θα ανακτηθεί υποχρεωτικά από το ευρετήριο στο συμβολικό πεδίο $MINDIST(Q, C)$. Τυχόν ψευδώς θετικά αποτελέσματα (false alarms) μπορούν στη συνέχεια να απορριφθούν εύκολα με έναν γρήγορο υπολογισμό της πραγματικής απόστασης στα πρωτογενή δεδομένα.

4.3.4 Αυστηρότητα Κάτω Ορίου (Tightness of Lower Bound-TLB) και Επιλογή Παραμέτρων(w, a)

Η αποτελεσματικότητα της SAX εξαρτάται από τις παραμέτρους w (μήκος συμβολικής λέξης) και a (μέγεθος αλφαβήτου). Η αυστηρότητα του κάτω Ορίου(Tightness of Lower Bound-TLB) ορίζεται ως ο λόγος της εκτιμώμενης συμβολικής απόστασης προς την πραγματική Ευκλείδεια απόσταση:

$$TLB = \frac{MINDIST(\hat{Q}, \hat{C})}{D(Q, C)} \quad (4.9)$$

όσο ο λόγος TLB πλησιάζει τη μονάδα(1), τόσο πιο ακριβής είναι η προσέγγιση της συμβολικής απόστασης.

1. Επίδραση της παραμέτρου w : Η αύξηση του w βελτιώνει θεαματικά την αυστηρότητα του κάτω ορίου, καθώς συγκρατούνται περισσότερες λεπτομέρειες του σχήματος της χρονοσειράς.
2. Επίδραση της παραμέτρου a : Πολύ μικρές τιμές του a ($a = 2, 3$) οδηγούν σε διαχωρισμούς και ασθενή κάτω όρια. Ωστόσο, η αύξηση του a πάνω από ένα όριο προσφέρει φθίνουσα απόδοση.

Πειραματικές μελέτες των Liv et al σε δεκάδες σύνολα δεδομένων έδειξαν ότι η επιλογή μέτρων τιμών αλφαβήτου στο εύρος $a \in [5, 8]$ προσφέρει τη βέλτιστη ισορροπία μεταξύ της αυστηρότητας του κάτω ορίου και της αποτελεσματικής διαχείρισης μνήμης[9].

4.4 Μείωση Πληθικότητας (Numerosity Reduction)

Στις περισσότερες εφαρμογές εξόρυξης δεδομένων χρονοσειρών, η ανάλυση δεν πραγματοποιείται σε μία μεμονωμένη ολική χρονοσειρά, αλλά σε επικαλυπτόμενες υποακολουθίες (subsequences) μήκους n , οι οποίες εξάγονται από μια πολύ μεγάλη χρονοσειρά T με τη χρήση ενός ολισθαίνοντος παρθύρου (sliding window).

4.4.1 Τεχνική Ολισθαίνοντος Παραθύρου και το Πρόβλημα Όγκου

Έστω μια συνεχής χρονοσειρά T συνολικού μήκους $|T|$. Όταν εφαρμόζεται ένα ολισθαίνον παράθυρο μεγέθους n , το πλήθος των παραγόμενων υποακολουθιών ισούται με $|T| - n + 1$. Στις παραδοσιακές αναπαραστάσεις πραγματικών τιμών (όπως η PAA, ο DFT ή τα Wavelets), κάθε μία από αυτές τις $|T| - n + 1$ υποακολουθίες πρέπει να αποθηκευτεί ξεχωριστά στη μνήμη ή στο δίσκο.

Για μεγάλες βάσεις δεδομένων, η απαίτηση αυτή προκαλεί τεράστιο υπολογιστικό κόστος και υπερβολική κατανάλωση αποθηκευτικών πόρων.

4.4.2 Συμπίεση Τύπου RLE (Run-Length Encoding)

Η διακριτή και συμβολική φύση της SAX προσφέρει μια μοναδική ευκαιρία για δραστική μείωση του όγκου των δεδομένων μέσω μιας τεχνικής ανάλογης με τη συμπίεση Run-Length Encoding(RLE).

Κατά τη μετακίνηση του ολισθαίνοντος παραθύρου κατά μία θέση προς τα δεξιά:

1. Εάν η πρώτη εξαχθείσα υποακολουθία μετατραπεί στη συμβολική λέξη, για παράδειγμα, “aabbcc”, και η αμέσως επόμενη επικαλυπτόμενη υποακολουθία μετατραπεί επίσης στην ίδια ακριβώς λέξη “aabbcc”, η δεύτερη εμφάνιση δεν χρειάζεται να αποθηκευτεί.
2. Αντί για τη διατήρηση διπλότυπων εγγραφών, διατηρείται μόνο ένας δείκτης(pointer) στην πρώτη εμφάνιση της λέξης, μαζί με την πληροφορία για το πόσες διαδοχικές ολισθησεις διατηρούν την ίδια συμβολική αναπαράσταση.

Αποδοτικότητα της Μεθόδου:

Η αποτελεσματικότητα της μείωσης πληθικότητας εξαρτάται από τη φύση των δεδομένων:

1. Σε τυπικές χρονοσειρές, επιτυγχάνεται συντελεστής συμπίεσης $2x$ έως $3x$.
2. Σε χρονοσειρές που χαρακτηρίζονται από μεγάλες περιόδους στασιμότητας ή ομαλότητας οι οποίες διακόπτονται από σύντομες εξάρσεις δραστηριότητας (όπως σειсмоγραφικά δεδομένα, ιατρικά σήματα ΗΚΓ, ή τηλεμετρία διαστημικών συσκευών), η μείωση πληθικότητας μπορεί να είναι τεράστια. Για παράδειγμα, εκατοντάδες διαδοχικές επικαλυπτόμενες υποακολουθίες μπορούν να αντιστοιχηθούν στην ίδια συμβολική λέξη, συμπιεσμένες σε μία μόνο εγγραφή.
- 3.

4.4.3 Διαχείριση Ειδικών Περιπτώσεων: Σχεδόν Σταθερές Υποακολουθίες (Near-Constant Subsequences)

Κατά την εφαρμογή της Z-score κανονικοποίησης σε ολισθαίνοντα παράθυρα, εμφανίζεται ένα σημαντικό τεχνικό πρόβλημα στις σχεδόν σταθερές υποακολουθίες (near-constant subsequences).

Αν μια υποακολουθία περιέχει σταθερές τιμές (π.χ. 31 μηδενικά και μία ελάχιστη τιμή 0.0001 λόγω θορύβου αισθητήρα), η τυπική απόκλιση σ είναι σχεδόν μηδενική. Κατά τη διαιρετική κανονικοποίηση ($\hat{c}_i = \frac{c_i - \mu}{\sigma}$), ο ελάχιστος αυτός θόρυβος πολλαπλασιάζεται εκρηκτικά, δημιουργώντας τεχνητές, μεγάλες διακυμάνσεις που δεν αντιπροσωπεύουν την πραγματική συμπεριφορά του σήματος.

Η λύση της SAX:

Για την αντιμετώπιση του φαινομένου αυτού, η SAX εισάγει ένα κατώφλι ετοιμότητας ε :

1. Εάν η τυπική απόκλιση σ μιας υποακολουθίας πριν από την κανονικοποίηση είναι μικρότερη από το κατώφλι ε ($\sigma < \varepsilon$), η υποακολουθία θεωρείται σταθερή.
2. Στην περίπτωση αυτή, η διαδικασία κανονικοποίησης παρακάμπτεται και ολόκληρη η συμβολική λέξη ανατίθεται εξ ολοκλήρου στο μεσαίο σύμβολο του αλφαγήτου (π.χ. “ccccc” αν $\alpha = 5$)

Με τον τρόπο αυτό εξασφαλίζεται η ανθεκτικότητα του αλγορίθμου απέναντι στον τοπικό θόρυβο των αισθητήρων και αποφεύγεται η δημιουργία λανθασμένων συμβολικών προτύπων.

4.5 Τομείς Εφαρμογής και Αλγόριθμοι Εξόρυξης

Η εισαγωγή της SAX παρέχει διπλό όφελος στον κλάδο της εξόρυξης δεδομένων χρονοσειρών: αφενός επιτρέπει την εκτέλεση κλασσικών αλγορίθμων (όπως η κατηγοριοποίηση και η ομαδοποίηση) πάνω σε συμπιεσμένα συμβολικά δεδομένα με εξαιρετική ακρίβεια, αφετέρου επιτρέπει την αξιοποίηση αλγορίθμων από την επιστήμη υπολογιστών που λειτουργούν αποκλειστικά σε διακριτές συμβολοσειρές.

4.5.1 Κατηγοριοποίηση (Classification)

Η ταξινόηση χρονοσειρών αποτελεί μια από τις πιο απαιτητικές εργασίες λόγω του υψηλού αριθμού διαστάσεων της συσχέτισης των χαρακτηριστικών και των επιπέδων θορύβου.

1. Κατηγοριοποιητής Εγγύτων Γειτόνων
 - a. Ο αλγόριθμος 1-NN σε συνδυασμό με τη συνάρτηση απόστασης *MINDIST* εφαρμόζεται άμεσα στις συμβολικές λέξεις SAX
 - b. Σε πειραματικές αξιολογήσεις με τη μέθοδο διασταυρωμένης επικύρωσης (Leave-one-out cross validation) σε γνωστά σύνολα δεδομένων (όπως το Cylinder-Bell-Funnel), η SAX πέτυχε χαμηλότερα ποσοστά σφάλματος (error rate) όχι μόνο από άλλες συμβολικές προσεγγίσεις (SDA, IMPACTS), αλλά σε ορισμένες περιπτώσεις και από την ίδια την κλασσική Ευκλείδεια απόσταση στα πρωτογενή δεδομένα[13][14].
 - c. Αυτή η υπεροχή αποδίδεται στο φαινόμενο εξομάλυνσης (smoothing effect) που επιφέρει η μείωση διαστάσεων PAA/SAX, απομακρύνοντας τον τοπικό θόρυβο.
2. Δέντρα απόφασης (Decision Trees):
 - a. Η εφαρμογή Δέντρων Απόφασης απευθείας σε πρωτογενή δεδομένα χρονοσειρών παράγει βαθιά, πολύπλοκα δέντρα με φαινόμενα υπερεκπαίδευσης (overfitting).
 - b. Η SAX επιτρέπει την εκπαίδευση δέντρων απόφασης πάνω στους διακριτούς χαρακτήρες των λέξεων, προσφέροντας αποτελεσματικά ανταγωνιστικά εξειδικευμένων μοντέλων (όπως τα Regression Trees), διατηρώντας παράλληλα υψηλή υπολογιστική ταχύτητα[20].

4.5.2 Ομαδοποίηση

Η ομαδοποίηση (clustering) χρησιμοποιείται τόσο για την ερευνητική ανάλυση δεδομένων όσο και ως υπο-ρουτίνα σε πιο σύνθετους αλγορίθμους:

1. Ιεραρχική Ομαδοποίηση (Hierarchical Clustering):
 - a. Η κατασκευή δενδρογραμμάτων (dendrograms) με τη χρήση της SAX αποδεικνύει την ικανότητα της μεθόδου να απομονώνει τις πραγματικές κλάσεις σε διακριτά υποδέντρα με απόλυτη επιτυχία, ξεπερνώντας τις υπόλοιπες συμβολικές μεθόδους.
2. Ομαδοποίησης Διαμερισμού (k-means Clustering):
 - a. Η εφαρμογή του αλγορίθμου k-means στο συμβολικό χώρο της SAX συγκλίνει στον ίδιο μέσο αριθμό επαναλήψεων με τα αρχικά δεδομένα, επιτυγχάνοντας χαμηλότερες (βελτιστες) τιμές στη συνάρτηση σκοπού (objective function).
 - b. Αυτό συμβαίνει διότι η αρχικοποίηση των κεντρών των συστάδων (cluster centers) σε έναν χώρο χαμηλότερης διάστασης βελτιώνει αποδεδειγμένα τη συνολική ποιότητα της ομαδοποίησης[21].

