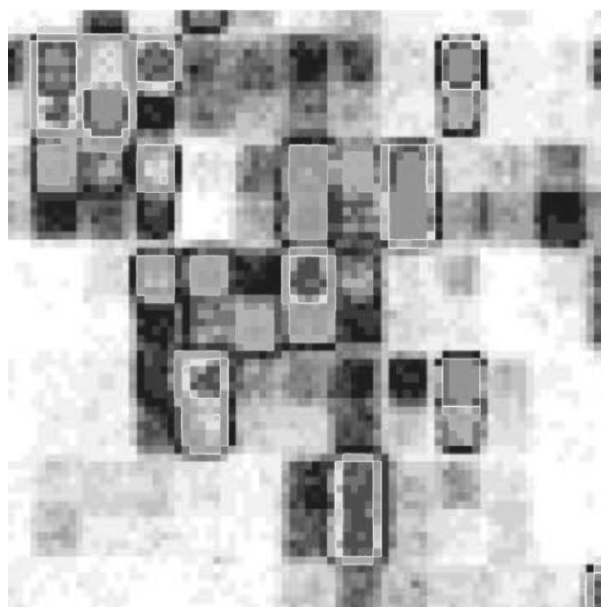


Νέοι Αλγόριθμοι Υπολογιστικής Νοημοσύνης και Ομαδοποίησης για την Εξόρυξη Πληροφορίας

Δημήτριος Κ. Τασουλής

Διδακτορική Διατριβή



Πανεπιστήμιο Πατρών
Σχολή Θετικών Επιστημών
Τμήμα Μαθηματικών
Πάτρα

Επιβλέπων: Καθηγητής Μιχαήλ Ν. Βραχάτης

(Απρίλιος 2006)

ΑΥΤΗ ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ ΣΤΟΙΧΕΙΟΘΕΤΗΘΗΚΕ ΜΕ ΤΟ ΠΡΟΓΡΑΜΜΑ $\text{\LaTeX}$ (ΔΙΑΝΟΜΗ $\text{\TeX}$), ΧΡΗΣΗΜΟΠΟΙΩΝΤΑΣ ΤΗ ΜΟΡΦΗ `vrr_thesis`. Η ΣΥΓΓΡΑΦΗ ΕΓΙΝΕ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ `vi` (ΣΤΟ ΛΕΙΤΟΥΡΓΙΚΟ ΣΥΣΤΗΜΑ `Debian GNU Linux`). ΓΙΑ ΤΗΝ ΑΝΑΠΤΥΞΗ ΚΑΙ ΤΟΝ ΕΛΕΓΧΟ ΤΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΕ ΚΥΡΙΩΣ Η ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ `C++`. ΟΙ ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ ΕΓΙΝΑΝ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ `Gnuplot`, ΚΑΙ Η ΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΣΧΗΜΑΤΩΝ ΜΕ ΤΟ ΠΡΟΓΡΑΜΜΑ `Gimp`.

Περίληψη

Η εκπαίδευση είναι μια προοδευτική ανακάλυψη
της άγνοιάς μας.

—William James Durant (1885–1981)

Αυτή η Διδακτορική Διατριβή πραγματεύεται το θέμα της ομαδοποίησης δεδομένων (clustering), καθώς και εφαρμογές των τεχνικών αυτών σε πραγματικά προβλήματα. Η παρουσίαση των επιμέρους θεμάτων και αποτελεσμάτων της διατριβής αυτής οργανώνεται ως εξής:

Στο Κεφάλαιο 1 παρέχουμε τον ορισμό της Υπολογιστικής Νοημοσύνης σαν τομέας ερευνάς, και αναλύουμε τα ξεχωριστά τμήματα που τον αποτελούν. Για κάθε ένα από αυτά παρουσιάζεται μια σύντομη περιγραφή.

Το Κεφάλαιο 2, ασχολείται με την ανάλυση του ερευνητικού πεδίου της ομαδοποίησης. Κάθε ένα από τα χαρακτηριστικά της αναλύεται ξεχωριστά και γίνεται μια επισκόπηση των σημαντικότερων αλγόριθμων ομαδοποίησης.

Το Κεφάλαιο 3, αφιερώνεται στη παρουσίαση του αλγορίθμου UKW, που κατά την εκτέλεση του έχει την ικανότητα να προσεγγίζει το πλήθος των ομάδων σε ένα σύνολο δεδομένων. Επίσης παρουσιάζονται πειραματικά αποτελέσματα με σκοπό τη μελέτη της απόδοσης του αλγορίθμου.

Στο Κεφάλαιο 4 προτείνεται μια επέκταση του αλγορίθμου UKW, σε μετρικούς χώρους. Η προτεινόμενη επέκταση διατηρεί όλα τα πλεονεκτήματα του αλγορίθμου UKW. Τα πειραματικά αποτελέσματα που παρουσιάζονται επίσης σε αυτό το κεφάλαιο, συγκρίνουν την προτεινόμενη επέκταση με άλλους αλγόριθμους.

Στο επόμενο κεφάλαιο παρουσιάζουμε τροποποιήσεις του αλγορίθμου με στόχο την βελτίωση των αποτελεσμάτων του. Οι προτεινόμενες τροποποιήσεις αξιοποιούν πληροφορία από τα τοπικά χαρακτηριστικά των δεδομένων, ώστε να κατευθύνουν όσο το δυνατόν καλύτερα την αλγοριθμική διαδικασία.

Το Κεφάλαιο 6, πραγματεύεται επεκτάσεις του αλγορίθμου σε κατανεμημένες Βάσεις δεδομένων. Για τις διάφορες υποθέσεις που μπορούν να γίνουν όσον αφορά τη φύση του περιβάλλοντος επικοινωνίας, παρουσιάζονται κατάλληλοι αλγόριθμοι.

Στο Κεφάλαιο 7, εξετάζουμε την περίπτωση δυναμικών βάσεων δεδομένων. Σε ένα τέτοιο μη στατικό περιβάλλον αναπτύσσεται μια επέκταση του αλγορίθμου UKW, που ενσωματώνει τη δυναμική δομή δεικτοδότησης Bkd-tree. Επιπλέον παρουσιάζονται θεωρητικά αποτελέσματα για την πολυπλοκότητα χειρότερης περίπτωσης του αλγορίθμου.

Το Κεφάλαιο 8, μελετά την εφαρμογή αλγορίθμων ομαδοποίησης σε δεδομένα γονιδιακών εκφράσεων. Επίσης προτείνεται και αξιολογείται ένα υβριδικό σχήμα που καταφέρνει να αυτοματοποιήσει την όλη διαδικασία επιλογής γονιδίων και ομαδοποίησης.

Τέλος, η παρουσίαση του ερευνητικού έργου αυτής της διατριβής ολοκληρώνεται στο Κεφάλαιο 9 που ασχολείται με την ανάπτυξη υβριδικών τεχνικών που συνδυάζουν την ομαδοποίηση και τα Τεχνητά Νευρωνικά Δίκτυα, και αναδεικνύει τις δυνατότητες τους σε δύο πραγματικά προβλήματα.

Synopsis

This Doctoral Dissertation appoints the issue of data Clustering, as well as the applications of these kind of methods in real world problems. The presentation of the individual results of this dissertation is organised as follows:

In Chapter 1, the definition of Computational Intelligence is provided as a research area. For each distinct part of this area a short description is supplied.

Chapter 2, deals with the analysis of the research area of Clustering per se, and its characteristics are analysed separably. Moreover, we provide a review of the most representative clustering algorithms.

Chapter 3, is devoted to the presentation of the UKW algorithm, that is able to endogenously provide approximations for the number of clusters in a dataset, during its execution. Furthermore, the included experimental results demonstrate the algorithm's efficiency.

In Chapter 4, an extension of the UKW algorithm to metric spaces is proposed. This extension preserves all the advantages of the original algorithm. The included experimental results compare the proposed extension to other approaches.

In the next chapter we present modifications of the UKW algorithm that scope to improve its efficiency. This is performed through the utilisation of information from the local characteristics of the data, so as to direct more efficiently the whole clustering procedure.

Chapter 6, deals with extensions of the algorithm in distributed data bases. For the various assumptions that can be postulated for the nature of the communication environment different algorithms are proposed.

In Chapter 7, we consider the case of dynamic databases. In such a non-static environment, an algorithm is developed that draws from the principles of the UKW algorithm, and embodies the dynamic indexing Bkd-tree data structure. Moreover, theoretical results are presented regarding the worst case complexity of the algorithm.

Chapter 8, studies the application of clustering algorithms in gene expression data. Besides, it is proposed and evaluated, a hybrid algorithmic scheme that manages to automate the whole procedure of gene selection and clustering.

Finally, the presentation of the research work of this dissertation is fulfilled in Chapter 9. This Chapter is devoted to the development of hybrid techniques that combine clustering methods and Artificial Neural Networks, and demonstrate their abilities in two real world problems.

Ευχαριστίες

Ολοκληρώνοντας την παρούσα διατριβή δεν θα μπορούσα να μην ευχαριστήσω μερικούς ανθρώπους που είχαν καίρια συμμετοχή στην εκπόνησή της.

Θα ήταν πολύ δύσκολο να υπερβάλω στην ευγνωμοσύνη μου, για δάσκαλο μου καθηγητή Μιχάλη Ν. Βραχάτη. Με τον ενθουσιασμό του, την έμπνευσή του, την υπομονή του και τις μεγάλες προσπάθειές του, έπαιξε το σημαντικότερο ρόλο στην ολοκλήρωση της διατριβής αυτής. Καθ' όλη τη διάρκεια της ερευνητικής μου πορείας, παρείχε ενθάρρυνση, υγιείς συμβουλές, καλή διδασκαλία και το σημαντικότερο ηθική και ψυχολογική στήριξη. Θα είχα χαθεί χωρίς αυτόν.

Θα ήθελα ακόμα να ευχαριστήσω τα άλλα δύο μέλη της Τριμελούς Συμβουλευτικής Επιτροπής τους καθηγητές κ.κ. Ε. Γαλλόπουλο, και Π. Αλεβίζο, για τη καθοριστική τους βοήθεια στο ερευνητικό μου έργο.

Επίσης θα ήθελα να ευχαριστήσω τον φίλο Ν.Γ. Παυλίδη, καθώς ήταν ο καλύτερος συνοδοιπόρος που μπορεί κάποιος να διαλέξει σε μια τέτοια πορεία. Οι φίλοι μου Μ. Μαυρικάκης και Σ. Σταυρόπουλος και ο αδερφός μου Σ. Τασουλής αξίζουν πολλά θερμά ευχαριστώ για τη βοήθεια τους στους δύσκολους περιόδους, και για όλη τη συναισθηματική υποστήριξη και συντροφιά, που μου παρείχαν.

Στη διάρκεια αυτής της πορείας, ήμουν στην ευχάριστη θέση να συνεργάζομαι με εξαιρετους επιστήμονες όπως ο Δρ. Β.Π. Πλαγιαννάκος, που συνεχώς κατάφερναν να μου δείχνουν το δρόμο όταν αυτός χανόταν. Ακόμα θα ήθελα να ευχαριστήσω τους υπόλοιπους συνεργάτες μου, Κ. Παρσόπουλο, Ε. Λάσκαρη, Ι. Πεταλά, Μ. Επιτροπάκη.

Θερμά ευχαριστώ θα ήθελα επίσης να εκφράσω στο προσωπικό του Εργαστηρίου Ηλεκτρονικών Υπολογιστών και Πληροφορικής του Τμήματος Μαθηματικών, Ι. Μαραματάκη, Δ. Ανυφαντή για την διάθεση των υπολογιστικών πόρων.

Τέλος, και πιο σημαντικά, θα ήθελα να ευχαριστήσω τους γονείς μου Κωνσταντίνο Δ. Τασουλή και Αθανασία Σ. Τασουλή, που με την αγάπη τους με γέννησαν, με ανέθρεψαν, με υποστήριξαν, και με δίδαξαν. Σε αυτούς αφιερώνω αυτή τη διατριβή.

Δημήτρης Κ. Τασουλής

Πάτρα, 2006.

Περιεχόμενα

Περίληψη	iii
Synopsis	v
Ευχαριστίες	vii
I Εισαγωγή και Βασικές Έννοιες	1
1 Εισαγωγή	3
1.1 Ασαφή Συστήματα	3
1.2 Μηχανική Μάθηση	3
1.3 Τεχνητά Νευρωνικά Δίκτυα	4
1.4 Ομαδοποίηση Δεδομένων	6
1.5 Εξελικτικοί Αλγόριθμοι	6
2 Ομαδοποίηση Δεδομένων	9
2.1 Τα στάδια μιας διαδικασίας ομαδοποίησης	10
2.2 Ορολογία και ορισμοί	11
2.3 Αναπαράσταση των Δεδομένων	12
2.4 Μέτρα Ομοιότητας	13
2.5 Μέθοδοι Ομαδοποίησης	14
2.5.1 Ιεραρχικοί Αλγόριθμοι Ομαδοποίησης	15
2.5.2 Διαχωριστικοί αλγόριθμοι	20
2.5.3 Density ή Mode-Seeking Τεχνικές	25
2.5.4 Ομαδοποίηση με Τεχνητά Νευρωνικά Δίκτυα	28
2.6 Σύνοψη	33
II Η Μέθοδος Ομαδοποίησης k-windows – Επεκτάσεις, Γενικεύσεις και Τροποποιήσεις	35
3 Ο Αλγόριθμος k-windows	37
3.1 Η Διαδικασία της Μετακίνησης	37
3.2 Η Διαδικασία της Διεύρυνσης	38
3.3 Αυτόματος Υπολογισμός του Αριθμού των Ομάδων	39
3.4 Ο Αλγόριθμος Unsupervised k -Windows	41
3.5 Υπολογιστική Πολυπλοκότητα	42
3.6 Πειραματικά Αποτελέσματα	42
3.7 Συμπεράσματα - Συνεισφορά	44

4	Γενίκευση του Αλγορίθμου k-windows σε Μετρικούς Χώρους	45
4.1	Παρόμοιες Προσεγγίσεις	45
4.2	Ο Γενικευμένος Αλγόριθμος k -windows	46
4.2.1	Η Γενικευμένη Διαδικασία Μετακίνησης	46
4.2.2	Η Γενικευμένη Διαδικασία Διεύρυνσης	47
4.2.3	Η Γενικευμένη Διαδικασία Συγχώνευσης	47
4.3	Θέματα Υπολογιστικής Πολυπλοκότητας	48
4.4	Πειραματικά Αποτελέσματα	48
4.5	Συμπεράσματα - Συνεισφορά	53
5	Τροποποιήσεις του αλγορίθμου k-windows	55
5.1	Ομαδοποίηση με τη Χρήση Συνόλων Μορφοκλασματικής Δομής	55
5.1.1	Ενσωμάτωση της fractal Διάστασης, στον Αλγόριθμο k -windows	56
5.1.2	Πειραματικά Αποτελέσματα	56
5.2	Προσανατολισμένος Αλγόριθμος k -windows	58
5.2.1	Ο αλγόριθμος k -windows καθοδηγούμενος από την τεχνική PCA	58
5.2.2	Υπολογιστικά Θέματα	61
5.2.3	Πειραματικά Αποτελέσματα	62
5.3	Συμπεράσματα - Συνεισφορά	63
6	Ομαδοποίηση Σε Κατανεμημένες Βάσεις Δεδομένων	65
6.1	Κατανεμημένη Ομαδοποίηση για Περιβάλλοντα Ελάχιστης Επικοινωνίας	66
6.2	Κατανεμημένη Ομαδοποίηση σε Υψηλής Ταχύτητας Δίκτυο	67
6.3	Ομαδοποίηση Διατηρώντας την Ιδιωτικότητα	68
6.3.1	Αρχικοποίηση των Παραθύρων	69
6.3.2	Μετακινήσεις και Διευρύνσεις	69
6.3.3	Συγχώνευση των Τελικών Παραθύρων	71
6.3.4	Υπολογιστική Πολυπλοκότητα	71
6.4	Πειραματικά Αποτελέσματα	71
6.5	Συμπεράσματα - Συνεισφορά	73
7	Ομαδοποίηση σε Δυναμικές Βάσεις Δεδομένων	75
7.1	Η Δομή Bkd-tree	76
7.2	Ο Αλγόριθμος UKW σε Δυναμικές Βάσεις Δεδομένων	77
7.2.1	Αντιμετώπιση της Εισαγωγής Νέων Στοιχείων	77
7.2.2	Αντιμετώπιση της Διαγραφής Στοιχείων	78
7.2.3	Η Δυναμική Επέκταση του Αλγορίθμου UKW	80
7.3	Πειραματικά Αποτελέσματα	81
7.4	Συμπεράσματα - Συνεισφορά	84
III	Εφαρμογές	87
8	Ομαδοποίηση σε Εκφράσεις mRNA	89
8.1	Σύντομη Βιβλιογραφική Επισκόπηση	90
8.2	Επιλογή των Χαρακτηριστικών	91
8.3	Πειραματικά Αποτελέσματα	91
8.3.1	Ομαδοποίηση Βασισμένη σε Επιβλεπόμενη Επιλογή Γονιδίων .	92
8.3.2	Ομαδοποίηση Βασισμένη σε Μη-Επιβλεπόμενη Επιλογή Γονιδίων	94

	xi
8.3.3 Αξιολόγηση του Υβριδικού Σχήματος	96
8.4 Συμπεράσματα - Συνεισφορά	97
9 Υβριδικές Εφαρμογές Υπολογιστικής Νοημοσύνης και Ομα- δοποίησης	99
9.1 Αυτόματος Εντοπισμός Ισχυαμικών Επεισοδίων	99
9.1.1 Προ-Επεξεργασία Δεδομένων	100
9.1.2 Η Μεθοδολογία Κατηγοριοποίησης	100
9.1.3 Αριθμητικά Αποτελέσματα	100
9.2 Αυτόματη Αναγνώριση Ανωμαλιών σε Κολονοσκοπήσεις	102
9.2.1 Προτεινόμενη Μεθοδολογία	102
9.2.2 Πειραματικά Αποτελέσματα	103
9.3 Συμπεράσματα - Συνεισφορά	105
 IV Συμπεράσματα – Βιβλιογραφία – Ευρετήριο	 107
10 Συμπεράσματα Διατριβής	109
Βιβλιογραφία	113
Κατάλογος Δημοσιεύσεων Υποψηφίου	125
Ευρετήριο	134

Κατάλογος Σχημάτων

2.1	Τα στάδια της ομαδοποίησης.	10
2.2	Μία ημικυκλική ομάδα από σημεία όπου κάθε σημείο απέχει προσεγγιστικά ίση απόσταση από την αρχή των αξόνων.	12
2.3	Μία ταξινόμηση των αλγορίθμων ομαδοποίησης.	15
2.4	Ομαδοποίηση χρησιμοποιώντας την μέθοδο απλής σύνδεσης (αριστερά), και της πλήρους σύνδεσης (δεξιά).	16
2.5	Το παραγόμενο δενδρόγραμμα.	18
2.6	Διαδοχικές διαμερίσεις του αλγορίθμου PDDP, για 2,3,4 και 5 τελικές ομάδες του ίδιου συνόλου δεδομένων.	20
2.7	Τέσσερις διαδοχικές επαναλήψεις του αλγορίθμου k -means.	22
2.8	Τέσσερις διαδοχικές επαναλήψεις του αλγορίθμου fuzzy c -means.	24
2.9	Το αποτέλεσμα του αλγορίθμου DBSCAN, για διαφορετικές τιμές τις παραμέτρου Eps , 5(α), 8(β), 10(γ) και 15(δ) αντίστοιχα.	28
2.10	Τέσσερις καταστάσεις του αλγορίθμου SOM, μετά από 0(α), 3(β), 10(γ) και 20(δ) επαναλήψεις.	30
2.11	Τέσσερις καταστάσεις του αλγορίθμου GNG, μετά από 0(α), 5(β), 15(γ) και 50(δ) επαναλήψεις.	33
3.1	Διαδοχικές μετακινήσεις ενός παραθύρου.	38
3.2	Διαδοχικές διευρύνσεις ενός παραθύρου.	39
3.3	Ένα παράδειγμα της διαδικασίας συγχώνευσης.	40
3.4	Ένα παράδειγμα εφαρμογής του αλγορίθμου UKW.	41
3.5	Το αποτέλεσμα του αλγορίθμου UKW στα σύνολα $Dset_1$ (αριστερά) και $Dset_2$ (δεξιά).	43
3.6	Το αποτέλεσμα του αλγορίθμου UKW στα σύνολα $Dset_3$ (αριστερά) και στην τρισδιάστατη προβολή του $Dset_4$ (δεξιά).	43
3.7	Το αποτέλεσμα του αλγορίθμου UKW στα σύνολα $Dset_5$ (αριστερά) και $Dset_6$ (δεξιά).	44
4.1	Ποσοστά αντιστοίχισης του αλγορίθμου GUKW με τυχαία αρχικοποίηση χρησιμοποιώντας 4,8,16 και 32 αρχικά παράθυρα.	50
4.2	Ποσοστά αντιστοίχισης του αλγορίθμου GUKW για διάφορα αρχικά μεγέθη των παραθύρων.	50
4.3	Ποσοστά αντιστοίχισης του αλγορίθμου GUKW με τυχαία αρχικοποίηση χρησιμοποιώντας 2,4,8,16,32 και 64 αρχικά παράθυρα.	52
4.4	Ποσοστά αντιστοίχισης του αλγορίθμου GUKW για διάφορα αρχικά μεγέθη των παραθύρων.	52
5.1	Ομάδες με περιοχές με διαφορετική πυκνότητα. Η προτεινόμενη τροποποίηση είναι ικανή να εντοπίσει διαφορετικές περιοχές ακόμη και μέσα στην ίδια ομάδα.	57

5.2	Το σύνολο δεδομένων $Dset_1$ με τα αποτελέσματα του αλγορίθμου UKW (αριστερά) και της προτεινόμενης τροποποίησης (δεξιά).	57
5.3	Το αποτέλεσμα του αλγορίθμου UKW στο σύνολο δεδομένων $Dset_{eq}$	59
5.4	Το αποτέλεσμα της προτεινόμενης τροποποίησης στο σύνολο δεδομένων $Dset_{eq}$	59
5.5	Τα αρχικά δεδομένα και το παράθυρο (α') μετακινούνται και υπολογίζονται οι κύριες κατευθύνσεις (β'). Υπολογίζουμε το παραλληλεπίπεδο προσανατολισμένο στις κύριες διευθύνσεις (γ') και τον αντίστοιχο κύβο (Ενότητα 5.2.2) (δ'). Τέλος, μετακινούμε τα δεδομένα στην αρχική τους θέση (ε').	60
5.6	Τα σύνολα δεδομένων $Dsets_{1,2}$ με τα αποτελέσματα των αλγορίθμων UKW (αριστερά) και OkW (δεξιά).	63
6.1	Υπολογισμός των σημείων που εσωκλείονται σε ένα παράθυρο όταν τα χαρακτηριστικά (x, y) είναι κατανομημένα σε 2 διαφορετικές οντότητες.	70
6.2	Το σύνολο δεδομένων $Dset_d$ με το αποτέλεσμα του αλγορίθμου DUKW, για 1 (αριστερά) και 4 (δεξιά) τύπους.	72
6.3	Το σύνολο δεδομένων $Dset_d$ με το αποτέλεσμα του αλγορίθμου DUKW, για 16 (αριστερά) και 32 (δεξιά) τύπους.	72
6.4	Επιτάχυνση του αλγορίθμου PUKW χρησιμοποιώντας μέχρι και 16 μηχανές.	73
7.1	Η αντιμετώπιση των εισαγωγών σημείων από τον αλγόριθμο.	78
7.2	(a) Η εφαρμογή του αλγορίθμου στα διαγραφόμενα σημεία. (b) Τα τελικά παράθυρα που παράγει ο δυναμικός αλγόριθμος.	79
7.3	Εφαρμόζοντας τον δυναμικό αλγόριθμο στις 4 διαφορετικές καταστάσεις του συνόλου $Dset1$	83
7.4	Εφαρμόζοντας τον δυναμικό αλγόριθμο στα 6 διαφορετικές καταστάσεις του συνόλου $Dset2$	83
7.5	Η επιτάχυνση (Speedup) του δυναμικού αλγορίθμου, για διάφορους αριθμούς εισαγωγής στοιχείων, και για διάφορα μεγέθη της βάσης δεδομένων.	84
7.6	Η επιτάχυνση (Speedup) του δυναμικού αλγορίθμου, για διάφορους αριθμούς διαγραφής στοιχείων, και για διάφορα μεγέθη της βάσης δεδομένων.	84
8.1	Το γράφημα των ιδιοτιμών σε φθίνουσα σειρά (αριστερά), και η απόδοση αντιστοίχισης με τη χρήση 2 ως 70 χαρακτηριστικών (δεξιά).	95
8.2	Γραφήματα των ιδιοτιμών σε φθίνουσα σειρά για τα σύνολα δεδομένων COLON (αριστερά), PROSTATE (μέση) και LYMPHOMA (δεξιά).	96
8.3	Η απόδοση αντιστοίχισης με τη χρήση όλων των διαθέσιμων χαρακτηριστικών για τα σύνολα COLON (αριστερά), PROSTATE (μέση) και LYMPHOMA (δεξιά).	97
9.1	Διαδοχικές εικόνες από μια κολονοσκόπηση όπου απεικονίζεται ένας πολυποειδής όγκος του παχέος εντέρου.	103

Κατάλογος Πινάκων

2.1	Παράμετροι για τον τύπο 2.1.	17
3.1	Μέθοδοι Ορθογώνιας Αναζήτησης Διαστήματος.	42
4.1	Πίνακας αντιστοίχισης των αποτελέσματα του αλγορίθμου GUKW, χρησιμοποιώντας Vantage Point Tree μέθοδο αρχικοποίησης και 4 αρχικά παράθυρα.	50
4.2	Πίνακας αντιστοίχισης για τους αλγορίθμους GDBSCAN και PDDP για το soybean disease σύνολο δεδομένων.	51
4.3	Πίνακας αντιστοίχισης των αποτελέσματα του αλγορίθμου GUKW, χρησιμοποιώντας Vantage Point Tree μέθοδο αρχικοποίησης και 32 αρχικά παράθυρα.	52
4.4	Πίνακας αντιστοίχισης για τους αλγορίθμους GDBSCAN και PDDP για το Voting Records Database σύνολο δεδομένων.	53
5.1	Πίνακες αντιστοίχισης για το σύνολο δεδομένων $Dset_{iris}$	58
5.2	Πίνακας αντιστοίχισης για το σύνολο $Dset_{UCI}$: Κάθε ζεύγος αριθμών αναπαριστά τα στοιχεία του πίνακα αντιστοίχισης για του δύο αλγορίθμους UKW και OkW, αντίστοιχα.	63
8.1	Η απόδοση του αλγορίθμου UKW για το σύνολο εκπαίδευσης με διαφορετικά σύνολα γονιδίων.	92
8.2	Η απόδοση του αλγορίθμου UKW για το σύνολο ελέγχου με διαφορετικά σύνολα γονιδίων.	93
8.3	Συγκριτικά αποτελέσματα για διάφορους αλγορίθμους με γονίδια από το σύνολο $GeneSet_3$	94
8.4	Η απόδοση του αλγορίθμου UKW όταν χρησιμοποιείται και για την επιλογή των γονιδίων ($GeneSet_4$) αλλά και για την ομαδοποίηση των δειγμάτων.	94
8.5	Η απόδοση του υβριδικού σχήματος εξαγωγής χαρακτηριστικών. . .	95
8.6	Η απόδοση του υβριδικού σχήματος στα 3 σύνολα δεδομένων.	97
9.1	Ποσοστά σωστής κατηγοριοποίησης των FNNs εκπαιδευμένων με διαφορετικές μεθόδους για όλα τα 100 πειράματα.	101
9.2	Αποτελέσματα σωστής ταξινόμησης του συνόλου ελέγχου για κάθε αλγόριθμο.	104

Μέρος I

Εισαγωγή και Βασικές Έννοιες

Εισαγωγή

Η αρχή είναι το πιο σημαντικό κομμάτι της δουλειάς.

—Πλάτωνας (427 π.Χ. – 347 π.Χ.)

Η Υπολογιστική Νοημοσύνη (ΥΝ) είναι ένα ταχέως αναπτυσσόμενο ερευνητικό πεδίο που ασχολείται με το σημαντικό θέμα της δημιουργίας πιο «έξυπνων» συστημάτων. Η ΥΝ περιλαμβάνει μεθόδους, όπως τα ασαφή συστήματα, τα νευρωνικά δίκτυα, τη μηχανική μάθηση, τις προηγμένες τεχνικές ομαδοποίησης, τα ευφυή προσαρμοστικά συστήματα, τους εξελικτικούς αλγόριθμους, και άλλες αναδυόμενες τεχνικές, που στοχεύουν στη διευκόλυνση της ανάπτυξης ευφυούς συμπεριφοράς σε πολύπλοκες και μεταβαλλόμενες καταστάσεις.

1.1 Ασαφή Συστήματα

Τα Ασαφή Συστήματα (ΑΣ) είναι ένας εναλλακτικός τρόπος αντιμετώπισης των παραδοσιακών εννοιών της λογικής. Η μελέτη των συστημάτων αυτών ξεκίνησε από την αρχή του προηγούμενου αιώνα όταν ο Bertrand Russell, ανακάλυψε ένα αρχαίο Ελληνικό παράδοξο που βρίσκεται στη βάση της σύγχρονης θεωρίας συνόλων και λογικής. Οι εφαρμογές των συστημάτων αυτών είναι πολλές και βρίσκονται στην αιχμή των τεχνικών Υπολογιστικής Νοημοσύνης. Παρά την πολύ παλαιά καταγωγή τους, είναι ένα σχετικά νέο πεδίο έρευνας, με πολλά περιθώρια ανάπτυξης. Τα ΑΣ, συμπεριλαμβανομένων της Ασαφούς Λογικής (ΑΛ) και της Ασαφούς Θεωρίας Συνόλων παρέχουν μια πλούσια και μεγάλης σημασίας προσθήκη στη τυπική λογική. Τα μαθηματικά που γεννιούνται από αυτές τις θεωρίες είναι σθεναρά, και η ΑΛ μπορεί να θεωρηθεί επέκταση της κλασικής λογικής. Οι εφαρμογές που εμπλέκονται τα ΑΣ έχουν ένα πολύ ευρύ φάσμα, καθώς πολλά συστήματα μπορούν να μοντελοποιηθούν, να εξομοιωθούν, και ακόμα να αναπαραχθούν με τη βοήθεια των ΑΣ, συμπεριλαμβάνοντας και την ίδια την ανθρώπινη λογική.

1.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση (ΜΜ) αναφέρεται στις αλλαγές που συμβαίνουν σε συστήματα που επιτελούν έργα συσχετισμένα με την Τεχνητή Νοημοσύνη. Σε αυτά περιλαμβάνονται η αναγνώριση προτύπων, η διάγνωση, ο σχεδιασμός, ο έλεγχος, η πρόβλεψη, κ.α.. Οι αλλαγές μπορεί να είναι είτε βελτιώσεις σε συστήματα που ήδη επιτελούν

μια εργασία, είτε καινοτόμες συνθέσεις νέων συστημάτων. Ένα τυπικό σύστημα MM, αντιλαμβάνεται και μοντελοποιεί το περιβάλλον του, και υπολογίζει τις κατάλληλες ενέργειες που πρέπει να επιτελέσει από τις μεταβολές που επιφέρουν αυτές στο ίδιο το περιβάλλον. Ως μάθηση θεωρούνται οι μεταβολές που συμβαίνουν σε οποιοδήποτε από τα μέρη ενός τέτοιου συστήματος με στόχο την βελτίωση της απόδοσής του.

1.3 Τεχνητά Νευρωνικά Δίκτυα

Τα *Τεχνητά Νευρωνικά Δίκτυα* ΤΝΔ είναι παράλληλα υπολογιστικά μοντέλα που αποτελούνται από πυκνά διασυνδεδεμένες, απλές και προσαρμοστικές υπολογιστικές μονάδες. Τα ΤΝΔ χαρακτηρίζονται από μια τάση να συγκεντρώνουν εμπειρική γνώση και να την κάνουν προσπελάσιμη προς χρήση. Τα ΤΝΔ προσομοιάζουν τον ανθρώπινο εγκέφαλο από δύο θεμελιώδεις πλευρές: Καταρχήν, το δίκτυο αποκτά γνώση για το περιβάλλον μέσα από μια διαδικασία μάθησης, και δεύτερο, η γνώση του δικτύου αποθηκεύεται στις διασυνδέσεις μεταξύ των νευρώνων που το αποτελούν [68, 69].

Πολλές μορφές τεχνητών νευρωνικών δικτύων έχουν αναπτυχθεί και χρησιμοποιηθεί για διαφορετικούς σκοπούς. Τα πιο διαδεδομένα από αυτά είναι τα δίκτυα Εμπρόσθιας Τροφοδότησης (Feed-Forward Neural Networks (FNN)). Η θεμελιώδης υπολογιστική μονάδα ενός FNN ονομάζεται *τεχνητός νευρώνας*, (artificial neuron). Ένας τεχνητός νευρώνας, αποτελείται από ένα αριθμό, n , από *συνάψεις*, ένα διάνυσμα βαρών w , μια *συνάρτηση ενεργοποίησης* (activation function), f , και τέλος από μια έξοδο y .

Η λειτουργία ενός νευρώνα συνοψίζεται στην επεξεργασία ενός διανύσματος εισόδου x_i , $i = 1, \dots, n$, ώστε να παραχθεί μια έξοδος y_i . Πιο συγκεκριμένα:

$$y_i = f(\langle x, y \rangle) = f\left(\sum_{i=0}^n w_i x_i\right).$$

Το εσωτερικό γινόμενο $v = \langle x, y \rangle$, γνωστό και ως *επίπεδο διέγερσης* (excitation level), αποτελεί το όρισμα της συνάρτησης ενεργοποίησης. Διάφορες συναρτήσεις ενεργοποίησης έχουν προταθεί αλλά οι πλέον διαδεδομένες είναι:

- Συνάρτηση κατώφλι (hard limit):

$$f(v) = \begin{cases} 1, & \text{αν } v \geq h, \\ 0, & \text{αν } v < h, \end{cases}$$

- Η τμηματικά γραμμική συνάρτηση (piecewise linear):

$$f(v) = \begin{cases} 1, & \text{αν } v \geq 1, \\ 0, & \text{αν } v \leq 0, \\ \xi, & \text{αν } 0 < v < 1, \end{cases}$$

- Η λογιστική συνάρτηση (logistic):

$$f(v) = \frac{1}{1 + e^{-v}},$$

- Η υπερβολική εφαπτομένη (hyperbolic tangent):

$$f(v) = \tanh(v).$$

Συνδυάζοντας νευρώνες, χρησιμοποιώντας τις εξόδους μερικών σαν εισόδους για άλλους, είναι δυνατόν να κατασκευαστούν ΤΝΔ με διαφορετικές τοπολογίες. Τα FNN αποτελούν ΤΝΔ στα οποία οι νευρώνες είναι οργανωμένοι σε επίπεδα και δεν υπάρχουν προς τα πίσω συνδέσεις. Στα FNN η είσοδος του δικτύου ανατίθεται στους νευρώνες που αποτελούν το επίπεδο εισόδου, ενώ η έξοδος του δικτύου προκύπτει από τους νευρώνες στο επίπεδο εξόδου. Οι υπόλοιποι νευρώνες είναι οργανωμένοι σε ενδιάμεσα επίπεδα, που ονομάζονται «κρυφά». Η έξοδος κάθε νευρώνα που ανήκει στο επίπεδο j αποτελεί είσοδο όλων των νευρώνων που ανήκουν στο επίπεδο $j + 1$.

Η λειτουργία αυτών των δικτύων βασίζεται σε μια σειρά από επαναλαμβανόμενα βήματα. Στην αρχή κάθε στοιχείο του διάνυσματος εισόδου αποδίδεται στην έξοδο ενός νευρώνα εισόδου. Στο επόμενο βήμα, οι νευρώνες του πρώτου κρυφού επιπέδου συλλέγουν το διάνυσμα εισόδου, το πολλαπλασιάζουν με το διάνυσμα βαρών, υπολογίζουν το άθροισμα και στη συνέχεια την τιμή της συνάρτησης ενεργοποίησης που ενεργεί σε αυτό. Αυτή η διαδικασία συνεχίζεται στα επόμενα επίπεδα του δικτύου και τελειώνει με τον υπολογισμό της εξόδου του δικτύου από τους νευρώνες του αντίστοιχου επιπέδου.

Η υπολογιστική ισχύς των FNNs πηγάζει από την ικανότητα τους να προσαρμόζονται σε ένα σύνολο εκπαίδευσης. Σύμφωνα με το *universal approximation theorem* [32], ένα FNN που αποτελείται από νευρώνες με μη-γραμμική συνάρτηση ενεργοποίησης, οργανωμένους σε ένα μόνο κρυφό επίπεδο, μπορεί να προσεγγίσει μια οποιαδήποτε συνεχή συνάρτηση.

Σε αυτά τα πλαίσια, ένα σύνολο εκπαίδευσης T είναι ένα σύνολο P από πρότυπα, (patterns), που ορίζεται ως:

$$T = \left\{ (x_k, d_k) \mid \begin{array}{l} x_k = (x_t, \dots, x_{t-n-1}) \in \mathbb{R}^M \\ d_k = x_{t+1} \in \mathbb{R} \end{array}, \quad k = 1, \dots, P \right\},$$

όπου x_k αναπαριστά το k -οστό πρότυπο εκπαίδευσης, και d_k την επιθυμητή απάντηση που αντιστοιχεί σε αυτό το πρότυπο. Ο στόχος της εκπαίδευσης είναι η ανάθεση τιμών στις ελεύθερες παραμέτρους του δικτύου, δηλαδή τα βάρη W , ώστε η μη αντιστοιχία μεταξύ των αποκρίσεων του δικτύου με βάση το διάνυσμα εισόδου και των επιθυμητών απαντήσεων να είναι όσο το δυνατό μικρότερη. Η διαδικασία εκπαίδευσης αρχίζει με την παρουσίαση όλων των προτύπων στο δίκτυο και τον υπολογισμό της συνάρτησης συνολικού σφάλματος E που ορίζεται ως:

$$E = \sum_{k=1}^P E_k,$$

όπου E_k είναι το μερικό σφάλμα του δικτύου ως προς το k -οστό πρότυπο:

$$E_k = \frac{1}{2} \left(y_i(W, x_k) - d_{ki} \right)^2.$$

Κάθε πλήρης παρουσίαση όλων των προτύπων ονομάζεται *εποχή εκπαίδευσης* (epoch). Αν η τιμή της συνάρτησης σφάλματος δεν υπερβαίνει το προκαθορισμένο κατώφλι, η εκπαίδευση τερματίζεται.

Εκπαιδεύοντας ένα FNN η πληροφορία που βρίσκεται στο σύνολο εκπαίδευσης αποθηκεύεται στα συναπτικά βάρη, μορφοποιώντας τη μακροχρόνια μνήμη του δικτύου. Μόλις η εκπαίδευση ενός δικτύου ολοκληρωθεί, η γνώση αυτή γίνεται αξιοποιήσιμη. Η ικανότητα ενός ΤΝΔ να ανταποκρίνεται σωστά σε πρότυπα τα οποία δεν έχει συναντήσει κατά την εκπαίδευσή του ονομάζεται *γενίκευση* (generalization). Η ικανοποιητική απόδοση ενός δικτύου στα πρότυπα που περιέχονται στο σύνολο εκπαίδευσης σε συνδυασμό με μια κακή απόδοση σε πρότυπα που δεν περιέχονται στο σύνολο αυτό, ονομάζεται *υπερπροσαρμογή* (overfitting). Το φαινόμενο αυτό υπονοεί ότι το δίκτυο έχει προσαρμοστεί υπερβολικά καλά στα δεδομένα του συνόλου εκπαίδευσης με αποτέλεσμα να έχει προσαρμοστεί σε ανακρίβειες, ή και λάθη, που μπορεί να περιέχονται σε αυτό.

1.4 Ομαδοποίηση Δεδομένων

Η *Ομαδοποίηση Δεδομένων* (data clustering) είναι μια από τις βασικές μεθόδους ΤΝ. Στόχος της είναι ο διαχωρισμός ενός συνόλου, σε ομάδες ώστε τα στοιχεία που ανήκουν στην ίδια ομάδα, να είναι περισσότερο όμοια μεταξύ τους από ότι τα στοιχεία που ανήκουν σε διαφορετικές ομάδες. Οι μέθοδοι ομαδοποίησης χρησιμοποιούνται σε πάρα πολλά επιστημονικά πεδία που περιλαμβάνουν, την αναγνώριση προτύπων [52], τη συμπίεση δεδομένων [119], τη βελτιστοποίηση [142], όπως επίσης και τις κοινωνικές επιστήμες [3]. Η ομαδοποίηση δεδομένων, που είναι και ο τομέας στον οποίο επικεντρώνεται η διατριβή αυτή, θα αναλυθεί εκτενέστερα στο επόμενο κεφάλαιο.

1.5 Εξελικτικοί Αλγόριθμοι

Ένας *Εξελικτικός Αλγόριθμος* (ΕΑ), είναι ένας γενικός όρος που χρησιμοποιείται για να χαρακτηρίσει οποιοδήποτε μέθοδο βελτιστοποίησης βασίζεται σε ένα πληθυσμό από πιθανές λύσεις και χρησιμοποιεί μηχανισμούς εμπνευσμένους από τη βιολογική εξέλιξη, όπως η αναπαραγωγή, η μετάλλαξη, ο επανασυνδυασμός, η φυσική επιλογή και η επιβίωση του καταλληλότερου. Οι πιθανές λύσεις στο πρόβλημα βελτιστοποίησης διαδραματίζουν το ρόλο των ατόμων σε έναν πληθυσμό, και η συνάρτηση κόστους, ή συνάρτηση προσαρμογής, καθορίζει το περιβάλλον μέσα στο οποίο οι λύσεις εξελίσσονται. Η εξέλιξη του πληθυσμού πραγματοποιείται μετά από την επαναλαμβανόμενη χρήση των ανωτέρω μηχανισμών (τελεστών). Συγκεκριμένα παραδείγματα ΕΑ δίνονται κατωτέρω. Οι περισσότερες από αυτές τις τεχνικές είναι παρόμοιες στις βασικές αρχές, αλλά διαφέρουν στις λεπτομέρειες της εφαρμογής τους και τη φύση του ιδιαίτερου προβλήματος στο οποίο έχουν εφαρμοστεί.

- Γενετικοί Αλγόριθμοι - Αυτός είναι ο δημοφιλέστερος τύπος ΕΑ, όπου η λύση ενός προβλήματος αναπαριστάνεται υπό την μορφή σειρών αριθμών (παραδοσιακά δυαδικών). και χρησιμοποιώντας τους τελεστές επανασυνδυασμού, επιλογής και μετάλλαξης, επιδιώκεται η εύρεση της βέλτιστης λύσης.
- Εξελικτικές Στρατηγικές - Όπου οι λύσεις αναπαρίστανται με διανύσματα πραγματικών αριθμών, και χρησιμοποιείται μόνο ο τελεστής μετάλλαξης ή ανασυνδυασμού για τη βελτιστοποίηση τους.

- Γενετικός προγραμματισμός - Σε αυτή την περίπτωση οι λύσεις αναπαρίστανται υπό τη μορφή προγραμμάτων υπολογιστών, και η ικανότητά τους καθορίζεται από τη δυνατότητά τους να λύσουν ένα υπολογιστικό πρόβλημα.
- Εξελικτικός Προγραμματισμός - Όπως στο γενετικό προγραμματισμό, με τη διαφορά ότι η δομή του προγράμματος είναι καθορισμένη και επιτρέπονται να εξελιχθούν μόνο οι αριθμητικές παράμετροί του.

Επειδή δεν κάνουν οποιαδήποτε υπόθεση για το πρόβλημα, γενικά θεωρείται ότι οι εξελικτικοί αλγόριθμοι χαρακτηρίζονται από καλή και συνεπή απόδοση σε όλους τους τύπους προβλημάτων. Αυτό αποδεικνύεται από την επιτυχία τους στους τομείς τόσο διαφορετικούς όσο η εφαρμοσμένη μηχανική, τέχνη, η βιολογία, οικονομικά, γενετική, έρευνα διαδικασιών, ρομποτική, κοινωνικές επιστήμες, φυσική, χημεία, και άλλοι.

Ομαδοποίηση Δεδομένων

Κάθε πρόβλημα που έλυσα έγινε κανόνας,
ο οποίος με εξυπηρέτησε για να λύσω άλλα.

—René Descartes (1596–1650)

Η ομαδοποίηση δεδομένων (data clustering) έχει τις ρίζες της στην στατιστική ανάλυση δεδομένων, αλλά πλέον χρησιμοποιείται σε πολλά πεδία που περιλαμβάνουν τη μηχανική μάθηση, την εξόρυξη δεδομένων, την αναγνώριση προτύπων, την ανάλυση εικόνων, την βιοπληροφορική και την βελτιστοποίηση.

Η πρώτη περιεκτική μελέτη της ομαδοποίησης, δημοσιεύθηκε το 1939 [143], αλλά οι πρωταρχικές αναφορές ξεκινούν από τον 4ο αιώνα π.Χ., από την προσπάθεια του Αριστοτέλη και του Θεόφραστου να ομαδοποιήσουν φυτά, σύμφωνα με τα μορφολογικά και δομικά χαρακτηριστικά τους.

Με τον όρο *ομαδοποίηση* (clustering) αποκαλούμε την διαδικασία διαμέρισης ενός συνόλου δεδομένων σε διαζευγμένα υποσύνολα (clusters) με στόχο στοιχεία που κατατάσσονται στο ίδιο υποσύνολο, να είναι περισσότερο όμοια μεταξύ τους από ότι στοιχεία κατατάσσονται σε διαφορετικά. Από αυτόν τον ορισμό είναι προφανές ότι προϋποτίθεται η ύπαρξη μια σχέσης ομοιότητας, ανάμεσα στα στοιχεία του συνόλου.

Είναι πολύ σημαντικό να γίνει ο διαχωρισμός ανάμεσα στην ομαδοποίηση που είναι μια τεχνική κατηγοριοποίηση χωρίς επίβλεψη, και τις τεχνικές κατηγοριοποίησης με επίβλεψη. Στης τεχνικές κατηγοριοποίησης με επίβλεψη δίνεται μια συλλογή αντικειμένων όπου κάθε αντικείμενο έχει μια επισήμανση «ετικέτα» που συμβολίζει την κατηγορία του. Το πρόβλημα σε αυτήν την περίπτωση ορίζεται στο να δοθεί μια ετικέτα σε ένα νέο αντικείμενο που δεν έχει ακόμα. Τυπικά χρησιμοποιείται πληροφορία από το ήδη υπάρχον σύνολο δεδομένων (εκπαίδευση) ώστε να ανακαλυφθούν διακρίσεις ανάμεσα στις κατηγορίες, οι οποίες χρησιμοποιούνται για να ανατεθεί το νέο αντικείμενο σε κάποια κατηγορία. Στην ομαδοποίηση το πρόβλημα ορίζεται σαν το διαχωρισμό μιας συλλογής αντικειμένων σε ξεχωριστές ομάδες (clusters) αλλά αυτή τη φορά η πληροφορία για το πως θα γίνει αυτό προέρχεται μόνο από τα δεδομένα τα ίδια (data driven) και όχι από εκ των προτέρων γνώση για την κατηγορία του καθενός.

Εκτός από τον όρο ομαδοποίηση δεδομένων (Data Clustering) ή απλά ομαδοποίηση (clustering) συχνά αναφέρεται και ως ανάλυση ομάδων (cluster analysis), αυτόματη κατηγοριοποίηση (automatic classification), αριθμητική ταξινόμηση (numerical taxonomy), και τυπολογική ανάλυση (typological analysis).

Η ευρύτητα των εφαρμογών της ομαδοποίησης και των επιστημονικών περιοχών που χρησιμοποιείται, αντικατοπτρίζει την μεγάλη σημασία της και τη χρησιμότητα της

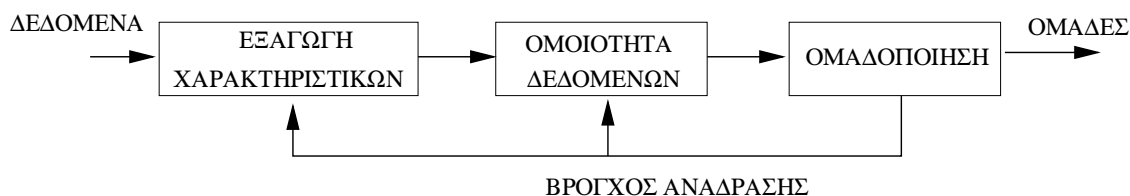
στην ανάλυση δεδομένων. Παρόλα αυτά οι διαφορές στις υποθέσεις και τα πλαίσια που χρησιμοποιούν οι διαφορετικές κοινότητες, δυσκολεύουν τη διάχυση των χρήσιμων γενικών εννοιών και μεθοδολογιών. Σε αυτήν την διατριβή η ομαδοποίηση αντιμετωπίζεται από το πλαίσιο της αναγνώρισης προτύπων, με στόχο την ανάδειξη νέων τεχνικών και μεθοδολογιών.

2.1 Τα στάδια μιας διαδικασίας ομαδοποίησης

Μια ολοκληρωμένη διαδικασία ομαδοποίησης τυπικά περιλαμβάνει τα ακόλουθα στάδια [75]:

1. Αναπαράσταση των δεδομένων (μπορεί να περιλαμβάνει την εξαγωγή ή/και επιλογή των χαρακτηριστικών).
2. Ορισμός του μέτρου ομοιότητας ανάμεσα στα στοιχεία.
3. Ομαδοποίηση.
4. Αφαίρεση δεδομένων.
5. Αξιολόγηση των αποτελεσμάτων.

Στο Σχήμα 2.1, απεικονίζεται η αλληλουχία των πρώτων τριών σταδίων, και περιλαμβάνει και το μονοπάτι ανάδρασης σύμφωνα με το οποίο το αποτέλεσμα της διαδικασίας επηρεάζει και τα προηγούμενα στάδια με στόχο την βελτιστοποίηση των αποτελεσμάτων.



Σχήμα 2.1: Τα στάδια της ομαδοποίησης.

Ο όρος *αναπαράσταση των δεδομένων* αναφέρεται στο πλήθος και το είδος των χαρακτηριστικών που περιγράφουν το κάθε ένα από τα δεδομένα. Η *επιλογή των χαρακτηριστικών* είναι η διαδικασία που επιχειρεί να επιλέξει από τα διαθέσιμα χαρακτηριστικά τα πιο καθοριστικά από αυτά. Αντίθετα η *εξαγωγή των χαρακτηριστικών*, αναφέρεται στο μετασχηματισμό των χαρακτηριστικών σε άλλα, περισσότερο καθοριστικά όσον αφορά την περαιτέρω ομαδοποίηση των δεδομένων.

Ένα μέτρο ομοιότητας που ορίζεται σε ζευγάρια από στοιχεία συνήθως χρησιμοποιείται για να αναπαραστήσει την ομοιότητα ανάμεσα στα δεδομένα. Τέτοια μέτρα έχουν προταθεί πολλά και ποικίλουν ανάλογα με την εφαρμογή [150]. Ένα απλό μέτρο απόστασης όπως η Ευκλείδεια απόσταση μπορεί να χρησιμοποιηθεί για την έκφραση της ομοιότητας ανάμεσα σε δύο στοιχεία, ενώ μπορούν να χρησιμοποιηθούν άλλα μέτρα απόστασης για να χαρακτηρίσουν την λογική ομοιότητα [101]. Στην Ενότητα 2.4, θα μελετηθούν εκτενέστερα τέτοιου είδους μέτρα.

Η τελική ομαδοποίηση των δεδομένων μπορεί να είναι αυστηρή (μια διαμέριση των στοιχείων σε ομάδες) ή ασαφής (κάθε στοιχείο έχει ένα μεταβλητό βαθμό συμμετοχής

σε κάθε ομάδα). Από την άλλη μεριά, οι Ιεραρχικοί αλγόριθμοι ομαδοποίησης παράγουν μία φωλιασμένη σειρά από διαμερίσεις βασισμένη σε ένα κριτήριο ενοποίησης ή διαχωρισμού ανάλογα με τις τιμές του μέτρου ομοιότητας. Οι Διαμεριστικοί αλγόριθμοι είναι σχεδιασμένοι για να ανακαλύψουν την διαμέριση που βελτιστοποιεί (συνήθως τοπικά) ένα κριτήριο ομαδοποίησης. Άλλες τεχνικές ομαδοποίησης περιλαμβάνουν τις *Πιθανοτικές* [40] και *Γράφο-Θεωρητικές* [155] μεθόδους. Οι διαφορετικές τεχνικές ομαδοποίησης θα μελετηθούν εκτενέστερα στην Ενότητα 2.5.

Με τον όρο *Αφαίρεση Δεδομένων* εννοούμε τη διαδικασία της εξαγωγής απλών και σθενερών αναπαραστάσεων των δεδομένων. Σε αυτή την περίπτωση η απλότητα μπορεί να αναφέρεται είτε από την σκοπιά της αυτόματης ανάλυσης (έτσι ώστε η μηχανή να μπορεί να εκτελέσει ευκολότερα περαιτέρω υπολογισμούς) είτε από την ανθρωπο-κεντρική σκοπιά (έτσι ώστε η παραγόμενη αναπαράσταση να είναι εύκολα κατανοήσιμη). Στο πλαίσιο της ομαδοποίησης, μια τυπική αφαίρεση δεδομένων είναι ο χαρακτηρισμός της κάθε ομάδας, συνήθως με όρους πρωτότυπων ομάδας (cluster prototypes), ή αντιπροσωπευτικών στοιχείων όπως ο μέσος όρος [35].

Η αξιολόγηση του αποτελέσματος ενός αλγόριθμου ομαδοποίησης είναι πολύπλευρη. Μια πλευρά περιλαμβάνει την αξιολόγηση των ίδιων των δεδομένων αντί του αποτελέσματος του αλγόριθμου. Δεδομένα που δεν περιέχουν μια δομή ομάδων δεν θα έπρεπε να επεξεργαστούν από ένα αλγόριθμο ομαδοποίησης. Η μελέτη της *Τάσης Ομαδοποίησης* (cluster tendency), όπου εξετάζονται τα δεδομένα πριν γίνει οποιαδήποτε προσπάθεια για την ομαδοποίηση τους, είναι μια σχετικά αδρανής επιστημονική περιοχή [75]. Περισσότερες πληροφορίες αναφέρονται στις εργασίες [27, 41].

Αντίθετα η *Ανάλυση Εγκυρότητας* επιχειρεί να αξιολογήσει το αποτέλεσμα της ομαδοποίησης. Συνήθως χρησιμοποιεί για το σκοπό αυτό ένα συγκεκριμένο κριτήριο βελτιστότητας. Ωστόσο, τα κριτήρια αυτά συνήθως δημιουργούνται υποκειμενικά. Επομένως, δεν υπάρχουν αναμφισβήτητοι κανόνες, εκτός από σε συγκεκριμένες υποπεριοχές. Όταν χρησιμοποιούνται στατιστικές προσεγγίσεις στην ομαδοποίηση, τότε εξετάζεται η εγκυρότητα χρησιμοποιώντας στατιστικές μεθόδους και έλεγχο υποθέσεων. Γενικότερα υπάρχουν τρεις κατηγορίες ελέγχων εγκυρότητας [75]:

1. *Εξωτερικής Εκτίμησης* εγκυρότητας, όπου συγκρίνεται η τελική ομαδοποίηση σύμφωνα με κάποια προκαθορισμένη δομή.
2. *Εσωτερικής Εκτίμησης* εγκυρότητας, με την οποία επιχειρείται να προσδιοριστεί εάν η δομή της ομαδοποίησης έχει μια ουσιαστική καταλληλότητα για τα δεδομένα.
3. *Σχετικού Ελέγχου*, που συγκρίνει δύο δομές μεταξύ τους και μετράει τη σχετική τους αξία.

Περισσότερες πληροφορίες για την ανάλυση εγκυρότητας υπάρχουν στις εργασίες [42, 74].

2.2 Ορολογία και ορισμοί

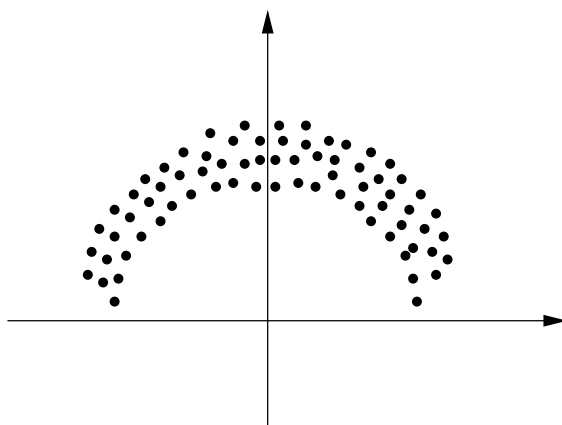
Σε αυτή την ενότητα αναφέρουμε τους όρους και τους ορισμούς που θα χρησιμοποιηθούν στη διατριβή:

- Ένα *στοιχείο* (ή εγγραφή, ή παρατήρηση, ή πρότυπο, ή σημείο), X αποτελεί ένα συγκεκριμένο δεδομένο, και τυπικά περιγράφεται σαν ένα d -διάστατο πραγματικό διάνυσμα, $X = (x_1, \dots, x_d)$, όπου d το πλήθος των χαρακτηριστικών.

- Τα στοιχεία x_i , της παρατήρησης x , καλούνται χαρακτηριστικά (ή ιδιότητες, ή μεταβλητές, ή διαστάσεις).
- Ένα σύνολο δεδομένων $\mathcal{X}$, n στοιχείων αναπαρίσταται σαν $\mathcal{X} = \{X_1, \dots, X_n\}$, όπου $X_i = (x_{i,1}, \dots, x_{i,d})$. Σε πολλές περιπτώσεις το σύνολο $\mathcal{X}$, αναπαρίσταται και σαν ένα μητρώο $n \times d$, στοιχείων.
- Ένα μέτρο ομοιότητας, είναι μία μετρική ορισμένη στο χώρο των χαρακτηριστικών, που χρησιμοποιείται για να μετρήσει την ομοιότητα ανάμεσα στα δεδομένα.

2.3 Αναπαράσταση των Δεδομένων

Γενικότερα δεν υπάρχουν, συγκεκριμένες θεωρητικές προδιαγραφές που να προτείνουν τον βέλτιστο τρόπο αναπαράστασης των δεδομένων. Μάλιστα τις περισσότερες φορές η δημιουργία των δεδομένων δεν μπορεί καν να ελεγχθεί. Ο ρόλος του χρήστη σε αυτή την περίπτωση είναι να συλλέξει τα δεδομένα και προαιρετικά να επιλέξει ή να παράγει τα χαρακτηριστικά που θα χρησιμοποιηθούν στα επόμενα βήματα της διαδικασίας. Εξαιτίας της δυσκολίας αυτού του βήματος, συνήθως γίνεται η παραδοχή ότι η αναπαράσταση των δεδομένων είναι εκ των προτέρων δεδομένη. Ωστόσο, αυτό το βήμα είναι μεγάλης σημασίας καθώς διαφορετικές αναπαραστάσεις μπορούν να επηρεάσουν σημαντικά το αποτέλεσμα της ομαδοποίησης. Για παράδειγμα στο Σχήμα 2.2, απεικονίζονται σημεία ενός δισδιάστατου Ευκλείδειου χώρου, οργανωμένα σε μία ημικυκλική ομάδα, με προσεγγιστικά ίση απόσταση από την αρχή των αξόνων. Επιλέγοντας το Καρτεσιανό τρόπο αναπαράστασης, πολλοί αλγόριθμοι ομαδοποίησης θα διασπάσουν την ομάδα σε περισσότερες εξαιτίας της μη κυρτής μορφής της. Αντίθετα, αν χρησιμοποιηθεί το πολικό σύστημα συντεταγμένων, είναι ευκολότερο να προσδιοριστεί η μονο-ομαδική μορφή των δεδομένων.



Σχήμα 2.2: Μία ημικυκλική ομάδα από σημεία όπου κάθε σημείο απέχει προσεγγιστικά ίση απόσταση από την αρχή των αξόνων.

Ένα στοιχείο μπορεί να αναπαριστά είτε ένα φυσικό αντικείμενο (π.χ. τραπέζι) είτε μία αφηρημένη έννοια (π.χ. τρόπος συμπεριφοράς). Ο παραδοσιακός τρόπος αναπαράστασης του κάθε στοιχείου είναι με ένα πολυ-διάστατο διάνυσμα, όπου κάθε συντεταγμένη αναπαριστά ένα συγκεκριμένο χαρακτηριστικό [43]. Κάθε χαρακτηριστικό μπορεί να είναι είτε ποσοτικό (quantitative) είτε ποιοτικό (qualitative). Για παράδειγμα, εάν

το χρώμα μαλλιών και η ηλικία χρησιμοποιούνται σαν χαρακτηριστικά για να αναπαραστήσουν μια ομάδα ανθρώπων, τότε το στοιχείο (ξανθά, 24), είναι η αναπαράσταση ενός ανθρώπου με ξανθά μαλλιά, 24 ετών. Τα χαρακτηριστικά μπορούν να χωριστούν περαιτέρω στις παρακάτω κατηγορίες:

1. Ποσοτικά χαρακτηριστικά:

- (α') Συνεχείς τιμές (π.χ. βάρος).
- (β') Διακριτές τιμές (π.χ. αριθμός τραπεζιών).
- (γ') Τιμές διαστημάτων (π.χ. διάρκεια ενός γεγονότος).

2. Ποιοτικά χαρακτηριστικά:

- (α') Ονομαστικά (κατηγορικά) ή χωρίς διάταξη (π.χ. χρώμα).
- (β') Διατεταγμένα (π.χ. ποιοτική αναπαράσταση της ηλικίας, νέος, μεσήλικας, γέρος).

Ένα πολύ σημαντικό βήμα της διαδικασίας αναπαράστασης είναι και η επιλογή των πιο αντιπροσωπευτικών χαρακτηριστικών. Οι τεχνικές επιλογής χαρακτηριστικών επιλέγουν το υποσύνολο των χαρακτηριστικών από τα διαθέσιμα, με στόχο συνήθως την βελτιστοποίηση κάποιου κριτηρίου εγγύτητας της ομαδοποίησης μέσω μιας διαδικασίας δοκιμής-λάθους [43]. Αντίθετα οι τεχνικές εξαγωγής χαρακτηριστικών όπως η Ανάλυση Πρωτευόντων Συνιστωσών (Principal Component Analysis), μετασχηματίζουν την αναπαράσταση των δεδομένων σε ένα χώρο μικρότερης διάστασης.

2.4 Μέτρα Ομοιότητας

Όπως έχει ήδη αναφερθεί ο καθορισμός του τρόπου που θα υπολογίζεται η ομοιότητα ανάμεσα στα δεδομένα είναι θεμελιώδης για την όλη διαδικασία της ομαδοποίησης. Εξαιτίας, της μεγάλης ποικιλίας των τύπων των χαρακτηριστικών και της διαφορετικής κλίμακας των μεγεθών τους, η επιλογή του μέτρου ομοιότητας πρέπει να γίνει με μεγάλη προσοχή. Είναι πολύ συνηθισμένο να μετράται η μη ομοιότητα δύο στοιχείων από το σύνολο δεδομένων μέσω μιας συνάρτησης απόστασης ορισμένη στο πεδίο τιμών των χαρακτηριστικών.

Η πιο δημοφιλής μετρική για συνεχή χαρακτηριστικά είναι η *Ευκλείδεια απόσταση*,

$$d_2(X_i, X_j) = \left(\sum_{k=1}^d |x_{i,k} - x_{j,k}|^2 \right)^{1/2} = \|X_i - X_j\|_2,$$

η οποία είναι η ειδική περίπτωση της μετρικής (για $p = 2$) *Minkowski*,

$$d_p(X_i, X_j) = \left(\sum_{k=1}^d |x_{i,k} - x_{j,k}|^p \right)^{1/p} = \|X_i - X_j\|_p.$$

Η Ευκλείδεια απόσταση είναι η περισσότερο διαισθητικά κατανοητή, καθώς χρησιμοποιείται για να μετρήσει την απόσταση των αντικείμενων στο επίπεδο και στον τρισδιάστατο χώρο. Γενικά, οι αλγόριθμοι που την χρησιμοποιούν αποδίδουν καλά όταν το σύνολο δεδομένων περιέχει «πυκνές» ή «απομονωμένες» ομάδες [99]. Το μειονέκτημα

της άμεσης χρήσης μετρικών Minkowski, είναι η τάση να επικρατεί στην μέτρηση της ομοιότητας το χαρακτηριστικό με τη μεγαλύτερη κλίμακα. Αυτό το πρόβλημα μπορεί να λυθεί εάν κανονικοποιηθούν τα δεδομένα (σε μία κοινή κλίμακα ή διασπορά), είτε εάν χρησιμοποιηθούν άλλες τεχνικές επιβολής βαρών στα χαρακτηριστικά. Η γραμμική αυτοσυσχέτιση ανάμεσα στα χαρακτηριστικά μπορεί επίσης να αλλοιώσει τέτοια μέτρα ομοιότητας· αυτή η αλλοίωση μπορεί να αποφευχθεί χρησιμοποιώντας ένα μετασχηματισμό βαρύτητας στα δεδομένα ή με τη χρήση της τετραγωνικής μετρικής *Mahalanobis*:

$$d_M(X_i, X_j) = (X_i - X_j)\Sigma^{-1}(X_i - X_j)^T,$$

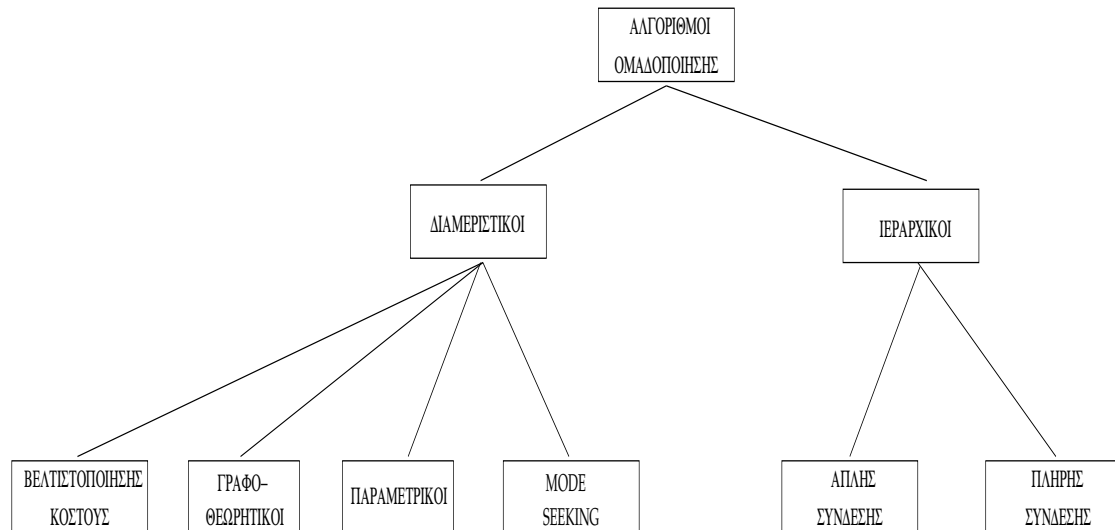
όπου τα στοιχεία X_i, X_j θεωρούνται διανύσματα γραμμής, και Σ είναι το μητρώο συνδιασποράς του δείγματος, ή το πραγματικό μητρώο συνδιασποράς της διαδικασίας δημιουργίας των δεδομένων. Η μετρική αυτή αναθέτει διαφορετικά βάρη στα χαρακτηριστικά ανάλογα με τις διασπορές τους και τις γραμμικές συσχετίσεις τους. Η κανονικοποιημένη μετρική *Mahalanobis* χρησιμοποιήθηκε στην εργασία [99], για την εξαγωγή υπερ-ελλειψοειδών ομάδων.

Ο υπολογισμός αποστάσεων δεν είναι αποδοτικός όταν μερικά ή όλα τα χαρακτηριστικά δεν είναι συνεχή, καθώς διαφορετικοί τύποι χαρακτηριστικών δεν είναι συγκρίσιμοι μεταξύ τους. Παρόλα αυτά, για το σκοπό αυτό έχουν προταθεί πολλές μετρικές. Περισσότερες πληροφορίες για αυτήν την περίπτωση θα δοθούν στο Κεφάλαιο 4.

2.5 Μέθοδοι Ομαδοποίησης

Οι περισσότεροι ερευνητές διαχωρίζουν τους αλγόριθμους ομαδοποίησης σύμφωνα με την κατάταξη που απεικονίζεται στο Σχήμα 2.3. Ωστόσο αυτή η ταξινόμηση μπορεί να συμπληρωθεί και από άλλα στοιχεία που εμπλέκονται σε διαφορετικούς αλγόριθμους ανεξάρτητα από την ταξινόμηση τους.

- **Συγχωνευτικοί ή Διαιρετικοί αλγόριθμοι:** Η ιδιότητα της συγχώνευσης αναφέρεται σε αλγόριθμος που λειτουργούν θεωρώντας αρχικά κάθε στοιχείο σαν μια ομάδα (μονο-ομάδα (singleton)), και σταδιακά ενοποιούν ομάδες μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού. Αντίθετα ένας διαιρετικός αλγόριθμος θεωρεί ότι όλα τα δεδομένα σχηματίζουν μία ομάδα την οποία έπειτα διαιρεί σε άλλες, έως ότου ικανοποιήσει το κριτήριο τερματισμού.
- **Αυστηροί ή Ασαφείς αλγόριθμοι:** Ένα αυστηρός αλγόριθμος αναθέτει το κάθε δεδομένο σε μία μόνο ομάδα κατά τη λειτουργία και έξοδο του. Ένας ασαφής αλγόριθμος αναθέτει βαθμούς συμμετοχής του κάθε στοιχείου στις ομάδες. Μία ασαφής ομαδοποίηση μπορεί να μετατραπεί σε αυστηρή αναθέτοντας τα στοιχεία στην ομάδα στην οποία έχουν το μεγαλύτερο βαθμό συμμετοχής.
- **Ντετερμινιστικοί ή Στοχαστικοί αλγόριθμοι:** Συνήθως αυτή η ιδιότητα αναφέρεται σε μεθόδους που επιχειρούν να βελτιστοποιήσουν μια συνάρτηση κόστους. Ο τρόπος βελτιστοποίησης καθορίζει το είδος του αλγορίθμου σε στοχαστικό ή ντετερμινιστικό.