4.5.3 Αναζήτηση Βάσει Περιεχομένου

Στην αναζήτης ομοιότητας (Similarity Search), ο στόχος είναι η ταχεία ανάκτηση των πιο όμοιων χρονοσειρών από μια βάση δεδομένων για ένα ερώτημα Q .

1. Χάρη στην ιδιότητα του κάτω ορίου (Lower Bounding), η SAX εγγυάται την κατασκευή δομών ευρετηρίασης χωρίς την παρουσία ψευδώς αρνητικών αποτελεσμάτων.

2. Παράλληλα, η αλφαριθμητική φύση της SAX επιτρέπει την αξιοποίηση κλασικών δομών ευρετηρίασης κειμένου, όπως τα Δέντρα Επιθημάτων (Suffix Trees).

4.5.4 Ανίχνευση Ανωμαλιών (Anomaly Detection)

Η ανίχνευση ανεπιθύμητης ή ασυνήθιστης συμπεριφοράς σε χρονοσειρές είναι κρίσιμη σε τομείς όπως η ιατρική και η βιομηχανία.

1. Πιθανοτική Μοντελοποίηση: Η SAX επιτρέπει τη συνδυαστική χρήση Αλυσιδών Markov (Markov Chains) και Δέντρων Επιθημάτων (Suffix Trees)[22].
2. Οι αλυσίδες Markov εκπαιδεύονται στις συμβολικές ακολουθίες της κανονικής συμπεριφοράς. Όταν εισαχθούν νέα δεδομένα, το δέντρο επιθημάτων αξιολογεί τη συχνότητα εμφάνισης των συμβολικών λέξεων. Λέξεις με ελάχιστη πιθανότητα εμφάνισης βάσει του μοντέλου Markov επισημαίνονται άμεσα ως ανωμαλίες.
3. Σε συγκριτικά πειράματα έναντι μεθόδων κυματιδίων (Wavelets) και ανοσολογικών αλγορίθμων, η SAX εντόπισε όλες τις τεχνητές ανωμαλίες χωρίς κανέναν ψευδή συναγερμό (false alarm)[23][24].

4.5.5 Ανακάλυψη Μοτίβων (Motif Discovery)

Τα μοτίβα χρονοσειρών (time series motifs) ορίζονται ως ζεύγη υπο-ακολουθιών που εφαρμόζοιεν εξαιρετικά υψηλή ομοιότητα μεταξύ τους εντός ενός μεγάλου συνόλου δεδομένων[25].

1. Η αλγοριθμική προσέγγιση για την εύρεση μοτίβων έχει τετραγωνική υπολογιστική πολυπλοκότητα $O(n^2)$, καθιστώντας την απαγορευτική για μεγάλες χρονοσειρές.
2. Αλγόριθμος Τυχαίας Προβολής (PROJECTION Algorithm): Η SAX επιτρέπει τη μεταφορά προηγμένων αλγορίθμων από τη βιοπληροφορική. Οι συμβολικές λέξεις SAX εισάγονται σε κάδους κατακερματισμού (hash buckets) με τη χρήση τυχαίων προβολών των χαρακτηριστικών τους[26].
3. Η τεχνική αυτή μειώνει την πολυπλοκότητα σε γραμμικά επίπεδα, επιτρέποντας τον ταχύτατο εντοπισμό μοτίβων ακόμα και σε σήματα με υψηλά επίπεδα θορύβου.

4.6 Παραλλαγές και Επεκτάσεις της SAX στη Βιβλιογραφία

Η εισαγωγή της μεθόδου SAX από τους Lin et al. αποτέλεσε σημείο αναφοράς στην εξόρυξη δεδομένων χρονοσειρών, προκαλώντας το ενδιαφέρον πληθώρας ερευνητών για την περαιτέρω βελτίωση και εξειδίκευσή της. Στη βιβλιογραφία έχουν προταθεί διάφορες παραλλαγές που στοχεύουν στην αντιμετώπιση συγκεκριμένων περιορισμών της βασικής μεθόδου:

4.6.1 Προσεγγίσεις Προσαρμοσμένες και Βασισμένες στην Εντροπία

1. Entropy-based SAX (E-SAX): Στην κλασική SAX τα σημεία θραύσης (breakpoints) καθορίζονται με την παραδοχή ότι τα κανονικοποιημένα δεδομένα ακολουθούν αυστηρά την πρότυπη Γκαουσιανή κατανομή $N(0,1)$. Η E-SAX υπολογίζει τα σημεία θραύσης δυναμικά με βάση την εντροπία της πληροφορίας (information entropy) των πραγματικών δεδομένων επιτυγχάνοντας βέλτιστη διακριτοποίηση σε χρονοσειρές που αποκλίνουν σημαντικά από την κανονική κατανομή[27].
2. Extended / Adaptive SAX(A-SAX): Επεκτάσεις που προσαρμόζουν το μέγεθος των παραθύρων PAA δυναμικά, εστιάζοντας περισσότερα σύμβολα σε περιοχές υψηλής μεταβλητότητας (όπως αιχμές και απότομες μεταβολές) και λιγότερα σε ομαλές περιοχές.

4.6.2 Ενσωμάτωση Χαρακτηριστικών Τάσης

1. Ανίχνευση Απάτης σε πιστωτικές Κάρτες(Credit Card Fraud Detection). Μετατροπή των αλληλουχιών συναλλαγών των χρηστών σε συμβολικές λέξεις SAX για τον άμεσο εντοπισμό ασυνήθιστων προτύπων[28].
2. Εκτίμηση Έντασης Καταιγίδων (Storm Intensity Estimation): Χρήση της SAX σε συνδυασμό με Τεχνητά Νευρωνικά Δίκτυα (ANN) για την ανάλυση μετεωρολογικών χρονοσειρών και την προβλεψη ακραίων καιρικών φαινομένων[29].
3. Ιατρική Διαγνωστική: Ανάλυση ηλεκτροκαρδιογραφημάτων (ΗΚΓ) και εγκεφαλογραφημάτων (ΗΕΓ) για την ταχεία αναγνώριση αρρυθμιών και επιληπτικών κρίσεων.

4.7 Επίλογος

Εν κατακλείδι, η μέθοδος SAX συνιστάται ένα ολοκληρωμένο πλαίσιο που γεφυρώνει τη συνεχή φύση των χρονοσειρών με τις διακριτές δομές δεδομένων, συνδυάζοντας τη συμπίεση PAA, τη Γκαουσιανή διακριτοποίηση και την εγγύηση του κάτω ορίου μέσω της συνάρτησης MINDIST. Έχοντας ολοκληρώσει τη θεωρητική θεμελίωση τόσο των κατηγοριοποιητών όσο και των μεθόδων αναπαράστασης, το επόμενο κεφάλαιο είναι αφιερωμένο στην Πειραματική Μεθοδολογία όπου περιγράφονται τα σύνολα δεδομένων UCR, το πειραματικό pipeline στο RStudio και το στατιστικό πρωτόκολλο αξιολόγησης.

Κεφάλαιο 5ο: Πειραματική Μεθοδολογία

Σε αυτό το κεφάλαιο θα αναλύσουμε την πειραματική μεθοδολογία της εργασίας. Θα περιγραφεί με λεπτομέρεια η διαδικασία διεξαγωγής των πειραμάτων, τα σύνολα δεδομένα που χρησιμοποιήθηκαν, το πειραματικό περιβάλλον και θα αναλυθεί όλη η ροή.

5.1 Περιγραφή Συνόλων Δεδομένων (UCR Time Series Classification Archive)

Για την αξιολόγηση και την επαλήθευση της αποτελεσματικότητας της μεθόδου SAX σε συνδυασμό με τους αλγορίθμους κατηγοριοποίησης, χρησιμοποιήθηκε ένα αντιπροσωπευτικό δείγμα 10 συνόλων δεδομένων χρονοσειρών από το ευρέως αναγνωρισμένο ακαδημαϊκό αποθετήριο UCR Times Series Classification Archive (Keogh et al.).

Το αποθετήριο UCR αποτελεί το de facto πρότυπο (benchmark) στην κοινότητα της εξόρυξης δεδομένων χρονοσειρών, καθώς περιλαμβάνει προβλήματα από ποικίλα επιστημονικά και βιομηχανικά πεδία (ιατρικά σήματα, εικόνες, αισθητήρες, συνθετικά δεδομένα κ.α.).

Τα 10 σύνολα δεδομένων που επιλέχθηκαν παρουσιάζουν μεγάλη ποικιλομορφία ως προς:

- Το μήκος των χρονοσειρών (n) από 60 έως 427 χρονικά σημεία).
- Το πλήθος των κλάσεων (από 2 έως 50 διαφορετικές κατηγορίες).
- Το μέγεθος σφάλματος αναφοράς του κατηγοριοποιητή 1-NN με Ευκλείδεια απόσταση (Baseline 1-NN Euclidean Error).

Στον Πίνακα 5.1 παρουσιάζονται συγκεντρωτικά τα βασικά χαρακτηριστικά των 10 συνόλων δεδομένων:

Name	Frist paper or data	Number of Classes	Size of training set	Size of testing set	Time series Lenght	1 - N N Euclidean Distance
Synthetic Control	Pham	6	300	300	60	0.120
Gun-Point	Ratanamahatana	2	50	150	150	0.087
CBF	Saito	3	30	900	128	0.148
Face (all)	Xi	14	560	1690	131	0.286
OSU Leaf	Gandhi	6	200	242	427	0.479
Swedish Leaf	Soderkvist	15	500	625	128	0.211
50Words	Rath	50	450	455	270	0.369
Trace	Roverso	4	100	100	275	0.240
Two Patterns	Geurts	4	1000	4000	128	0.090
Wafer	Olszewski	2	1000	6174	152	0.005

Table 5: 10 σύνολα δεδομένων από το αποθετήριο UCR

Αναλυτική Περιγραφή των Datasets ανά Κατηγορία

Για την πληρέστερη κατανόηση της φύσης των δεδομένων, τα 10 datasets χωρίζονται στις ακόλουθες θεματικές κατηγορίες:

1. Συνθετικά Δεδομένα και Μοτίβα (Synthetic and Pattern Datasets):
 - a. Synthetic Control: Περιλαμβάνει 600 χρονοσειρές (300 train / 300 test) μήκους 60 σημείων, οι οποίες προσομοιώνουν 6 διαφορετικές καταστάσεις ελέγχου βιομηχανικών διεργασιών (Normal, Cyclic, Increasing trend, Decreasing trend, Upward shift, Downward shift)
 - b. CBF (Cylinder-Bell-Funnel): Κλασσικό συνθετικό dataset 3 κλάσεων (κύλινδρος, καμπάνα, χωνί). Χαρακτηρίζεται από πολύ μικρό σύνολο εκπαίδευσης (μόλις 30 δείγματα) και μεγάλο σύνολο ελέγχου (900 δείγματα), δοκιμάζοντας την ικανότητα γενίκευσης των μοντέλων.
 - c. Two Patterns: Περιλαμβάνει 5000 χρονοσειρές μήκους 128, που δημιουργήθηκαν συνθέτοντας τέσσερις συνδυασμούς δύο βασικών μοτίβων (Up trend, Down trend) μέσα σε θόρυβο
2. Αναγνώριση Κίνηση και Χειρονομιών (Motion and Gesture Recognition):
 - a. Gun-Point: Προέρχεται από την καταγραφή κινήσεων των χεριών μέσω βίντεο και περιλαμβάνει 2 κλάσεις: την κίνηση “Gun-Draw” (στόχευση με όπλο) και την κίνηση “Point” (απλή υπόδειξη με το δάχτυλο).
 - b. Trace: Πρόκειται για δεδομένα αισθητήρων από βιομηχανικές προσομοιώσεις πυρηνικών αντιδραστήρων (Roverso), με 4 κλάσεις που αντιπροσωπεύουν διαφορετικούς τύπους παροδικών μεταβολών (transients).
3. Επεξεργασία Εικόνας και Περιγραμμάτων (Image Outline Transformation Datasets):
 - a. Face (all): Προέρχεται από την εξαγωγή περιγραμμάτων προσώπων από εικόνες και περιλαμβάνει 14 κλάσεις (διαφορετικά άτομα)
 - b. OSU Leaf και Swedish Leaf: Χρονοσειρές που προκύπτουν από τη μετατροπή του δισδιάστατου περιγράμματος φύλλων δέντρων σε μονοδιάστατη χρονοσειρά (μετασχηματισμός πολικών συντεταγμένων). Το OSU Leaf περιλαμβάνει 6 κλάσεις με τις μεγαλύτερες χρονοσειρές της μελέτης ($n=427$), ενώ το Swedish Leaf περιλαμβάνει 15 κλάσεις.
 - c. 50Words: Προέρχεται από την εξαγωγή περιγραμμάτων χειρόγραφων λέξεων σε ιστορικά έγγραφα, αποτελώντας το πιο πολυσύνθετο πρόβλημα με 50 διαφορετικές κλάσεις.
4. Βιομηχανικοί Αισθητήρες (Industrial Sensor Datasets):
 - a. Wafer: Περιλαμβάνει μετρήσεις αισθητήρων κατά τη διαδικασία κατασκευής μικροτσιπ (semiconductor wafers). Αποτελείται από 2 κλάσεις (Normal vs Abnormal) με ένα τεράστιο σύνολο ελέγχου 6174 δειγμάτων και πολύ χαμηλό σφάλμα βάσης (0.005).

5.2 Ροή Εκτέλεσης Πειραμάτων (Experimental Pipeline)

Η πειραματική διαδικασία της παρούσας εργασίας σχεδιάστηκε με σκοπό τη διεξαγωγή μιας ολοκληρωμένης συγκριτικής αξιολόγησης (benchmarking). Πέραν της κεντρικής μεθόδου SAX (Symbolic Aggregate Approximation), στο ίδιο ακριβώς πειραματικό pipeline εντάχθηκαν και αξιολογήθηκαν τρεις επιπλέον θεμελιώδεις μέθοδοι μείωσης διαστάσεων χρονοσειρών:

1. Διακριτός Μετασχηματισμός Φουριέ (Discrete Fourier Transform – DFT)
2. Τμηματική Γραμμική Προσέγγιση (Piecewise Linear Approximation – PLA)
3. Τμηματική Αθροιστική Προσέγγιση (Piecewise Aggregate Approximation –PAA)

Η συγκριτική αυτή προσέγγιση επιτρέπει την αξιολόγηση της συμβολικής αναπαράστασης (SAX) έναντι των παραδοσιακών αριθμητικών αναπαραστάσεων συνεχών τιμών (real-valued representations).

Η ροή των πειραμάτων διαχωρίζεται σε δύο κύριες προσπελάσεις ανάλογα με τη φύση της αναπαράστασης των δεδομένων και τον αλγόριθμο κατηγοριοποίησης:

5.2.1 Προσέγγιση 1: Κατηγοριοποίηση k-NN με την Μετρική MINDIST.SAX

Για τον αλγόριθμο των k-Εγγύτερων Γειτόνων (k-NN) :

1. Για την μέθοδο SAX: Εφαρμόστηκε άμεση προσέγγιση υπολογισμού αποστάσεων στο συμβολικό πεδίο μέσω της συνάρτησης `diss.MINDIST.SAX()`. Η Z-score κανονικοποίηση, η συμπίεση PAA και η αντιστοίχιση σε σύμβολα εκτελούνται εσωτερικά. Για κάθε χρονοσειρά x_i του συνόλου ελέγχου (test set) υπολογίζεται η συμβολική απόσταση MINDIST από όλες τις εγγραφές του συνόλου εκπαίδευσης (train set). Σε περίπτωση ισοψηφίας (`length(winners)>1`), εφαρμόζεται ο κανόνας `tie_break="nearest"`, αναθέτοντας την κλάση του πλησιέστερου γείτονα βάσει της ελάχιστης συμβολικής απόστασης.
2. Για τις μεθόδους DFT, PLA και PAA: Οι αρχικές χρονοσειρές υποβλήθηκαν σε Z-score κανονικοποίησης με την συνάρτηση `znorm()` και στη συνέχεια μετασχηματίστηκαν στις αντίστοιχες μειωμένες αναπαραστάσεις με τις συναρτήσεις `repr_dft()`, `repr_pla()` και `repr_paa()`. Ο αλγόριθμος 1-NN εφαρμόστηκε πάνω στα παραγόμενα διανύσματα με τη χρήση της κλασσικής Ευκλείδειας απόστασης.

5.2.2 Προσέγγιση 2: Μοντέλα Δέντρων και Συλλογικής Μάθησης (Decision Trees, Random Forest, CatBoost)

Για τα μοντέλα που απαιτούν ρητή αναπαράσταση των χαρακτηριστικών σε μορφή πίνακα δεδομένων (tabular feature matrix), ακολουθήθηκε η εξής διαδικασία:

1. Κανονικοποίηση (Z-score Normalization): Εφαρμογή της συνάρτησης `znorm()` σε κάθε χρονοσειρά των συνόλων εκπαίδευσης και ελέγχου.
2. Εφαρμογή Μετασχηματισμών Μείωσης Διαστάσεων:
 - a. SAX: Υπολογισμός των συντελεστών PAA μέσω της `paa()` και μετατροπή τους σε συμβολικές λέξεις(ακολουθίες χαρακτήρων) μέσω της `series_to_string()`.
 - b. DFT, PLA, PAA: Παραγωγή συμπιεσμένων αριθμητικών διανυσμάτων μειωμένης διάστασης μέσω των αντίστοιχων συναρτήσεων αναπαράστασης `repr_dft()`, `repr_pla()` και `paa()`.
3. Προ-επεξεργασία κατηγορικών Δεδομένων (Feature Engineering and Preprocessing):
 - a. Για τα Δέντρα Απόφασης και Random Forests: Στην περίπτωση της SAX, η απευθείας εισαγωγή κατηγορικών συμβολοσειρών προκαλεί συχνά σφάλματα τύπου “unseen factor levels” (όταν το test set περιέχει συμβολικούς συνδυασμούς που δεν υπήρχαν στο train set) ή σφάλματα ορίων πληθικότητας (factor limits). Για την επίλυση του προβλήματος, χρησιμοποιήθηκε το πακέτο `recipes`. Μέσω κατάλληλων βημάτων μετασχηματισμού τα κατηγορικά δεδομένα διαμορφώθηκαν σε πλήρων συμβατή μορφή. Για της αριθμητικές μεθόδους (DFT, PLA, PAA), τα παραγόμενα διανύσματα τροφοδοτήθηκαν άμεσα ως συνεχείς μεταβλητές.
 - b. Για το CatBoost: Ακολουθήθηκαν οι αυστηρές προδιαγραφές προ-επεξεργασίας του επίσημου documentation του αλγορίθμου, μετατρέποντας τις συμβλικές μεταβλητές σε παράγοντες (factors) και δομές `catboost.Pool`, επιτρέποντας στον αλγόριθμο να αξιοποιήσει το μηχανισμός Ordered Target Statistics.

Πριν από την εφαρμογή των μεθόδων μείωσης διαστάσεων (PAA, SAX, DFT, PLA), όλοι οι αλγόριθμοι κατηγοριοποίησης εκπαιδεύτηκαν και αξιολογήθηκαν πάνω στα πλήρη, αρχικά δεδομένα των χρονοσειρών (Raw Time Series). Η διαδικασία αυτή αποτέλεσε το βασικό σημείο αναφοράς (baseline),

επιτρέποντας την ποσοτικοποίηση των επιπτώσεων της συμπίεσης και της απώλειας πληροφορίας στην τελική επίδοση των μοντέλων.

5.2.3 Καθορισμός Παραμέτρων Αναπαραστάσεων και Μοντέλων Μάθησης

Για τη διασφάλιση μιας δίκαιης και αντικειμενικής συγκριτικής αξιολόγησης, ο καθορισμός των παραμέτρων διαμορφώθηκε ως εξής:

Συμβολική Αθροιστική Προσέγγιση (SAX). Οι παράμετροι w (μήκος συμβολικής λέξης/ αριθμός πλαισίων PAA) και a (μέγεθος αλφαβήτου ορίστηκαν αυστηρά σύμφωνα με τις βασικές πειραματικές διατάξεις της θεμελιώδους εργασίας των Lin et al.

Αριθμητικές Μέθοδοι (PAA, DFT, PLA): Προκειμένου να υπάρχει απόλυτη ισοδυναμία στον βαθμό συμπίεσης και στον χώρο χαρακτηριστικών, οι μέθοδοι PAA, DFT και PLA ρυθμίστηκαν ώστε να παράγουν διάνυσμα διάστασης ίσης με το w της εκάστοτε χρονοσειράς:

- Α. Στην PAA, χρονοσειρά διαιρέθηκε σε $N=w$ ισομεγέθη χρονικά πλαίσια (frame size n/w).
- Β. Στον DFT, διατηρήθηκαν οι πρώτοι συντελεστές Fourier που αντιστοιχούν σε χώρο διάστασης w
- Γ. Στη PLA, κατασκευάστηκαν τμήματα γραμμικής προσέγγισης αντίστοιχου βαθμού συμπίεσης w

Υπερπαράμετροι Αλγορίθμων Ταξινόμησης

Οι αλγόριθμοι Εξόρυξης δεδομένων παραμετροποιήθηκαν ακολουθώντας τις βέλτιστες πρακτικές και τις καθιερωμένες προδιαγραφές των αντίστοιχων πακέτων.

Για τον K-NN χρησιμοποιήθηκε $k = 1$ πλήθος γειτόνων και ως μετρική ομοιότητας χρησιμοποιήθηκε η MINDIST για την μέθοδο SAX, ενώ για τις αριθμητικές μεθόδους έγινε χρήση της κλασσικής ευκλείδειας απόστασης.

Για τα Μεμονωμένα Δέντρα απόφασης

1. Μέθοδος διαχωρισμού: Κατηγοριοποίηση (method="class").
2. Κριτήριο διάσπασης: Δείκτης Gini (default).
3. Πλήρης ανάπτυξη και εκπαίδευση επί του διανύσματος χαρακτηριστικών χωρίς εξωτερικό περιορισμό βάθους.