Σχήμα 2.3: Μία ταξινόμηση των αλγορίθμων ομαδοποίησης.

2.5.1 Ιεραρχικοί Αλγόριθμοι Ομαδοποίησης

Οι Ιεραρχικοί Αλγόριθμοι Ομαδοποίησης εκτελούν συναθροίσεις, *hierarchical agglomerative*, ή διαχωρισμούς των δεδομένων, *hierarchical divisive*.

Χαρακτηριστικό των ιεραρχικών μεθόδων είναι πως η ανάθεση ενός αντικειμένου σε μια ομάδα είναι οριστική. Δηλαδή μόλις ένα αντικείμενο ενωθεί με μια ομάδα ποτέ δεν απομακρύνεται και δεν συγχωνεύεται με άλλα αντικείμενα που ανήκουν σε κάποια άλλη ομάδα.

Οι συναθροιστικές ιεραρχικές (agglomerative) μέθοδοι σχηματίζουν μια σειρά από συγχωνεύσεις των n αντικειμένων σε ομάδες καταλήγοντας σε μια ομάδα η οποία περιλαμβάνει το σύνολο των αντικειμένων. Αντίθετα, οι διαιρετικές ιεραρχικές (divisive) μέθοδοι χωρίζουν το σύνολο των n αντικειμένων σε πιο εκλεπτυσμένες διαμερίσεις και τελικά καταλήγουν στον εντοπισμό n ομάδων, που η καθεμία περιέχει ένα αντικείμενο.

Τα αποτελέσματα των συναθροιστικών και των διαιρετικών μεθόδων παριστάνονται σε ένα δισδιάστατο διάγραμμα γνωστό ως *δενδρόγραμμα* (*dendrogram*).

Οι παραγόμενες ομάδες είναι φωλιασμένες και καθεμία μπορεί να θεωρηθεί ως ένα μέρος μίας ευρύτερης και πιο περιεκτικής ομάδας που χαρακτηρίζεται από ένα υψηλότερο επίπεδο ομοιότητας [3, 36, 50, 125].

Ιεραρχικές Συναθροιστικές Μέθοδοι

Γενικά, τα βήματα ενός συναθροιστικού αλγορίθμου για την ομαδοποίηση n αντικειμένων είναι τα εξής:

Βήμα 1^ο Αρχικά, τα n αντικείμενα χωρίζονται σε n ομάδες. Κατασκευάζεται το $n \times n$ συμμετρικό μητρώο αποστάσεων $D = d_{ij}$.

Βήμα 2^ο Στο μητρώο D αναζητείται το ζεύγος των πλησιέστερων ομάδων. Έστω ότι η απόσταση μεταξύ των κοντινότερων ομάδων U και V είναι d_{UV} .

Βήμα 3^ο Συγχωνεύονται οι ομάδες U και V . Η νέα ομάδα ονομάζεται (UV) . Ενημερώνονται τα πεδία του μητρώου αποστάσεων, α) διαγράφοντας

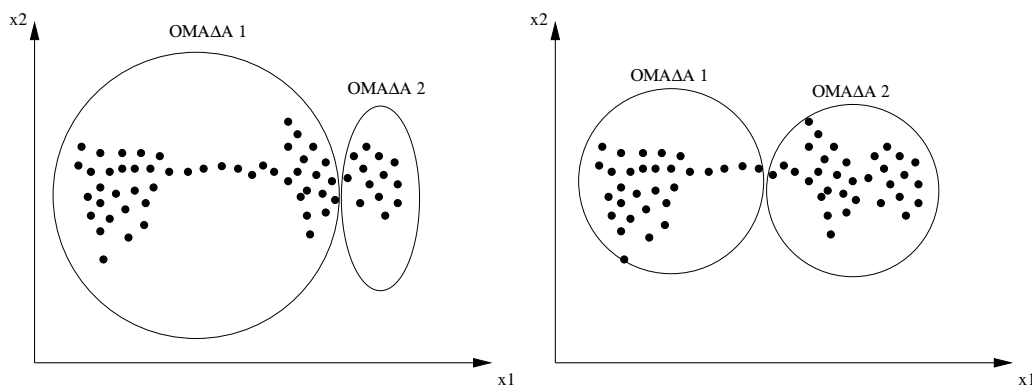
τις γραμμές και τις στήλες που αντιστοιχούν στις ομάδες U και V και β) προσθέτοντας μια γραμμή και στήλη που δείχνει την απόσταση της ομάδας (UV) από τις υπόλοιπες ομάδες.

Βήμα 4^ο Επαναλαμβάνονται τα βήματα 2 και 3 συνολικά $n - 1$ φορές. Στον τερματισμό του αλγορίθμου όλα τα αντικείμενα ανήκουν σε μία ομάδα. Οι ομάδες που προέκυψαν κατά την εκτέλεση του αλγορίθμου καταγράφονται μαζί με το επίπεδο ομοιότητας κάθε συγχώνευσης.

Οι ιεραρχικοί συνανθροιστικοί μέθοδοι διακρίνονται και από τους διαφορετικούς τρόπους «σύνδεσης» (linkage) που εφαρμόζονται κατά τον σχηματισμό των ομάδων. Ενώ η μέτρηση της απόστασης ανάμεσα σε δύο στοιχεία είναι αρκετά σαφής, η μέτρηση της απόστασης ανάμεσα σε δύο ομάδες που περιέχουν περισσότερα από ένα στοιχεία μπορεί να γίνει με πολλούς τρόπους. Οι πιο σημαντικοί τρόποι μέτρησης (ή σύνδεσης) είναι οι παρακάτω [3, 13, 36, 42, 50, 75, 125, 140]:

- Απλή σύνδεση ή κοντινότερου γείτονα (Single linkage - The nearest neighbour)
- Πλήρης σύνδεση ή απώτατου γείτονα (Complete linkage - The furthest neighbour)
- Σύνδεση μέσου όρου (Average linkage)
- Μέθοδος του Ward
- Σύνδεση κεντροειδούς (Centroid)
- Σύνδεση διαμέσου (Median)

Η μέθοδος πλήρους σύνδεσης υπολογίζει την απόσταση δύο ομάδων χρησιμοποιώντας την απόσταση των πιο απόμακρων στοιχείων τους. Αυτός ο τρόπος υπολογισμού της απόστασης, παράγει στενά συνδεδεμένες ή πυκνές ομάδες. Αντίθετα η μέθοδος της απλής σύνδεσης όπου υπολογίζει την απόσταση ανάμεσα στις ομάδες από την απόσταση των πιο κοντινών τους σημείων υποφέρει από το πρόβλημα της αλυσίδας (chaining effect). Στο Σχήμα 2.4, απεικονίζονται αριστερά το αποτέλεσμα της ομαδοποίησης χρησιμοποιώντας την μέθοδο απλής σύνδεσης ενώ στα δεξιά της πλήρους σύνδεσης.



Σχήμα 2.4: Ομαδοποίηση χρησιμοποιώντας την μέθοδο απλής σύνδεσης (αριστερά), και της πλήρους σύνδεσης (δεξιά).

Οι τρόποι σύνδεσης που χρησιμοποιούνται από τις προαναφερθείσες μεθόδους ικανοποιούν έναν επαναληπτικό τύπο για την απόσταση, $d_{k(ij)}$, μεταξύ της ομάδας k και

της ομάδας (ij) , που προέκυψε από την συγχώνευση των ομάδων (ή των στοιχείων) i και j κατά τη διάρκεια ενός συναθροιστικού βήματος. Ο τύπος αυτός είναι ο ακόλουθος:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}|, \quad (2.1)$$

όπου d_{ij} είναι η απόσταση μεταξύ των ομάδων i και j και α , β , γ αναφέρονται ως παράμετροι, οι τιμές των οποίων καθορίζουν τον τύπο 2.1, όπως φαίνεται στον Πίνακα 2.1.

Κοντινότερος Γείτονας	$\alpha_i = \alpha_j = \frac{1}{2}$	$\beta = 0$	$\gamma = -\frac{1}{2}$
Απώτατος Γείτονας	$\alpha_i = \alpha_j = \frac{1}{2}$	$\beta = 0$	$\gamma = \frac{1}{2}$
Σύνδεση Μέσου Όρου	$\alpha_i = \frac{n_i}{n_i+n_j}, \alpha_j = \frac{n_j}{n_i+n_j}$	$\beta = 0$	$\gamma = 0$
Μέθοδος του Ward	$\alpha_i = \frac{n_k+n_i}{n_k+n_i+n_j}, \alpha_j = \frac{n_k+n_j}{n_k+n_i+n_j}$	$\beta = \frac{-n_k}{n_k+n_i+n_j}$	$\gamma = 0$
Centroid ομαδοποίηση	$\alpha_i = \frac{n_i}{n_i+n_j}, \alpha_j = \frac{n_j}{n_i+n_j}$	$\beta = -\alpha_i \alpha_j$	$\gamma = 0$
Median ομαδοποίηση	$\alpha_i = \alpha_j = \frac{1}{2}$	$\beta = -\frac{1}{4}$	$\gamma = 0$

Πίνακας 2.1: Παράμετροι για τον τύπο 2.1.

Για την καλύτερη κατανόηση των αλγορίθμων αυτών παραθέτουμε ένα παράδειγμα ιεραρχικής συγχωνευτικής ομαδοποίησης. Έστω λοιπόν το σύνολο δεδομένων X , με 5 στοιχεία ως ακολούθως:

$$X = \left\{ \begin{matrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{matrix} \right\} = \left\{ \begin{matrix} 1 & 1 \\ 1 & 1.5 \\ 3.5 & 1 \\ 4 & 1.5 \\ 3 & 1 \end{matrix} \right\}$$

Το μητρώο αποστάσεων μεταξύ των σημείων για το σύνολο αυτό χρησιμοποιώντας την Ευκλείδεια απόσταση γίνεται:

$$\left\{ \begin{matrix} & x_0 & x_1 & x_2 & x_3 & x_4 \\ x_0 & 0 & 0.25 & 6.25 & 9.25 & 4 \\ x_1 & 0.25 & 0 & 6.5 & 9 & 4.25 \\ x_2 & 6.25 & 6.5 & 0 & 0.5 & 0.25 \\ x_3 & 9.25 & 9 & 0.5 & 0 & 1.25 \\ x_4 & 4 & 4.25 & 0.25 & 1.25 & 0 \end{matrix} \right\}$$

Εξετάζοντας τις τιμές αυτού του πίνακα ο αλγόριθμος εντοπίζει ότι μια από τις μικρότερες αποστάσεις βρίσκεται ανάμεσα στα στοιχεία x_2 και x_4 , και έτσι ενοποιεί τις μονο-ομάδες $\{x_2\}$ και $\{x_4\}$ κατασκευάζοντας την ομάδα $\{x_2, x_4\}$. Στην επόμενη επανάληψη το μητρώο αποστάσεων γίνεται:

$$\left\{ \begin{matrix} & x_0 & x_1 & x_3 & \{x_2, x_4\} \\ x_0 & 0 & 0.25 & 9.25 & 4 \\ x_1 & 0.25 & 0 & 9 & 4.25 \\ x_3 & 9.25 & 9 & 0 & 0.5 \\ \{x_2, x_4\} & 4 & 4.25 & 0.5 & 0 \end{matrix} \right\}$$

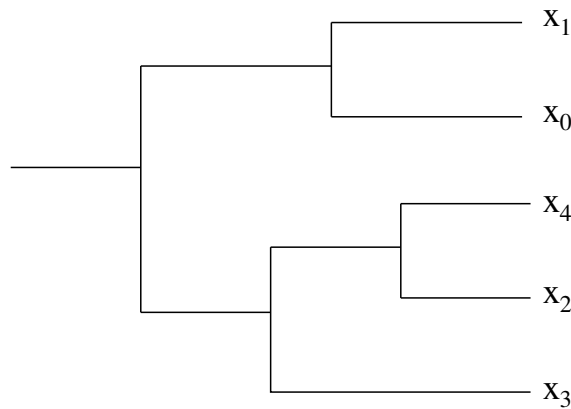
Ομοίως ο αλγόριθμος στο επόμενο βήμα κατασκευάζει την ομάδα $\{x_0, x_1\}$, και το αντίστοιχο μητρώο αποστάσεων γίνεται:

$$\begin{pmatrix} & \{x_0, x_1\} & x_3 & \{x_2, x_4\} \\ \{x_0, x_1\} & 0 & 9 & 4 \\ x_3 & 9 & 0 & 0.5 \\ \{x_2, x_4\} & 4 & 0.5 & 0 \end{pmatrix}$$

Στο επόμενο βήμα η μικρότερη απόσταση βρίσκεται ανάμεσα στη μονο-ομάδα $\{x_3\}$ και την ομάδα $\{x_2, x_4\}$, οπότε και ενοποιούνται. Το μητρώο αποστάσεων παίρνει και την τελική του μορφή:

$$\begin{pmatrix} & \{x_0, x_1\} & \{x_3, x_2, x_4\} \\ \{x_0, x_1\} & 0 & 4 \\ \{x_3, x_2, x_4\} & 4 & 0 \end{pmatrix}$$

Σε αυτό το σημείο ο αλγόριθμος δεν έχει παρά μόνο μια επιλογή να ενώσει και τις δύο ομάδες σε μια. Το παραγόμενο δενδρόγραμμα απεικονίζεται στο Σχήμα 2.5.



Σχήμα 2.5: Το παραγόμενο δενδρόγραμμα.

Ιεραρχικές Μέθοδοι Διαίρεσης

Μία ιεραρχική μέθοδος διαίρεσης (divisive method) στην εκκίνησή της θεωρεί ότι όλες οι οντότητες ανήκουν σε μία ομάδα και στη συνέχεια διαχωρίζει αυτή την ομάδα. Αν το σύνολο που θέλουμε να ομαδοποιήσουμε αποτελείται από n οντότητες τότε το πλήθος των υποσυνόλων μεγέθους 2 είναι $2^{n-1} - 1$. Μόλις πραγματοποιηθεί ο αρχικός διαχωρισμός, τα αντικείμενα μετακινούνται από μία ομάδα σε μία άλλη ή εκτελούνται πιο εκλεπτυσμένες υποδιαίρεσεις των ήδη σχηματιζόμενων ομάδων [36, 50, 125].

Υπάρχουν δύο στρατηγικές διαίρεσης [3, 50, 125]:

- **Μονοθετικές (Monothetic):** Μονοθετική χαρακτηρίζεται μια ομάδα στην οποία όλες οι οντότητες έχουν προσεγγιστικά την ίδια τιμή για μια συγκεκριμένη μεταβλητή. Δηλαδή, οι μονοθετικές ομάδες καθορίζονται από συγκεκριμένες μεταβλητές στις οποίες συγκεκριμένες τιμές είναι *απαραίτητες* για να γίνουν οι οντότητες μέλη των ομάδων.

- **Πολυθετικές (Polythetic):** Μια πολυθετική ομάδα είναι μια ομάδα στην οποία όλες οι οντότητες έχουν προσεγγιστικά τις ίδιες τιμές για ένα υποσύνολο συγκεκριμένων μεταβλητών. Δηλαδή, οι πολυθετικές ομάδες καθορίζονται από συγκεκριμένο υποσύνολο μεταβλητών για τις οποίες συγκεκριμένες τιμές είναι απαραίτητες για να γίνουν οι οντότητες μέλη των ομάδων.

Η πιο συνηθισμένη μορφή των αλγορίθμων αυτών, ξεχωρίζει επαναληπτικά από τις ομάδες το στοιχείο το οποίο είναι περισσότερο αταίριαστο με την αντίστοιχη ομάδα. Για την επιλογή της ομάδας κάθε φορά, χρησιμοποιεί το μέτρο της διαμέτρου, που είναι η μεγαλύτερη απόσταση ανάμεσα σε κάθε ζευγάρι στοιχείων. Με αυτό τον τρόπο κατασκευάζεται το δενδρόγραμμα που αναπαριστά και το τελικό αποτέλεσμα του αλγορίθμου.

Ο Αλγόριθμος PDDP

Ο αλγόριθμος PDDP (Principal Direction Divisive Partitioning) [17], είναι μια ιεραρχική μέθοδος διαίρεσης με βασικό του πλεονέκτημα την ταχύτητα εκτέλεσης σε σχέση με αντίστοιχους αλγορίθμους. Ο αλγόριθμος αυτός, βασίζεται στην τεχνική Ανάλυσης σε Πρωτεύουσες Συνιστώσες (Principal Component Analysis (PCA)) [78], και αξιοποιεί τη σποραδικότητα των δεδομένων. Βέβαια σαν ιεραρχικός αλγόριθμος ομαδοποίησης το αποτέλεσμα του είναι ένα ιεραρχικό δενδρόγραμμα των ομάδων.

Πιο συγκεκριμένα, αν θεωρήσουμε το σύνολο δεδομένων να αναπαριστάται από ένα μητρώο $A \in \mathbb{R}^{m \times n}$ όπου κάθε γραμμή αναπαριστά ένα στοιχείο τότε το κέντρο βάρους $w \in \mathbb{R}^{1 \times n}$ ανά γραμμή είναι:

$$w = \frac{1}{m} \sum_{i=1}^m \alpha_i^\top = \frac{e^\top A}{m},$$

όπου το e αντιστοιχεί στο μοναδιαίο διάνυσμα κατάλληλης διάστασης. Το μητρώο συνδιασποράς C των δεδομένων ορίζεται ως εξής:

$$C = (A - ew)^\top (A - ew) = B^\top B.$$

Έστω $d_1, d_2, \dots, d_n$ ένα σύνολο των ορθογώνιων ιδιοδιανυσμάτων του C . Αυτά τα διανύσματα ονομάζονται και πρωτεύουσες συνιστώσες (*principal components*) ή πρωτεύουσες κατευθύνσεις (*principal directions*) του A , και έχουν την ιδιότητα να μεγιστοποιούν την διασπορά με την διασπορά των προβολών των δεδομένων σε αυτά. Το μητρώο C είναι συμμετρικό θετικά ημι-ορισμένο, έτσι έχει πραγματικές μη αρνητικές ιδιοτιμές και τα ιδιο-διανύσματα που αντιστοιχούν σε διαφορετικές ιδιοτιμές είναι ορθογώνια. Ο υπολογισμός των ιδιο-διανυσμάτων μπορεί να γίνει χρησιμοποιώντας την τεχνική Διάσπασης σε Ιδιάζουσες Τιμές (Singular Value Decomposition (SVD)) του μητρώου B , όπου $B = U\Sigma V^\top$, αφού:

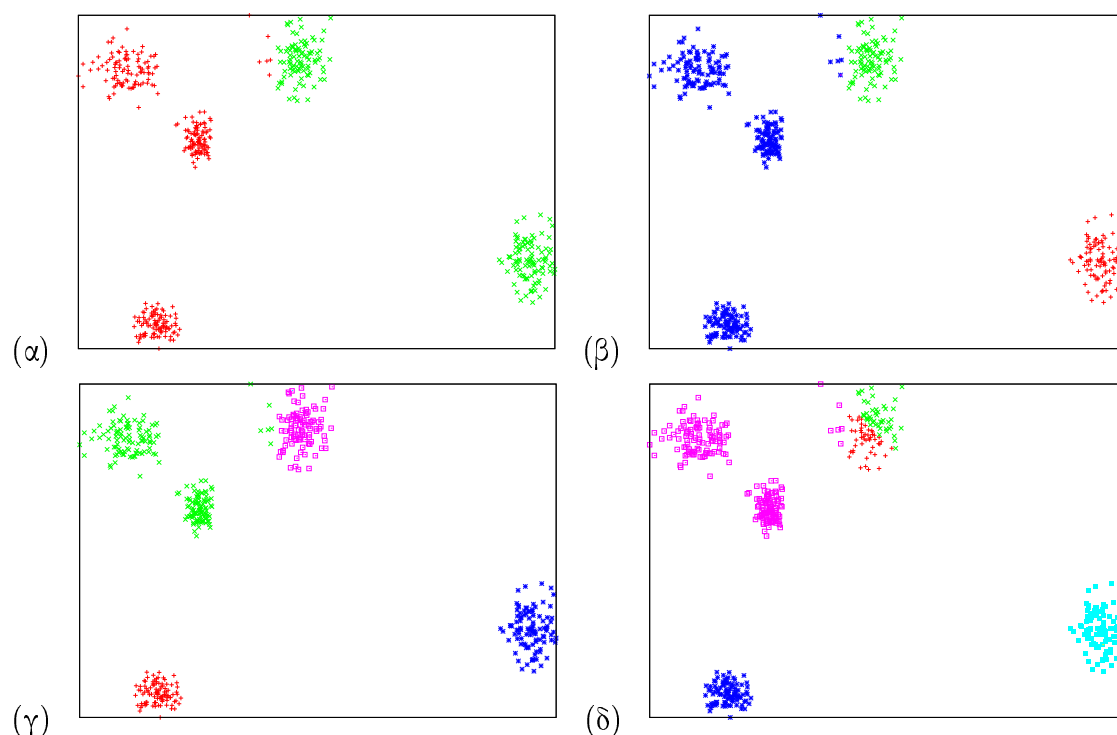
$$C = V\Sigma^\top U^\top U\Sigma V^\top = V\Sigma^2 V^\top.$$

Ο αλγόριθμος χρησιμοποιεί τη προβολή του κάθε στοιχείου στο ιδιο-διάνυσμα u που αντιστοιχεί στη μέγιστη ιδιοτιμή του μητρώου συνδιασποράς C . Η προβολή αυτή u_i του στοιχείου x_i δίνεται από τον τύπο:

$$u_i = u^\top (x_i - w), \quad (2.2)$$

Με αυτό το τρόπο όλα τα στοιχεία χωρίζονται σε δύο ομάδες ανάλογα με το πρόσημο της προβολής τους. Η μία ομάδα περιέχει αυτά με θετική προβολή ενώ η άλλη με αρνητική. Έπειτα για κάθε μια από τις δύο ομάδες υπολογίζεται η τιμή της Frobenius στάθμης του μητρώου των στοιχείων της. Η ομάδα με την μεγαλύτερη τιμή αυτής της στάθμης επιλέγεται για περαιτέρω διαχωρισμό και η διαδικασία επαναλαμβάνεται. Ο αλγόριθμος τερματίζει όταν ικανοποιηθεί κάποιο κριτήριο τερματισμού (συνήθως το πλήθος των τελικών ομάδων). Ο αλγόριθμος PDDP καθώς και οι γενικεύσεις του [156], δεν παρέχουν ένα τρόπο καθορισμού του πλήθους των ομάδων. Διάφορες τεχνικές που έχουν προταθεί για το σκοπό αυτό συνήθως υπερεκτιμούν το πραγματικό πλήθος.

Ένα παράδειγμα εκτέλεσης του αλγορίθμου σε ένα σύνολο δεδομένων αποτελούμενο από σημεία 2 διαστάσεων οργανωμένα σε 5 ομάδες φαίνεται στο Σχήμα 2.6. Ο αλγόριθμος καταφέρνει να διασπάσει σωστά το σύνολο μέχρι και το διαχωρισμό σε 4 ομάδες (Σχήματα 2.6(α),(β),(γ)). Ωστόσο από εκεί και έπειτα επιλέγει για διαχωρισμό τη λάθος ομάδα.



Σχήμα 2.6: Διαδοχικές διαμερίσεις του αλγορίθμου PDDP, για 2,3,4 και 5 τελικές ομάδες του ίδιου συνόλου δεδομένων.

2.5.2 Διαχωριστικοί αλγόριθμοι

Οι μέθοδοι που ανήκουν στην κατηγορία διαχωριστικών αλγορίθμων δημιουργούν μια μοναδική ομαδοποίηση των δεδομένων αντί να κατασκευάζουν μια δομή ομαδοποίησης όπως το δένδρογραμμα που κατασκευάζουν οι ιεραρχικοί αλγόριθμοι. Το κύριο τους πλεονέκτημα είναι η δυνατότητα χρήσης τους σε πολύ μεγάλα σύνολα δεδομένων. Συνήθως αυτοί οι αλγόριθμοι επιχειρούν να βελτιστοποιήσουν μία αντικειμενική συνάρτηση που ορίζεται είτε τοπικά (σε ένα υποσύνολο των δεδομένων) είτε ολικά

(σε όλα τα δεδομένα). Σε αυτή την κατηγορία ανήκει και ο k -means που είναι ο πιο διαδεδομένος αλγόριθμος ομαδοποίησης.

Ο αλγόριθμος k -means

Ο αλγόριθμος k -means [67], είναι ένας από τους απλούστερους αλγορίθμους ομαδοποίησης αλλά και ο πιο διαδεδομένος. Η βασική ιδέα του αλγορίθμου είναι η αναπαράσταση της κάθε ομάδας από ένα στοιχείο το οποίο αποκαλείται και κέντρο της. Με αυτό τον τρόπο για τον διαχωρισμό ενός συνόλου σε k ομάδες, αρχικοποιούνται με τυχαίο τρόπο k κέντρα $P_j, j = 1, \dots, k$, το κάθε ένα από τα οποία αντιπροσωπεύει και μία ομάδα. Στο επόμενο βήμα ο αλγόριθμος αναθέτει κάθε στοιχείο του συνόλου δεδομένων στην ομάδα της οποίας το κέντρο απέχει τη μικρότερη απόσταση από αυτό. Η ανάθεση αυτή γίνεται χρησιμοποιώντας την ακόλουθη συνάρτηση συμμετοχής για κάθε στοιχείο X_i :

$$\mu_j(X_i) = \begin{cases} 1 & \text{εάν } \|X_i - P_j\| \leq \|X_i - P_l\| \quad \forall l \neq j \\ 0 & \text{διαφορετικά} \end{cases}, \quad (2.3)$$

Συνήθως για τον υπολογισμό της απόστασης $\|X_i - P_l\|$, χρησιμοποιείται η Ευκλείδεια απόσταση. Αφού όλα τα στοιχεία έχουν ανατεθεί σε ομάδες ο αλγόριθμος υπολογίζει νέες θέσεις για τα κέντρα των ομάδων, από το κέντρο βάρους των στοιχείων που ανήκουν σε κάθε μία:

$$P^j = \frac{\sum_{i=1}^n \mu_j(X_i) X_i}{\sum_{i=1}^n \mu_j(X_i)}. \quad (2.4)$$

Τα τελευταία δύο βήματα εκτελούνται επαναληπτικά μέχρι να επιτευχθεί σύγκλιση, ενός μέτρου διαστρέβλωσης D που ορίζεται ως εξής:

$$D = \sum_{j=1}^k \sum_{i=1}^n \mu_j(X_i) \|X_i - P_j\|^2. \quad (2.5)$$

Όταν επιτευχθεί σύγκλιση ο αλγόριθμος αναφέρει τις αναθέσεις των σημείων σε ομάδες καθώς και τις θέσεις των τελικών κέντρων. Συνοπτικά ο αλγόριθμος μπορεί να περιγραφεί ως εξής:

1. Αρχικοποίηση k κέντρων στο σύνολο των δεδομένων
2. Ανάθεση του κάθε σημείου στη ομάδα με το κοντινότερο κέντρο
3. Υπολογισμός των νέων κέντρων των ομάδων
4. Επανάληψη των βημάτων 2,3, μέχρι να επιτευχθεί σύγκλιση.

Ουσιαστικά ο αλγόριθμος είναι μια διαδικασία βελτιστοποίησης της αντικειμενικής συνάρτησης:

$$J = \sum_{j=1}^k \sum_{i=1}^n \|X_i - P_j\|. \quad (2.6)$$

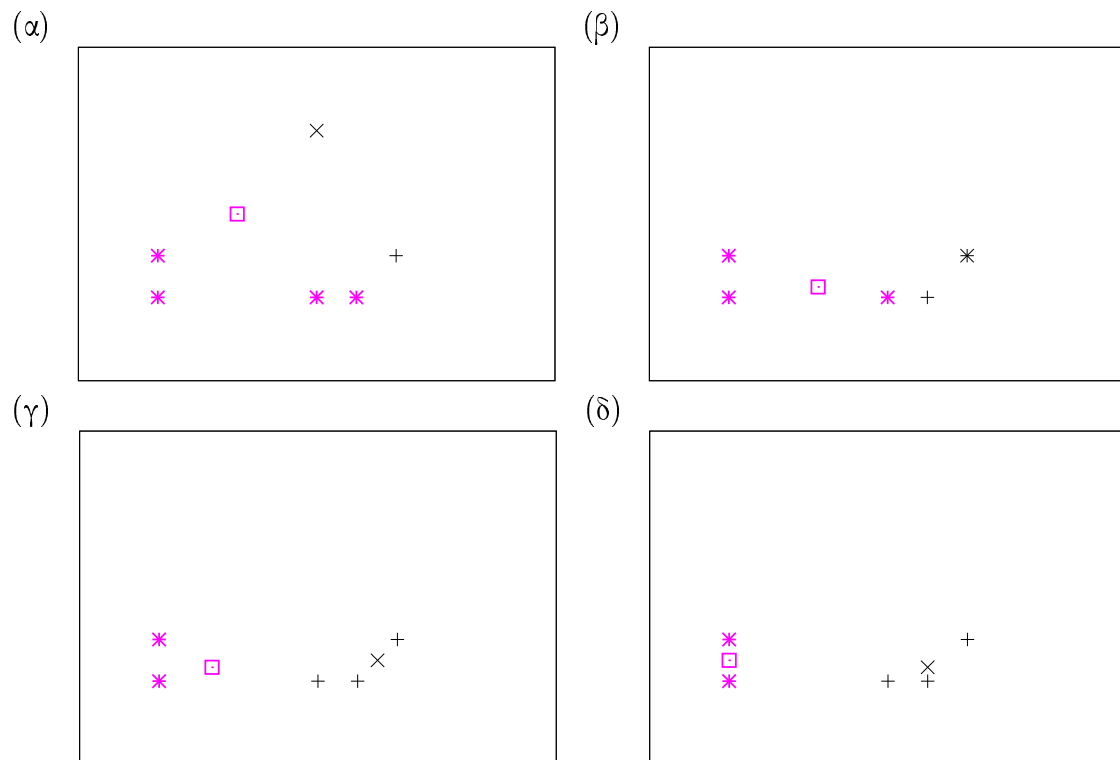
Αν και μπορεί να αποδειχθεί ότι η διαδικασία θα συγκλίνει πάντα, ο αλγόριθμος δεν βρίσκει απαραίτητως τη βέλτιστη διαμέριση που αντιστοιχεί στο ολικό ελάχιστο της αντικειμενικής συνάρτησης. Επίσης είναι ιδιαίτερα ευαίσθητος στα αρχικά τυχαία επιλεγμένα κέντρα των ομάδων καθώς και το πλήθος τους. Η επαναλαμβανόμενη

εκτέλεση του αλγορίθμου και η επιλογή του καλύτερου αποτελέσματος είναι συνήθως η τακτική που επιλέγουν οι ερευνητές για την αντιμετώπιση αυτού του προβλήματος.

Ένα παράδειγμα εκτέλεσης του αλγορίθμου σύνολο δεδομένων X :

$$X = \begin{Bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 1 & 1 \\ 1 & 1.5 \\ 3.5 & 1 \\ 4 & 1.5 \\ 3 & 1 \end{Bmatrix},$$

φαίνεται στο Σχήμα 2.7. Σε αυτήν την εκτέλεση δύο τυχαία αρχικοποιημένα κέντρα, που σημειώνονται με το τετράγωνο και το σύμβολο « $\times$ » αντίστοιχα, εξελίσσονται σύμφωνα με τον αλγόριθμο. Σε κάθε βήμα με το σύμβολο « $*$ » απεικονίζονται τα στοιχεία που αντιστοιχούν στο πρώτο κέντρο και με το « $+$ » αυτά που αντιστοιχούν στο δεύτερο. Όπως φαίνεται, στο Σχήμα 2.7(α), μετά την αρχικοποίηση των κέντρων στο δεύτερο κέντρο αντιστοιχεί μόνο ένα στοιχείο από τα δεδομένα. Αυτό έχει ως αποτέλεσμα την νέα θέση των κέντρων που απεικονίζεται στο Σχήμα 2.7(β). Από αυτό το σημείο και έπειτα, το κέντρο που πριν είχε μόνο ένα στοιχείο αποκτά ένα ακόμα, και ο αλγόριθμος καταλήγει στη φάση του Σχήματος 2.7(γ). Από αυτό το βήμα και έπειτα η αντιστοίχιση των σημείων είναι και η τελική, οπότε οι τελικές θέσεις των κέντρων απεικονίζονται στο Σχήμα 2.7(δ).



Σχήμα 2.7: Τέσσερις διαδοχικές επαναλήψεις του αλγορίθμου k -means.

Ο αλγόριθμος fuzzy c -means

Ο αλγόριθμος fuzzy c -means (FCM), παρόλο που λειτουργικά είναι όμοιος με τον k -means, στηρίζεται στην Ασαφή Λογική. Πιο συγκεκριμένα θεωρεί κάθε ομάδα σαν

ένα ασαφές σύνολο. Όμοια με τον k -means, αναπαριστά κάθε ομάδα με ένα στοιχείο το οποίο αποκαλείται και *κέντρο* της. Αρχικά λοιπόν για τον υπολογισμό μια ομαδοποίησης σε c ομάδες, επιλέγει τυχαία c κέντρα, ένα για κάθε ομάδα. Έπειτα για κάθε ένα στοιχείο υπολογίζει τον βαθμό συμμετοχής του σε όλες τις ομάδες σύμφωνα με τη συνάρτηση:

$$\mu_j(X_i) = \left(\sum_{l=1}^c \left(\frac{\|X_i - P_j\|}{\|X_i - P_l\|} \right)^{1/r-1} \right)^{-1}. \quad (2.7)$$

Η συνάρτηση αυτή παίρνει τιμές στο διάστημα $[0, 1]$, όπου το $r \in (1, \infty)$, καθορίζει την «ασάφεια» της διαμέρισης. Όταν το r τείνει στο 1_+ , τότε η διαμέριση ασυμπτωτικά προσεγγίζει μια αυστηρή διαμέριση. Αντίθετα όσο το r τείνει στο άπειρο η διαμέριση γίνεται μέγιστα ασαφής. Ο υπολογισμός, των νέων κέντρων γίνεται σύμφωνα με τον τύπο:

$$P_j = \frac{\sum_{i=1}^n [m_j(X_i)]^r X_i}{\sum_{i=1}^n [m_j(X_i)]^r}. \quad (2.8)$$

Το μέτρο διαστρέβλωσης που χρησιμοποιείται σαν κριτήριο τερματισμού γίνεται:

$$D = \sum_{j=1}^c \sum_{i=1}^n [m_j(X_i)]^r \|X_i - P_j\|^2. \quad (2.9)$$

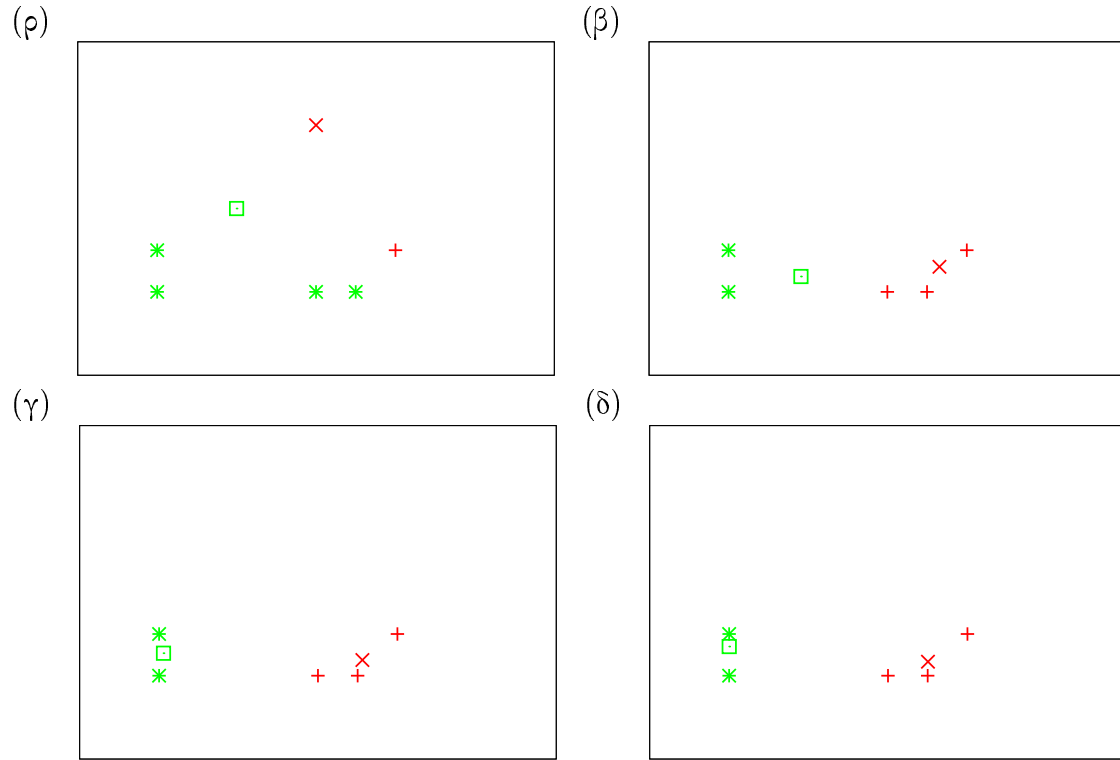
Ένα παράδειγμα εκτέλεσης του αλγορίθμου στο ακόλουθο σύνολο δεδομένων X , φαίνεται στο Σχήμα 2.8.

$$X = \begin{Bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 1 & 1 \\ 1 & 1.5 \\ 3.5 & 1 \\ 4 & 1.5 \\ 3 & 1 \end{Bmatrix}.$$

Σε αυτήν την εκτέλεση, όμοια με την περίπτωση του k -means, δύο τυχαία αρχικοποιημένα κέντρα στις ίδιες αρχικές θέσεις απεικονίζονται με το τετράγωνο και το σύμβολο «x», αντίστοιχα στο Σχήμα 2.8(α). Τα διαφορετικά σχήματα των σημείων απεικονίζουν την αντιστοίχιση στο κέντρο με την μεγαλύτερη τιμή της συνάρτησης συμμετοχής. Όπως φαίνεται, στο Σχήμα 2.8(β), μετά την αρχικοποίηση των κέντρων παρόλο που στο ένα από αυτά αντιστοιχεί μόνο ένα στοιχείο από τα δεδομένα, η θέση του καθορίζεται και από το βαθμό συμμετοχής των άλλων οπότε και έχουμε τις θέσεις του σχήματος 2.8(β). Όπως φαίνεται ακόμα και από αυτό το πρώτο βήμα ο αλγόριθμος έχει καταλήξει στην τελική αντιστοίχιση των σημείων. Βέβαια τα κέντρα συνεχίζουν να εξελίσσουν τις θέσεις τους (Σχήματα 2.8(β) και 2.8(γ)), καθώς αυτές δεν έχουν συγχλίνει.

Ο Αλγόριθμος Mean Shift

Ο αλγόριθμος Mean Shift, είναι μια τεχνική ομαδοποίησης που δεν προϋποθέτει τον προ-καθορισμό του πλήθους των ομάδων και είναι ικανή να αναγνωρίσει ομάδες με διαφορετικά σχήματα. Έχει χρησιμοποιηθεί κυρίως στην ανάλυση εικόνων [58]. Είναι μια επαναληπτική τεχνική η οποία στοχεύει στην εντοπισμό των μεγίστων μια συνάρτησης πυκνότητας ορισμένης στο σύνολο δεδομένων. Οι περιοχές έλξης των μεγίστων καθορίζουν και το σχήμα της ομάδας.



Σχήμα 2.8: Τέσσερις διαδοχικές επαναλήψεις του αλγορίθμου fuzzy c -means.

Πιο συγκεκριμένα το σύνολο δεδομένων $\mathcal{X} = \{X_i, i = 1, \dots, n\}$, θεωρείται ότι αποτελείται από μετρήσεις μια κατανομής πιθανότητας $f(X)$. Έπειτα κατασκευάζεται ένας μη-παραμετρικός εκτιμητής $\hat{f}$ της f , χρησιμοποιώντας μια συνάρτηση πυρήνα $K(X)$ (kernel). Ο εκτιμητής αυτός για ένα προκαθορισμένο εύρος h (bandwidth), με $\mathcal{X} \subset \mathbb{R}^d$, ορίζεται ως εξής:

$$\hat{f}(X) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{X - X_i}{h}\right). \quad (2.10)$$

Ο πυρήνας $K(X)$, είναι μια συνάρτηση που πρέπει να ικανοποιεί τις παρακάτω συνθήκες:

$$\begin{aligned} \sup_{X \in \mathbb{R}^d} |K(X)| &< \infty, \\ \lim_{\|x\| \rightarrow \infty} \|X\| K(X) &= 0, \\ \int_{\mathbb{R}^d} |K(X)| dX &< \infty, \\ \int_{\mathbb{R}^d} K(X) dX &= 1, \end{aligned} \quad (2.11)$$

όπου $\|\cdot\|$, είναι η L_2 στάθμη (Ευκλείδεια απόσταση).

Ακόμη ορίζεται η συνάρτηση $k(x)$, $0 \leq x \leq 1$, (προφίλ πυρήνα) για την οποία να ισχύει:

$$K(X) = c_{k,d} k(\|X\|^2) > 0, \quad \|X\| \leq 1, \quad (2.12)$$

με τη σταθερά $c_{k,d}$, να βεβαιώνει ότι ο πυρήνας $K(X)$, ολοκληρώνεται στη μονάδα. Χρησιμοποιώντας παραγωγίσιμους πυρήνες μπορεί πάντα να κατασκευαστεί η συνάρτηση $g(x) = -k'(x)$, όποτε να οριστεί και ο πυρήνας $G(X) = c_{g,d} g(\|X\|^2)$. Χρησιμο-

ποιώντας την παράγωγο της σχέσης (2.10), μπορεί να προκύψει η ακόλουθη σχέση [29]:

$$m_G(X) = \frac{\sum_{i=1}^n X_i g\left(\left\|\frac{X-X_i}{h}\right\|\right)}{\sum_{i=1}^n g\left(\left\|\frac{X-X_i}{h}\right\|\right)} - X. \quad (2.13)$$

Το διάνυσμα $m_G(X)$, ονομάζεται διάνυσμα Mean Shift. Με τη χρήση αυτού ο αλγόριθμος ανανεώνει επαναληπτικά, ξεκινώντας από κάθε σημείο του συνόλου X_i , τις θέσεις τους σύμφωνα με την σχέση:

$$Y^{j+1} = Y^j + m_G(Y^j).$$

Οι τελικές θέσεις των σημείων είναι και τα ακρότατα του εκτιμητή της συνάρτησης πυκνότητας, που καθορίζουν και τις τελικές ομάδες. Επίσης έχουν προταθεί και διάφορες τεχνικές για τη χρήση μεταβλητού εύρους σε κάθε σημείο ώστε να γίνεται ομαλότερα η σύγκλιση [30].

2.5.3 Density ή Mode-Seeking Τεχνικές

Οι αλγόριθμοι ομαδοποίησης που βασίζονται στην πυκνότητα (Density-Based) αναγνωρίζουν μια ομάδα βασιζόμενοι στην υπόθεση πως η πυκνότητα (ή το πλήθος των σημείων) σε μια γειτονιά ξεπερνά κάποιο κατώφλι, σε αντίθεση με τους Distance-Based αλγόριθμους ομαδοποίησης που αναθέτουν ένα στοιχείο σε μια ομάδα με βάση την απόσταση του από το κέντρο της ομάδας. Οι Distance-Based αλγόριθμοι είναι ιδιαίτερα ικανοί στον εντοπισμό ομάδων σφαιρικού σχήματος, ενώ αντιμετωπίζουν δυσκολίες όταν οι ομάδες έχουν τυχαίο σχήμα. Αντίθετα, οι Density-Based αλγόριθμοι δεν υπόκεινται σε αυτό τον περιορισμό. Ένας από τους πιο γνωστούς αλγόριθμους της οικογένειας αυτής είναι ο DBSCAN [123].

Αλγόριθμος DBSCAN

Ο αλγόριθμος DBSCAN (Density Based Spatial Clustering of Application with Noise) στηρίζεται στην υπόθεση πως η πυκνότητα των σημείων στις περιοχές του χώρου στις οποίες βρίσκεται μια ομάδα είναι σημαντικά υψηλότερη σε σχέση με την πυκνότητα των σημείων σε χώρους που δεν αντιστοιχούν σε ομάδα. Ο αλγόριθμος αυτός σχηματίζει ομάδες βασιζόμενος σε μια έννοια γειτονικότητας των σημείων. Για κάθε σημείο μιας ομάδας η γειτονιά του, σχηματίζεται με βάση το επιλεγμένο μέτρο απόστασης γύρω από το σημείο. Για να θεωρηθεί ότι το συγκεκριμένο σημείο δεν είναι θόρυβος, πρέπει η πυκνότητα της γειτονιάς του να υπερβαίνει ένα ορισμένο κατώφλι. Τόσο το σχήμα όσο και μέγεθος της γειτονιάς είναι παράμετροι που καθορίζονται από τον χρήστη. Το σχήμα της γειτονιάς καθορίζεται από την επιλογή της συνάρτησης απόστασης (στάθμης) η οποία συμβολίζεται με $dist(.,.)$. Για παράδειγμα, όταν χρησιμοποιείται η Manhattan απόσταση στον δισδιάστατο χώρο το σχήμα της γειτονιάς είναι ρόμβος ενώ αν χρησιμοποιηθεί η ℓ_∞ στάθμη τετράγωνο. Για την περιγραφή της λειτουργίας του αλγόριθμου απαιτούνται μερικές θεμελιώδεις έννοιες.

Ορισμός 2.1. (*Eps-γειτονιά ενός σημείου*). Η *Eps-γειτονιά ενός σημείου* p , συμβολίζεται με $N_{Eps}(p)$, ορίζεται ως $N_{Eps}(p) = \{q \in D | dist(p, q) \leq Eps\}$.

Αρχικά όπως ήδη ειπώθηκε απαιτείται από κάθε σημείο μιας ομάδας να γειτονεύει με ένα ελάχιστο πλήθος (*MinPts*) σημείων σε μια *Eps-γειτονιά*. Έπειτα τα σημεία

αυτά διαχωρίζονται σε σημεία μέσα στον πυρήνα της ομάδας (core points) και σημεία στο σύνορο της ομάδας (border points). Γενικά, ένα συνοριακό σημείο βρίσκεται στην ίδια γειτονιά με λιγότερα σημεία σε αντίθεση με ένα σημείο που βρίσκεται στον πυρήνα της ομάδας. Συνεπώς, απαιτείται για κάθε σημείο p σε μια ομάδα C να υπάρχει ένα σημείο q στην C έτσι ώστε το p να είναι εσωτερικό σημείο της Eps -γειτονιά του q και $N_{Eps}(q)$ να περιέχει τουλάχιστον $MinPts$ σημεία.

Ορισμός 2.2. (*directly density-reachable*). Ένα σημείο p είναι *directly density-reachable* από ένα σημείο q σε σχέση με τις παραμέτρους Eps και $MinPts$, αν

1. $p \in N_{Eps}(q)$ και
2. $|N_{Eps}(q)| \geq MinPts$ (συνθήκη εσωτερικού σημείου).

Είναι φανερό πως η ιδιότητα *directly density-reachable* είναι συμμετρική για ζεύγη σημείων κοντά στον πυρήνα της ομάδας. Δεν ισχύει το ίδιο όμως αν ένα από τα δύο σημεία είναι συνοριακό. Επέκταση της *directly density-reachable* ιδιότητας είναι η παρακάτω.

Ορισμός 2.3. (*density-reachable*). Ένα σημείο p είναι *density-reachable* από ένα σημείο q σε σχέση με τις παραμέτρους Eps και $MinPts$, αν υπάρχει μια αλυσίδα σημείων $p_1, p_2, \dots, p_n$, όπου $p_1 = q, p_n = p$, έτσι ώστε p_{i+1} να είναι *directly density-reachable* από το p_i .

Δύο συνοριακά σημεία της ίδιας ομάδας C είναι πιθανόν να μην είναι *density-reachable* το ένα από το άλλο, διότι μπορεί να μην ισχύει η συνθήκη εσωτερικού σημείου για αυτά. Ωστόσο, πρέπει να υπάρχει ένα εσωτερικό σημείο της C από το οποίο και τα δύο συνοριακά σημεία της C να είναι *density-reachable*.

Ορισμός 2.4. (*density-connected*). Ένα σημείο p είναι *density-connected* με ένα σημείο q σε σχέση με τις παραμέτρους Eps και $MinPts$, αν υπάρχει ένα σημείο o τέτοιο ώστε μαζί τα p και q να είναι *density-reachable* από το σημείο o σε σχέση με τις παραμέτρους Eps και $MinPts$.

Η *density-connectivity* είναι μια συμμετρική σχέση. Για *density-reachable* σημεία η σχέση αυτή είναι επίσης ανακλαστική (reflexive). Διαισθητικά, μια ομάδα ορίζεται ως το σύνολο των *density-connected* σημείων, το οποίο είναι μέγιστο σε σχέση με την *density-reachability* ιδιότητα.

Ορισμός 2.5. (Ομάδα). Έστω D είναι η βάση δεδομένων των σημείων. Μια ομάδα C σε σχέση με τις παραμέτρους Eps και $MinPts$, είναι ένα μη κενό υποσύνολο του D που ικανοποιεί τις ακόλουθες συνθήκες:

1. $\forall p, q$ τέτοια ώστε αν $p \in C$ και q είναι *density-reachable* από το p σε σχέση με τις παραμέτρους Eps και $MinPts$, τότε $q \in C$ (*Maximality*) και
2. $\forall p, q \in C$, το σημείο p είναι *density-connected* με το σημείο q σε σχέση με τις παραμέτρους Eps και $MinPts$ (*Connectivity*).

Ο θόρυβος ορίζεται σχετικά με ένα δοσμένο σύνολο ομάδων που έχουν εντοπιστεί. Ο θόρυβος σε αυτό το πλαίσιο δεν είναι παρά το σύνολο των σημείων του D που δεν ανήκουν σε κάποια από τις ομάδες.

Ορισμός 2.6. (Θόρυβος). Έστω $C_1, \dots, C_k$ ομάδες μιας βάσης δεδομένων που έχουν προκύψει από τις παραμέτρους Eps και $MinPts$, $i = 1, \dots, k$. Τότε ορίζεται ο θόρυβος ως το σύνολο των σημείων στη βάση δεδομένων D που δεν ανήκουν σε κάποια ομάδα C_i , δηλαδή, $\text{θόρυβος} = \{p \in D \mid \forall i : p \notin C_i\}$.

Δεδομένων των παραμέτρων Eps και $MinPts$, είναι εφικτός ο εντοπισμός μιας ομάδας με τη χρήση μιας προσέγγισης δύο βημάτων. Πρώτα, επιλέγεται ένα τυχαίο σημείο από τη βάση δεδομένων, που ικανοποιεί την συνθήκη εσωτερικού σημείου, ως ένα κέντρο (σπόρο). Δεύτερον, ανακτώνται όλα τα σημεία που είναι density-reachable από το κεντρικό σημείο. Με αυτόν τον τρόπο είναι δυνατόν να αποδειχθεί το ακόλουθο λήμμα για αυτόν τον αλγόριθμο [123]:

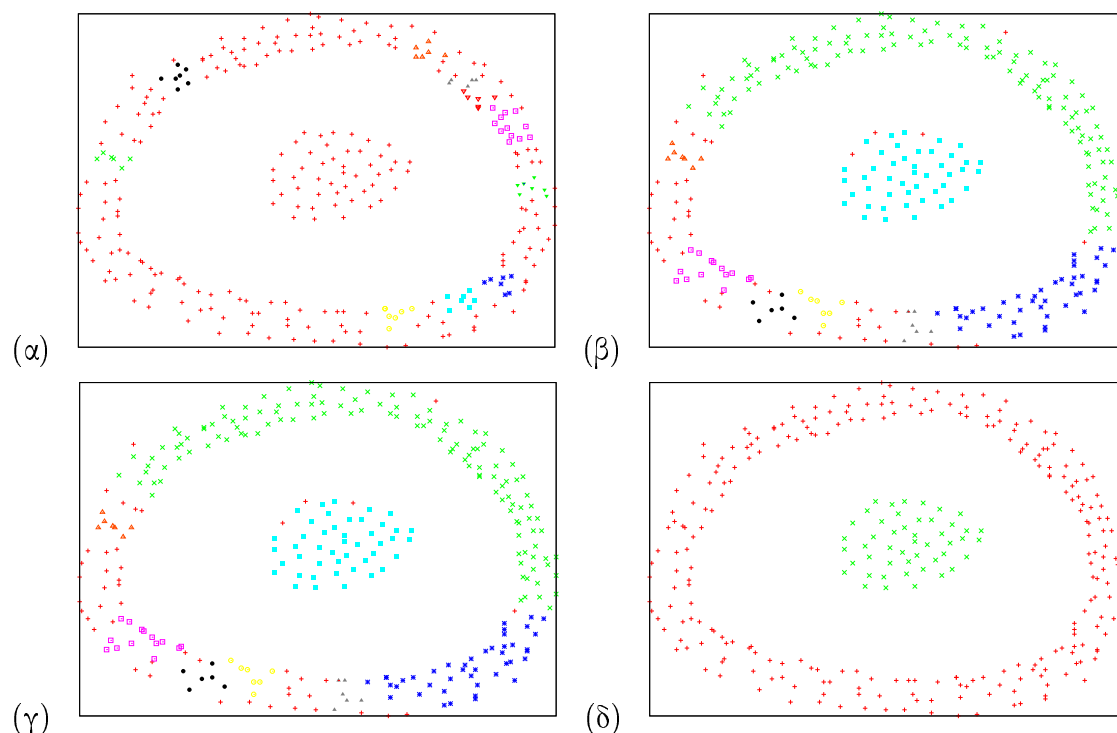
Λήμμα 2.1. Έστω p είναι ένα σημείο της D και $|N_{Eps}(p)| \geq MinPts$. Τότε το σύνολο $O = \{o \mid o \in D \text{ όπου } o \text{ είναι density-reachable από το } p \text{ σε σχέση με τις παραμέτρους } Eps \text{ και } MinPts\}$ είναι μια ομάδα σε σχέση με τις παραμέτρους Eps και $MinPts$.

Είναι φανερό πως μια ομάδα είναι μοναδικά ορισμένη μέσω οποιουδήποτε από τα εσωτερικά της σημεία. Ωστόσο, κάθε σημείο της C είναι density-reachable από οποιοδήποτε εσωτερικό σημείο της και άρα η ομάδα C περιέχει ακριβώς τα σημεία τα οποία είναι density-reachable από κάθε εσωτερικό σημείο της C .

Λήμμα 2.2. Έστω C είναι μια ομάδα σε σχέση με τις παραμέτρους Eps και $MinPts$, και έστω p είναι ένα οποιοδήποτε σημείο της C με $|N_{Eps}(p)| \geq MinPts$. Τότε η ομάδα C ισοδυναμεί με το σύνολο $O = \{o \mid o \text{ είναι density-reachable από το } p \text{ σε σχέση με τις παραμέτρους } Eps \text{ και } MinPts\}$.

Η επιτυχής λειτουργία του αλγορίθμου DBSCAN απαιτεί τη γνώση των κατάλληλων τιμών των παραμέτρων Eps και $MinPts$ για κάθε ομάδα και τουλάχιστον ένα σημείο από κάθε ομάδα. Υπό αυτές τις προϋποθέσεις είναι εφικτή η ανάκτηση όλων των σημείων που είναι density-reachable από κάθε δοσμένο σημείο και επομένως είναι εφικτός ο εντοπισμός όλων των ομάδων. Δυστυχώς η ιδανική αυτή κατάσταση δύσκολα επιτυγχάνεται στην πράξη, αφού γενικά δεν υπάρχει τρόπος για τη σωστή εκτίμηση των παραμέτρων του αλγορίθμου από μια τυχαία βάση δεδομένων.

Τελειώνοντας την ενότητα αυτή, θα πρέπει να επισημανθεί πως ο αλγόριθμος GDBSCAN είναι μια γενίκευση του αλγορίθμου DBSCAN, προκειμένου να χρησιμοποιηθεί στην ομαδοποίηση χωρικών βάσεων δεδομένων (spatial databases) [123]. Οι χωρικές βάσεις περιέχουν χωρικά δεδομένα, δηλαδή αντικείμενα στον δισδιάστατο, τρισδιάστατο ή μεγαλύτερης διάστασης χώρο. Ένα παράδειγμα τέτοιων αντικειμένων είναι τα πολύγωνα. Έτσι λοιπόν, μπορεί να χρησιμοποιηθεί οποιαδήποτε έννοια της γειτονιάς ενός αντικειμένου αν ο ορισμός της γειτονιάς βασίζεται σε ένα δυαδικό κατηγορημα (binary predicate), το οποίο είναι συμμετρικό και ανακλαστικό (reflexive). Για παράδειγμα, στην ομαδοποίηση πολυγώνων η γειτονιά μπορεί να οριστεί από το κατηγορημα της μη-κενής τομής των πολυγώνων. Επιπλέον, αντί για την απλή μέτρηση των αντικειμένων σε μία γειτονιά ενός αντικειμένου, μπορεί να χρησιμοποιηθούν άλλα μέτρα, για παράδειγμα η μελέτη των μη χωρικών χαρακτηριστικών όπως είναι ο μέσος όρος εισοδήματος μιας πόλης, ώστε να καθοριστεί ο πληθάρημος (cardinality) εκείνης της γειτονιάς. Με τον τρόπο αυτό επιτυγχάνεται η γενίκευση του αλγορίθμου DBSCAN ώστε να πραγματοποιηθεί η ομαδοποίηση χωρικά εκτεταμένων αντικειμένων σύμφωνα με τα χωρικά και τα μη χωρικά τους χαρακτηριστικά.



Σχήμα 2.9: Το αποτέλεσμα του αλγορίθμου DBSCAN, για διαφορετικές τιμές τις παραμέτρου Eps , 5(α), 8(β), 10(γ) και 15(δ) αντίστοιχα.

Στη συνέχεια παραθέτουμε ένα παράδειγμα εφαρμογής του αλγορίθμου, σε ένα σύνολο δεδομένων αποτελούμενο από σημεία 2 διαστάσεων οργανωμένα σε δύο ομάδες ομόκεντρων κυκλικών δίσκων. Το σύνολο αυτό επιλέχθηκε γιατί αναδεικνύει τη δυνατότητά του αλγορίθμου να εντοπίζει ομάδες με μη κυρτό σχήμα. Στο Σχήμα 2.9, απεικονίζονται τα αποτελέσματα του αλγορίθμου, για διαφορετικές τιμές τις παραμέτρου Eps , 5(α), 8(β), 10(γ) και 15(δ) αντίστοιχα. Η παράμετρος $MinPts$, είχε τη σταθερή τιμή 5. Όπως φαίνεται ο αλγόριθμος είναι εξαιρετικά ευαίσθητος στην τιμή της παραμέτρου αυτής, και δεν έχει τη δυνατότητα να προσαρμόζει την τιμή της κατά την διάρκεια της εκτέλεσης. Στο Σχήμα 2.9(α), ο αλγόριθμος καταλήγει σε 10 διαφορετικές ομάδες και θεωρεί ένα μεγάλο πλήθος σημείων σαν θόρυβο. Αντίθετα στο Σχήμα 2.9(β), λιγότερα σημεία αναγνωρίζονται σαν θόρυβος αλλά το πλήθος των ομάδων αυξάνεται σε 15. Για την τιμή 8 της παραμέτρου Eps , οι ομάδες αρχίζουν και μειώνονται σε αριθμό και συγχωνεύονται φτάνοντας τις 8 (Σχήμα 2.9(γ)), και τέλος για την τιμή 15 ο αλγόριθμος αναγνωρίζει σωστά τις δύο ομάδες όπως φαίνεται στο Σχήμα 2.9(δ).

2.5.4 Ομαδοποίηση με Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα μπορούν να οριστούν σαν ένα μαθηματικό μοντέλο που έχει την ικανότητα να προσαρμόζεται, να γενικεύει και να ομαδοποιεί δεδομένα. Όλες αυτές οι λειτουργίες βασίζονται στη παράλληλη επεξεργασία των δεδομένων και μπορούν να είναι αρκετά γρήγορες και αποδοτικές σε σύγκριση με άλλες τεχνικές. Ανταγωνιστικά Νευρωνικά Δίκτυα έχουν χρησιμοποιηθεί πολλές φορές για την ομαδοποίηση δεδομένων [75]. Η πιο αντιπροσωπευτική μέθοδος από αυτήν την κατηγορία

είναι οι *Χάρτες Αυτο-οργάνωσης* (Self-Organizing Maps, (SOM)) [84].

Χάρτες Αυτο-οργάνωσης (SOM)

Οι Χάρτες Αυτο-οργάνωσης, χρησιμοποιούν ένα πλέγμα από κόμβους–νευρώνες για να μοντελοποιήσουν το σύνολο δεδομένων και να προσφέρουν μια οπτική απεικόνιση της δομής τους. Σε κάθε κόμβο i , αντιστοιχείται και ένα διάνυσμα $m = [m_{i,1}, m_{i,2}, \dots, m_{i,n}]$, το οποίο αρχικά είναι τυχαία αρχικοποιημένο. Το διάνυσμα αυτό ονομάζεται και βάρος και αναπαριστά την υποτιθέμενη θέση του νευρώνα στο χώρο των δεδομένων. Ο αλγόριθμος, για κάθε στοιχείο X του συνόλου δεδομένων $\mathcal{X}$, μετράει την απόσταση του από όλους τους κόμβους (σύμφωνα με το διάνυσμα βαρών) και υπολογίζει τον κοντινότερο κόμβο m_c , ο οποίος και ονομάζεται κόμβος νικητής:

$$\|X - m_c\| = \min_i \{\|X - m_i\|\}. \quad (2.14)$$

Στη συνέχεια προσαρμόζει, το διάνυσμα του κόμβου νικητή, καθώς και των κόμβων που είναι είναι τοπογραφικά κοντά του στο πλέγμα. Με αυτό τον τρόπο, καταφέρνει να προσαρμόσει όλο το πλέγμα στη κατανομή των δεδομένων. Αυτή η διαδικασία (μάθηση) επαναλαμβάνεται για κάθε στοιχείο του συνόλου δεδομένων μέχρι την ικανοποίηση κάποιου κριτηρίου τερματισμού.

Πιο συγκεκριμένα, σε κάθε επανάληψη t (χρόνος), κατά τη διάρκεια της μάθησης, αυτοί οι κόμβοι που είναι τοπογραφικά κοντά στον πίνακα σε μια ορισμένη γεωμετρική απόσταση θα ενεργοποιήσει ο ένας τον άλλο έτσι ώστε να μάθουν κάτι από την ίδια είσοδο X . Έστω ότι το σύνολο των γειτονικών κόμβων του κόμβου νικητή συμβολίζεται με N_c (μπορούμε να ορίσουμε το $N_c = N_c(t)$ σαν μια συνάρτηση του χρόνου). Τότε σε κάθε κόμβο τα διανύσματα βαρών m_i , για ένα στοιχείο X του συνόλου δεδομένων, μεταβάλλονται ως εξής:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x - m_i(t)],$$

όπου $h_{ci}(t) = a(t)$ αν $i \in N_c$ και $h_{ci}(t) = 0$ διαφορετικά. Η τιμή του $a(t)$ αντιπροσωπεύει ένα παράγοντα ρυθμού μάθησης ($0 < a(t) < 1$). Συνήθως ορίζουμε την γειτονιά N_c σαν μια συνάρτηση του χρόνου $N_c = N_c(t)$, και της θέσης των κόμβων στο χάρτη $N_c(t) = \{i : \|r_c - r_i\| < r_t\}$, όπου r_c και r_i είναι τα διανύσματα θέσης των κόμβων c και i αντίστοιχα, στον χάρτη. Το r_t αντιπροσωπεύει το μέγεθος της γειτονιάς (ακτίνα) στην επανάληψη t . Το $a(t)$ όπως και η ακτίνα του $N_c(t)$ συνήθως μειώνονται μονοτονικά στο χρόνο κατά τη διάρκεια της διαδικασίας μάθησης.

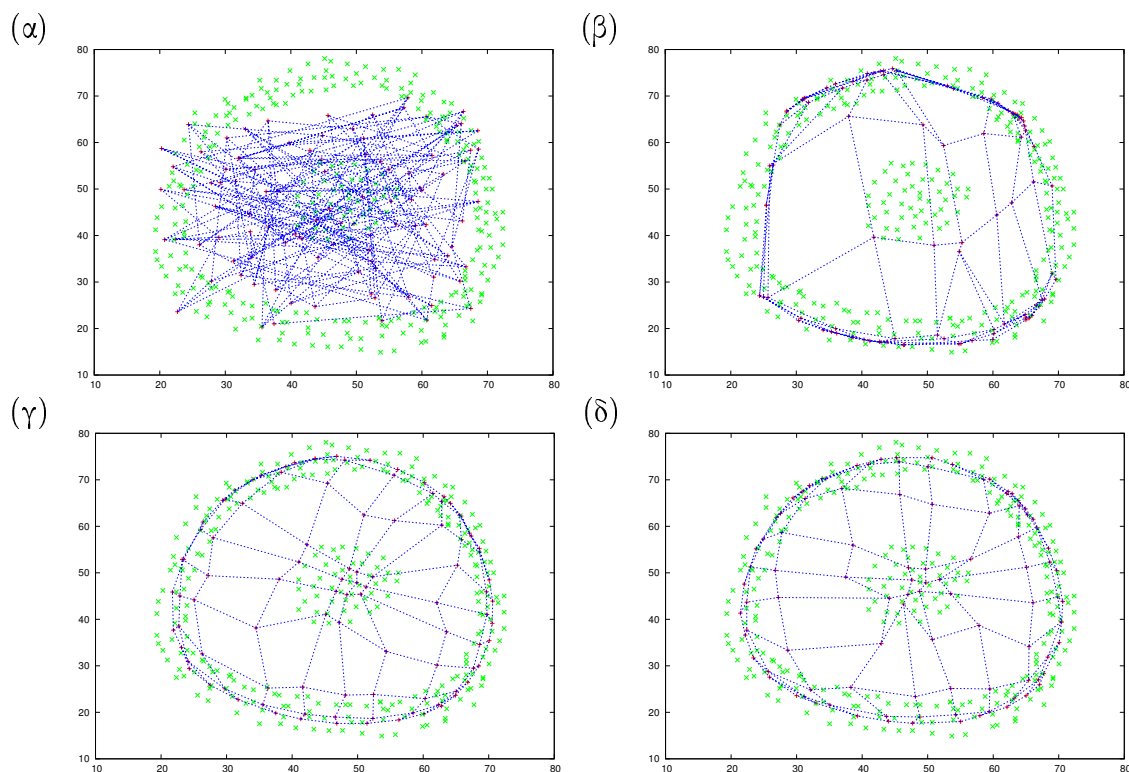
Αν το δίκτυο SOM δεν είναι πολύ μεγάλο, η επιλογή των παραμέτρων της διαδικασίας δεν είναι πολύ σημαντική. Μπορεί να χρησιμοποιηθεί ο απλός ορισμός του συνόλου γειτονιάς για το $h_{ci}(t)$. Όμως ιδιαίτερη προσοχή χρειάζεται στην επιλογή του μεγέθους του $N_c = N_c(t)$. Αν η γειτονιά με την οποία αρχίζει ο αλγόριθμος είναι πολύ μικρή οι κόμβοι του πίνακα δεν θα ταξινομηθούν συνολικά. Αυτό το φαινόμενο μπορεί να αποφευχθεί ξεκινώντας με μια αρκετά πλατιά γειτονιά $N_c = N_c(0)$, η οποία όμως θα συρρικνώνεται με το πέρασμα του χρόνου. Η αρχική ακτίνα του N_c θα μπορούσε να είναι περισσότερη από τη μισή ακτίνα του δικτύου. Κατά τη διάρκεια περίπου των πρώτων 1000 βημάτων του αλγορίθμου, εκεί όπου οι κόμβοι του πλέγματος ταξινομούνται σωστά και ο ρυθμός μάθησης είναι αρκετά μεγάλος, η ακτίνα του N_c μπορεί να συρρικνώνεται γραμμικά στη μονάδα.

Περίπου για τα πρώτα 1000 βήματα, η συνάρτηση $a(t)$ θα πρέπει να ξεκινάει με μια τιμή που να είναι κοντά στη μονάδα κι έπειτα να μειώνεται μονοτονικά. Η $a(t)$ μπορεί

να είναι γραμμική, εκθετική ή αντιστρόφως ανάλογη του χρόνου t . Η διάταξη των m_i συμβαίνει κατά την αρχική περίοδο ενώ τα υπόλοιπα βήματα χρειάζονται για την τελική προσαρμογή του χάρτη. Μετά από τη φάση της διάταξης, η συνάρτηση $a(t)$ πρέπει να παίρνει μικρές τιμές (της τάξης ή και λιγότερου του 0.02) για μια μεγάλη περίοδο. Και για την τελική φάση επίσης δεν είναι σημαντικό πως θα μειώνεται η συνάρτηση $a(t)$ (γραμμικά ή εκθετικά). Βέβαια για πολύ μεγάλους χάρτες, η επιλογή ενός βέλτιστου κανόνα για την αλλαγή του ρυθμού μάθησης μπορεί να είναι καθοριστική για τη σύγκλιση.