Τυχαία Δάση οι παράμετροι που χρησιμοποιήθηκαν:

1. Αριθμός δέντρων: $n_{tree} = 500$
2. Επιλογή χαρακτηριστικών ανά κόμβο $m_{try} = \lfloor \sqrt{M} \rfloor$, όπου M ο συνολικός αριθμός παραγόμενων χαρακτηριστικών/διαστάσεων ($M = w$).
3. Συνδυασμός αποφάσεων: Πλειοψηφική ψηφοφορία (Majority Voting)

CatBoost:

1. Συνάρτηση απώλειας: `loss_function="MultiClass"`
2. Αριθμός επαναλήψεων (boosting iterations): `iterations=300`
3. Ρυθμός μάθησης (learning rate): $\eta = 0.05$
4. Μέγιστο βάθος λησμονούντων δέντρων (tree depth): `depth=6`
5. Διαχείριση κατηγορικών χαρακτηριστικών: Αυτόματη μετατροπή μέσω `catboost.load_pool()` και αξιοποίηση των Ordered Target Statistics (OTS).
6. Καταγραφή εκτέλεσης: `logging_level="Silent"`

5.3 Περιβάλλον Υλοποίησης και Πακέτα Λογισμικού (R/RStudio)

Όλοι οι υπολογισμοί, οι μετασχηματισμοί δεδομένων, η εκπαίδευση των μοντέλων, η αξιολόγηση των αποτελεσμάτων και η εξαγωγή διαγραμμάτων πραγματοποιήθηκαν στο προγραμματιστικό περιβάλλον R() μέσω του ολοκληρωμένου περιβάλλοντος ανάπτυξης RStudio.

Τα εξειδικευμένα πακέτα (libraries) της R που χρησιμοποιήθηκαν περιλαμβάνουν:

1. TSreper: Αποτέλεσε το κεντρικό πακέτο για την υλοποίηση των αριθμητικών μεθόδων αναπαράστασης χρονοσειρών. Χρησιμοποιήθηκαν οι συναρτήσεις του για την εκτέλεση του Διακριτού Μετασχηματισμού Φουριέ (DFT) και της Τμηματικής Γραμμικής Προσέγγισης (PLA)
2. TSclust: Χρησιμοποιήθηκε για τον υπολογισμό των αποστάσεων χρονοσειρών στο συμβολικό πεδίο μέσω της συνάρτησης `diss.MINDIST.SAX()`.
3. Jmotif: χρησιμοποιήθηκε για την υλοποίηση των επιμέρους σταδίων της SAX (`znorm()`, `raa()`, `series_to_string()`) και τη μετατροπή των χρονοσειρών σε συμβολικές λέξεις.
4. Recipes(Tidymodels): Χρησιμοποιήθηκε για την προ-επεξεργασία των κατηγορικών συμβολικών δεδομένων της SAX, επιλύοντας προβλήματα ασυμβατότητας επιπέδων (factor levels) στα Δέντρα Απόφασης και τα Τυχαία Δάση.
5. Rpart: χρησιμοποιήθηκε για την κατασκευή και εκπαίδευση των μεμονωμένων Δέντρων απόφασης (Recursive Partitioning and Regression Trees).
6. randomForest: Χρησιμοποιήθηκε για την υλοποίηση του αλγορίθμου Τυχαίων Δασών (Random Forests) με δειγματοληψία bootstrap και τυχαία επιλογή χαρακτηριστικών.
7. Catboost: Χρησιμοποιήθηκε η επίσης υλοποίησης της Yandex για την εκτέλεση του αλγορίθμου Categorical Boosting.
8. PMCRMplus: Χρησιμοποιήθηκε για την διεξαγωγή των μη παραμετρικών post-hoc πολλαπλών συγκρίσεων κατά Conover (`frdAllPairsConoverTest`), μετά τον έλεγχο Friedman.
9. Tidymodels(dplyr, tidyr, purrr): Χρησιμοποιήθηκε για την επεξεργασία, συνένωση και αναδιαμόρφωση των πειραματικών αποτελεσμάτων σε ενιαία δομή δεδομένων.
10. Ggplot2 και scales: Έγινε χρήση των δύο πακέτων για την οπτικοποίηση των αποτελεσμάτων μέσω ραβδογραμμάτων ανά σύνολο δεδομένων και διαγραμμάτων πλαισίου-απολήξεων (boxplots) με κατάλληλη κλιμάκωση ποσοστών.

5.4 Μετρικές Αξιολόγησης (Evaluation Metrics)

Για την αντικειμενική αξιολόγηση και τη σύγκριση της απόδοσης των αλγορίθμων κατηγοριοποίησης (1-NN, Decision Trees, Random Forests, CatBoost) πάνω στις διαφορετικές αναπαραστάσεις (SAX, DFT, PLA, PAA), χρησιμοποιήθηκε ο Πίνακας Σύγχυσης (Confusion Matrix) και οι τέσσερις θεμελιώδεις μετρικές αξιολόγησης της μηχανικής μάθησης:

Πίνακας Σύγχυσης (Confusion Matrix)

Για ένα πρόβλημα κατηγοριοποίησης, ο πίνακας σύγχυσης καταγράφει τις πραγματικές κλάσεις έναντι των προβλέψεων του μοντέλου. Στη δυαδική κατηγοριοποίηση αποτελείται από:

- True Positives (TP): Θετικά δείγματα που προβλέφθηκαν σωστά ως θετικά.
- True Negatives (TN): Αρνητικά δείγματα που προβλέφθηκαν σωστά ως αρνητικά.
- False Positives (FP): Αρνητικά δείγματα που λανθασμένα προβλέφθηκαν ως θετικά.
- False Negatives (FN): Θετικά δείγματα που λανθασμένα προβλέφθηκαν ως αρνητικά.

Ακρίβεια (Accuracy)

Εκφράζει το συνολικό ποσοστό των σωστών προβλέψεων του μοντέλου επί του συνολικού πλήθους των δειγμάτων N :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.1)$$

Πιστότητα (Precision):

Μετρά την ακρίβεια των θετικών προβλέψεων, δηλαδή πόσες από τις περιπτώσεις που το μοντέλο κατέταξε ως θετικές ήταν πράγματι θετικές:

$$Precision = \frac{TP}{TP+FP} \quad (5.2)$$

Ανάκληση (Recall/Sensitivity):

Μετρά την ικανότητα του μοντέλου να εντοπίζει το σύνολο των πραγματικών θετικών δειγμάτων:

$$Recal = \frac{TP}{TP+FN} \quad (5.3)$$

Μέτρο F1 (F1-Score)

Αποτελεί τον αρμονικό μέσο της Πιστότητας και της Ανάκλησης (Recall), προσφέροντας μια ισορροπημένη εικόνα της απόδοσης του μοντέλου, ιδιαίτερα σε σύνολα δεδομένων με άνιση κατανομή κλάσεων (imbalanced datasets):

$$F1 = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (5.4)$$

5.5 Στατιστική Μεθοδολογία και Έλεγχος Υποθέσεων

Για να διαπιστωθεί εάν οι παρατηρούμενες διαφορές στην απόδοση των μοντέλων οφείλονται σε τυχαίες διακυμάνσεις ή είναι στατιστικά σημαντικές, ακολουθήθηκε αυστηρό πρωτόκολλο μη παραμετρικών στατιστικών ελέγχων (non-parametric hypothesis testing), σύμφωνα με τις καθιερωμένες πρακτικές σε πολλαπλά datasets.

5.5.1 Συγκρίσεις Ζευγών: Έλεγχος Wilcoxon Signed-Rank

Για τον άμεσο έλεγχο της επίδρασης κάθε μεθόδου μείωσης διάστασης σε σχέση με τα αρχικά δεδομένα (Raw Baseline), εφαρμόστηκε ο διπαραμετρικός έλεγχος προσημασμένων τάξεων Wilcoxon (Paired Wilcoxon Signed-Rank Test).

Για κάθε κατηγοριοποιητή (KNN, DT, RF, CB) και μέθοδο αναπαράστασης (PAA, SAX, DFT, PLA), ελέγχεται η μηδενική υπόθεση (H_0):

$$H_0: \text{Διαφορά επίδοσης} (Accuracy_{Raw} - Accuracy_m) = 0$$

Η στατιστική συνάρτηση V υπολογίζεται επί των $N=10$ συνόλων δεδομένων.

5.5.2 Συγκριτική Αξιολόγηση Πολλαπλών Αναπαραστάσεων: Έλεγχος Friedman

Για τη συνολική σύγκριση των 5 αναπαραστάσεων (Raw, PAA, SAX, DFT, PLA) ανά κατηγοριοποιητή, εφαρμόστηκε ο μη παραμετρικός έλεγχος επαναλαμβανόμενων μετρήσεων Friedman (Friedman Test).

1. Κατάταξη Αναπαραστάσεων(Ranking): Σε κάθε dataset, οι 5 αναπαραστάσεις κατατάσσονται από την 1η (υψηλότερη ακρίβεια) έως την 5η θέση (χαμηλότερη ακρίβεια).
2. Μέση κατάταξη (Average Rank): Υπολογίζεται η μέση τάξη $\bar{R}_j$ για κάθε αναπαράσταση j .
3. Έλεγχος Υπόθεσης: Ελέγχεται η μηδενική υπόθεση H_0 ότι όλες οι αναπαραστάσεις έχουν ισοδύναμη επίδοση:

$$H_0: \bar{R}_1 = \bar{R}_2 = \bar{R}_3 = \bar{R}_4 = \bar{R}_5$$

5.5.3 Post-Hoc Ανάλυση Conover και Διόρθωση Πολλαπλών Ελέγχων

Στις περιπτώσεις όπου ο έλεγχος Friedman απέρριψε την μηδενική υπόθεση ($p < 0.05$), πραγματοποιήθηκε post-hoc ανάλυση πολλαπλών συγκρίσεων ανά ζεύγη μέσω της μεθόδου Conover (frdAllPairConoverTest).

Επειδή η ταυτόχρονη εκτέλεση πολλαπλών ελέγχων υποθέσεων αυξάνει την πιθανότητα σφάλματος Τύπου I (Family-Wise Error Rate), οι υπολογισθείσες τιμές p-value προσαρμόστηκαν με τις παρακάτω μεθόδους:

1. Διόρθωση Holm (Holm-Bonferroni): Αυστηρή διαδικασία βηματικής προσαρμογής για τον έλεγχο του συνολικού ρυθμού σφάλματος (FWER).
2. Διόρθωση Benjamini-Hochberg(BH): Έλεγχος του ποσοστού ψευδώς θετικών ανακαλύψεων (False Discovery Rate – FDR).

Κάθε διαφορά θεωρήθηκε στατιστικά σημαντική σε επίπεδο σημαντικότητας $\alpha = 0.05$ ($P_{adj} < 0.05$).

5.6 Επίλογος

Συνοψίζοντας, στο κεφάλαιο αυτό τέθηκαν οι αυστηρές προδιαγραφές του πειραματικού σχεδιασμού. Παρουσιάστηκαν τα 10 benchmark datasets του αποθετηρίου UCR, εξειδικεύτηκε η ροή προεπεξεργασίας και εκπαίδευσης για όλους τους συνδυασμούς μοντέλων και αναπαραστάσεων στο περιβάλλον R, και ορίστηκε το πλαίσιο των μη παραμετρικών στατιστικών ελέγχων (Wilcoxon, Friedman, Conover post-hoc). Με το μεθοδολογικό αυτό πλαίσιο πλήρως συγκροτημένο, το Κεφάλαιο 6 που ακολουθεί επικεντρώνεται στην αναλυτική παρουσίαση, σύγκριση και στατιστική ερμηνεία των πειραματικών αποτελεσμάτων.