Εφόσον η μάθηση είναι μια στοχαστική διαδικασία, η τελική στατιστική ακρίβεια εξαρτάται από τον αριθμό των βημάτων στην τελική φάση σύγκλισης που λογικά πρέπει να είναι μεγάλος. Δεν υπάρχει κάποιος τρόπος για να αποτραπεί αυτή η απαίτηση. Ένας πρακτικός κανόνας είναι ότι για καλή στατιστική ακρίβεια, ο αριθμός των βημάτων πρέπει να είναι τουλάχιστον 500 φορές τον αριθμό των μονάδων του δικτύου. Από τη άλλη μεριά ο αριθμός των διαστάσεων του συνόλου δεδομένων δεν έχει καμιά επίδραση στον αριθμό των βημάτων επανάληψης. Ας σημειωθεί τέλος ότι ο αλγόριθμος που παρουσιάστηκε είναι υπολογιστικά πολύ λιτός κι ότι αν τα διαθέσιμα δεδομένα εισόδου είναι λίγα τότε αυτά θα πρέπει να ανακυκλωθούν για να συμπληρωθεί ο επιθυμητός αριθμός βημάτων του αλγορίθμου.

Για τη καλύτερη κατανόηση της λειτουργίας του αλγορίθμου παραθέτουμε ένα παράδειγμα εκτέλεσης του που φαίνεται στο Σχήμα 2.10. Σε αυτό το παράδειγμα το σύνολο δεδομένων αποτελείται από δύο ομάδες σημείων οργανωμένες σε ομόκεντρους κυκλικούς δίσκους. Στο σχήμα με το σύμβολο «x» σημειώνονται τα σημεία του συνόλου, ενώ τα σημεία με σταυρούς αναπαριστούν τα βάρη των κόμβων του SOM.



Σχήμα 2.10: Τέσσερις καταστάσεις του αλγορίθμου SOM, μετά από 0(α), 3(β), 10(γ) και 20(δ) επαναλήψεις.

Οι γραμμές αναπαριστούν τις συνδέσεις μεταξύ των κόμβων. Το συνολικό μέγεθος του χάρτη είναι 10×10 , με μορφή τετραγώνου. Στο Σχήμα 2.10(α), απεικονίζεται η τυχαία αρχικοποίηση του αλγορίθμου. Έπειτα από 3 επαναλήψεις χρησιμοποιώντας όλα τα σημεία του συνόλου δεδομένων κάθε φορά, ο χάρτης βρίσκεται στη κατάσταση που απεικονίζεται στο Σχήμα 2.10(β), ακόμη και από αυτό το βήμα βλέπουμε την ικανότητα του να προσεγγίσει την κατανομή των δεδομένων. Στα επόμενα βήματα (Σχήματα 2.10(β),(γ),(δ)), ο αλγόριθμος παρόλο που αλλάζει τις θέσεις των κόμβων δεν αλλάζει σημαντικά την τοπολογική τους διάταξη.

Το κυριότερο μειονέκτημά των (SOM), είναι το σταθερό μέγεθος του αρχικού πλέγματος. Για την επίλυση αυτού του προβλήματος στις εργασίες [55,56], προτάθηκαν αλγόριθμοι που προσαρμόζουν το μέγεθος το πλέγματος αλλά και την τοπολογική του διάταξη.

Ο αλγόριθμος Growing Neural Gas (GNG)

Αυτή η μέθοδος παρόλο που βασίζεται στην γενική ιδέα των Χαρτών Αυτο-Οργάνωσης, έχει την ικανότητα να προσαρμόζει το πλήθος των κόμβων καθώς και την τοπολογική τους διάταξη κατά την εκτέλεση του αλγορίθμου. Αρχικά ο αλγόριθμος, χρησιμοποιεί πολύ λίγους κόμβους και νέοι κόμβοι εισέρχονται σταδιακά. Ο καθορισμός του σημείου εισαγωγής των νέων κόμβων γίνεται, χρησιμοποιώντας τοπικά μέτρα λάθους, και επιλέγεται το μέγιστο από αυτά.

Ο GNG αλγόριθμος είναι ουσιαστικά ένα επαυξητικό νευρωνικό δίκτυο. Αυτό μπορεί να περιγραφεί σαν ένα γράφημα που αποτελείται από k κόμβους, κάθε ένας από τους οποίους έχει ένα διάνυσμα βάρους, w_i , και ένα σύνολο ακμών προς άλλους κόμβους γείτονες. Κατά τη διάρκεια της διαδικασίας ομαδοποίησης, νέοι κόμβοι εισάγονται στο δίκτυο μέχρι να επιτευχθεί ένας μέγιστος αριθμός.

Αρχικά η διαδικασία, ξεκινά με δύο κόμβους, τυχαία τοποθετημένους και συνδεδεμένους με μια ακμή. Η προσαρμογή των βαρών των κόμβων, εκτελείται επαναληπτικά. Για κάθε στοιχείο καθορίζεται ο πιο κοντινός κόμβος (νικητής), s_1 και ο δεύτερος κοντινότερος κόμβος s_2 . Αυτοί οι δύο κόμβοι συνδέονται με μια ακμή.

Σε κάθε ακμή είναι ανατιθεμένη μια μεταβλητή ηλικίας. Σε κάθε βήμα του αλγορίθμου, στις ακμές που προέρχονται από τον κόμβο νικητή, αυξάνεται η μεταβλητή της ηλικίας κατά 1. Όταν δημιουργείται η ακμή που συνδέει τους κόμβους s_1 και s_2 , η ηλικία της τίθεται ίση με 0. Επισημαίνοντας τις αλλαγές των τιμών της ηλικίας, εντοπίζονται κόμβοι που δεν ενεργοποιούνται. Ακμές που έχουν πολύ μεγάλη ηλικία διαγράφονται, και με αυτό τον τρόπο διαγράφονται και κόμβοι που δεν έχουν ακμές.

Η γειτονιά του κόμβου νικητή είναι περιορισμένη στους τοπολογικούς γείτονές του. Ο νικητής και οι τοπολογικοί γείτονες του κινούνται στο χώρο, προς το εκάστοτε στοιχείο με σταθερό μέρος της απόστασης τους, καθορισμένο χωριστά για το νικητή και τους τοπολογικούς γείτονές του. Καθώς, δεν υπάρχει κάποια συνάρτηση γειτονιάς, όλοι τοπολογικοί γείτονες ενημερώνονται με τον ίδιο τρόπο. Πιο αυστηρά τα βήματα του αλγορίθμου είναι τα εξής:

1. Αρχικοποίηση του συνόλου των κόμβων A , να περιέχει δύο κόμβους c_1 , και c_2 , με τυχαία διανύσματα βαρών w_1 , w_2 . Επίσης το σύνολο των συνδέσεων $C \subset A \times A$, αρχικοποιείται στο κενό.
2. Για κάθε ένα στοιχείο του συνόλου X_i επανέλαβε:

3. Εντόπισε τους δύο κοντινότερους κόμβους s_1, s_2 σε αυτό:

$$s_1 = \operatorname{argmin}_{c \in A} \|c - X_i\|, \quad s_2 = \operatorname{argmin}_{c \in A - \{s_1\}} \|c - X_i\|.$$

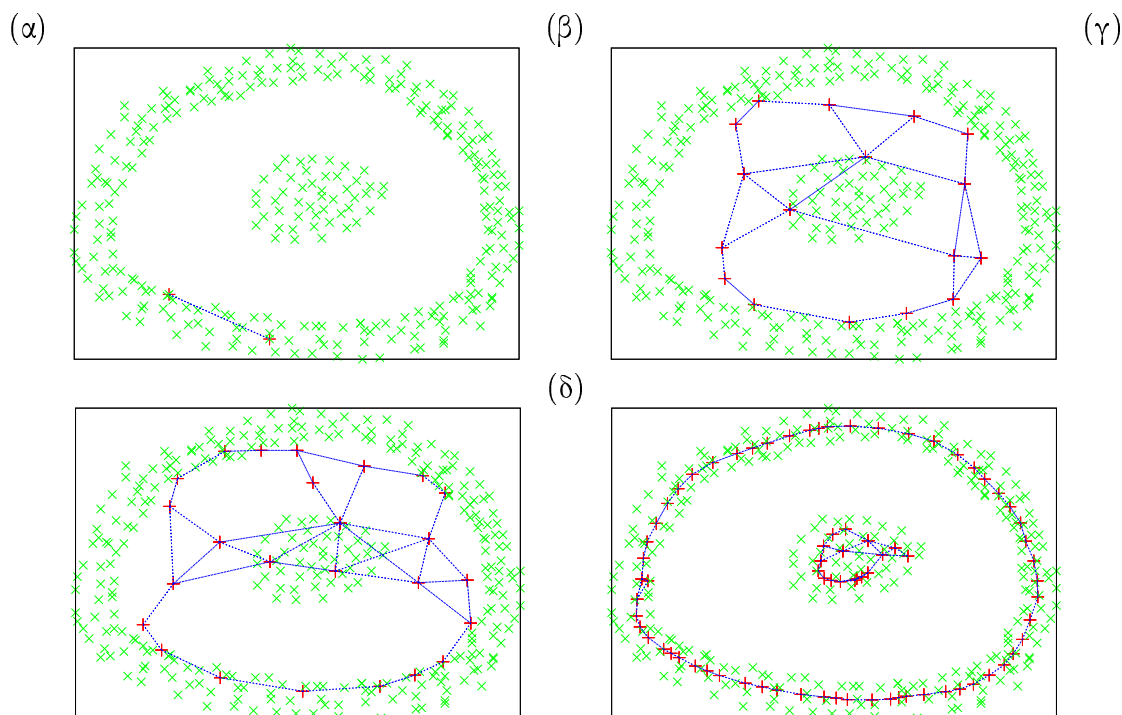
4. Εάν δεν υπάρχει ακμή μεταξύ τους, τότε δημιουργείται: $C = C \cup \{(s_1, s_2)\}$.
5. Η ηλικία της ακμής των κόμβων s_1, s_2 τίθεται στο 0.
6. Η τετραγωνική απόσταση του στοιχείου X_i από το κόμβο s_1 , προστίθεται στην μεταβλητή λάθους του κόμβου s_1 .
7. Τα διανύσματα βαρών του s_1 , και των τοπολογικών του γειτόνων N_{s_1} ανανεώνονται κατά e_b , και e_n αντίστοιχα:

$$\Delta w_{s_1} = e_b \|w_{s_1} - X_i\|, \quad \Delta w_{s_i} = e_n \|w_{s_i} - X_i\|, \quad i \in N_{s_1}.$$

8. Σε όλες τις ακμές που προέρχονται από τον κόμβο s_1 , αυξάνεται η ηλικία κατά 1.
9. Όλες οι ακμές με ηλικία μεγαλύτερη του προκαθορισμένου αριθμού $Age_{\max}$ διαγράφονται. Αν αυτό δημιουργεί κόμβους χωρίς ακμές τότε διαγράφονται και αυτοί.
10. Εάν τα στοιχεία που έχουν επεξεργαστεί μέχρι τώρα είναι πολλαπλάσιο της προκαθορισμένης παραμέτρου λ , τότε γίνεται εισαγωγή νέου κόμβου ως ακολούθως:
 - Υπολογίζεται ο κόμβος q με τη μεγαλύτερη τοπική μεταβλητή λάθους.
 - Από τους γείτονες του q υπολογίζεται ο κόμβος f με τη μεγαλύτερη τοπική μεταβλητή λάθους.
 - Ένας νέος κόμβος r εισάγεται $A = A \cup \{r\}$, με διάνυσμα βάρους $w_r = 1/2(w_q + w_f)$, και τοπικό σφάλμα $E_r = 1/2(E_q + E_f)$.
 - Η ακμή ανάμεσα στους κόμβους q και f , διαγράφεται, και στη θέση της εισέρχονται οι ακμές $\{(r, q)\}$ και $\{(r, f)\}$.
11. Όλες οι τοπικές μεταβλητές λάθους μειώνονται κατά ένα προκαθορισμένο ποσοστό β .
12. Εάν ικανοποιηθεί ένα κριτήριο τερματισμού ο αλγόριθμος τερματίζει.

Όπως είναι φανερό ο παραπάνω αλγόριθμος απαιτεί την ύπαρξη πολλών παραμέτρων οι οποίες δεν έχουν φυσική αναλογία, οπότε είναι δύσκολο να ανατεθούν από το χρήστη. Ωστόσο ο αλγόριθμος, έχει την ικανότητα να δημιουργεί καλές ομαδοποιήσεις των δεδομένων και να παρέχει και τρόπους οπτικοποίησης τους [56].

Για την καλύτερη κατανόηση της λειτουργίας του αλγορίθμου GNG παραθέτουμε ένα παράδειγμα εκτέλεσης του που φαίνεται στο Σχήμα 2.10. Σε αυτό το παράδειγμα το σύνολο δεδομένων είναι το ίδιο με αυτό που χρησιμοποιήθηκε στην περίπτωση του αλγορίθμου SOM. Στο σχήμα με το σύμβολο «x», σημειώνονται τα σημεία του συνόλου, ενώ με σταυρούς σημειώνονται τα σημεία που αναπαριστούν τα βάρη των κόμβων. Οι συνδέσεις μεταξύ των κόμβων σημειώνονται με ακμές. Στο Σχήμα 2.10(α), απεικονίζεται η τυχαία αρχικοποίηση του αλγορίθμου χρησιμοποιώντας δύο κόμβους, συνδεδεμένους μεταξύ τους. Έπειτα από 3 επαναλήψεις χρησιμοποιώντας όλα τα σημεία του συνόλου δεδομένων κάθε φορά, το γράφημα βρίσκειται στη κατάσταση που



Σχήμα 2.11: Τέσσερις καταστάσεις του αλγορίθμου GNG, μετά από 0(α), 5(β), 15(γ) και 50(δ) επαναλήψεις.

απεικονίζεται στο Σχήμα 2.10(β), όπου φαίνεται η σταδιακή προσαρμογή του στη μορφή των δεδομένων. Στο επόμενο βήματα (Σχήμα 2.10(γ)), ο αλγόριθμος προσεγγίζει ακόμα καλύτερα την κατανομή των δεδομένων, και καταφέρνει να σπάσει το γράφο σε δύο μέρη, όπως απεικονίζεται και στο Σχήμα 2.10(δ), που αποτελεί και την τελική κατάσταση του.

2.6 Σύνοψη

Η λήψη αποφάσεων και η αναγνώριση προτύπων σε διάφορες εφαρμογές, συνήθως περιλαμβάνει μεγάλα σύνολα δεδομένων. Είναι δυνατόν λοιπόν για το σκοπό αυτό, να χρησιμοποιηθεί μια αναπαράσταση των δεδομένων αντί για ολόκληρο το σύνολο τους. Αυτή η αναπαράσταση στοχεύει στην γρηγορότερη επεξεργασία των δεδομένων από τους διαθέσιμους υπολογιστικούς πόρους αλλά και στην πιο άμεση κατανόηση των δεδομένων από το ανθρώπινο δυναμικό. Οι αλγόριθμοι ομαδοποίησης είναι σχεδιασμένοι για να κατασκευάζουν τέτοιες αναπαραστάσεις δεδομένων.

Η ομαδοποίηση είναι ο διαχωρισμός ενός συνόλου δεδομένων σε υπο-ομάδες βασισμένη σε κάποιο μέτρο ομοιότητας. Γενικά είναι μια υποκειμενική διαδικασία, καθώς το ίδιο σύνολο συνήθως χρειάζεται να διαχωριστεί με διαφορετικό τρόπο σε διαφορετικές εφαρμογές. Αυτή ακριβώς η υποκειμενικότητα δημιουργεί διάφορες δυσκολίες στην ομαδοποίηση. Μια πιθανή λύση σε αυτό το πρόβλημα είναι η αναπαράσταση της υποκειμενικότητας αυτής με γνώση. Αυτή η γνώση έμμεσα ή άμεσα εισάγεται σε διάφορα στάδια της διαδικασίας, με στόχο να διαμορφώσει το επιθυμητό αποτέλεσμα. Σε αυτό το κεφάλαιο, παρουσιάστηκαν αναλυτικά τα διαφορετικά βήματα που περιλαμβάνονται σε μια διαδικασία ομαδοποίησης. Επίσης, αναλύθηκαν διαφορετικές προσεγγίσεις στην

κατασκευή αλγορίθμων ομαδοποίησης, και παρουσιάστηκαν παραδείγματα που αναδεικνύουν του περιορισμούς της κάθε μίας.

Παρόλο που έχουν προταθεί διάφοροι αλγόριθμοι ομαδοποίησης, ο συνεχώς αυξανόμενος όγκος των βάσεων δεδομένων και οι νέες εξαιρετικά πολύπλοκες εφαρμογές που αναδύονται καθημερινά, δημιουργούν την ανάγκη για κατασκευή νέων αλγορίθμων. Οι νέοι αλγόριθμοι θα πρέπει να έχουν την ικανότητα να αξιοποιούν τις δυνατότητες των σύγχρονων βάσεων δεδομένων, και να μπορούν να ενσωματώνουν γνώση για τα διαφορετικά πεδία εφαρμογής εύκολα.

Στο επόμενο μέρος της διατριβής θα παρουσιαστεί μια νέα προσέγγιση στους αλγόριθμους ομαδοποίησης που αξιοποιεί στοιχεία από διαφορετικές μεθόδους.

Μέρος II

Η Μέθοδος Ομαδοποίησης
k-windows – Επεκτάσεις,
Γενικεύσεις και Τροποποιήσεις

Ο Αλγόριθμος k -windows

‘Τα πάντα πρέπει να γίνονται όσο το δυνατόν απλά,
αλλά όχι απλούστερα από αυτό.’

—Albert Einstein (1879–1955)

Ο αλγόριθμος k -windows [147] είναι μια καινοτόμος μέθοδος ομαδοποίησης, που στοχεύει στην κάλυψη της κάθε ομάδας από ένα πολυδιάστατο παράθυρο. Για αυτόν τον λόγο, αφού αρχικοποιήσει τυχαία ένα παράθυρο συγκεκριμένου μεγέθους a , αξιοποιεί δύο θεμελιώδεις διαδικασίες: τη «μετακίνηση» και τη «διεύρυνση». Η διαδικασία της μετακίνησης στοχεύει στον προσδιορισμό της θέσης του παραθύρου όσο το δυνατόν πιο κοντά κέντρο μιας ομάδας, ενώ η διαδικασία της διεύρυνσης προσπαθεί να διευρύνει το σύνορο του παραθύρου, ώστε να περιλάβει όσο το δυνατόν περισσότερα στοιχεία από την αντίστοιχη ομάδα.

Αν θεωρήσουμε λοιπόν ένα σύνολο δεδομένων X , n στοιχείων, υποσύνολο του d -διάστατου χώρου $\mathbb{R}^d$, τότε ένα παράθυρο W σε αυτόν ορίζεται ως εξής:

Ορισμός 3.1. Έστω δύο σημεία $a, b \in \mathbb{R}^d$, ώστε να ισχύει $a_j \leq b_j$ για $j = 1, \dots, d$. Παράθυρο W , με άκρες τα σημεία a, b , ορίζεται σαν το d -διάστατο διάστημα $W = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_d, b_d] \subset \mathbb{R}^d$. Κέντρο του παραθύρου ορίζεται το σημείο $c_W = (1/2(a_1 + b_1), \dots, 1/2(a_d + b_d)) \in \mathbb{R}^d$.

3.1 Η Διαδικασία της Μετακίνησης

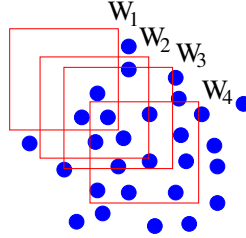
Η διαδικασία της μετακίνησης είναι μια επαναληπτική διαδικασία, που όπως έχει ήδη ειπωθεί στοχεύει στην προσδιορισμό της θέσης του παραθύρου όσο το δυνατόν πιο κοντά στο κέντρο μιας ομάδας. Έστω λοιπόν, ένα παράθυρο $W = [a, b]^d$, με κέντρο c_W . Το πρώτο βήμα της διαδικασίας είναι ο υπολογισμός του συνόλου $Q_W \subset X$, των σημείων του X που περιέχονται στο W , δηλαδή $Q_W = \{x \in X : a_i \leq x_i \leq b_i, i = 1, \dots, d\}$. Αφού υπολογιστεί το σύνολο Q_W , υπολογίζεται το κέντρο βάρους του:

$$g = \left(\frac{1}{|Q_W|} \sum_{x \in Q_W} x_1, \dots, \frac{1}{|Q_W|} \sum_{x \in Q_W} x_d \right). \quad (3.1)$$

Έπειτα κατασκευάζεται το παράθυρο:

$$W_t = [g_i - 1/2(b_i - a_i), g_i + 1/2(b_i - a_i)]^d, \quad i = 1, \dots, d, \quad (3.2)$$

το οποίο ουσιαστικά είναι μια πιθανή μετακίνηση του W στο σημείο g , και υπολογίζεται το σύνολο Q_{W_t} , των σημείων που περιέχει. Αν $|Q_{W_t}| > |Q_W|$, τότε το W αντικαθίσταται από το W_t , και η διαδικασία επαναλαμβάνεται, διαφορετικά τερματίζεται. Ένα παράδειγμα αυτής της διαδικασίας παρουσιάζεται στο Σχήμα 3.1.



Σχήμα 3.1: Διαδοχικές μετακινήσεις ενός παραθύρου.

Μια υψηλού επιπέδου αλγοριθμική περιγραφή, της διαδικασίας της μετακίνησης είναι η ακόλουθη:

Function μετακίνηση(W) {

Επανάλαβε:

Υπολόγισε το σύνολο $Q_W \subset X$ των σημείων του X ,
 που περιέχονται στο W .

Υπολόγισε το κέντρο βάρους g από τον τύπο (3.1).

Κατασκεύασε το παράθυρο W_t από τον τύπο (3.2).

Υπολόγισε το σύνολο $Q_{W_t} \subset X$ των σημείων του X ,
 που περιέχονται στο W_t .

Εάν $|Q_{W_t}| > |Q_W|$, τότε θέσε $W \leftarrow W_t$.

Μέχρι $|Q_{W_t}| \geq |Q_W|$.

Επέστρεψε το παράθυρο W .

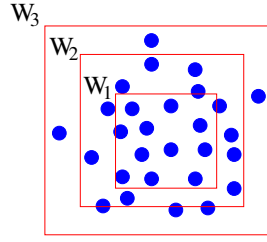
}

3.2 Η Διαδικασία της Διεύρυνσης

Η διαδικασία της διεύρυνσης προσπαθεί να αυξήσει το εύρος για κάθε μια συντεταγμένη του παραθύρου ώστε να περιληφθούν όσο το δυνατόν περισσότερα στοιχεία από την ομάδα, που εγκλωβίζει το κάθε παράθυρο. Κατά συνέπεια, για ένα παράθυρο $W = [a, b]^d$ με κέντρο c_W , η διαδικασία για κάθε συντεταγμένη $i, 1 \leq i \leq d$, χωριστά, διευρύνει, το παράθυρο W , κατά ένα ποσοστό θ_e , το οποίο είναι προκαθορισμένο από τον χρήστη, κατασκευάζοντας το παράθυρο:

$$W_t = [a_1, b_1] \times \cdots \times [a_i - \theta_e(b_i - a_i), b_i + \theta_e(b_i - a_i)] \times \cdots \times [a_d, b_d]. \quad (3.3)$$

Το παράθυρο αυτό, υπόκειται στη διαδικασία της μετακίνησης κατασκευάζοντας το παράθυρο W'_t . Για αυτό το παράθυρο υπολογίζεται ο πληθύνισμός του συνόλου $Q_{W'_t}$, των σημείων του X που περιέχει. Αν $\theta_e |Q_W| < Q_{W'_t}$, όπου $\theta_e \geq 1$, είναι μια παράμετρος καθορισμένη από το χρήστη, τότε $W \leftarrow W'_t$. Διαφορετικά η προσπάθεια διεύρυνσης



Σχήμα 3.2: Διαδοχικές διευρύνσεις ενός παραθύρου.

θεωρείται ανεπιτυχής, και εξετάζεται για διεύρυνση η επόμενη συντεταγμένη. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να μην συμβαίνουν επιτυχημένες διευρύνσεις σε καμία συντεταγμένη. Ένα παράδειγμα αυτής της διαδικασίας φαίνεται στο Σχήμα 3.2.

Παρακάτω δίνεται σε υψηλού επιπέδου περιγραφή ολόκληρη η διαδικασία της διεύρυνσης:

Function διεύρυνση(θ_e, θ_c, W) {
 Επανάλαβε
 Για κάθε συντεταγμένη i
 Επανάλαβε:
 Υπολόγισε το σύνολο $Q_W \subset X$ των σημείων του X ,
 που περιέχονται στο W .
 Διεύρυνε κατά θ_e το W κατασκευάζοντας το W_t από τον τύπο (3.3).
 Εκτέλεσε $W'_t \leftarrow \text{μετακίνηση}(W_t)$.
 Υπολόγισε το σύνολο $Q_{W'_t} \subset X$ των σημείων του X ,
 που περιέχονται στο W'_t .
 Εάν $\theta_c |Q_W| < Q_{W'_t}$, τότε θέσε $W \leftarrow W'_t$.
 Διαφορετικά θεώρησε ανεπιτυχή την προσπάθεια διεύρυνσης.
 Μέχρι $\theta_c |Q_W| \geq Q_{W'_t}$.
 Μέχρι οι προσπάθειες διεύρυνσης να είναι ανεπιτυχείς για όλα τα $i = 1, \dots, d$.
 Επέστρεψε το παράθυρο W .
}

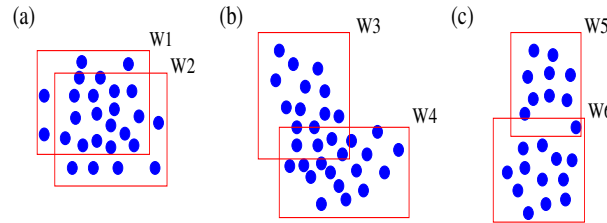
3.3 Αυτόματος Υπολογισμός του Αριθμού των Ομάδων

Ένα κρίσιμο ζήτημα στην ομαδοποίηση γενικότερα είναι προσδιορισμός του αριθμού ομάδων σε ένα σύνολο δεδομένων. Για το σκοπό αυτό έχει αναπτυχθεί μια διαδικασία [18], που επιτρέπει στον αλγόριθμο k -windows, να παρέχει μια προσέγγιση στον αριθμό των ομάδων. Η βασική ιδέα είναι να αρχικοποιηθεί και να εφαρμοστεί ο αλγόριθμος με ένα μεγάλο αριθμό παραθύρων. Οπότε, μετά το τέλος της εκτέλεσης των διαδικασιών μετακίνησης και διεύρυνσης όλων των παραθύρων, εξετάζοντας τα επικαλυπτόμενα παράθυρα μπορεί να υπολογιστεί μια προσέγγιση του πλήθους των ομάδων. Για το σκοπό αυτό, χρησιμοποιούνται δύο παράμετροι δοσμένοι από το χρήστη: θ_s (παράμετρος ομοιότητας) και θ_m (παράμετρος ενοποίησης), με $0 < \theta_s, \theta_m \leq 1$. Πιο συγκεκριμένα για κάθε ζεύγος επικαλυπτόμενων παραθύρων εξετάζεται η αναλογία

του πλήθους των κοινών τους σημείων προς το συνολικό πλήθος σημείων που περιέχει το καθένα. Εάν αυτή η αναλογία υπερβαίνει την παράμετρο ομοιότητας θ_s , τότε τα εξεταζόμενα παράθυρα θεωρούνται όμοια και αυτό με το μικρότερο πλήθος σημείων διαγράφεται. Αντίθετα αν η αναλογία των σημείων είναι μόνο μεγαλύτερη από την παράμετρο ενοποίησης θ_m , τότε και τα δύο παράθυρα διατηρούνται στο τελικό σύνολο αλλά θεωρούνται ότι καλύπτουν περιοχές της ίδιας ομάδας. Τέλος, αν η αναλογία των κοινών σημείων είναι μικρότερη και από την τιμή θ_m , τότε τα παράθυρα θεωρούνται ότι καλύπτουν διαφορετικές ομάδες.

Αυτή ακριβώς η διαδικασία, επιτρέπει στον αλγόριθμο, να ανακαλύψει και ομάδες σημείων με μη κυρτή μορφολογία. Αν ωστόσο, ο χρήστης δεν ενδιαφέρεται για μη κυρτές μορφολογικά ομάδες και θέσει την τιμή της παραμέτρου αυτής μεγαλύτερη της μονάδας τότε ανακαλύπτονται μόνο κυρτές μορφολογικά ομάδες από τον αλγόριθμο. Μάλιστα όσο η τιμή της παραμέτρου τείνει στο 0 τόσο ενισχύεται η ανακάλυψη μη-κυρτών μορφολογικά ομάδων ενώ αντίθετα όσο τείνει στην μονάδα τόσο προωθείται η ανακάλυψη κυρτών μορφολογικά ομάδων. Πρέπει επίσης να τονιστεί, ότι η διαδικασία αυτή είναι τελείως ξεχωριστή από τις προηγούμενες, οπότε και μπορεί να εφαρμοστεί για διάφορες τιμές, των θ_s και θ_m , χωρίς να χρειάζεται επανεκτέλεση του αλγορίθμου.

Ένα παράδειγμα της διαδικασίας απεικονίζεται στο Σχήμα 3.3. Η αναλογία των κοινών σημείων των παραθύρων $W1$ και $W2$, ως προς τα συνολικά τους σημεία ξεπερνά την τιμή της παραμέτρου θ_s , οπότε και θεωρούνται ότι καλύπτουν την ίδια ομάδα, με αποτέλεσμα το παράθυρο $W1$ να διαγραφεί. Αντίθετα, η αντίστοιχη αναλογία, για τα παράθυρα $W3$ και $W4$, ξεπερνά μόνο την τιμή της παραμέτρου θ_m , οπότε και θεωρούνται ότι καλύπτουν τμήματα της ίδιας ομάδας. Τα παράθυρα, $W5$ και $W6$, έχουν πολύ μικρή αναλογία κοινών σημείων, οπότε και θεωρούνται ότι καλύπτουν διαφορετικές ομάδες.



Σχήμα 3.3: Ένα παράδειγμα της διαδικασίας συγχώνευσης.

Παρακάτω δίνεται ολόκληρη η διαδικασία της συγχώνευσης παραθύρων σε υψηλού επιπέδου περιγραφή, όπου το $\mathcal{W}$, αποτελεί ένα σύνολο παραθύρων:

function συγχώνευση($\theta_s, \theta_m, \mathcal{W}$) {

 Για **κάθε** παράθυρο W_j στο $\mathcal{W}$ που δεν έχει ετικέτα τότε **εκτέλεσε**:

 Θέσε στο W_j την ετικέτα w_j .

 Εάν $\exists W_i \neq W_j$ στο $\mathcal{W}$, που επικαλύπτεται με το W_j τότε:

 Υπολόγισε το πλήθος των κοινών τους σημείων $n_{i,j}$

 Εάν $\frac{n_{i,j}}{n_i} \geq \theta_s$ και $n_i < n_j$ τότε:

 Θέσε σε όλα τα παράθυρα στο $\mathcal{W}$, με την ετικέτα w_i
 την ετικέτα w_j .

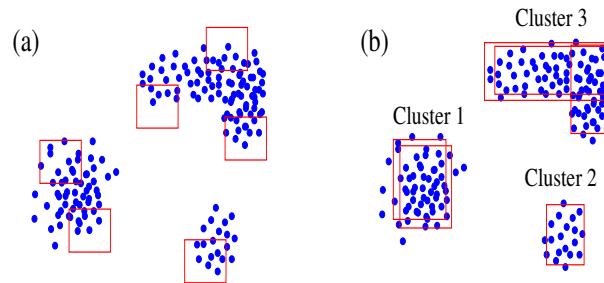
 Εάν $0.5(\frac{n_{i,j}}{n_j} + \frac{n}{n_i}) \geq \theta_m$ τότε:

 Θέσε σε όλα τα παράθυρα στο $\mathcal{W}$, με την ετικέτα w_i την ετικέτα w_j .

Επέστρεψε το σύνολο παραθύρων με ετικέτες $\mathcal{W}$.
}

3.4 Ο Αλγόριθμος Unsupervised k -Windows

Χρησιμοποιώντας τις διαδικασίες μετακίνησης, διεύρυνσης και συγχώνευσης, που παρουσιάστηκαν στις προηγούμενες ενότητες, σε αυτήν την ενότητα προτείνεται ένα ολοκληρωμένο αλγοριθμικό σχήμα που ονομάζεται Unsupervised k -Windows (UKW). Το προτεινόμενο σχήμα ξεκινά αρχικοποιώντας ένα σύνολο παραθύρων στα δεδομένα. Μια απλή μέθοδος αρχικοποίησης, βασίζεται στην επιλογή τυχαίων στοιχείων από το σύνολο δεδομένων και την κατασκευή παραθύρων με κέντρα τα σημεία αυτά. Έπειτα ο αλγόριθμος εκτελεί επαναληπτικά τις διαδικασίες μετακίνησης και διεύρυνσης, μέχρι να οριστικοποιηθούν οι θέσεις των παραθύρων. Το τελικό σύνολο παραθύρων δίνεται στην διαδικασία της συγχώνευσης, η οποία κατασκευάζει την τελική δομή και το πλήθος των ομάδων, αναθέτοντας ετικέτες στα παράθυρα. Παράθυρα με την ίδια ετικέτα επικαλύπτουν περιοχές της ίδιας ομάδας. Όπως είναι προφανές ο αλγόριθμος δεν καταλήγει σε μια διαμέριση του συνόλου δεδομένων, αλλά μόνο σε τελικές θέσεις παραθύρων. Εύκολα όμως μπορεί να διαμεριστεί το σύνολο δεδομένων, αναθέτοντας το κάθε στοιχείο στην ομάδα που ανήκει το κοντινότερο σε αυτό παράθυρο. Ένα παράδειγμα, της λειτουργίας του αλγορίθμου παρουσιάζεται στο Σχήμα 3.4. Στο τμήμα (a) απεικονίζεται το σύνολο δεδομένων αποτελούμενο από 3 ομάδες, με 6 παράθυρα τυχαία αρχικοποιημένα. Τελικά, ο αλγόριθμος καταφέρνει να ανακαλύψει τόσο τις κυρτές όσο και τις μη κυρτές από αυτές (Σχήμα 3.4(b)).



Σχήμα 3.4: Ένα παράδειγμα εφαρμογής του αλγορίθμου UKW.

Στη συνέχεια παρουσιάζεται σε μια υψηλού επιπέδου περιγραφή ο συνολικός αλγόριθμος, όπως αυτός χρησιμοποιείται, όπου αναφέρεται η εφαρμογή του, στο υπόλοιπο της διατριβής.

Αλγόριθμος UKW $(k, a, \theta_e, \theta_c, \theta_s, \theta_m)$ {

Εκτέλεσε $\mathcal{W}$ = αρχικοποίηση(k, a).

Για κάθε παράθυρο W_j στο $\mathcal{W}$ εκτέλεσε:

Επανάλαβε:

Εκτέλεσε $W_j \leftarrow$ μετακίνηση(W_j).

Εκτέλεσε $W_j \leftarrow$ διεύρυνση(θ_e, θ_c, W_j).

Μέχρι να μη αλλάζουν τα όρια και το κέντρο του παραθύρου W_j .

Εκτέλεσε συγχώνευση($\theta_m, \theta_s, \mathcal{W}$)

Επέστρεψε τις ομάδες $\{c_{l1}, c_{l2}, \dots$ όπου:
 $c_{li} = \{i : i \in W_j, \text{label}(W_j) = li\}$ }

}

function αρχικοποίηση(k, a) {
Αρχικοποίησε k παράθυρα $w_1, \dots, w_k$ μεγέθους a σε κάθε διάσταση.
Επέλεξε k τυχαία στοιχεία από το σύνολο δεδομένων και θέσε το κέντρο των παραθύρων ίσο με τις συντεταγμένες των σημείων αυτών.
Επέστρεψε το σύνολο $\mathcal{W}$ των k αυτών παραθύρων.
}

3.5 Υπολογιστική Πολυπλοκότητα

Η υπολογιστική πολυπλοκότητα του αλγορίθμου UKW, ουσιαστικά καθορίζεται από το βήμα υπολογισμού των σημείων που βρίσκονται μέσα σε ένα συγκεκριμένο παράθυρο. Αυτό είναι το γνωστό πρόβλημα της ορθογώνιας αναζήτησης διαστήματος της Υπολογιστικής Γεωμετρίας [115]. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί πολλοί αλγόριθμοι [57], που επιτρέπουν την υπο-γραμμική χρονική πολυπλοκότητα της ανακάλυψης των σημείων ενός συνόλου που περιέχονται σε ένα παράθυρο, κατασκευάζοντας μια κατάλληλη δομή δεικτοδότησης για την αποθήκευση των δεδομένων. Στον Πίνακα 3.1 αναφέρεται η πολυπλοκότητα διάφορων τέτοιων προσεγγίσεων. Πιο συγκεκριμένα για εφαρμογές πολύ μεγάλης διάστασης, δομές σαν το Multi-Dimensional Binary Tree [115], και των Bentley και Maurer [12] δείχνουν πιο κατάλληλες. Αντίθετα, για εφαρμογές μικρής διάστασης με μεγάλο πλήθος στοιχείων, είναι πιο ελκυστικές προσεγγίσεις σαν την Alevizos [4].

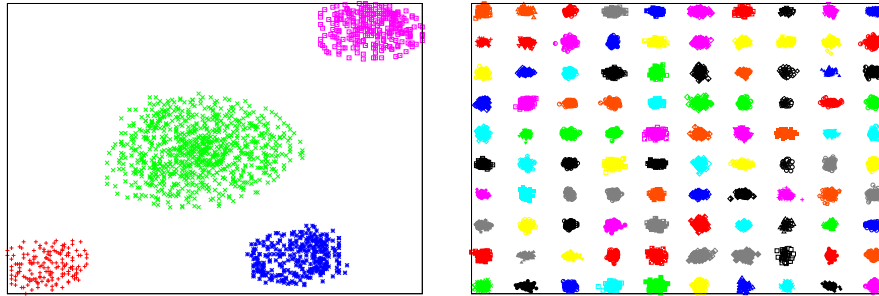
Μέθοδος	Προεπεξεργασία		Χρόνος Αναζήτησης
	Χρόνος	Χώρος	
Multidim. Binary Tree [115]	$\theta(dn \log n)$	$\theta(dn)$	$O(s + dn^{1-1/d})$
Range Tree [115]	$O(n \log^{d-1} n)$	$O(n \log^{d-1} n)$	$O(s + \log^d n)$
Wilard και Luecker [115]	$O(n \log^{d-1} n)$	$O(n \log^{d-1} n)$	$O(s + \log^{d-1} n)$
Chazelle [25]	$O(n \log^{d-1} n)$	$O\left(n \frac{\log^{d-1} n}{\log \log n}\right)$	$O(s + \log^{d-1} n)$
Chazelle και Guibas [26]	$O(n \log^{d+1} n)$	$O(n \log^d n)$	$O(s + \log^{d-2} n)$
Alevizos [4]	$O(n \log^{d-1} n)$	$O(n \log^{d-1} n)$	$O(s + \log^{d-2} n)$
Bentley και Maurer [12]	$O(n^{2d-1})$	$O(n^{2d-1})$	$O(s + d \log n)$
Bentley και Maurer [12]	$O(n^{1+\epsilon})$	$O(n^{1+\epsilon})$	$O(s + \log n)$
Bentley και Maurer [12]	$O(n \log n)$	$O(n)$	$O(n^\epsilon)$

Πίνακας 3.1: Μέθοδοι Ορθογώνιας Αναζήτησης Διαστήματος.

3.6 Πειραματικά Αποτελέσματα

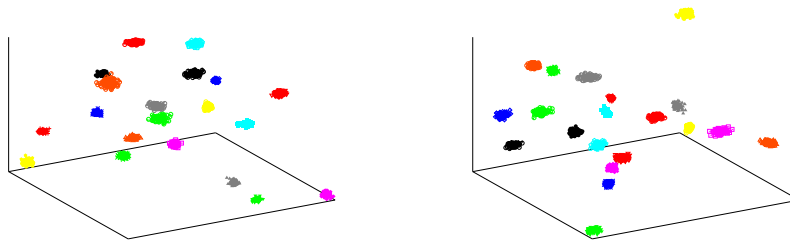
Για την αξιολόγηση της απόδοσης του αλγορίθμου UKW, θα χρησιμοποιηθούν 6 τεχνητά σύνολα δεδομένων και ένα πραγματικό. Τα πρώτα δύο σύνολα είναι δύο δια-

στάσεων, και περιέχουν 1600 και 100000 σημεία, αντίστοιχα. Στο σύνολο $Dset_1$ τα σημεία είναι οργανωμένα σε 4 ομάδες διαφορετικού μεγέθους. Αντίθετα, το σύνολο $Dset_2$ περιέχει 100 ομάδες του ίδιου μεγέθους (1000 στοιχεία σε κάθε ομάδα). τα κέντρα της κάθε ομάδας είναι ευθυγραμμισμένα σε ένα πλέγμα στο χώρο $[10, 200]^2$, και τα αντίστοιχα σημεία έχουν ληφθεί με τυχαία δειγματοληψία από μια κανονική κατανομή για την κάθε ομάδα, με διασπορά σε κάθε διάσταση ένα τυχαίο αριθμό στο διάστημα $[1, 2]$. Οι συνδιασπορές ανάμεσα στις διαφορετικές συντεταγμένες ήταν 0. Ο αλγόριθμος εφαρμόστηκε με 16 και 256, αρχικά παράθυρα στα δυο σύνολα αντίστοιχα. Οι τιμές των παραμέτρων $\{a, \theta_e, \theta_c, \theta_s, \theta_v\}$ ήταν $\{5, 0.1, 0.1, 0.7, 0.1\}$ και στις δύο περιπτώσεις. Τα σύνολα καθώς και τα αποτελέσματα απεικονίζονται στο Σχήμα 3.5.



Σχήμα 3.5: Το αποτέλεσμα του αλγορίθμου UKW στα σύνολα $Dset_1$ (αριστερά) και $Dset_2$ (δεξιά).

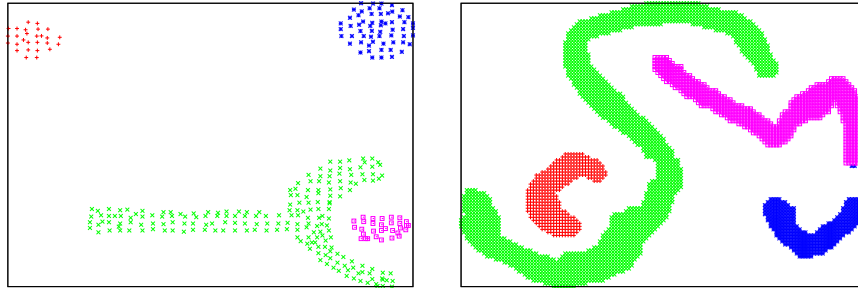
Το επόμενο σύνολο δεδομένων $Dset_3$ είναι τριών διαστάσεων, και δημιουργήθηκε επιλέγοντας τυχαία 20 κέντρα για ομάδες στο διάστημα $[10, 200]^3$. Γύρω από κάθε κέντρο έχουν ληφθεί με τυχαία δειγματοληψία 100 σημεία από μια κανονική κατανομή με διασπορά σε κάθε διάσταση έναν τυχαίο αριθμό στο διάστημα $[1, 2]$. Οι συνδιασπορές ανάμεσα στις διαφορετικές συντεταγμένες ήταν 0. Το σύνολο δεδομένων $Dset_4$, δημιουργήθηκε με τον ίδιο τρόπο αλλά είναι διάστασης 50. Και στα δύο σύνολα ο αλγόριθμος χρησιμοποίησε 128 παράθυρα, με τιμές των παραμέτρων ίδιες με αυτές των προηγούμενων πειραμάτων. Και στις δύο περιπτώσεις ο αλγόριθμος ήταν ικανός να εντοπίσει με ακρίβεια και τις 20 ομάδες. Το σύνολο $Dset_3$, και μια τρισδιάστατη προβολή του $Dset_4$, φαίνονται στο Σχήμα 3.6, μαζί με τα αποτελέσματα του αλγορίθμου.



Σχήμα 3.6: Το αποτέλεσμα του αλγορίθμου UKW στα σύνολα $Dset_3$ (αριστερά) και στην τρισδιάστατη προβολή του $Dset_4$ (δεξιά).

Τα τελευταία δύο τεχνητά σύνολα δεδομένων, $Dset_5$ και $Dset_6$ αποτελούνται από 319 και 3651 σημεία αντίστοιχα. Και τα δύο περιέχουν 4 μη κυρτές με ανώμαλο σχήμα

ομάδες από ομοιόμορφα κατανεμημένα σημεία. Η εφαρμογή του αλγορίθμου UKW, με τις ίδιες τιμές των παραμέτρων και 32 και 256 παράθυρα στην κάθε περίπτωση αντίστοιχά, καταλήγει σε ομαδοποιήσεις που αναγνωρίζουν σωστά όλες τις ομάδες όπως φαίνεται στο Σχήμα 3.7. Τέτοια σύνολα έχουν χρησιμοποιηθεί και για την αξιολόγηση άλλων παρόμοιων αλγορίθμων [65, 82, 123].



Σχήμα 3.7: Το αποτέλεσμα του αλγορίθμου UKW στα σύνολα $Dset_5$ (αριστερά) και $Dset_6$ (δεξιά).

Το σύνολο δεδομένων $Dset_7$ αποτελεί ένα μέρος του διαγωνισμού KDD 1999 [83], και έχει κατασκευαστεί το 1998 από το DARPA Intrusion Detection Evaluation Program στο MIT Lincoln Labs. Ο σκοπός εκείνου του προγράμματος ήταν η επισκόπηση της έρευνας στην ανίχνευση επιθέσεων σε υπολογιστικά συστήματα. Από αυτό το σύνολο χρησιμοποιήθηκαν τα πρώτα 100000 στοιχεία. Τα 77888 από αυτά έχουν χαρακτηριστεί σαν normal ενώ τα υπόλοιπα 22112, σαν επιθέσεις στα συστήματα τύπου Denial of Service (DoS). Από τα 42 χαρακτηριστικά χρησιμοποιήθηκαν τα 37 αριθμητικά. Εφαρμόζοντας τον αλγόριθμο UKW σε αυτό το σύνολο δεδομένων με 16 αρχικά παράθυρα, αυτός καταλήγει σε 7 ομάδες, από τις οποίες μία περιέχει 22087 στοιχεία τύπου DoS. Οι υπόλοιπες έξι περιέχουν αποκλειστικά στοιχεία τύπου normal, εκτός από μία που περιέχει και 24 DoS, στοιχεία. Από την φυσική σημασία των εξαγομένων ομάδων είναι προφανές ότι ο αλγόριθμος καταφέρνει να καταλήξει σε ομαδοποίηση που αναπαριστά την πραγματική δομή των δεδομένων.

3.7 Συμπεράσματα - Συνεισφορά

Στο κεφάλαιο αυτό, παρουσιάστηκε το πλήρες αλγοριθμικό σχήμα του UKW, που κατά την εκτέλεση του έχει την ικανότητα να προσεγγίζει και το πλήθος των ομάδων σε ένα σύνολο δεδομένων. Βασισμένο στον αλγόριθμο k -windows, το σχήμα αυτό καταφέρνει με μια πολύ μικρή χρονική πολυπλοκότητα σε σχέση με άλλους αντίστοιχους αλγόριθμους, να καταλήξει σε ακριβείς ομαδοποιήσεις. Οι παράμετροι που χρησιμοποιεί έχουν άμεση σύνδεση με το πρόβλημα και μπορούν να γίνουν εύκολα κατανοητοί από τον χρήστη. Η γενίκευση του αλγορίθμου σε κατηγορικά δεδομένα καθώς και η έμφυτη παραλληλία του θα αναλυθούν σε επόμενα κεφάλαια.

Παρόλο που πολλά μπορούν να δοκιμαστούν ακόμη, για τη βελτίωση του αλγορίθμου, η μορφή που παρουσιάζεται σε αυτό το κεφάλαιο είναι αποδοτική, και έχει χρησιμοποιηθεί ως τώρα σε μεγάλο πλήθος προβλημάτων που περιλαμβάνουν την βιοπληροφορική [130, 131], τις ιατρικές διαγνώσεις [93, 133], την πρόβλεψη τιμών σε χρονοσειρές [107–109], και την εξατομίκευση των προτιμήσεων χρηστών στην δικτυακή πλοήγηση [121].

Γενίκευση του Αλγορίθμου k -windows σε Μετρικούς Χώρους

Η κατεύθυνση του μυαλού
είναι πιο σημαντική από την πρόοδο του.

—Joseph Joubert (1754–1824)

Ο συνεχώς αυξανόμενος αριθμός εφαρμογών βάσεων δεδομένων σε πεδία όπως η ανάκτηση πληροφορίας πολυμέσων (εικόνων, ήχων, video, κ.α.), η ανάλυση και μοντελοποίηση χρονοσειρών και η γονιδιακή στοιχειοθέτηση, συμπεριλαμβάνουν δεδομένα που αναπαρίστανται με διακριτές τιμές χωρίς καμία διάταξη. Η ομοιότητα σε τέτοια δεδομένα μετράται μέσω μιας συνάρτησης απόστασης που τις περισσότερες φορές αποτελεί μετρική ορισμένη στον εκάστοτε χώρο. Για αυτό το λόγο είναι πολύ σημαντική η κατασκευή αλγορίθμων ομαδοποίησης ικανών να εξάγουν πληροφορία από τέτοια δεδομένα.

Μέχρι τώρα σε αυτή τη διατριβή, παρουσιάστηκε ο αλγόριθμος k -windows για συνεχή αριθμητικά δεδομένα. Σε αυτό το κεφάλαιο παρουσιάζουμε μια γενίκευση του σε μετρικούς χώρους [136]. Η προτεινόμενη γενίκευση διατηρεί την δυνατότητα του αλγορίθμου να προσεγγίζει το πλήθος των ομάδων σε κάθε σύνολο δεδομένων. Επίσης αναλύεται και η απόδοση του στην ομαδοποίηση γνωστών συνόλων δεδομένων.

4.1 Παρόμοιες Προσεγγίσεις

Παρόμοιες προσεγγίσεις με αυτήν που παρουσιάζεται σε αυτό το κεφάλαιο, περιλαμβάνουν αλγόριθμους, που έχουν την ικανότητα να ομαδοποιήσουν μεγάλα σύνολα δεδομένων με συνεχή και κατηγορικά χαρακτηριστικά. Πιο αυστηρά, θεωρούμε ότι κάθε στοιχείο του συνόλου δεδομένων αναπαρίσταται από το σύνολο των χαρακτηριστικών $A_1, A_2, \dots, A_m$. Κάθε χαρακτηριστικό έχει το δικό του πεδίο τιμών $DOM(A_i)$. Όπως έχει ήδη αναφερθεί (Ενότητα 2.3), ένα χαρακτηριστικό είναι κατηγορικό εάν έχει πεπερασμένο και χωρίς διάταξη πεδίο τιμών. Ενώ θεωρείται συνεχές εάν το πεδίο τιμών είναι υποσύνολο του $\mathbb{R}$.

Έχουν προταθεί διάφορες μετρικές [62, 63] που επιτρέπουν την χρήση ιεραρχικών αλγορίθμων σε κατηγορικά σύνολα δεδομένων. Όμως ο μη γραμμικός χρόνος που απαιτούν αυτοί οι αλγόριθμοι κάνει απαγορευτική την χρήση τους σε μεγάλα σύνολα.

Βασισμένοι στη διαδικασία του k -means έχουν προταθεί αλγόριθμοι [71, 118], που επιχειρούν να εφαρμοστούν σε κατηγορικά χαρακτηριστικά. Το βασικό μειονέκτημα

αυτών των προσεγγίσεων είναι η απαίτηση του προκαθορισμού του πλήθους των ομάδων πριν την εκτέλεση του αλγορίθμου.

Πιο ανεπτυγμένες τεχνικές όπως ο CLARANS [105], ενώ είναι ικανές να παράγουν αποδοτικές ομαδοποιήσεις, απαιτούν υπολογιστικό χρόνο που πλησιάζει τον κυβικό [123]. Επίσης για να προσεγγίσουν τον αριθμό των ομάδων χρειάζονται διαδοχικές επανεκτελέσεις που διογκώνουν τον υπολογιστικό χρόνο περαιτέρω.

Προσεγγίσεις σαν των OPTICS και GDBSCAN [8, 123] δείχνουν πιο ελκυστικές. Ο υπολογιστικός όμως χρόνος που χρειάζονται είναι της τάξης $O(n \log(n))$, και έχουν το επιπρόσθετο μειονέκτημα των μεγάλων απαιτήσεων σε μνήμη που μεγαλώνει ανάλογα με το μέγεθος των ομάδων.

4.2 Ο Γενικευμένος Αλγόριθμος k -windows

Στη ενότητα αυτή παρουσιάζουμε τον *Γενικευμένο αλγόριθμο k -windows* (*Generalized Unsupervised k -windows Algorithm* (GUKW), που είναι σχεδιασμένος να επεκτείνει τις λειτουργίες του αλγορίθμου k -windows σε μετρικούς χώρους. Η προτεινόμενη γενίκευση βασίζεται στην αλγοριθμική διαδικασία του UKW, αλλά χρησιμοποιεί δομές που είναι κατάλληλες για το πεδίο των μετρικών χώρων. Με αυτόν τον τρόπο εάν δηλώσουμε με (X, d) ένα μετρικό χώρο, όπου X είναι το σύνολο των στοιχείων και $d : X \rightarrow \mathbb{R}$ είναι μια μετρική συνάρτηση απόστασης, τότε ένα παράθυρο W μεγέθους r , ορίζεται ως εξής:

Ορισμός 4.1. Έστω στοιχείο $c \in X$, και αριθμός $r \in \mathbb{R}$. Τότε ορίζουμε παράθυρο μεγέθους r γύρω από το c το σύνολο $W = \{y \in X : d(c, y) \leq r\}$. Το σημείο c λέγεται και κέντρο του παραθύρου.

Επομένως οι τρεις διαδικασίες του αλγορίθμου UKW, *μετακίνηση*, *διεύρυνση* και *συγχώνευση*, πρέπει να γενικοποιηθούν στις νέες συνθήκες. Κάθε μία από αυτές θα αναλυθεί ξεχωριστά στις επόμενες παραγράφους.

4.2.1 Η Γενικευμένη Διαδικασία Μετακίνησης

Όπως έχει ειπωθεί και νωρίτερα, η διαδικασία της μετακίνησης στοχεύει στην τοποθέτηση του κάθε παραθύρου όσο το δυνατόν πλησιέστερα στο κέντρο της κάθε ομάδας. Η υπόθεση του μετρικού χώρου όμως αναγκάζει τον αλγόριθμο να επιλέξει για κέντρο βάρους οποιαδήποτε υποσυνόλου του, ένα από τα σημεία του. Με αυτό τον τρόπο, το κέντρο βάρους $g \in W$ των σημείων ενός παραθύρου ορίζεται σαν το σημείο που ελαχιστοποιεί την ποσότητα:

$$D(g, W) = \sum_{\forall x \in W, x \neq g} d(x, g). \quad (4.1)$$

Αυτό το στάδιο επηρεάζει σημαντικά την πολυπλοκότητα του αλγορίθμου. Ο υπολογισμός του ελαχιστοποιητή της ποσότητας (4.1), μέσω μιας «ωμής δύναμης» (brutal force) τεχνικής απαιτεί τετραγωνικό χρόνο ως προς τον πληθύνει του W . Στην εργασία [71], προτάθηκε ένα μέτρο ομοιότητας που επιτρέπει τον υπολογισμό του ελαχιστοποιητή (*mode*, χρησιμοποιώντας την αντίστοιχη ορολογία) σε γραμμικό χρόνο. Εξαιτίας αυτού του χαρακτηριστικού αυτή η προσέγγιση χρησιμοποιείται και στην προτεινόμενη γενίκευση.

Πιο συγκεκριμένα, σύμφωνα με την εργασία [71], χρησιμοποιούμε την ακόλουθη μετρική για τα κατηγορικά χαρακτηριστικά:

$$d(x, y) = \sum_{i=1}^m \delta(x_i, x_j),$$

όπου,

$$\delta(x_i, x_j) = \begin{cases} 0 & (x_i = y_j) \\ 1 & (x_i \neq y_j) \end{cases}.$$

Έπειτα, αντικαθιστούμε τα κέντρα των παραθύρων με modes, που ορίζονται ως εξής:

Ορισμός 4.2. (mode ενός συνόλου) Mode του συνόλου $Q = \{Q_1, \dots, Q_n\}$, είναι το διάνυσμα $G = [g_1, \dots, g_m]$ που ελαχιστοποιεί την ποσότητα $D(G, Q) = \sum_{i=1}^n d(Q_i, G)$.

Στον παραπάνω ορισμό, δεν περιορίζουμε το mode ενός συνόλου να περιέχεται σε αυτό. Για τον υπολογισμό του mode ο Huang απέδειξε το παρακάτω θεώρημα [71]:

Θεώρημα 4.1. Έστω $n_{v_{i,j}}$, ο αριθμός των αντικειμένων ενός συνόλου Q , που έχουν την i -στή τιμή $v_{i,j}$, για το χαρακτηριστικό A_j , και $f_r(A_j = v_{i,j}|Q) = \frac{n_{v_{i,j}}}{|Q|}$ η σχετική συχνότητα της τιμής $v_{i,j}$ στο Q . Η συνάρτηση $D(c, Q)$ ελαχιστοποιείται αν και μόνο εάν $f_r(A_j = c_{i,j}|Q) \geq f_r(A_j = v_{i,j}|Q)$ για $c_j \neq v_{i,j}$ για κάθε $j = 1, \dots, m$.

Το θεώρημα αυτό επιτρέπει τον υπολογισμό, του mode, ενός συνόλου σε γραμμικό χρόνο, και μπορεί εύκολα να ενσωματωθεί στον αλγόριθμο k -windows. Πιο συγκεκριμένα, ένα παράθυρο σε κάθε μετακίνηση κεντροποιείται στο mode των σημείων που περιέχει. Αν αυτή η μετακίνηση δεν οδηγεί σε παράθυρο με περισσότερα περισσότερα σημεία, τότε το παράθυρο επαναφέρεται στην αρχική του θέση.

4.2.2 Η Γενικευμένη Διαδικασία Διεύρυνσης

Η διαδικασία διεύρυνσης έχει μια άμεση γενίκευση στην περίπτωση των μετρικών χώρων. Μία διεύρυνση ενός παραθύρου με κέντρο c μεγέθους r , είναι εφικτή μόνο αν μεγενθύνουμε το r . Πιο λεπτομερώς, το r αυξάνεται κατά ένα προκαθορισμένο από το χρήστη ποσοστό θ_c . Έπειτα, εκτελείται η διαδικασία της μετακίνησης. Αν η νέα δοκιμαστική θέση του παραθύρου έχει συμπεριλάβει περισσότερα από θ_c παραπάνω από την προηγούμενη θέση σημεία, τότε η διεύρυνση θεωρείται επιτυχής. Διαφορετικά το παράθυρο επανατοποθετείται στη αρχική του θέση και η διαδικασία της διεύρυνσης τερματίζεται.

4.2.3 Η Γενικευμένη Διαδικασία Συγχώνευσης

Ο αλγόριθμος, έχει την δυνατότητα κατά την διάρκεια τις διαδικασίας να υπολογίσει μια προσέγγιση του πλήθους των ομάδων σε ένα σύνολο δεδομένων. Αυτό επιτυγχάνεται όπως και στην περίπτωση των συνεχών χαρακτηριστικών εξετάζοντας το πλήθος των κοινών σημείων ανάμεσα σε κάθε ζευγάρι παραθύρων που επικαλύπτονται. Η πιθανή επικάλυψη δύο παραθύρων μπορεί εύκολα να επαληθευτεί μέσω της τριγωνικής ανισότητας εφόσον αναφερόμαστε σε μετρικούς χώρους. Έτσι, δύο παράθυρα W_i, W_j με κέντρα c_i, c_j και μεγέθη r_i, r_j , επικαλύπτονται εάν $d(c_i, c_j) \leq r_i + r_j$. Σε κάθε ζευγάρι από επικαλυπτόμενα παράθυρα, υπολογίζεται το πλήθος των κοινών τους στοιχείων

από το σύνολο δεδομένων, και χρησιμοποιώντας δύο προκαθορισμένες από το χρήστη παραμέτρους θ_s (παράμετρος ομοιότητας) και θ_m (παράμετρος ενοποίησης), λαμβάνεται η απόφαση της ενοποίησης. Αν το ποσοστό των κοινών τους σημείων ως προς το πλήθος των συνολικών τους σημείων είναι μεγαλύτερο από την παράμετρο θ_s , τότε διαγράφεται το παράθυρο με τα λιγότερα σημεία. Διαφορετικά εάν το ποσοστό είναι μεγαλύτερο από την παράμετρο θ_m , τα παράθυρα θεωρούνται ότι καλύπτουν μέρος της ίδιας ομάδας. Τέλος εάν το ποσοστό είναι μικρότερο και από το θ_m , τότε τα παράθυρα θεωρούνται ότι εγκλωβίζουν δύο διαφορετικές ομάδες.

4.3 Θέματα Υπολογιστικής Πολυπλοκότητας

Η υπολογιστική πολυπλοκότητα του αλγορίθμου GUKW, καθορίζεται από το κόστος του εντοπισμού των σημείων που βρίσκονται σε ένα συγκεκριμένο παράθυρο. Οι προσεγγίσεις υπολογιστικής γεωμετρίας για την περίπτωση των συνεχών χαρακτηριστικών δεν μπορούν να εφαρμοστούν άμεσα σε κατηγορικά δεδομένα. Αυτό συμβαίνει γιατί μερικές αναγκαίες γεωμετρικές έννοιες, όπως η ελάχιστη καλύπτουσα περιοχή και το μέγεθος μιας περιοχής, δεν ισχύουν στους μετρικούς χώρους [117]. Ωστόσο έχουν προταθεί διάφορες μέθοδοι που ειδικεύονται στην περίπτωση των μετρικών χώρων [70]. Αυτές οι μέθοδοι, μπορούν να διαχωριστούν σε δύο κατηγορίες. Στην πρώτη κατηγορία, μια δομή δεικτοδότησης κατασκευάζεται απευθείας από τις αποστάσεις ανάμεσα στα στοιχεία, ενώ στη δεύτερη κατηγορία τα στοιχεία απεικονίζονται σε έναν διανυσματικό χώρο. Εμπειρικά αποτελέσματα [19, 31, 37, 104, 117] έχουν δείξει ότι ερωτήματα διαστημάτων (range queries), μπορούν να απαντηθούν σε υπο-γραμμικό χρόνο ως προς το μέγεθος του συνόλου δεδομένων.

4.4 Πειραματικά Αποτελέσματα

Η εκτίμηση των αποτελεσμάτων ενός αλγορίθμου ομαδοποίησης, βασίζεται στη υπόθεση της ύπαρξης μιας δομής ομάδων στα δεδομένα. Ο βαθμός στον οποίο το αποτέλεσμα της ομαδοποίησης επαληθεύει αυτήν την υπόθεση, και ανακτά την δομή αυτή, μετρά και την αποτελεσματικότητα της. Σε αυτή την ενότητα, χρησιμοποιούμε το κριτήριο *Εξωτερικής Εκτίμησης* που αναφέρεται στην Ενότητα 2.1, για την μέτρηση της εγχυρότητας μιας ομαδοποίησης.

Αυτό το κριτήριο που επίσης χρησιμοποιείται στην εργασία [71], μετρά το βαθμό αντιστοιχίας ανάμεσα στις ανακαλυφθείσες ομάδες και στις κλάσεις στις οποίες έχουν ανατεθεί τα στοιχεία εκ των προτέρων. Με αυτό τον τρόπο, η ποιότητα της ομαδοποίησης r , ορίζεται ως:

$$r = \frac{\sum_{i=1}^{\#c} a_i}{n},$$

όπου το $\#c$ αναπαριστά το πλήθος των ομάδων· το a_i τον αριθμό των αντικειμένων της κλάσης με τη μεγαλύτερη αντιπροσώπευση στην ομάδα i · και το n το συνολικό πλήθος των στοιχείων.

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι το *Soybean Disease Data Set* και το *1984 United States Congressional Voting Records Database*. Και τα δύο αυτά σύνολα, προέρχονται από το UCI Machine Learning Repository [15], και έχουν χρησιμοποιηθεί ελεύως για την αξιολόγηση παρόμοιων αλγορίθμων [53, 71, 101].

Η Βάση Δεδομένων «The Soybean Disease Data Set»

Αυτό το σύνολο περιέχει 47 στοιχεία αναπαριστώμενα με 35 χαρακτηριστικά. Κάθε στοιχείο, έχει αντιστοιχισμένη μια προκαθορισμένη κλάση ασθένειας: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, και Phytophthora Rot. Υπάρχουν 10 στοιχεία αντιστοιχισμένα στην κάθε μια από τις κλάσεις ασθένειας με την εξαίρεση της κλάσης Phytophthora Rot, στην οποία αντιστοιχούν 17 στοιχεία. Από τα 35 χαρακτηριστικά μόνο τα 21 έχουν παραπάνω από μια διαφορετικές τιμές, και έτσι μόνο αυτά λαμβάνονται υπόψη στην διαδικασία ομαδοποίησης.

Ο αλγόριθμος εφαρμόστηκε σε αυτό το σύνολο δεδομένων, με ένα αρχικό παράθυρο μεγέθους 4. Οι τιμές των άλλων παραμέτρων ήταν 0.1, 0.1, 0.2, 0.8 για τις θ_e , θ_c , θ_m , και θ_s αντίστοιχα.

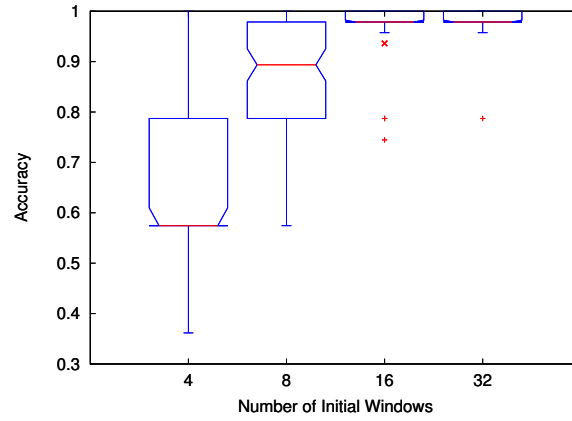
Ο καθορισμός των αρχικών θέσεων των παραθύρων μπορεί να γίνει με διαφορετικούς τρόπους. Ο πιο απλός είναι η τυχαία επιλογή στοιχείων από το σύνολο δεδομένων, και η ανάθεση τους σαν κέντρα του κάθε παραθύρου. Είναι όμως επίσης δυνατόν να επιλεγούν σημεία, από τη δομή δεικτοδότησης, που χρησιμοποιείται για τα ερωτήματα διαστήματος. Στη προτεινόμενη υλοποίηση, χρησιμοποιήσαμε την δομή *Vantage Point Tree* [154]. Με αυτό τον τρόπο είναι δυνατόν να επιλέξουμε αρχικά παράθυρα, χρησιμοποιώντας μια breadth-first επιλογή κόμβων από τη δένδροειδή δομή, σαν κέντρα για τα αρχικά παράθυρα. Επιπρόσθετα, εκτελέσαμε 100 διαφορετικά πειράματα, με την τυχαία μέθοδο αρχικοποίησης, ώστε να συγκρίνουμε τις δύο προσεγγίσεις. Στον πίνακα αντιστοίχησης 4.4, αναφέρονται τα αποτελέσματα του αλγορίθμου GUKW, χρησιμοποιώντας την μέθοδο αρχικοποίησης που χρησιμοποιεί το *Vantage Point Tree*, με 4 αρχικά παράθυρα. Οι ετικέτες PR, DSC, CR, και RRR, αντιστοιχούν στις τέσσερις κλάσεις Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, και Phytophthora Rot.

Αυτό το σχήμα αρχικοποίησης, κατέληξε σε ομαδοποίηση που αντιστοιχεί σε 100%, αντιστοίχιση των κλάσεων, ποσοστό που παρέμεινε σταθερό ακόμα και όταν αυξήθηκε το πλήθος των παραθύρων. Στα boxplots που απεικονίζονται στο Σχήμα 4.1, συνοψίζονται τα αποτελέσματα για την ακρίβεια που επιτεύχθηκε στα 100 πειράματα με την τυχαία αρχικοποίηση. Το κάθε κουτί, έχει γραμμές στο μικρότερο, μεσαίο και υψηλότερο τεταρτημόριο των τιμών. Οι γραμμές που εκτείνονται πέρα από τα όρια των του κουτιού (whiskers), αναδεικνύουν το πλήρες φάσμα των τιμών. Οι τιμές που θεωρούνται κατασκευάσματα (outliers), επισημαίνονται με σταυρούς. Οι τομές σε κάθε κουτί αναπαριστούν μια σθεναρή εκτίμηση γύρω από τη διάμεσο.

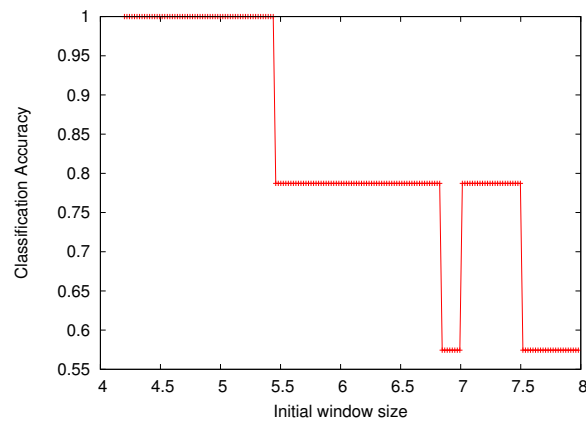
Όπως απεικονίζεται και στο Σχήμα 4.1, χρησιμοποιώντας τυχαία αρχικοποίηση ο αλγόριθμος GUKW, ήταν ικανός να πετύχει το 100%, αντιστοίχησης, όπως με την *Vantage Point Tree*, αρχικοποίηση. Επίσης, καθώς το πλήθος των αρχικών παραθύρων αυξάνει σε 16 και 32, η πιθανότητα να επιτευχθεί ποσοστό αντιστοίχησης 100%, γίνεται πολύ μεγάλη.