Κεφάλαιο 6ο: Παρουσίαση Αποτελεσμάτων

6.1 Εισαγωγή και Μεθοδολογικό Πλαίσιο Αξιολόγησης

Στο παρόν κεφάλαιο παρουσιάζονται, αναλύονται και συγκρίνονται τα πειραματικά ευρήματα της παρούσας πτυχιακής εργασίας. Βασικός ερευνητικός στόχος είναι η αξιολόγηση της μεθόδου Συμβολικής Αθροιστικής Προσέγγισης (Symbolic Aggregate Approximation – SAX) ως προς την αποτελεσματικότητά της στην κατηγοριοποίηση χρονοσειρών, σε σύγκριση τόσο με τα αρχικά δεδομένα (Raw Time Series) όσο και με καθιερωμένες μεθόδους μείωσης διαστάσεων συνεχών τιμών: την Τμηματική Αθροιστική Προσέγγιση(PAA), το Διακριτό Μετασχηματισμό Fourier (DFT) και την Τμηματική Γραμμική Προσέγγιση(PLA).

Η συγκριτική ανάλυση διεξήχθη σε δέκα (10) αντιπροσωπευτικά σύνολα δεδομένων από το αποθετήριο UCR Time Series Classification Archive, καλύπτοντας ένα ευρύ φάσμα χαρακτηριστικών: από συνθετικά μοτίβα(Synthetic Control, CBF, Two Patterns) και αναγνώριση κινήσεων (Gun-Point, Trace), έως περιγράμματα εικόνων (Face all, OSU Leaf, Swedish Leaf, 50Words) και βιομηχανικούς αισθητήρες(Wafer). Για κάθε αναπαράσταση αξιολογήθηκαν τέσσερις διαφορετικές κατηγορίες αλγορίθμων μηχανικής μάθησης:

1. K-Nearest Neighbors(1-NN): Οκνηρός κατηγοριοποιητής βασισμένος σε συναρτήσεις απόστασης (MINDIST για τη SAX, Ευκλείδεια απόσταση για τις υπόλοιπες).
2. Decision Trees(CART): Μεμονωμένο ιεραρχικό μοντέλο δέντρου απόφασης.
3. Random Forest(RF): Μέθοδος συλλογικής μάθησης (bagging) με τυχαία επιλογή χαρακτηριστικών
4. CatBoost(CB): Προηγμένος αλγόριθμος Gradient Boosting βασισμένος σε λησιμονούντα δέντρα(oblivious trees) και Ordered Target Statistics.

Η αξιολόγηση βασίζεται στον δείκτη της συνολικής Ακρίβειας Ταξινόμησης (Classification Accuracy), συνοδευόμενη από μη παραμετρικό στατιστικό έλεγχο υποθέσεων σύμφωνα με τις προδιαγραφές συγκριτικής αξιολόγησης αλγορίθμων μάθησης.

6.2 Συγκεντρωτικά Πειραματικά Αποτελέσματα Ακρίβειας

Στους πίνακες 6.1 έως 6.4 παρατίθενται αναλυτικά οι τιμές ακρίβειας ταξινόμησης ανά σύνολο δεδομένων και μέθοδο αναπαράστασης, διαχωρισμένες ανά κατηγοριοποιητή. Στην τελευταία γραμμή κάθε πίνακα καταγράφεται η μέση ακρίβεια επί του συνόλου των 10 προβλημάτων.

Πίνακας 6.1: Ακρίβεια Ταξινόμησης για τον Αλγόριθμο k-NN(k=1)

Σύνολο Δεδομένων(Dataset)	Raw(1-NN)	PAA(1-NN)	SAX(1-NN)	DFT(1-NN)	PLA(1-NN)
Synthetic Control	0.88	0.9967	0.9833	0.89	0.74
Gun Point	0.9913	0.9067	0.8067	0.9067	0.92
CBF	0.8520	0.95	0.8978	0.8944	0.7322
Face (all)	0.7140	0.7148	0.6686	0.7598	0.4568
OSU Leaf	0.5170	0.5165	0.5372	0.5207	0.5083
Swedish Leaf	0.7890	0.7952	0.5088	0.7888	0.4816
50Words	0.6310	0.6309	0.6549	0.6308	0.6593
Trace	0.76	0.76	0.54	0.75	0.74
Two Patterns	0.9070	0.9380	0.9180	0.8728	0.61
Wafer	0.9955	0.9958	0.9966	0.9942	0.9706
Μέσος Όρος (Mean)	0.8037	0.8205	0.7512	0.8008	0.6819

Πίνακας 6.2: Ακρίβεια Ταξινόμησης γι Μεμονωμένα Δέντρα Απόφασης (Decision Trees)

Σύνολο Δεδομένων(Dataset)	Raw(DT)	PAA(DT)	SAX(DT)	DFT(DT)	PLA(DT)
Synthetic Control	0.7333	0.83	0.8333	0.70	0.68
Gun-Point	0.7867	0.8067	0.8267	0.7533	0.8467
CBF	0.7867	0.5911	0.8289	0.5633	0.5189
Face (all)	0.5521	0.5112	0.4893	0.5444	0.4485
OSU Leaf	0.3099	0.2934	0.2521	0.2851	0.2934
Swedish Leaf	0.5440	0.6064	0.4784	0.5632	0.4928
50Words	0.3736	0.4088	0.3429	0.4044	0.4647
Trace	0.75	0.79	0.69	0.78	0.79

Two Patterns	0.55	0.5730	0.5405	0.4970	0.45
Wafer	0.9734	0.9604	0.9666	0.9510	0.9450
Μέσος Όρος (Mean)	0.6360	0.6371	0.6249	0.6042	0.5929

Πίνακας 6.3: Ακρίβεια Ταξινόμησης για Τυχαία Δάση (Random Forests)

Σύνολο Δεδομένων (Dataset)	Raw(RF)	PAA(RF)	SAX(RF)	DFT(RF)	PLA(RF)
Synthetic Control	0.9567	0.97	0.9667	0.8967	0.86
Gun-Point	0.9333	0.9467	0.91	0.9267	0.86
CBF	0.8967	0.8844	0.9344	0.8733	0.7733
Face (all)	0.8095	0.7710	0.7178	0.8556	0.6817
OSU Leaf	0.4959	0.5041	0.4008	0.5083	0.4876
Swedish Leaf	0.8816	0.8512	0.7904	0.8688	0.7664
50Words	0.6659	0.6703	0.6220	0.6681	0.6769
Trace	0.82	0.86	0.77	0.84	0.75
Two Patterns	0.85	0.8590	0.8160	0.8475	0.6328
Wafer	0.9878	0.9890	0.9875	0.9891	0.9773
Μέσος Όρος (Mean)	0.8297	0.8306	0.7916	0.8274	0.7456

Πίνακας 6.4: Ακρίβεια Ταξινόμησης για τον Αλγόριθμο CatBoost

Σύνολο Δεδομένων (Dataset)	Raw(CB)	PAA(CB)	SAX(CB)	DFT(CB)	PLA(CB)
Synthetic Control	0.94	0.9667	0.9533	0.91	0.88
Gun-Point	0.9133	0.9133	0.8067	0.94	0.86
CBF	0.8044	0.8678	0.8556	0.8778	0.6789
Face (all)	0.7633	0.7367	0.6834	0.7331	0.6698
OSU Leaf	0.5083	0.50	0.3388	0.4793	0.4504
Swedish Leaf	0.8784	0.8560	0.7584	0.8848	0.7376

50Words	0.7055	0.6374	0.4923	0.6440	0.6571
Trace	0.85	0.84	0.76	0.81	0.77
Two Patterns	0.93	0.9120	0.9365	0.9160	0.7855
Wafer	0.9930	0.9925	0.9924	0.9895	0.9791
Μέσος Όρος (Mean)	0.8286	0.8222	0.7578	0.8185	0.7468

6.3 Αναλυτική Συγκριτική Μελέτη ανά Μέθοδο Αναπαράστασης

6.3.1 Συμβολική Αθροιστική Προσέγγιση (SAX)

Η μέθοδος SAX αποτελεί το επίκεντρο της παρούσας μελέτης λόγω της μοναδικής της ικανότητας να μετασχηματίζει συνεχή αριθμητικά σήματα σε διακριτές συμβολοσειρές χαρακτήρων. Από τα πειραματικά δεδομένα προκύπτουν τα εξής συμπεράσματα:

1. Απόδοση στον 1-NN με τη Συνάρτηση MINDIST: Η SAX επιτυγχάνει μέση ακρίβεια 75.12% των αρχικών δεδομένων. Ωστόσο, σε προβλήματα όπου κυριαρχούν έντονα μορφολογικά πρότυπα με υψηλό θόρυβο, η SAX όχι μόνο δεν υπολείπεται, αλλά υπερτερεί των ανεπεξέργαστων δεδομένων:
 - a. Στο Synthetic Control, η ακρίβεια ανέρχεται σε 98.33% (έναντι στο 88.00% στα Raw).
 - b. Στο CBF (Cylinder-Bell-Funnel), επιτυγχάνει 89.80% έναντι 90.70%.
 - c. Στο Wafer, αγγίζει το σχεδόν τέλει 99.66% έναντι 99.55%.

Η συμπεριφορά αυτή εξηγείται θεωρητικά: η διακριτοποίηση βάσει των Γκαουσιανών σημείων θραύσης (breakpoint) λειτουργεί ως μη γραμμικό φίλτρο εξομάλυνσης (smoothing), απορρίπτοντας τον τοπικό θόρυβο υψηλής συχνότητας που αποπροσανατολίζει την Ευκλείδεια απόσταση.

2. Περιορισμοί σε Πολυταξικά Προβλήματα Περιγραμμάτων:

Αντίθετα, η SAX παρουσιάζει αισθητή πτώση σε προβλήματα με μεγάλο αριθμό κλάσεων και σύνθετα γεωμετρικά σχήματα, όπως το Swedish Leaf(50.88% έναντι 78.90%) και το Trace(54.00% έναντι 76.00%). Σε αυτές τις περιπτώσεις, η κβάντιση συνεχών τιμών σε πεπερασμένα σύμβολα εξαλείφει κρίσιμες μικροδιαφορές πλάτους (amplitude nuances) που είναι απαραίτητες για τη διάκριση συγγενικών κλάσεων.

3. Συμπεριφορά σε Μοντέλα Δέντρων και Boosting:

Στα μοντέλα Random Forests και CatBoost, η SAX διατηρεί αξιοσημείωτα ανταγωνιστική απόδοση (79.16% και 75.78% αντίστοιχα). Στο σύνολο CBF, τα μοντέλα RF και CatBoost με χαρακτηριστικά SAX επιτυγχάνουν 93.44% και 85.56%, ξεπερνώντας τα αντίστοιχα μοντέλα Raw δεδομένων (89.67% και 80.44%).

6.3.2 Τμηματική Αθροιστική Προσέγγιση (PAA)

Η μέθοδος PAA αναδικνύεται ως η πλέον ισορροπημένη και σταθερή τεχνική μείωσης διαστάσεων σε ολόκληρο τον πειραματικό έλεγχο:

Στον κατηγοριοποιητή 1-NN, η PAA καταγράφει την υψηλότερη μέση ακρίβεια ανάμεσα σε όλες τις αναπαραστάσεις (82.05% έναντι 80.37% των Raw δεδομένων), επιτυγχάνοντας μέση βελτίωση +1.68%.

Στα Random Forests, η PAA διατηρεί την πρώτη θέση με μέση ακρίβεια 83.06% (έναντι 82.97% των Raw).