Τα αποτελέσματα του αλγορίθμου ήταν σχεδόν ίδια αλλάζοντας τις παραμέτρους, με την εξαίρεση του μεγέθους του αρχικού παραθύρου, η επιλογή του οποίου επηρεάζει σημαντικά τα αποτελέσματα. Για αυτό το σκοπό στο Σχήμα 4.2, παρουσιάζονται τα ποσοστά αντιστοίχησης, για όλες τις τιμές αυτής της παραμέτρου στο διάστημα [4, 8] με βήμα 0.2.

Για συγκριτικούς σκοπούς, εκτελέσαμε πειράματα, για τους αλγορίθμους GDBSCAN, και PDDP [17]. Οι παράμετροι του GDBSCAN είχαν τιμές $Epts = 2.0$, $Minpts = 4$, μια και αυτές έδωσαν τα καλύτερα αποτελέσματα μετά από εκτενή αναζή-



Σχήμα 4.1: Ποσοστά αντιστοίχισης του αλγορίθμου GUKW με τυχαία αρχικοποίηση χρησιμοποιώντας 4,8,16 και 32 αρχικά παράθυρα.



Σχήμα 4.2: Ποσοστά αντιστοίχισης του αλγορίθμου GUKW για διάφορα αρχικά μεγέθη των παραθύρων.

τηση. Καθώς ο αλγόριθμος PDDP δεν μπορεί να υπολογίζει μόνος του το πλήθος των ομάδων δόθηκε η σταθερή τιμή 4 που είναι και η πραγματική (τα αποτελέσματα δεν βελτιώθηκαν για μεγαλύτερες τιμές). Στον Πίνακα 4.4, παρουσιάζονται τα αποτελέσματα αντιστοίχισης, για τους δύο αλγορίθμους. Ο αλγόριθμος GDBSCAN, κατέληξε σε

	Ασθένεια			
	PR	DSC	CR	RRR
Ομάδα 1	17	0	0	0
Ομάδα 2	0	10	0	0
Ομάδα 3	0	0	10	0
Ομάδα 4	0	0	0	10
Ποσοστό Αντιστοίχισης: 100%				

Πίνακας 4.1: Πίνακας αντιστοίχισης των αποτελέσματα του αλγορίθμου GUKW, χρησιμοποιώντας Vantage Point Tree μέθοδο αρχικοποίησης και 4 αρχικά παράθυρα.

μια ομαδοποίηση με 5 ομάδες, επιτυγχάνοντας αντιστοίχιση 100%, αλλά μόνο για το 91%, των δεδομένων καθώς θεωρεί 4 αντικείμενα σαν θόρυβο.

	GDBSCAN				PDDP			
	Ασθένεια				Ασθένεια			
	PR	DSC	CR	RRR	PR	DSC	CR	RRR
Ομάδα 1	0	10	0	0	9	0	0	5
Ομάδα 2	0	0	10	0	8	0	0	5
Ομάδα 3	0	0	8	0	0	0	10	0
Ομάδα 4	11	0	0	0	0	10	0	0
Ομάδα 5	4	0	0	0				
Θόρυβος	2	0	0	2				
Ποσοστό αντιστοίχισης: 100%					Ποσοστό αντιστοίχισης: 78.72%			

Πίνακας 4.2: Πίνακας αντιστοίχισης για τους αλγορίθμους GDBSCAN και PDDP για το soybean disease σύνολο δεδομένων.

Ο αλγόριθμος GUKW, κατάφερε να καταλήξει σε ομαδοποίησης με μεγάλο ποσοστό αντιστοίχισης. Επιπρόσθετα, η αρχικοποίηση με τη μέθοδο vantage point tree, δίνει μια σθεναρότητα στα αποτελέσματα, που για να επιτευχθεί με την τυχαία αρχικοποίηση απαιτεί μεγάλο πλήθος παραθύρων.

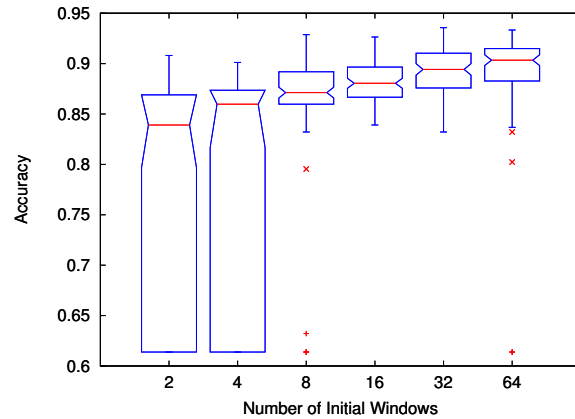
Η Βάση Δεδομένων «The 1984 United States Congressional Voting Records Database»

Στη βάση αυτή το σύνολο δεδομένων περιλαμβάνει, τους ψήφους των U.S. House of Representatives Congressmen που διακρίνονται με 16 διαφορετικά χαρακτηριστικά. Συνολικά υπάρχουν 435 στοιχεία, από τα οποία 267 έχουν αντιστοιχηθεί σαν *democrats*, ενώ τα υπόλοιπα 168 σαν *republicans*.

Παρόμοια με την προηγούμενη περίπτωση, εκτελέστηκαν 100 πειράματα με την μέθοδο της τυχαίας αρχικοποίησης. Τα αποτελέσματα συνοψίζονται στα boxplots του Σχήματος 4.3. Όπως φαίνεται για περισσότερα από 16 αρχικά παράθυρα το ποσοστό αντιστοίχισης σταθεροποιείται γύρω στο 90%. Στον πίνακα αντιστοίχισης 4.4, αναφέρονται τα αποτελέσματα του αλγορίθμου GUKW, χρησιμοποιώντας την vantage point tree αρχικοποίηση με 32 αρχικά παράθυρα. Ο αλγόριθμος επιτυγχάνει να ανακαλύψει τις δύο ομάδες που περιέχουν το μεγαλύτερο πληθυσμό από τις δύο κλάσεις. Τρεις ακόμη ομάδες αναγνωρίζονται που περιέχουν κυρίως στοιχεία από την κλάση *democrats*. Ακόμη μία από τις ομάδες περιέχει μόνο 3 στοιχεία και μπορεί να θεωρηθεί σαν θόρυβος.

Στο Σχήμα 4.4, απεικονίζεται η επίδραση του αρχικού μεγέθους του παραθύρου, στο ποσοστό αντιστοίχισης, για μεγέθη στο διάστημα [3, 6] με βήμα 0.15.

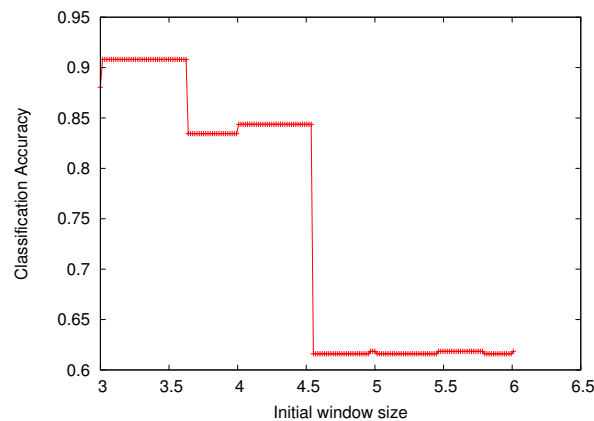
Τέλος, για λόγους σύγκρισης στον Πίνακα 4.4, παρουσιάζουμε τα αποτελέσματα του ποσοστού αντιστοίχισης για τους αλγορίθμους GDBSCAN και PDDP. Οι παράμετροι για τον GDBSCAN πήραν τιμές $Epts = 3.0$, $Minpts = 20$, καθώς αυτές επέφεραν τα καλύτερα αποτελέσματα μετά από εκτενή αναζήτηση. Εκτελώντας πειράματα με διαφορετικές τιμές του πλήθους των ομάδων για τον PDDP αλγόριθμο, παρατηρήθηκε ότι το καλύτερο ποσοστό αντιστοίχισης υπήρχε για 5 τελικές ομάδες, οπότε και χρησιμοποιήθηκε αυτό. Το αποτέλεσμα και των δύο αυτών αλγορίθμων



Σχήμα 4.3: Ποσοστά αντιστοίχησης του αλγορίθμου GUKW με τυχαία αρχικοποίηση χρησιμοποιώντας 2,4,8,16,32 και 64 αρχικά παράθυρα.

	Κλάση	
	Democrat	Republican
Ομάδα 1	28	6
Ομάδα 2	26	7
Ομάδα 3	8	1
Ομάδα 4	21	148
Ομάδα 5	1	2
Ομάδα 6	183	4
Ποσοστό αντιστοίχησης: 90.84%		

Πίνακας 4.3: Πίνακας αντιστοίχησης των αποτελέσματα του αλγορίθμου GUKW, χρησιμοποιώντας Vantage Point Tree μέθοδο αρχικοποίησης και 32 αρχικά παράθυρα.



Σχήμα 4.4: Ποσοστά αντιστοίχησης του αλγορίθμου GUKW για διάφορα αρχικά μεγέθη των παραθύρων.

επαληθεύουν την εύρεση της πραγματικής δομής των δεδομένων από τον αλγόριθμο GUKW. Σε κάθε κλάση, υπάρχει μία ομάδα, με το μεγαλύτερο ποσοστό στοιχείων από αυτήν. Επίσης υπάρχουν και ένα πλήθος διάσπαρτων ομάδων που κυρίως περιέχουν

στοιχεία από την κλάση democrat.

	GDBSCAN		PDDP	
	Κλάση		Κλάση	
	Democrat	Republican	Democrat	Republican
Ομάδα 1	193	6	61	2
Ομάδα 2	27	144	34	51
Ομάδα 3			18	103
Ομάδα 4			16	7
Ομάδα 5			138	5
Ομάδα 6				
Θόρυβος	47	18		
	Ποσοστό αντιστοίχισης: 90.11%		Ποσοστό αντιστοίχισης: 84.83%	

Πίνακας 4.4: Πίνακας αντιστοίχισης για τους αλγορίθμους GDBSCAN και PDDP για το Voting Records Database σύνολο δεδομένων.

4.5 Συμπεράσματα - Συνεισφορά

Καθώς οι τεχνολογίες των βάσεων δεδομένων αναπτύσσονται, δημιουργείται ένας αυξανόμενος αριθμός από δύσκολα και πολύπλοκα προβλήματα εξόρυξης πληροφορίας. Ένα συχνό χαρακτηριστικό αυτών των δεδομένων είναι η περιγραφή του κάθε στοιχείου από ανάμειξη συνεχών και κατηγορικών τιμών. Οι τυπικοί αλγόριθμοι ομαδοποίησης θεωρούν ότι τα δεδομένα αποτελούνται από συνεχείς τιμές. Για αυτό το λόγο αλγόριθμοι που μπορούν να εφαρμοστούν απευθείας σε αυτά τα δεδομένα είναι αναγκαίοι.

Σε αυτό το κεφάλαιο παρουσιάσαμε μια γενίκευση του αλγορίθμου k -windows σε μετρικούς χώρους. Το πολύ σημαντικό χαρακτηριστικό του υπολογισμού του πλήθους των ομάδων διατηρήθηκε κατά την γενίκευση. Για να επιτευχθεί υπό γραμμικός χρόνος εκτέλεσης, χρησιμοποιήθηκαν τεχνικές που προτάθηκαν στις εργασίες [70,71]. Τα πειραματικά αποτελέσματα έδειξαν ότι ο αλγόριθμος είναι ικανός να καταλήξει σε ομαδοποιήσεις υψηλής ακρίβειας. Επιπλέον αν χρησιμοποιηθεί για την καθοδήγηση των θέσεων των αρχικών παραθύρων, η δομή δεικτοδότησης, η απόδοση του αλγορίθμου γίνεται πιο σθεναρή με μικρό αριθμό παραθύρων.

Τροποποιήσεις του αλγορίθμου k -windows

Η θεωρία σου είναι τρελή,
αλλά όχι αρκετά τρελή για να είναι αληθινή.

—Niels Bohr (1885–1962)

Σε αυτό το κεφάλαιο παρουσιάζουμε δύο τροποποιήσεις του αλγορίθμου UKW, που στοχεύουν στην ανακάλυψη καλύτερων ομαδοποιήσεων. Και οι δύο αυτές τροποποιήσεις επιχειρούν να αξιοποιήσουν πληροφορία από τα τοπικά χαρακτηριστικά των δεδομένων. Στις επόμενες ενότητες μελετάται κάθε μία ξεχωριστά, και παρουσιάζονται πειραματικά αποτελέσματα που αναδεικνύουν τα προτερήματα και μειονεκτήματα τους. Το κεφάλαιο ολοκληρώνεται με τα συμπεράσματα που προέκυψαν και την συνεισφορά τους.

5.1 Ομαδοποίηση με τη Χρήση Συνόλων Μορφοκλασματικής Δομής

Παρατηρώντας απλά τα αντικείμενα γύρω μας, μπορούμε εύκολα να δούμε ότι τα περισσότερα από αυτά έχουν δομές πολύπλοκες και ακανόνιστες [64, 110, 124]. Παρόλο που τα περισσότερα ανθρώπινα κατασκευάσματα μπορούν να περιγράφουν από απλές γεωμετρικές δομές, η κλασική γεωμετρία αδυνατεί να περιγράψει τα περισσότερα αντικείμενα που φτιάχνει η φύση. Η ανάγκη ενός νέου μοντέλου για να περιγράψει τέτοια ακανόνιστη συμπεριφορά αντιμετωπίστηκε πρώτα από τον Mandelbrot [98], ο οποίος εισήγαγε τη έννοια των Συνόλων Μορφοκλασματικής Δομής (*fractals*). Ένα σύνολο λέγεται fractal, εάν η Hausdorff-Besicovitch διάσταση αυτού είναι αυστηρά μεγαλύτερη από την τοπολογική του διάσταση [124]. Τα σύνολα fractal, χαρακτηρίζονται από την διάσταση fractal, η οποία μετράει την πολυπλοκότητα τους. Ο Mandelbrot, πρώτος περιέγραψε μια προσέγγιση για τον υπολογισμό της fractal διάστασης, κατά τον υπολογισμό του μήκους μιας ακτογραμμής.

Ένα καθιερωμένος τρόπος υπολογισμού της διάστασης fractal ενός συνόλου είναι η μέθοδος box-counting [88]. Αυτή η μέθοδος, για ένα σύνολο δεδομένων n στοιχείων υποσύνολο του $\mathbb{R}^d$, και μια διαμέριση του χώρου σε κελιά εύρους l , υπολογίζει την fractal διάσταση D από τον τύπο:

$$D = -\lim_{l \rightarrow 0} \frac{\log_{10} N(l)}{\log_{10} l},$$

όπου το $N(l)$ αναπαριστά το πλήθος των κελιών που περιλαμβάνουν τουλάχιστον ένα σημείο του συνόλου.

Αλγόριθμοι ομαδοποίησης που χρησιμοποιούν την διάσταση fractal των συνόλων, έχουν προταθεί και στο παρελθόν. Οι Barbarä και Chen [10], πρότειναν έναν αλγόριθμο ομαδοποίησης πλέγματος, που χρησιμοποιεί την διάσταση fractal για να ομαδοποιήσει δεδομένα. Ο αλγόριθμος χρησιμοποιεί μια ευριστική τεχνική για την αρχικοποίηση των ομάδων, και σταδιακά προσθέτει στοιχεία στις ομάδες, όσο δεν αλλάζει η fractal διάσταση της κάθε ομάδας. Επίσης, στην εργασία [114], προτείνεται μια μέθοδος βασισμένη στη διάσταση fractal, για την ομαδοποίηση δισδιάστατων δεδομένων [114]. Και οι δύο αυτοί αλγόριθμοι έχουν το βασικό μειονέκτημα του καθορισμού από το χρήστη του πλήθους των ομάδων.

Σε αυτό το κεφάλαιο, ενσωματώνουμε τον υπολογισμό της διάστασης fractal, στον αλγόριθμο UKW [137], και δείχνουμε πως μπορεί να χρησιμοποιηθεί για την εξαγωγή ποιοτικής πληροφορίας από τα δεδομένα. Στη συνέχεια του κεφαλαίου αφού παρουσιάσουμε την προτεινόμενη επέκταση, θα δώσουμε μερικά ενδιαφέροντα πειραματικά αποτελέσματα και θα κλείσουμε το κεφάλαιο σχολιάζοντας την όλη προσέγγιση.

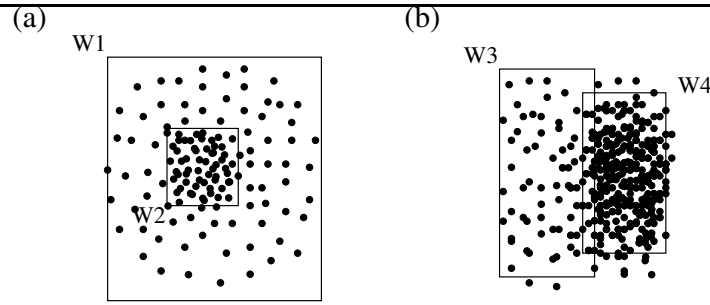
5.1.1 Ενσωμάτωση της fractal Διάστασης, στον Αλγόριθμο k -windows

Σε αυτή την ενότητα περιγράφουμε, την τροποποίηση του αλγορίθμου UKW, που καθοδηγεί τις διαδικασίες μετακίνησης, διεύρυνσης και συγχώνευσης, χρησιμοποιώντας την προσέγγιση της fractal διάστασης των σημείων που περιέχονται σε κάθε παράθυρο. Αυτό γίνεται με τη χρήση της μεθόδου box-counting σε κάθε βήμα. Η μέθοδος αυτή ταιριάζει απόλυτα με τη φιλοσοφία του αλγορίθμου UKW καθώς, χρησιμοποιεί κουτιά τα οποία μπορούν να προέλθουν άμεσα από τον τεμαχισμό των παραθύρων του αλγορίθμου UKW.

Πιο συγκεκριμένα, η διαδικασία τις μετακίνησης και της διεύρυνσης, αλλάζουν το παράθυρο μόνο αν αυτό οδηγεί σε αμελητέα αλλαγή της fractal διάστασης των σημείων. Επίσης η διαδικασία συγχώνευσης, δεν ενοποιεί παράθυρα με διαφορετική διάσταση fractal, ακόμη και αν το ένα περιλαμβάνεται μέσα στο άλλο. Αυτό δίνει την δυνατότητα στον αλγόριθμο να εντοπίσει ομάδες, όπου η πυκνότητα των σημείων στο κέντρο της ομάδας είναι σημαντικά μεγαλύτερη από ότι στις περιοχές πιο μακριά από το κέντρο. Με αυτό τον τρόπο, ο αλγόριθμος εντοπίζει το κέντρο της ομάδας πιο αποδοτικά, και είναι ικανός να αναγνωρίσει ομάδες με διαφορετικά ποιοτικά χαρακτηριστικά ακόμη και μέσα στην ίδια ομάδα. Για παράδειγμα στο Σχήμα 5.1, οι διαδικασίες μετακίνησης και μεγέθυνσης εμποδίζουν το παράθυρο $W3$, να συμπεριλάβει, το δεξί μέρος της ομάδας, καθώς η διάστασης fractal, αυτής της περιοχή είναι πολύ μεγαλύτερη (Σχήμα 5.1(b)). Όμοια, το παράθυρο $W4$, εμποδίζεται από να συμπεριλάβει το αριστερό μέρος της ομάδας. Η προτεινόμενη τροποποίηση του αλγορίθμου, επίσης αναγνωρίζει ότι παρόλο που τα παράθυρα έχουν πολλά κοινά σημεία (Σχήμα 5.1(a)), η διαφορά στην τιμή της διάστασης fractal, είναι αρκετά σημαντική, και έτσι τα θεωρεί σαν δύο διαφορετικές περιοχές της ίδιας ομάδας.

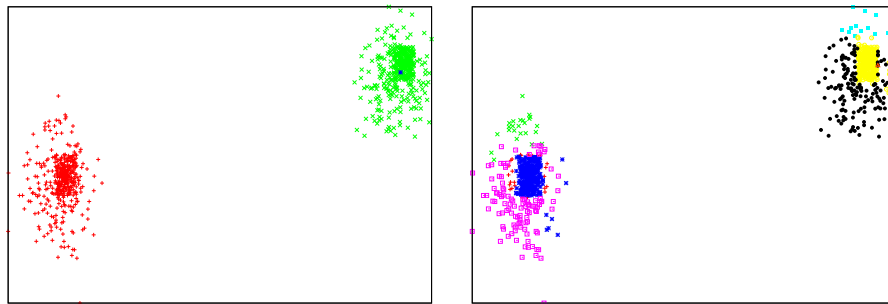
5.1.2 Πειραματικά Αποτελέσματα

Για να παρουσιάσουμε το αποτέλεσμα της προτεινόμενης τροποποίησης, αρχικά χρησιμοποιούμε το τεχνητό δισδιάστατο σύνολο δεδομένων $Dset_1$, που περιέχει 1200



Σχήμα 5.1: Ομάδες με περιοχές με διαφορετική πυκνότητα. Η προτεινόμενη τροποποίηση είναι ικανή να εντοπίσει διαφορετικές περιοχές ακόμη και μέσα στην ίδια ομάδα.

σημεία. Το σύνολο αυτό καθώς και το αποτέλεσμα, του αλγορίθμου UKW και της προτεινόμενης τροποποίησης, απεικονίζεται στο Σχήμα 5.2 (διαφορικά χρώματα αντιστοιχούν σε διαφορετικές ομάδες). Όπως είναι προφανές και οι δύο αλγόριθμοι ανακαλύπτουν τα κέντρα των δύο κύριων ομάδων. Ωστόσο η προτεινόμενη τροποποίηση καταφέρνει και εντοπίζει και τις περιοχές γύρω από το κέντρο των ομάδων με μικρότερη πυκνότητα.



Σχήμα 5.2: Το σύνολο δεδομένων $Dset_1$ με τα αποτελέσματα του αλγορίθμου UKW (αριστερά) και της προτεινόμενης τροποποίησης (δεξιά).

Στο επόμενο βήμα χρησιμοποιούμε το σύνολο δεδομένων Iris, $Dset_{iris}$, από το UCI machine learning repository [15]. Αυτό το σύνολο δεδομένων είναι από τα πιο διαδεδομένα, και περιέχει 150 στοιχεία, αναπαριστούμενα από 4 χαρακτηριστικά. Το κάθε χαρακτηριστικό είναι μέτρηση μήκους ή πλάτους του πέταλου ή του σέπαλου των τριών τύπων του φυτού iris. Τα 150 στοιχεία έχουν αντιστοιχηθεί ομοιόμορφα στις 3 διαφορετικές κλάσεις Setosa, Versicolour και Virginica, του φυτού.

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιούμε πάλι το κριτήριο *Εξωτερικής Εκτίμησης* που αναφέρεται στην Ενότητα 2.1. Οι πίνακες αντιστοίχισης 5.1.2, αναγράφουν το πλήθος των στοιχείων από την κάθε κλάση που περιλαμβάνονται στην κάθε ομάδα. Και οι δύο αλγόριθμοι, καταφέρνουν να αναγνωρίσουν, τις τρεις ομάδες οπότε η φυσική δομή των δεδομένων ανακαλύπτεται. Η μόνη διαφορά βρίσκεται στο ότι η τροποποιημένη μέθοδος αντιστοιχεί σε λάθος ομάδες, 2 λιγότερα στοιχεία από τον αλγόριθμο UKW.

Το τελευταίο σύνολο δεδομένων, $Dset_{eq}$, είναι ένα διδιάστατο σύνολο των γεωγραφικών μηκών και πλατών, των σεισμών με ισχύ μεγαλύτερη από 4 στην κλίμακα Richter, που συνέβησαν στην Ελλάδα την περίοδο 1983–2003. Το σύνολο δεδομένων

	Αλγόριθμος UKW			Προτεινόμενη τροποποίηση		
	Κλάση			Κλάση		
	Setosa	Versicolour	Virginica	Setosa	Versicolour	Virginica
Ομάδα 1	50	0	0	50	0	0
Ομάδα 2	0	46	4	0	46	4
Ομάδα 3	0	4	46	0	2	48

Πίνακας 5.1: Πίνακες αντιστοίχισης για το σύνολο δεδομένων $D_{set_{iris}}$.

αυτό παρέχεται από το Εθνικό Παρατηρητήριο Αθηνών [21], και δεν χρησιμοποιήθηκε για να δώσει νέα στοιχεία για το φαινόμενο των σεισμών, αλλά για να αναδείξει τη χρήση του αλγορίθμου σε ένα σύνολο δεδομένων πραγματικού κόσμου.

Στο Σχήμα 5.3 απεικονίζονται τα αποτελέσματα του αλγορίθμου UKW, στο σύνολο $D_{set_{eq}}$, ενώ στο Σχήμα 5.4, απεικονίζονται τα αποτελέσματα της προτεινόμενης τροποποίησης. Και στα δύο σχήματα αντιστοιχίζονται διαφορετικά χρώματα σε διαφορετικές ομάδες. Όπως και στις προηγούμενες περιπτώσεις ο τροποποιημένος αλγόριθμος, καταφέρνει να διαχωρίσει ομάδες που είχαν συγχωνευτεί από τον αρχικό αλγόριθμο. Στο Σχήμα 5.4, χαρακτηριστικά τέτοια παραδείγματα σημειώνονται με μαύρα τετράγωνα. Επίσης η τοποθέτηση της ακτογραμμής της Ελλάδας είναι προσεγγιστική και σκοπός της είναι να βοηθήσει στην καλύτερη οπτικοποίηση του αποτελέσματος. Αυτή είναι μια πρωταρχική διερεύνηση της εφαρμογής αλγορίθμων ομαδοποίησης σε δεδομένα σεισμών. Μια εκτενή διερεύνηση απαιτεί τη χρήση περαιτέρω πληροφορίας όπως ισχύς, εστιακό βάθος και χρόνο.

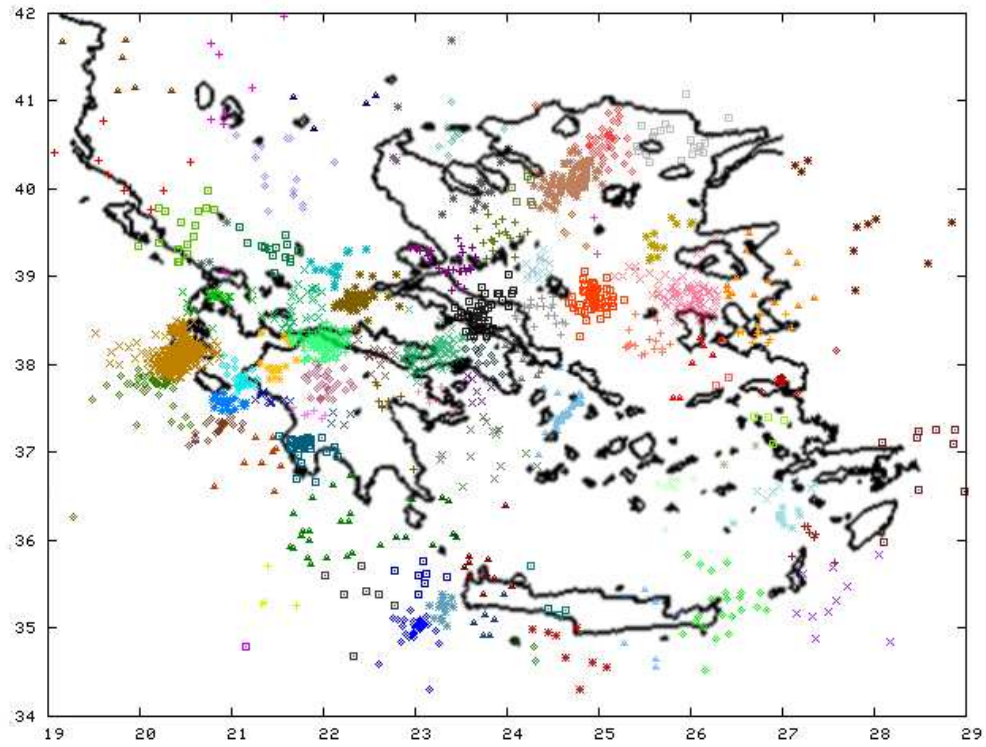
5.2 Προσανατολισμένος Αλγόριθμος k -windows

Σε αυτή την ενότητα παρουσιάζουμε ένα γενικό πλαίσιο που μας προσφέρει την δυνατότητα να τροποποιήσουμε τον αλγόριθμο UKW, με στόχο την βελτίωση της αποτελεσματικότητας του [138, 139]. Η τεχνική που χρησιμοποιείται για το σκοπό αυτό είναι η PCA [78], υλοποιημένη μέσω της διάσπασης σε ιδιάζουσες τιμές (Singular Value Decomposition (SVD)) [60].

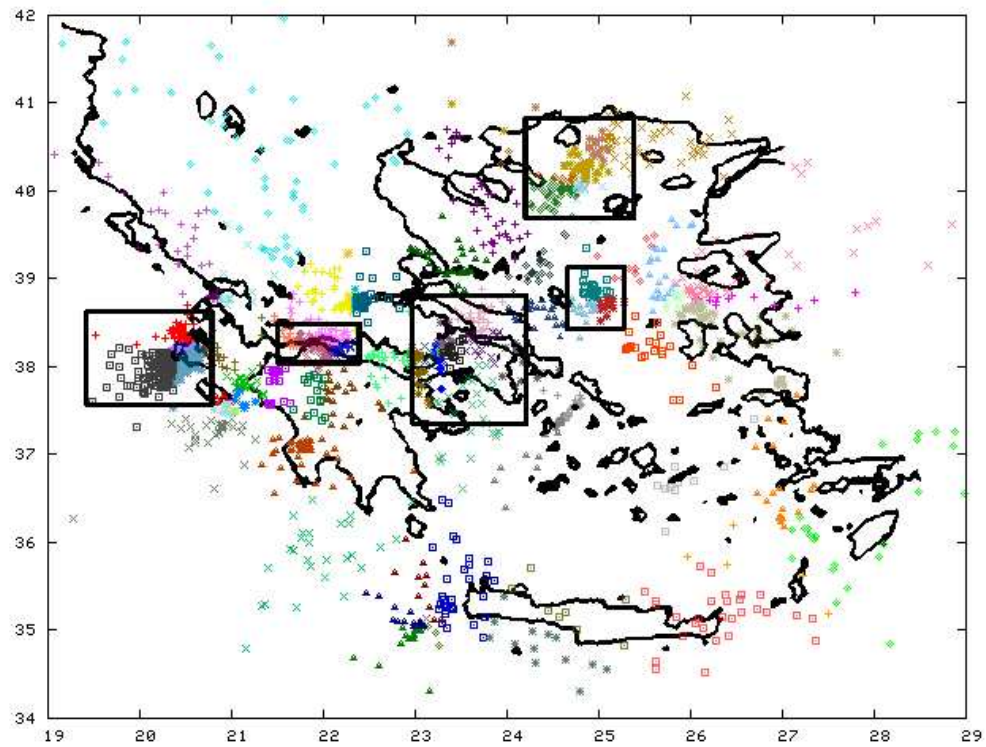
5.2.1 Ο αλγόριθμος k -windows καθοδηγούμενος από την τεχνική PCA

Παρόλο που η γενική ιδέα του αλγορίθμου k -windows, να μετακινούνται και να διευρύνονται τα παράθυρα, είναι πολύ αποτελεσματική (Κεφάλαιο 3), αυτές οι μετακινήσεις και διευρύνσεις περιορίζονται στο να είναι παράλληλες με τους Καρτεσιανούς άξονες. Σε αυτή την ενότητα, διερευνούμε την δυνατότητα των παραθύρων να προσαρμόζουν όχι μόνο τη θέση και το μέγεθος τους αλλά και τον προσανατολισμό τους, με στόχο την επιτυχέστερη κάλυψη των ομάδων.

Ας υποθέσουμε για παράδειγμα ότι k d -διάστατα παράθυρα έχουν αρχικοποιηθεί από τον αλγόριθμο. Μετά την εκτέλεση, της διαδικασίας της μετακίνησης, έχουν αλλάξει τόσο το κέντρο όσο και τα σημεία που εσωκλείονται στο παράθυρο. Θα ήταν λοιπόν κατάλληλο, να επάνα-προσανατολίσουμε τα παράθυρα από την αρχική τους θέση (παράλληλη με τους άξονες), έτσι ώστε να ληφθούν υπόψη οι τοπικές διαφοροποιήσεις των δεδομένων. Αυτό μπορεί να επιτευχθεί μέσω της PCA τεχνικής, εφαρμοσμένης σε



Σχήμα 5.3: Το αποτέλεσμα του αλγορίθμου UKW στο σύνολο δεδομένων $Dset_{eq}$.

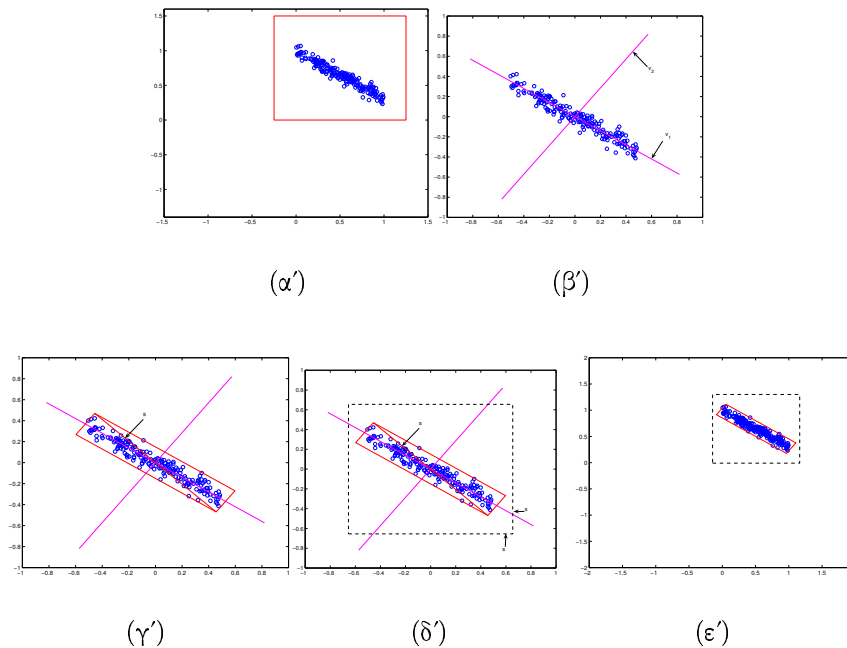


Σχήμα 5.4: Το αποτέλεσμα της προτεινόμενης τροποποίησης στο σύνολο δεδομένων $Dset_{eq}$.

μια μετάφραση του συνόλου των σημείων που περιλαμβάνονται στο εκάστοτε παράθυρο.