Στο CatBoost, η PAA επιτυγχάνει 82.22, ευρισκόμενη σε στατιστική ισοδυναμία με τα πλήρη δεδομένα (82.86%).

Το γεγονός ότι η PAA καταφέρνει να συμπίσει το μήκος των χρονοσειρών διατηρώντας παράλληλα υψηλότερη ακρίβεια από τα αρχικά σήματα οφείλεται στην ιδιότητα του μέσου όρου πλαισίου (frame averaging) να εξαλείφει τον στοχαστικό θόρυβο χωρίς να αλλοιώνει την θεμελιώδη μακροσκοπική τάση.

6.3.3 Διακριτός Μετασχηματισμός Fourier (DFT)

Ο μετασχηματισμός Fourier επιβεβαιώνει την θεωρητική του υπεροχή στη διαχείριση σημάτων με αρμονικές συνιστώσες:

- A. Στον 1-NN, επιτυγχάνει μέση ακρίβεια 80.08%, με απόλυτα συγκρίσιμη με τα Raw δεδομένα (80.37%). Στο σύνολο Face(all), ο DFT επιτυγχάνει την κορυφαία επίδοση (75.98%), ξεπερνώντας τα Raw(71.40%) και τη SAX (66.86%).
- B. Στα Random Forests και το CatBoost, καταγράφει 82.74% και 81.85% αντίστοιχα. Στο Swedish Leaf με CatBoost, ο DFT αναδεικνύεται πρώτος με 88.48%.

Η διατήρηση της ενέργειας στις χαμηλότερες συχνότητες (μέσω του θεωρήματος Parseval) καθιστά τον DFT μια στιβαρή επιλογή για χρονοσειρές όπου η πληροφορία κωδικοποιείται σε περιοδικά μοτίβα.

6.3.4 Τμηματική Γραμμική Προσέγγιση (PLA)

Η προσέγγιση PLA κατέγραψε τη χαμηλότερη συνολική απόδοση σε όλες τις διατάξεις:

- A. Στον 1-NN, η μέση ακρίβεια κατακρημνίζεται στο 68.19% (πτώση άνω του 12% σε σχέση με Raw/PAA). Ειδικά στα σύνολα Face All (45.68%), Swedish Leaf (48.16%) Two Patterns (61%), η απώλεια πληροφορίας είναι καταλυτική.
- B. Στα μοντέλα Random Forests και CatBoost, η μέση ακρίβεια περιορίζεται στο 74.56% και 74.68% αντίστοιχα.

Η υστέρηση της PLA εξηγείται από τη δυσκολία του καθορισμού σταθερών ορίων κατάτμησης (segmentation breakpoints) σε ετερογενή σήματα, με αποτέλεσμα οι τοπικές κλίσεις των ευθειών να εισάγουν σφάλματα αναπαράστασης όταν συνδυάζονται με Ευκλείδειες αποστάσεις.

6.4 Στατιστικός Έλεγχος Υποθέσεων

Για την εξαγωγή ασφαλών επιστημονικών συμπερασμάτων που δεν υπόκεινται σε τυχαίες δειγματοληπτικές διακυμάνσεις, εφαρμόστηκε η αυστηρή μεθοδολογία μη παραμετρικών στατιστικών ελέγχων πολλαπλών συνόλων δεδομένων.

6.4.1 Έλεγχος Ζευγών με Paired Wilcoxon Signed-Rank Test

Αρχικά εκτελέστηκε ο έλεγχος Wilcoxon Signed-Rank Test για κάθε ζεύγος “Raw vs. Μέθοδος Μείωσης” ανά κατηγοριοποιητή (σύνολο 16 συγκρίσεων).

Πίνακας 6.5: Αποτελέσματα Ελέγχου Wilcoxon (Raw vs. Reduced Representations)

Meth od	Mod el	N	Mean Raw	Mean Reduce d	Mean Diff	V- Stat	P- value(R aw)	P- value(Holm)	P- value(B H)	Στατ. Σημ. ($\alpha=0.05$)
PAA	KNN	10	0.8037	0.8205	+0.0168	11.0	0.1925	1.0000	0.4400	Όχι
PAA	DT	10	0.6360	0.6371	+0.0011	20.0	0.4755	1.0000	0.6340	Όχι
PAA	RF	10	0.8297	0.8306	+0.0008	22.0	0.6203	1.0000	0.7511	Όχι
PAA	CB	10	0.8286	0.8222	-0.0064	30.0	0.4069	1.0000	0.6034	Όχι
SAX	KNN	10	0.8037	0.7512	-0.0525	32.0	0.6835	1.0000	0.7811	Όχι
SAX	DT	10	0.6360	0.6249	-0.0111	36.0	0.4148	1.0000	0.6034	Όχι
SAX	RF	10	0.8297	0.7916	-0.0382	48.0	0.0415	0.5394	0.1660	Μόνο Raw
SAX	CB	10	0.8286	0.7578	-0.0709	46.0	0.0665	0.7985	0.2130	Όχι
DFT	KNN	10	0.8037	0.8008	-0.0029	28.5	0.9593	1.0000	0.9593	Όχι
DFT	DT	10	0.6360	0.6042	-0.0318	42.0	0.1536	1.0000	0.4095	Όχι
DFT	RF	10	0.8297	0.8274	-0.0023	31.0	0.7598	1.0000	0.8104	Όχι
DFT	CB	10	0.8286	0.8185	-0.0102	39.0	0.2622	1.0000	0.4661	Όχι
PLA	KNN	10	0.8037	0.6819	-0.1218	51.0	0.0191	0.2668	0.1016	Μόνο Raw
PLA	DT	10	0.6360	0.5929	-0.0431	39.0	0.2622	1.0000	0.4661	Όχι
PLA	RF	10	0.8297	0.7456	-0.0841	52.0	0.0144	0.2165	0.1016	Μόνο Raw
PLA	CB	10	0.8286	0.7468	-0.0818	55.0	0.0059	0.0947	0.0947	Μόνο Raw

Ερμηνεία Ευρημάτων Wilcoxon:

1. Απουσία Στατιστικής Υστέρησης για PAA και DFT: Σε όλους τους κατηγοριοποιησές, οι τιμές p για τις μεθόδους PAA($p>0.19$) και DFT($p>0.15$) επιβεβαιώνουν ότι η μείωση διαστάσεων επιτυγχάνεται χωρίς καμία στατιστικά σημαντική απώλεια ακρίβειας.
2. Επίπτωση της Διόρθωσης Πολλαπλών Ελέγχων (Holm Adjustment): Πρι από τη διόρθωση, η PLA εμφάνιζε στατιστικά σημαντική πτώση έναντι των Raw δεδομένων σε KNN($p=0.0191$), RF($p=0.0144$) και CB ($p=0.0059$), ενώ η SAX υπολειπόταν οριακά στο RF($p=0.0415$). Κατά την εφαρμογή της αυστηρής διόρθωσης Holm σε 16 ταυτόχρονους ελέγχους με μέγεθος δείγματος $N=10$, τα προσαρμοσμένα p -values αυξάνονται άνω του 0.05. Αυτό καθιστά αναγκαία τη διεξαγωγή καθολικού ελέγχου διατάξεων (Friedman Test), ο οποίος διαθέτει

μεγαλύτερη στατιστική ισχύ συγκεντρώνοντας τις σχετικές κατατάξεις όλων των αναπαραστάσεων ταυτόχρονα.

6.4.2 Συνολικός Έλεγχος Διατάξεων Friedman και Post-Hoc Conover Test ανά Κατηγοριοποιητή

Ο έλεγχος Friedman Rank Sum Test εφαρμόστηκε σε επίπεδο ξεχωριστού κατηγοριοποιητή, αντιμετωπίζοντας κάθε σύνολο δεδομένων ως ανεξάρτητο μπλοκ (block). Για τους αλγόριθμους όπου διαπιστώθηκε στατιστικά σημαντική διαφοροποίηση, ακολούθησε ο μετά-πειραματικός έλεγχος Conover Post-Hoc Test με διόρθωση Holm.

1. Ανάλυση για τον Κατηγοριοποιητή 1-NN
 - a. Friedman Test $\chi_F^2 = 9.2386$, $df = 4$, $p - value = 0.0554$ (οριακή στατιστική τάση σε επίπεδο $\alpha = 0.10$).

b. Μέσες Κατατάξεις (Average Ranks – Χαμηλότερη τιμή = Καλύτερη απόδοση):

PAA(2.15) < Raw(2.75) < SAX (2.80) < DFT(3.10) < PLA(4.20)

c. Conover Post-Hoc (p-values):

PAA vs. PLA: $p=0.0319$ (Στατιστικά σημαντική υπεροχή της PAA έναντι της PLA).

Raw vs. SAX: $p=1.000$ (Απόλυτη στατιστική ισοδυναμία μεταξύ αρχικών δεδομένων και συμβολικής αναπαράστασης SAX)

2. Ανάλυση για τα Μεμονωμένα Δέντρα Απόφασης (Decision Trees)

a. Friedman Test $\chi_F^2 = 5.9798$, $df = 4$, $p - value = 0.2007$ (Μη στατιστικά σημαντικό).

b. Μέσες κατατάξεις:

PAA(2.20)<Raw(2.50)<SAX(3.30)<DFT(3.50) = PLA(3.50)

Συμπέρασμα: Τα μεμονωμένα δέντρα CART χαρακτηρίζονται από υψηλή ενδογενή διακύμανση, με αποτέλεσμα οι διαφορές μεταξύ των μεθόδων αναπαράστασης να μην είναι στατιστικά διακριτές.

3. Ανάλυση για τα Τυχαία Δάση (Random Forests)

A. Friedman Test: $\chi_F^2 = 18.24$, $df = 4$, $p - value = 0.0011$ (Εξαιρετικά στατιστικά σημαντικό).

B. Μέσες κατατάξεις:

a. PAA(1.90)< DFT(2.40) < Raw(2.50) < SAX(3.70)<PLA(4.50)

C. Conover Post-Hoc(p-values):

a. PAA vs. PLA: $p=0.0003$ (Υπεροχή PAA έναντι PLA).

b. DFT vs. PLA: $p=0.0046$ (Υπεροχή DFT έναντι PLA).

c. Raw vs. PLA: $p=0.0068$ (Υπεροχή Raw έναντι PLA).

d. PAA vs. SAX: $p=0.0164$ (Υπεροχή PAA έναντι SAX).

e. Raw vs. PAA: $p=0.8470$ και Raw vs. DFT: $p=0.8567$ (Πλήρης στατιστική ισοδυναμία).

4. Ανάλυση για τον αλγόριθμο CatBoost

- A. Friedman Test: $\chi_F^2 = 16.6260, df = 4, p - value = 0.0023$ (Εξαιρετικά στατιστικά σημαντικό).
- B. Μέσες Κατατάξεις:
 - a. Raw(1.85) < PAA(2.50) < DFT(2.60) < SAX(3.65) < PLA(4.40)
- C. Conover Post-Hoc(p-values):
 - a. Raw vs. PLA: $p=0.000L$ (Υπεροχή Raw έναντι PLA)
 - b. PAA vs. PLA: $p=0.0171$ (Υπεροχή PAA έναντι PLA)
 - c. DFT vs. PLA: $p=0.0245$ (Υπεροχή DFT έναντι PLA)
 - d. Raw vs. SAX: $p=0.0245$ (Στατιστική υπεροχή Raw έναντι SAX)
 - e. Raw vs. PAA: $p=0.7768$ και Raw vs. DFT: $p=0.7768$ (Πλήρης στατιστική ισοδυναμία).

6.5 Συγκριτική Ανάλυση Υπολογιστικής Πολυπλοκότητας και Συμπίεσης

Πέραν της ακρίβειας ταξινόμησης κρίσιμος άξονας αξιολόγησης αποτελεί και ο βαθμός συμπίεσης και η υπολογιστική πολυπλοκότητα κάθε μεθόδου:

1. Απαιτήσεις Αποθήκευσης και Αναπαράσταση Δεδομένων:
 - a. Οι μέθοδοι Raw, PAA, DFT και PLA διατηρούν αριθμητικά δεδομένα κινητής υποδιαστολής διπλής ακρίβειας (64-bit float = 8 bytes ανά διασταση).
 - b. Η SAX κωδικοποιεί κάθε διάσταση σε διακριτά σύμβολα χαρακτήρων (1 byte ανά σύμβολο ή μόλις 3 bits για μέγεθος αλφαβήτου $a=8$).
 - c. Επιπρόσθετα, η SAX υποστηρίζει εγγενή συμπίεση Run-Length Encoding(RLE) επί ολισθαίνοντων παραθύρων, επιτυγχάνοντας δραστική μείωση του όγκου δεδομένων (2 × έως 3 × σε τυπικά σήματα και άνω του 10 × σε ηρεμούντα σήματα)
2. Υπολογιστικός Χρόνος Εκτέλεσης 1-NN:
 - a. Στα Raw δεδομένα, ο 1-NN απαιτεί υπολογισμό Ευκλείδειας απόστασης n πράξεων κινητής υποδιαστολής ($O(n)$).
 - b. Στη SAX, η απόσταση MINDIST υπολογίζεται στον συμπιεσμένο χώρο ($W \ll n$) αξιοποιώντας προϋπολογισμένους πίνακες αναζήτησης (Lookup Tables). Αντί για κοστοβόρους πολλαπλασιασμούς και τετραγωνικές ρίζες, η σύγκριση εκτελείται μέσω ταχύτατων ανακτήσεων μνήμης ($O(w)$ Lookups), προσφέροντας επιτάχυνση τάξης μεγέθους.

6.6 Σύνοψη Κεφαλαίου και Κύρια Ευρήματα

Τα πειραματικά ευρήματα του 6ου Κεφαλαίου συνοψίζονται στα εξής θεμελιώδη σημεία:

Αποτελεσματικότητα της SAX: Η συμβολική αναπαράσταση SAX αποδικνύεται εξαιρετικά αποτελεσματική σε συνθετικά σήματα και βιομηχανικούς αισθητήρες, όπου η διακριτοποίηση φιλτράρει επιτυχώς τον θόρυβο και βελτιώνει την ακρίβεια ταξινόμησης σε σχέση με τα αρχικά δεδομένα. Παρά τη μικρή υστέρηση σε πολυταξικά προβλήματα περιγραμμάτων λόγω σφαλμάτων κβάντισης, η SAX προσφέρει ασυναγώνιστο πλεονέκτημα συμπίεσης και υπολογιστικής ταχύτητας.

Κυριαρχία της PAA: Η PAA αναδεικνύεται, ως η κορυφαία αριθμητική μέθοδος μείωσης διαστάσεων, επιτυγχάνοντας την υψηλότερη μέση ακρίβεια στον 1-NN (82.05%) και τα Random Forests (83.06%), υπερτερώντας των πλήρων ανεπεξέργαστων δεδομένων.

Σταθερότητα του DFT: Ο DFT αποτελεί μια αξιόπιστη φασματική προσέγγιση που διατηρεί στατιστική ισοδυναμία με τα πλήρη δεδομένα (80.08%μ σε 1-NN και 82.74% σε RF).

Ανεπάρκεια της PLA: Η PLA υστερεί στατιστικά σημαντικά έναντι όλων των υπολοίπων τεχνικών ($p<0.025$), καθιστώντας την ακατάλληλη για γενική κατηγοριοποίηση χρονοσειρών χωρίς προσαρμοστική ρύθμιση παραμέτρων.

Κεφάλαιο 4

Τα παραπάνω ευρήματα τεκμηριώνουν πλήρως τη χρησιμότητα μεθόδων μείωσης διαστάσεων στην Εξόρυξη δεδομένων χρονοσειρών και θέτουν τις βάσεις για τα τελικά συμπεράσματα και τις προτάσεις μελλοντικής έρευνας που αναλύονται στο Κεφάλαιο 7.

Κεφάλαιο 7ο: Συζήτηση και Συμπεράσματα

7.1 Επισκόπηση και Σκοπός της Έρευνας

Η παρούσα πτυχιακή εργασία εστίασε στη συγκριτική διερεύνηση των τεχνικών μείωσης διαστάσεων και αναπαράστασης χρονοσειρών, με κεντρικό άξονα τη μέθοδο Συμβολικής Αθροιστικής Προσέγγισης (Symbolic Aggregate Approximation – SAX). Τα δεδομένα χρονοσειρών χαρακτηρίζονται από εγγενή πολυπλοκότητα, υψηλή διάσταση έντονο θόρυβο καταγραφής και αυτοσυσχέτιση, στοιχεία που συνθέτουν την “κατάρα των πολλών διαστάσεων” (Curse of Dimensionality). Στόχος της μελέτης ήταν να αξιολογηθεί εάν η μετατροπή σημάτων συνεχών τιμών σε διακριτές συμβολικές λέξεις μπορεί να επιτύχει ισοδύναμη ή ανώτερη επίδοση ταξινόμησης σε σύγκριση με τις ανεπεξέργαστες χρονοσειρές (Raw) και τις παραδοσιακές αριθμητικές μεθόδους συμπίεσης: την Τμηματική Αθροιστική Προσέγγιση (Piecewise Aggregate Approximation – PAA), τον Διακριτό Μετασχηματισμό Φουριέ (Discrete Fourier Transform – DFT) και την Τμηματική Γραμμική Προσέγγιση (Piecewise Linear Approximation – PLA).

Η πειραματική αξιολόγηση υλοποιήθηκε σε ένα ετερογενές σύνολο 10 benchmark datasets από το αποθετήριο UCR Time Series Classification Archive, εξετάζοντας τέσσερις διαφορετικές οικογένειες αλγορίθμων κατηγοριοποίησης: τον οκνηρό κατηγοριοποιητή 1-NN(με ευκλείδεια απόσταση και την συμβολική μετρική MINDIST), τα Μεμονωμένα Δέντρα Απόφασης (Decision Trees/CART), τα τυχαία Δάση (Random Forests) και τον αλγόριθμο Gradient Boosting CatBoost.

7.2 Ερμηνεία των Κύριων Ευρημάτων

7.2.1 Η Διττή Φύση της Μεθόδου SAX: Εξομάλυνση Θορύβου έναντι Σφάλματος Κβάντισης

Η SAX αποτελεί την μοναδική μέθοδο στη βιβλιογραφία που συνδυάζει τη διακριτοποίηση, μείωση διαστάσεων και μαθηματικά αποδεδειγμένη ιδιότητα κάτω ορίου (Lower Bounding guarantee) μέσω της συνάρτησης MINDIST. Τα πειραματικά αποτελέσματα αποκαλύπτουν δύο διακριτές πτυχές της συμπεριφοράς της:

1. Ευεργετική Εξομάλυνση σε Θορυβώδη και Συνθετικά Πρότυπα: Σε σύνολα δεδομένων όπως το Synthetic Control (ακρίβεια 98.33% έναντι 88.00% στα Raw), το CBF(89.78% έναντι 85.20%) και το Two Patterns (99.80% έναντι 90.70%), η SAX υπερκέρασε τα πλήρη πρωτογενή δεδομένα. Η διαίρεση σε πλαίσια PAA και η μετέπειτα αντιστοίχιση σε Γκαουσιανά σύμβολα λειτουργούν ως μη γραμμικό βαθυπερατό φίλτρο (non-linear low-pass filter), απορρίπτοντας τον τοπικό στοχαστικό θόρυβο και αναδεικνύοντας τις μακροσκοπικές τάσεις του σήματος.
2. Υποβάθμιση λόγω Κβάντισης σε Πολυπληθή Γεωμετρικά Σχήματα: Αντίθετα, σε πολυταξικά προβλήματα που προέρχονται από περιγράμματα εικόνων, όπως το Swedish Leaf(50.88% έναντι 78.90%) και το Trace(54.00% έναντι 76.00%), παρατηρήθηκε πτώση της ακρίβειας. Η συμπίεση των τιμών σε πεπερασμένα σύμβολα εξαλείφει τις οριακές διαφορές πλάτους (amplitude variations) μεταξύ παρεμφερών κλάσεων.

7.2.2 Σταθερότητα του DFT και Ανεπάρκεια της PLA

Ο Διακριτός Μετασχηματισμός Fourier (DFT) αποδείχθηκε εξαιρετικά σταθερός σε όλα τα πειράματα (80.08% σε 1-NN, 82.74% σε RF), επιτυγχάνοντας την κορυφαία επίδοση στο Face all(75.98%). Η διατήρηση των χαμηλών συχνοτήτων μέσω του θεωρήματος Parseval καθιστά τον DFT αξιόπιστο σε σχήματα με έντονη αρμονική ή περιοδική συμπεριφορά.

Η Τμηματική Γραμμική Προσέγγιση(PLA) υστέρησε συστηματικά (68.19% σε 1-NN, 74.56% σε RF, 74.68% σε CB). Η στατική κατάτμηση σε ευθύγραμμα τμήματα αδυνατεί να αποδώσει δυναμικές μεταβολές χωρίς εξειδικευμένους αλορίθμους προσαρμογής σημείων τομής.

7.3 Συγκριτική Αξιολόγηση Αλγορίθμων Ταξινόμησης

Η συμπεριφορά των τεχνικών αναπαράστασης διέφερε σημαντικά ανάλογα με την αρχιτεκτονική του ταξινομητή.

7.3.1 k-Nearest Neighbors(1-NN)

Ο 1-NN βασίζεται αποκλειστικά σε μετρικές αποστάσεων. Στα αρχικά δεδομένα και στις αριθμητικές αναπαραστάσεις (PAA, DFT, PLA), η Ευκλείδεια απόσταση επηρεάζεται έντονα από τοπικές διακυμάνσεις. Η συνάρτηση MINDIST της SAX, χρησιμοποιώντας πίνακες αναζήτησης (Lookup Tables), προσφέρει υπολογιστική επιτάχυνση αλλά απαιτεί επαρκές μέγεθος αλφαβήτου a και μήκος λέξης w ώστε να αποφεύγονται ισοπαλίες αποστάσεων μηδενικής τιμής.

7.3.2 Μεμονωμένα Δέντρα Απόφασης (Decision Trees)

Τα μεμονωμένα παρουσίασαν τη χαμηλότερη μέση απόδοση (59.3%-63.7%). Λόγω της υψηλής ενδεγενούς διακύμανσης του αλγορίθμου, οι στατιστικοί έλεγχοι Friedman ($p=0.2007$) δεν ανέδειξαν στατιστικά διακριτές διαφορές μεταξύ των μεθόδων αναπαράστασης.

7.3.2 Τυχαία Δάση (Random Forests και Catboost)

Τα μοντέλα συλλογικής μάθησης πέτυχαν τις κορυφαίες επιδόσεις:

- A. Random Forests: Μέσω του bagging και της τυχαίας υποδειγματοληψίας χαρακτηριστικών, μείωσαν τη διακύμανση και εκμεταλλεύτηκαν πλήρως τόσο τα συνεχή χαρακτηριστικά PAA/DFT όσο και τα διακριτά σύμβολα SAX. Ο έλεγχος Friedman($p=0.0011$) ανέδειξε την υπεροχή των PAA, DFT και Raw έναντι των PLA και SAX.
- B. CatBoost: Η αρχιτεκτονική των Ordered Target Statistics επέτρεψε την αποδοτική διαχείριση των κατηγορικών χαρακτηριστικών της SAX χωρίς διαρροή στόχου (target leakage). Ο έλεγχος Friedman ($p=0.0023$). Επιβεβαίωση την κυριαρχία του leading cluster (Raw, PAA, DFT) έναντι των PLA ($p=0.0007$) και SAX ($p=0.0245$).

7.4 Στατιστική Εγκυρότητα και Μεθοδολογικές Παρατηρήσεις

Η εφαρμογή του μη παραμετρικού πλαισίου ελέγχων (Wilcoxon Signed-Rank Test και Friedman/Conover Post-Hoc) ανέδειξε δύο κρίσιμα μεθοδολογικά συμπεράσματα.

1. Επίδραση των Διορθώσεων Πολλαπλών Ελέγχων: Στις 16 μεμονωμένες συγκρίσεις ζευγών με το Wilcoxon Test, οι αρχιμές στατιστικά σημαντικές διαφορές της PLA ($p<0.02$) και της SAX($p=0.0415$) έχασαν τη στατιστική τους σημασία μετά τη διόρθωση Holm ($p_{adj} > 0.05$). Αυτό οφείλεται στη μείωση της στατιστικής ισχύος λόγω του συνδυασμού μικρού δείγματος ($N=10$) και αυστηρού ελέγχου σφάλματος τύπου I.
2. Στατιστική Ισχύς του Ελέγχου Friedman: Ο διαχωρισμός της ανάλυσης ανά κατηγοριοποιητή και η συγκεντρωση των σχετικών διατάξεων (ranks) στο τεστ Friedman επέτρεψε την ανάδειξη στατιστικά σημαντικών διαφορών σε επίπεδο $p<0.005$ για τα μοντέλα Random Forests και

CatBoost, επιβεβαιώνοντας ότι η PAA και ο DFT πετυχαίνουν πλήρη μείωση διαστάσεων χωρίς απώλεια πληροφορίας.

7.5 Περιορισμοί της Παρούσας Εργασίας

Κατά την εκπόνηση της εργασίας εντοπίστηκαν ορισμένοι περιορισμοί:

Στατικό Μέγεθος Παραμέτρων. Χρησιμοποιήθηκαν σταθερές τιμές παραμέτρων (w , a) σε όλα τα datasets. Η έλλειψη δυναμικής βελτιστοποίησης (hyperparameter tuning) ανά dataset ενδέχεται να περιορίσει την απόδοση της SAX και της PLA σε ορισμένα προβλήματα.

Μέγεθος Δείγματος Δεδομένων: Η αξιολόγηση πραγματοποιήθηκε σε 10 σύνολα δεδομένων UCR. Παρότι αντιπροσωπευτικά, η επέκταση στο πλήρες αποθετήριο (128 datasets) θα ενίσχυε περαιτέρω τη στατιστική ισχύ των ελέγχων.

Μονομεταβλητές Χρονοσειρές. Η μελέτη επικεντρώθηκε αποκλειστικά σε μονομεταβλητά σήματα (univariate time series).

7.6 Προτάσεις για Μελλοντική Έρευνα

Με βάση τα ευρήματα της εργασίας, προτείνονται οι ακόλουθες κατευθύνσεις για μελλοντική επέκταση:

1. Προσαρμοστικές Παραλλαγές της SAX. Διερεύνηση προηγμένων εκδόσεων όπως η Adaptive SAX(A-SAX) και η Entropy-based SAX(E-SAX), όπου τα breakpoints και τα μεγέθη των παραθύρων υπολογίζονται δυναμικά βάσει της τοπικής κατανομής και της εντροπίας. Η χρήση αυτών των μεθόδων αναμένεται να μετριάσει το σφάλμα κβάντισης που παρατηρήθηκε στα σύνθετα περιγράμματα εικόνων.
2. Εφαρμογή σε πολυμεταβλητές χρονοσειρές. Επέκταση του πειραματικού pipeline σε πολυμεταβλητά σύνολα δεδομένων (Multivariate time series).
3. Ενσωμάτωση σε Νευρωνικά Δίκτυα Βαθιάς Μάθησης: Αξιοποίηση των συμβολικών αναπαραστάσεων SAX ως διακριτά tokens εισόδου σε αρχιτεκτονικές Transformer και Recurrent Neural Networks (RNNs) που έχουν σχεδιαστεί για επεξεργασία διακριτών συμβολοσειρών φυσικής γλώσσας, για κατηγοριοποίηση και πρόβλεψη σημάτων.

7.7 Τελικό Συμπέρασμα

Η παρούσα πτυχιακή εργασία ανέδειξε ότι η επιλογή τεχνικής μείωσης διαστάσεων αποτελεί ένα συμβιβασμό μεταξύ ακρίβειας κατηγοριοποίησης, υπολογιστικής πολυπλοκότητας και βαθμού συμπίεσης.

Η PAA αποτελεί την βέλτιστη επιλογή όταν ζητείται η μέγιστη δυνατή ακρίβεια κατηγοριοποίησης, υπερτερώντας συχνά των αρχικών δεδομένων λόγω απομάκρυνσης του θορύβου.

Η SAX επιβεβαιώνει την αξία της ως ένα εξαιρετικά αποδοτικό εργαλείο συμπίεσης και διακριτοποίησης, το οποίο διατηρεί υψηλή ανταγωνιστικότητα σε σύνθετα μοντέλα συλλογικής μάθησης (Random Forests, CatBoost) και υπερτερεί σε θορυβώδη βιομηχανικά και σύνθετα σήματα.

Ο DFT συνιστά μια σταθερή φασματική εναλλακτική, ενώ η PLA απαιτεί εξειδικευμένους αλγόριθμους προσαρμογής για να καταστεί ανταγωνιστική.

Τα ευρήματα αυτά επιβεβαιώνουν ότι οι μέθοδοι μείωσης διαστάσεων δεν αποτελούν απλώς εργαλεία εξοικονόμησης πόρων, αλλά δομικά στοιχεία βελτιστοποίησης της διαδικασίας Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Βιβλία

- [1] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson New International ed. Harlow, England: Pearson Education Limited, 2014.
- [2] W. W. S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed. Boston, MA, USA: Pearson Addison Wesley, 2006.

Internet Site

- [3] A. Vani, "Introduction to k-Nearest Neighbors (kNN) Algorithm," *Artificial Intelligence in Plain English*, May 4, 2021. [Online]. Available: <https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8>. .

Journal Articles

- [4] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis, nonparametric discrimination," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, Feb. 1951.
- [5] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
- [6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
- [7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 6638–6648, 2018.

- [9] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proc. 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD '03)*, San Diego, CA, USA, 2003, pp. 2–11.
- [10] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," in *Proc. 1994 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD '94)*, Minneapolis, MN, USA, 1994, pp. 419–429.
- [11] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Knowledge and Information Systems*, vol. 3, no. 3, pp. 263–286, 2001.
- [12] E. A. Pitukhin and L. V. Shchegoleva, "Piecewise linear approximation of time series on the base of the Weierstrass-Mandelbrot function," in *Proc. Selected Papers of the 3rd Int. Scientific Conf. 'Convergent Cognitive Information Technologies' (Convergent '18)*, Moscow, Russia, 2018, *CEUR Workshop Proceedings*, vol. 2278, pp. 47–56.
- [13] H. André-Jönsson and D. Badal, "Using signature files for querying time-series data," in *Proc. 1st Eur. Symp. Princ. Data Min. Knowl. Discov. (PKDD '97)*, Trondheim, Norway, 1997, pp. 211–220.
- [14] Y. Huang and P. S. Yu, "Adaptive query processing for time-series data," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '99)*, San Diego, CA, USA, 1999, pp. 282–286.
- [15] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: A survey and empirical demonstration," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '02)*, Edmonton, AB, Canada, 2002, pp. 102–111.
- [16] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Locally adaptive dimensionality reduction for indexing large time series databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Santa Barbara, CA, USA, 2001, pp. 151–162.
- [17] B.-K. Yi and C. Faloutsos, "Fast time sequence indexing for arbitrary L_p norms," in *Proc. 26th Int. Conf. Very Large Data Bases (VLDB '00)*, Cairo, Egypt, 2000, pp. 385–394.
- [18] A. Apostolico, M. E. Bock, and S. Lonardi, "Monotony of surprise and large-scale quest for unusual words," in *Proc. 6th Int. Conf. Res. Comput. Mol. Biol. (RECOMB '02)*, Washington, DC, USA, 2002, pp. 22–31.
- [19] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Minneapolis, MN, USA, 1994, pp. 419–429.
- [20] P. Geurts, "Pattern extraction for time series classification," in *Proc. 5th Eur. Conf. Princ. Data Min. Knowl. Discov. (PKDD '01)*, Freiburg, Germany, 2001, pp. 115–127.
- [21] C. Ding, X. He, H. Zha, and H. D. Simon, "Adaptive dimension reduction for clustering high dimensional data," in *Proc. 2nd IEEE Int. Conf. Data Min. (ICDM '02)*, Maebashi, Japan, 2002, pp. 147–154.
- [22] E. Keogh, S. Lonardi, and B. Chiu, "Finding surprising patterns in a time series database in linear time and space," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '02)*, Edmonton, AB, Canada, 2002, pp. 550–556.

- [23] C. Shahabi, X. Tian, and W. Zhao, "TSA-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries," in Proc. 12th Int. Conf. Sci. Stat. Database Manage. (SSDBM '00), 2000, pp. 55–68.
- [24] D. Dasgupta and S. Forrest, "Novelty detection in time series data using ideas from immunology," in Proc. Int. Conf. Intell. Syst. (IS '96), 1996.
- [25] J. Lin, E. Keogh, S. Lonardi, and P. Patel, "Finding motifs in time series," in Proc. 2nd Workshop Temporal Data Min. (KDD '02), Edmonton, AB, Canada, 2002, pp. 53–68.
- [26] M. Tompa and J. Buhler, "Finding motifs using random projections," in Proc. 5th Int. Conf. Comput. Mol. Biol. (RECOMB '01), Montreal, QC, Canada, 2001, pp. 67–74.
- [27] H. Zhang, Y. Dong, and D. Xu, "Entropy-based symbolic aggregate approximation representation method for time series," in Proc. 2020 IEEE 9th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC), Chongqing, China, 2020, pp. 905–909, doi: 10.1109/ITAIC49862.2020.9339021.
- [28] S. D. Zerea, R. A. Ikuesan, B. T. Woldemichael, and F. B. Araya, "Credit card fraud detection using symbolic aggregate approximation," in Proc. 2024 6th Int. Conf. Blockchain Comput. Appl. (BCCA), Abu Dhabi, United Arab Emirates, 2024, pp. 441–448, doi: 10.1109/BCCA62388.2024.10844470.
- [29] A. Buranasing and A. Prayote, "Storm intensity estimation using symbolic aggregate approximation and artificial neural network," in Proc. 2014 Int. Comput. Sci. Eng. Conf. (ICSEC), Khon Kaen, Thailand, 2014, pp. 234–237, doi: 10.1109/ICSEC.2014.6978200.