Έστω για παράδειγμα, $A^{(i)} \in \mathbb{R}^{n_i \times d}$ το μητρώο των στοιχείων-χαρακτηριστικών που αντιστοιχεί σε ένα υποσύνολο στοιχείων $\mathcal{P}$ του συνόλου δεδομένων $\mathcal{Q}$. Το μητρώο $B^{(i)} := A^{(i)} - \frac{1}{n_i} e e^T A^{(i)}$, όπου $e \in \mathbb{R}^{n_i}$ είναι ένα διάνυσμα αποτελούμενο από μονάδες, περιέχει τα σημεία του $\mathcal{P}$, σε σχέση με τους Καρτεσιανούς άξονες μετακινούμενους στο κέντρο $g = \frac{1}{n_i} e^T A^{(i)}$ του $\mathcal{P}$. Έστω επίσης η ανάλυση SVD του $B^{(i)}$, να είναι $B^{(i)} = U^{(i)} \Sigma^{(i)} (V^{(i)})^T$, και τα $\{u_j, \sigma_j, v_j\}, j = 1, \dots, d$, οι μη τετριμμένες ιδιάζουσες τριπλέτες του $B^{(i)}$, σε φθίνουσα σειρά, δηλαδή $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d \geq 0$. Χάριν ευκολίας παραλείπουμε τον δείκτη i . Ένας τρόπος κατασκευής ενός υπερ-παραλληλεπίπεδου, που να περιλαμβάνει τα στοιχεία του $\mathcal{P}$, είναι να χρησιμοποιηθεί το γεγονός, ότι οι στήλες του V_i είναι οι κύριες κατευθύνσεις των στοιχείων του $\mathcal{P}$. Το υπερ-ελλειψοειδές με ημιάξονες σ_j περιλαμβάνει όλα τα στοιχεία του $\mathcal{P}$. Ωστόσο, μπορεί να περιλαμβάνει ένα πολύ μεγαλύτερο χώρο, από αυτό που θα περιλάμβανε το ελάχιστο υπερ-ελλειψοειδές που περιέχει όλα αυτά τα σημεία.

Από την άλλη μεριά, ο υπολογισμός τω ελάχιστων αυτών σωμάτων είναι δύσκολος [86]. Αντίθετα, εάν προτιμήσουμε να παραμείνουμε στα παραλληλεπίπεδα, μπορούμε να χρησιμοποιήσουμε την σχέση $B^{(i)} V^{(i)} = U^{(i)} \Sigma^{(i)}$. Αυτή η σχέση δείχνει ότι οι προβολές των n_i γραμμών του $B^{(i)}$ σε οποιαδήποτε κύριο άξονα, έστω v_j , δίνονται από τα στοιχεία των στηλών j του $U^{(i)}$ πολλαπλασιασμένα με το σ_j . Με αυτόν τον τρόπο, μπορούμε να κατασκευάσουμε ένα υπερ-παραλληλεπίπεδο, με κέντρο το g και μέγεθος άκμων $2\sigma_j \|u_j\|_\infty$, που περιέχει όλα τα σημεία του $\mathcal{P}$.



Σχήμα 5.5: Τα αρχικά δεδομένα και το παράθυρο (α') μετακινούνται και υπολογίζονται οι κύριες κατευθύνσεις (β'). Υπολογίζουμε το παραλληλεπίπεδο προσανατολισμένο στις κύριες διευθύνσεις (γ') και τον αντίστοιχο κύβο (Ενότητα 5.2.2) (δ'). Τέλος, μετακινούμε τα δεδομένα στην αρχική τους θέση (ε').

Ένα παράδειγμα αυτής της διαδικασίας, απεικονίζεται στο Σχήμα 5.5. Τα αρχικά τα σημεία κεντροποιούνται, και υπολογίζονται οι κύριες διευθύνσεις του αντίστοιχου μητρώου. Έπειτα, υπολογίζεται το προσανατολισμένο στις κύριες διευθύνσεις υπερ-

παραλληλεπίπεδο, με μήκος ακμών $2\sigma_j \|u_j\|_\infty, j = 1, \dots, d$. Τέλος, υπολογίζουμε ένα κατάλληλο υπερκύβο (Ενότητα 5.2.2), και μετακινούμε τα δεδομένα στην αρχική τους θέση. Μετά από αυτό το βήμα εκκινείται η διαδικασία της διεύρυνσης, σε κάθε διάσταση του προσανατολισμένου όμως τώρα παραθύρου. Ονομάζουμε τον αλγόριθμο UKW, με αυτήν την τροποποίηση *Oriented k -Windows (OkW)* ο οποίος συνοψίζεται στα ακόλουθα βήματα:

Αλγόριθμος OkW $(k, a, \theta_e, \theta_c, \theta_s, \theta_m,) \{$
Εκτέλεσε $\mathcal{W} = \text{αρχικοποίηση}(k, a)$.
Για κάθε παράθυρο W_j στο $\mathcal{W}$ **εκτέλεσε**:
 Επανέλαβε:
 Εκτέλεσε $W_j \leftarrow \text{μετακίνηση}(W_j)$.
 Εκτέλεσε $W_j \leftarrow \text{SVD-διεύρυνση}(\theta_e, \theta_c, W_j)$.
 Μέχρι να μη αλλάζουν τα όρια και το κέντρο του παραθύρου W_j .
Εκτέλεσε συγχώνευση $(\theta_m, \theta_s, \mathcal{W})$
Επέστρεψε τις ομάδες $\{c_{l1}, c_{l2}, \dots \text{ όπου:}$
 $c_{li} = \{i : i \in W_j, \text{label}(W_j) = li\} \}$.
Function SVD-διεύρυνση $(\theta_e, \theta_c, W) \{$
 Επανέλαβε
 Για κάθε συντεταγμένη i
 Επανέλαβε:
 Υπολόγισε το σύνολο $Q_W \subset X$ των σημείων του X ,
 που περιέχονται στο W .
 Υπολόγισε τις κύριες κατευθύνσεις V του Q_W .
 Υπολόγισε το προσανατολισμένο παράθυρο W_o .
 Διεύρυνε κατά θ_e το W_o κατασκευάζοντας το W_t .
 Εκτέλεσε $W'_t \leftarrow \text{μετακίνηση}(W_t)$.
 Υπολόγισε το σύνολο $Q_{W'_t} \subset X$ των σημείων του X ,
 που περιέχονται στο W'_t .
 Εάν $\theta_c |Q_W| < Q_{W'_t}$, τότε θέσε $W \leftarrow W'_t$.
 Διαφορετικά θεώρησε ανεπιτυχή την προσπάθεια διεύρυνσης.
 Μέχρι $\theta_c |Q_W| \geq Q_{W'_t}$.
 Μέχρι οι προσπάθειες διεύρυνσης να είναι ανεπιτυχείς για όλα τα $i = 1, \dots, d$.
 Επέστρεψε το παράθυρο W .
}

5.2.2 Υπολογιστικά Θέματα

Μια σημαντική πρόκληση που πρέπει να αντιμετωπίσει ο αλγόριθμος OkW, είναι η αποδοτική διαχείριση των μη-παράλληλων με τους άξονες παραθύρων. Αντίθετα με την περίπτωση των παράλληλων με τους άξονες παραθύρων, όπου η συμμετοχή ή όχι ενός d -διάστατου σημείου στο παράθυρο μπορεί να αντιμετωπιστεί με την ικανοποίηση $2d$ ανισοτήτων, στην περίπτωση των προσανατολισμένων παραθύρων, επιβάλεται η αποθήκευση 2^d κορυφών. Για την αντιμετώπιση αυτού του θέματος, χρησιμοποιούμε μια τεχνική της Υπολογιστικής Γεωμετρίας σχεδιασμένη να αντιμετωπίσει ερωτήματα για μη παράλληλα με τους άξονες ευθύγραμμα τμήματα, όπου κάθε τμήμα εμβυθίζεται σε ένα παράλληλο με τους άξονες παραλληλεπίπεδο [33]. Όμοια και στην περίπτωση

του OkW, κάθε υπερ-παραλληλεπίπεδο εμβυθίζεται σε ένα παράλληλο με τους άξονες υπερκύβο, με το ίδιο κέντρο c , και μήκος ακμών s , ίσο με τη μέγιστη διαγώνιο του υπερ-παραλληλεπίπεδου. Αυτό το μήκος ακμών εγκυάται ότι ο υπερκύβος περιέχει πλήρως το προσανατολισμένο υπερ-παραλληλεπίπεδο. Με αυτόν τον τρόπο, ο υπερκύβος που κωδικοποιείται μόνο με το ζεύγος (c, s) , χρησιμοποιείται για τα ερωτήματα Ορθογώνιας Αναζήτησης Διαστήματος, αναδεικνύοντας πιθανά σημεία από το σύνολο δεδομένων, που μπορεί να περιέχονται στο προσανατολισμένο υπερ-παραλληλεπίπεδο. Για να υπολογίσουμε, αν κάθε ένα από αυτά τα σημεία ανήκει στο προσανατολισμένο υπερ-παραλληλεπίπεδο, αρκεί να ελέγξουμε την ορθογώνια προβολή του στις d κύριες κατευθύνσεις. Ένα τα $l_j, j = 1, \dots, d$, αναπαριστούν, τα μήκη των ακμών, του υπερ-παραλληλεπίπεδου, τότε το σημείο x εσωκλείεται σε αυτό εάν ικανοποιεί όλες τις ανισότητες $|x^\top v_j| \leq \frac{l_j}{2}$ για $j = 1, \dots, d$. Με αυτό τον τρόπο, χρειάζεται μόνο να αποθηκεύονται $O(d)$ στοιχεία για κάθε ένα προσανατολισμένο παράθυρο, τα $(c, s, l_1, \dots, l_d)$.

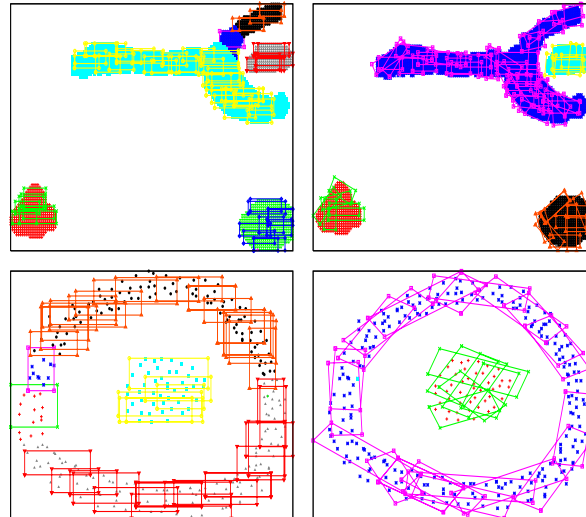
Ο αλγόριθμος OkW, απαιτεί τον υπολογισμό του SVD, σε κάθε κεντροποιημένο υποσύνολο δεδομένων, στη διάρκεια της διαδικασίας διεύρυνσης. Για αυτό το λόγο, είναι σημαντική η χρήση μιας γρήγορης ρουτίνας SVD, όπως η LAPACK [49], για παράθυρα μέτρου μεγέθους χωρίς κάποια συγκεκριμένη δομή, ή σαν αυτή που αναφέρετε στην εργασία [14], για αραιά δεδομένα. Σε κάθε περίπτωση είναι σημαντικό να γίνεται εκμετάλλευση της δομής των δεδομένων ώστε να μειωθεί το υπολογιστικό κόστος. Η εκτέλεση της ρουτίνας SVD, για προβλήματα χωρίς κάποια συγκεκριμένη δομή, έχει υπολογιστικό κόστος $O(n_i d^2)$, το οποίο μπορεί να μη είναι είναι αμελητέο. Για εφαρμογές βέβαια που ενδιαφέρουν περισσότερο τον αλγόριθμο OkW, η διάσταση των δεδομένων d είναι τυπικά πολύ μικρότερη όχι μόνο του n , αλλά και του n_i .

5.2.3 Πειραματικά Αποτελέσματα

Σε αυτήν την ενότητα θα παρουσιάσουμε πειραματικά αποτελέσματα, που συγκρίνουν τα αποτελέσματα των αλγορίθμων UKW και OkW, για να αναδείξουμε τις δυνατότητες της προτεινόμενης τροποποίησης.

Αρχικά θα χρησιμοποιηθούν δύο τεχνητά σύνολα δεδομένων, που θεωρούνται δύσκολα για διάφορους αλγορίθμους. Το πρώτο $Dset_1$, αποτελείται από 2761 σημεία οργανωμένα σε 4 ομάδες (3 κυρτές και μία μη κυρτή). Το δεύτερο $Dset_2$, αποτελείται από 299 στοιχεία, οργανωμένα σε δύο ομάδες που σχηματίζουν ομό-κεντρους κυκλικούς δίσκους. Οι τιμές των παραμέτρων $\{\theta_e, \theta_m, \theta_c, \theta_v\}$, ήταν $\{0.8, 0.1, 0.2, 0.02\}$, και χρησιμοποιήθηκαν 64 αρχικά παράθυρα για το σύνολο $Dset_1$ και 32 για το σύνολο $Dset_2$. Τα αποτελέσματα απεικονίζονται στο Σχήμα 5.6. Ο αλγόριθμος OkW, καταφέρνει και στις δύο περιπτώσεις να εντοπίσει τόσο τις κυρτές όσο και τις μη κυρτές μορφολογικά ομάδες. Αντίθετα, ο αλγόριθμος UKW, είναι ικανός να εντοπίσει με ευκρίνεια τις κυρτές ομάδες, αλλά τείνει να διασπά τις μη κυρτές. Αν βέβαια αυξηθεί το πλήθος των παραθύρων τα αποτελέσματα του UKW, είναι ίδια με αυτά του OkW, αλλά σε μερικές περιπτώσεις χρειάζεται πολύ μεγάλος αριθμός παραθύρων. Για παράδειγμα στο σύνολο $Dset_2$, για να αναγνωρίσει με ακρίβεια ο αλγόριθμος UKW, τη εξωτερική κυκλική ομάδα χρειάζεται πάνω από 200 αρχικά παράθυρα.

Στη συνέχεια χρησιμοποιούμε δύο σύνολα δεδομένων πραγματικού κόσμου τα $Dset_{UCI}$ και $Dset_{DARPA}$. Το πρώτο είναι το γνωστό Iris σύνολο δεδομένων του UCI Machine Learning Repository [15], που αποτελείται από 150 στοιχεία με 4 χαρακτηριστικά το καθένα και είναι χαρακτηρισμένα σε 3 κλάσεις, τις Setosa, Versicolour, και Virginica (Ενότητα 5.1.2). Ο πίνακας αντιστοίχισης 5.2 αναφέρει τα αποτελέσμα-



Σχήμα 5.6: Τα σύνολα δεδομένων $Dsets_{1,2}$ με τα αποτελέσματα των αλγορίθμων UKW (αριστερά) και OkW (δεξιά).

τα των δύο αλγορίθμων αν χρησιμοποιηθούν 32 αρχικά παράθυρα, όπου φαίνεται ότι 8 στοιχεία για τον αλγόριθμο UKW και 7 για τον OkW έχουν διαφορετική αντιστοίχιση ομάδας-κλάσης.

Τέλος, το σύνολο δεδομένων $Dset_{DARPA}$, (Ενότητα 3.6), έχει κατασκευαστεί το 1998 από το DARPA Intrusion Detection Evaluation Program στο MIT Lincoln Labs. Από αυτό το σύνολο χρησιμοποιήθηκαν τα πρώτα 100000 στοιχεία. Τα 77888 από αυτά έχουν χαρακτηριστεί σαν normal ενώ τα υπόλοιπα 22112, σαν επιθέσεις στα συστήματα τύπου Denial of Service (DoS). Από τα 42 χαρακτηριστικά χρησιμοποιήθηκαν τα 37 αριθμητικά. Σε αυτό το σύνολο και οι δύο αλγόριθμοι εφαρμόζονται με 32 αρχικά παράθυρα, και καταλήγουν σε ομάδες που περιέχουν εξολοκλήρου είτε στοιχεία normal είτε DoS. Στη περίπτωση του αλγορίθμου UKW υπάρχει μία ομάδα που αντιστοιχεί σε normal στοιχεία αλλά περιέχει και 37 DoS στοιχεία. Αντίθετα στην περίπτωση του αλγορίθμου OkW, μια ομάδα με 747 normal στοιχεία περιέχει μόνο 2 DoS. Συνεπώς, η απόδοση του αλγορίθμου OkW, δείχνει καλύτερη αφού κατέληξε σε πολύ λιγότερες λάθος αντιστοιχίσεις.

Πίνακας 5.2: Πίνακας αντιστοίχισης για το σύνολο $Dset_{UCI}$: Κάθε ζεύγος αριθμών αναπαριστά τα στοιχεία του πίνακα αντιστοίχισης για του δύο αλγορίθμους UKW και OkW, αντίστοιχα.

Ομάδα	Iris Κλάση		
	Setosa	Versicolour	Virginica
1	50, 50,	0, 0	0, 0
2	0, 0	46, 48	4, 5
3	0, 0	4, 2	46, 45

5.3 Συμπεράσματα - Συνεισφορά

Σε αυτό το κεφάλαιο, παρουσιάστηκαν δύο τροποποιήσεις του αλγορίθμου k -windows, που ακολουθούν μια πρόσφατη τάση στον τομέα των αλγορίθμων ομαδοποίησης, προς

την χρήση όσο το δυνατόν περισσότερης πληροφορίας από την τοπική δομή των δεδομένων για την παραγωγή καλύτερων ομαδοποιήσεων [16, 22, 38].

Αρχικά παρουσιάσαμε τα πρωταρχικά αποτελέσματα μια νέας προσέγγισης στην ομαδοποίηση, που χρησιμοποιεί την διάσταση Fractal, των σημείων για να κατευθύνει την όλη διαδικασία. Η προτεινόμενη διαδικασία καθοδηγεί την αλγοριθμική διαδικασία του αλγορίθμου UKW, υπολογίζοντας σε κάθε βήμα μια προσέγγιση της Fractal διάστασης με τη χρήση της μεθόδου box-counting. Η προσέγγιση που προτείνεται επιτρέπει την ανακάλυψη περιοχών με διαφορετική Fractal διάσταση, ακόμα και μέσα στην ίδια ομάδα. Αλγόριθμοι που μπορούν να εξάγουν τέτοιου είδους πληροφορία, είναι πολύτιμοι σε διάφορες εφαρμογές όπως ανάλυση χρονοσειρών, ανάλυση εικόνας, και ιατρικές εφαρμογές.

Έπειτα προτείναμε τη συνεργασία μεταξύ της τεχνικής PCA, και του αλγορίθμου UKW, κατασκευάζοντας τον αλγόριθμο OkW [138, 139]. Με αυτόν τον τρόπο αναδείξαμε πως τεχνικές υπολογιστικής γραμμικής άλγεβρας μπορούν να βοηθήσουν αλγόριθμους ομαδοποίησης να παράγουν πιο αποδοτικά αποτελέσματα. Ενδιαφέρουσα μελλοντική ερευνητική πρόκληση, φαίνονται να είναι η γενίκευση αυτού του σχήματος σε εφαρμογές που η διάσταση των δεδομένων είναι πολύ μεγάλη.

Ομαδοποίηση Σε Κατανεμημένες Βάσεις Δεδομένων

Δώστε μου ένα μοχλό αρκετά μακρύ
και ένα υπομόχλιο για να το τοποθετήσω
και θα μετακινήσω τον κόσμο.

—Αρχιμήδης (287 π.Χ. – 212 π.Χ.)

Τις τελευταίες δεκαετίες η ικανότητα συγκέντρωσης, αποθήκευσης και ανάκτησης δεδομένων αυξάνεται συνεχώς. Αυτό το γεγονός έχει κάνει την ανάπτυξη αλγορίθμων ομαδοποίησης που μπορούν να εξορύξουν πληροφορία, από διαφορετικές βάσεις δεδομένων ταυτόχρονα, περισσότερο από αναγκαία. Για το λόγο αυτό οι κατανεμημένοι αλγόριθμοι ομαδοποίησης, προσπαθούν να ενοποιήσουν τον υπολογισμό με την επικοινωνία ώστε να αξιοποιήσουν όσο το δυνατόν καλύτερα όλες τις πλευρές ενός κατανεμημένου υπολογιστικού περιβάλλοντος.

Παρόλο που έχουν προταθεί αρκετές προσεγγίσεις για την παράλληλη και κατανεμημένη εξόρυξη πληροφορίας [23, 80, 87], δεν υπάρχουν πολλές προσεγγίσεις στον τομέα της ομαδοποίησης. Στην εργασία [152] προτείνεται μια παράλληλη εκδοχή του αλγορίθμου DBSCAN, και στην εργασία [34], μια παράλληλη εκδοχή του k -means. Οι δύο αυτοί αλγόριθμοι ξεκινούν με όλο το σύνολο δεδομένων να βρίσκεται σε ένα κεντρικό εξυπηρετητή, και έπειτα κατανέμουν τα δεδομένα σε διαφορετικές μηχανές πελάτες.

Συνήθως, σε ένα κατανεμημένο υπολογιστικό περιβάλλον, το σύνολο δεδομένων είναι διάσπαρτο σε έναν αριθμό, από διαφορετικούς τόπους. Έτσι, ως υποθέσουμε ότι το σύνολο δεδομένων X είναι κατανεμημένο σε m τόπους, όπου κάθε ένας αποθηκεύει το δικό του σύνολο X_i για $i = 1, \dots, m$:

$$X = \bigcup_{i=1, \dots, m} X_i.$$

Ακόμη ως υποθέσουμε και την ύπαρξη ενός κεντρικού τόπου C , που θα περιέχει τα τελικά αποτελέσματα της ομαδοποίησης.

Σε αυτό το σημείο μπορούν να γίνουν διαφορετικές υποθέσεις για τη φύση του περιβάλλοντος επικοινωνίας ανάμεσα στους τόπους. Αρχικά μπορούμε να υποθέσουμε ότι οι τόποι συνδέονται με ένα υψηλής ταχύτητας δίκτυο, και επιτρέπεται η αποκάλυψη των δεδομένων του καθενός τόπου. Αντίθετα, θα μπορούσε να γίνει και η υπόθεση ότι, το περιβάλλον υπαγορεύει ελάχιστη επικοινωνία ανάμεσα στους τόπους. Αυτό μπορεί

να οφείλεται είτε στο πολύ υψηλό κόστος της επικοινωνίας είτε σε θέματα ασφάλειας και προσωπικών δεδομένων.

Στις επόμενες ενότητες, παρουσιάζονται δύο εκδοχές του αλγορίθμου UKW, για τις δύο διαφορετικές υποθέσεις του περιβάλλοντος επικοινωνίας.

Ένα ακόμη σενάριο που συναντάται σε καθημερινές εφαρμογές [145], είναι η αποθήκευση των διάφορων ιδιοτήτων των εγγραφών σε ξεχωριστούς οργανισμούς, όπως τράπεζες, εταιρείες πιστωτικών καρτών, καθώς και φορολογικές υπηρεσίες. Δηλαδή σε αυτήν την περίπτωση, η βάση δεδομένων είναι διαχωρισμένη «κάθετα», καθώς είναι κατανεμημένες οι ιδιότητες της κάθε εγγραφής, και όχι οι ίδιες οι εγγραφές. Πιο αυστηρά, σε αυτήν την περίπτωση το σύνολο δεδομένων, Q , αποτελεί το καρτεσιανό γινόμενο των συνόλων X_i , του κάθε τύπου $i = 1, \dots, m$:

$$X = X_1 \times \dots \times X_m.$$

Εάν μια νομική υπηρεσία, προσπαθούσε να ομαδοποιήσει άτομα, σύμφωνα με τις οικονομικές τους συναλλαγές, θα ήταν αναγκαίο από τους οργανισμούς να αποκαλύψουν τις τιμές των ιδιοτήτων που κατέχουν, για το κάθε άτομο. Αυτό όμως αναδεικνύει δικαιολογημένα διάφορα θέματα ασφάλειας της ιδιωτικότητας των δεδομένων.

Το παραπάνω παράδειγμα, σκιαγραφεί το πως ανησυχίες για την ασφάλεια και ιδιωτικότητα, μπορούν να αποτρέψουν την αναγκαστική αποκάλυψη των τιμών από κάθε οργανισμό, με αποτέλεσμα να αποθαρρύνουν την διαδικασία ομαδοποίησης. Με αυτήν την έννοια, η ομαδοποίηση και η ιδιωτικότητα, συχνά θεωρούνται να είναι ασύμβατες. Το πρόβλημα όμως δεν βρίσκεται στη ίδια την ομαδοποίηση, άλλα στον τρόπο που αυτή πραγματοποιείται. Για το λόγο αυτό υπάρχει μια συνεχώς αυξανόμενη ανάγκη για την ανάπτυξη μεθόδων, που έχουν ενδογενείς μηχανισμούς να προστατεύουν την εμπιστευτικότητα ευαίσθητων τιμών.

Παρόλα αυτά, η διαδικασία της ομαδοποίησης μπορεί να διαμορφωθεί κατάλληλα, ώστε να είναι συμβατή με τις απαιτήσεις ιδιωτικότητας και ασφάλειας. Για το σκοπό αυτό πρέπει να επιτευχθούν δύο στόχοι. Αρχικά οι αλγόριθμοι ομαδοποίησης χρειάζεται να είναι εφαρμόσιμοι χωρίς να απαιτούν το διαμοιρασμό των δεδομένων ανάμεσα στους ιδιοκτήτες των δεδομένων. Κατά δεύτερο λόγο, δεν πρέπει να είναι εφικτή η εξαγωγή συμπερασμάτων από τα αποτελέσματα της ομαδοποίησης. Εάν αυτοί οι δύο στόχοι επιτευχθούν, τότε η ομαδοποίηση δεν απαιτεί κανένα συμβιβασμό της ασφάλειας και της ιδιωτικότητας των δεδομένων, ενώ ταυτόχρονα συνεισφέρει στην εξαγωγή καθολικών συμπερασμάτων. Στην Ενότητα 6.3, θα παρουσιαστεί μια τροποποίηση του αλγορίθμου UKW [129], που επιτυγχάνει τους δύο αυτούς στόχους.

6.1 Κατανεμημένη Ομαδοποίηση για Περιβάλλοντα Ελάχιστης Επικοινωνίας

Αν στηριχθούμε στις υποθέσεις που υπαγορεύει ένα περιβάλλον ελάχιστης επικοινωνίας, είναι δυνατόν να μετατρέψουμε τον αλγόριθμο UKW, ώστε να κατανείμει τοπικά σε κάθε τόπο την διαδικασία ομαδοποίησης και κεντρικά να γίνει μόνο η διαδικασία της συγχώνευσης. Πιο συγκεκριμένα, σε κάθε τόπο i , εκτελείται ο αλγόριθμος UKW, σε κάθε σύνολο X_i ξεχωριστά. Αυτό το βήμα καταλήγει σε μερικά σύνολα από παράθυρα W_i σε κάθε τόπο. Για να κατασκευαστεί το τελικό αποτέλεσμα, μόνο τα σύνολα παραθύρων μεταφέρονται στο κεντρικό τόπο C . Αυτός, ο κεντρικός κόμβος είναι υπεύθυνος, για τη διαδικασία της συγχώνευσης παραθύρων, ώστε να κατασκευαστεί

το τελικό σύνολο παραθύρων που αντιπροσωπεύει το τελικό αποτέλεσμα της ομαδοποίησης. Αυτό το σύνολο μεταδίδεται πίσω σε κάθε τόπο ώστε όλοι να αποκτήσουν το πλήρες ομαδοποιημένο σχήμα, για τα δεδομένα τους.

Ωστόσο για να αποφασίσει ο κεντρικός τόπος την συγχώνευση του κάθε ζευγαριού παραθύρων χρειάζεται την πληροφορία τους πλήθους των σημείων που βρίσκεται στην τομή των παραθύρων. Οι υποθέσεις του περιβάλλοντος όμως, μπορεί να απαγορεύουν την αποκάλυψη αυτής της πληροφορίας λόγω ασφάλειας ή υψηλού κόστους επικοινωνίας. Υπό αυτό τον περιορισμό δύο επικαλυπτόμενα παράθυρα θεωρούνται πάντα ότι ανήκουν στην ίδια ομάδα, χωρίς να εξετάζεται ο αριθμός των κοινών τους στοιχείων. Έτσι οι παράμετροι θ_m και θ_s παύουν να ισχύουν. Ακολουθεί μια υψηλού επιπέδου περιγραφή του αλγορίθμου.

Κατανεμημένος Αλγόριθμος Ομαδοποίησης DUKW

Για κάθε τόπο i , με $i = 1, \dots, m$

Εκτέλεσε τον αλγόριθμο UKW στο σύνολο X_i

Στείλε το τελικό σύνολο παραθύρων W_i στον κεντρικό τόπο C .

Στον κεντρικό τόπο C

Για κάθε τόπο i

Μάζεψε τα σύνολα των παραθύρων W_i

Θέσε $W \leftarrow W \cup W_i$

Ενοποίησε τα επικαλυπτόμενα παράθυρα

Στείλε σε κάθε κόμβο i το τελικό σύνολο W

6.2 Κατανεμημένη Ομαδοποίηση σε Υψηλής Ταχύτητας Δίκτυο

Υποθέτοντας ότι οι διάφοροι τόποι είναι συνδεδεμένοι με ένα δίκτυο υψηλής ταχύτητας, ο αλγόριθμος UKW μπορεί να διαμορφωθεί, ώστε να κατανείμει το υπολογιστικό κόστος χωρίς να μειώνει την αποτελεσματικότητά του. Όπως έχει ήδη ειπωθεί το δύσκολο υπολογιστικό βήμα του αλγορίθμου UKW, είναι τα ερωτήματα διαστήματος για την ανακάλυψη του πλήθους των σημείων και το κέντρο βάρους τους, που περιλαμβάνονται σε κάθε παράθυρο. Για αυτό το λόγο κατανέμοντας αυτό το βήμα στην ουσία έχει κατανεμηθεί όλο το υπολογιστικό κόστος του αλγορίθμου.

Πιο συγκεκριμένα, ο αλγόριθμος UKW, εκτελείται στον κεντρικό τόπο C , και κάθε φορά που γίνεται μια ερώτηση διαστήματος για ένα παράθυρο W , στέλνεται στον κάθε τόπο i , το παράθυρο αυτό, και οι τόποι επιστρέφουν το πλήθος και το κέντρο βάρους των σημείων που περιέχονται στο παράθυρο από το δικό τους σύνολο. Σε μια υψηλού επιπέδου περιγραφή ο αλγόριθμος έχει ως εξής:

Παράλληλος Αλγόριθμος Ομαδοποίησης PUKW

Εκτέλεσε τον αλγόριθμο UKW στον κεντρικό τόπο C

Σε κάθε ερώτημα διαστήματος για παράθυρο W **εκτέλεσε** παράλληλα

Στείλε σε κάθε τόπο i το παράθυρο W

Σε κάθε τόπο i ,

Ανακάλυψε το πλήθος σημείων $n_{i,W}$ και το κέντρο βάρους τους $g_{i,W}$ από το σύνολο X_i που περιέχονται στο W .

Στείλε στον κεντρικό τόπο C τα $n_{i,W}$, και $g_{i,w}$
Υπολόγισε τα $n_W = \sum_{i=1}^m n_{i,W}$ και $g_W = \frac{1}{n_W} \sum_{i=1}^m n_{i,W} * g_{i,W}$ που αποτελούν
το τελικό πλήθος και το κέντρο βάρους των σημείων,
που περιλαμβάνονται στο W .

Για να αναλύσουμε την πολυπλοκότητα του αλγορίθμου, θεωρούμε ότι η δομή δεικτοδότησης που χρησιμοποιείται σε κάθε κόμβο είναι ένα Multidimensional Binary Tree (kd-tree) [115]. Σε αυτή την περίπτωση μια ερώτηση διαστήματος έχει πολυπλοκότητα $O(dn^{1-1/d} + s)$ [115], όπου d το πλήθος των χαρακτηριστικών, n το συνολικό πλήθος στοιχείων, και s , το πλήθος των στοιχείων που περιλαμβάνονται στο διάστημα. Οπότε η πολυπλοκότητα με τη χρήση m τόπων γίνεται:

$$O(d(n/m)^{1-1/d} + s + \epsilon(d, m)),$$

όπου $\epsilon(d, m)$ είναι μια συνάρτηση που αντιπροσωπεύει το χρόνο που χρειάζεται για την επικοινωνία μεταξύ του κεντρικού τόπου και των υπόλοιπων m τόπων. Αυτή η συνάρτηση, εξαρτάται μόνο από το πλήθος των χαρακτηριστικών d , και το πλήθος των τόπων m , καθώς μόνο το πλήθος των σημείων και το κέντρο βάρους τους χρειάζεται να μεταφέρεται και όχι τα ίδια τα σημεία. Αν λάβουμε υπόψιν και τα χαρακτηριστικά του ίδιου του δικτύου η τιμή της συνάρτησης μπορεί σε κάθε περίπτωση να υπολογιστεί με ακρίβεια. Οπότε μπορούμε να υπολογίσουμε αν ο αλγόριθμος μπορεί να επιτύχει επιτάχυνση σε σχέση με την σειριακή του έκδοση. Για να συμβεί κάτι τέτοιο θα πρέπει να ισχύει:

$$O\left(\frac{d\left(\frac{n}{m}\right)^{1-\frac{1}{d}} + s + \epsilon(d, m)}{dn^{1-\frac{1}{d}} + s}\right) \leq 1,$$

το οποίο γίνεται [128]:

$$O(\epsilon(d, m)) \leq O\left(d\left(n^{1-\frac{1}{d}} - \left(\frac{n}{m}\right)^{1-\frac{1}{d}}\right)\right). \quad (6.1)$$

Αν η ανισότητα (6.1) ισχύει, τότε η παράλληλη εκτέλεση του αλγορίθμου είναι ταχύτερη της σειριακής έκδοσης. Σε κάθε άλλη περίπτωση η δομή του δικτύου εμποδίζει, την όλη διαδικασία.

6.3 Ομαδοποίηση Διατηρώντας την Ιδιωτικότητα

Η πρώτη φορά που μελετήθηκε σε γενικά πλαίσια ο ασφαλής υπολογισμός συναρτήσεων ανάμεσα δύο οντότητες ήταν από τον Yao [153]. Αργότερα επεκτάθηκε σε περισσότερες από δύο, στην εργασία [59].

Μια πρόσφατη προσέγγιση σε αυτό το θέμα, ήταν η προσθήκη «θορύβου» στα δεδομένα, πριν την διαδικασία εξόρυξης γνώσης [1, 2, 51]. Έπειτα με τη χρήση μεθόδων που δεν είναι ευαίσθητες στο θόρυβο, εκτελείται η διαδικασία της εξόρυξης γνώσης. Στις εργασίες [2, 89], γίνεται μελέτη για την ποσότητα της πληροφορίας που μπορεί να υπολογιστεί ή να αποκαλυφθεί για τα δεδομένα κατά την εκτέλεση αλγορίθμων ταξινόμησης, και προτείνονται τρόποι για την ελαχιστοποίηση αυτής της πληροφορίας.

Η εξαγωγή κανόνων συσχέτισης διατηρώντας την ασφάλεια σε κατανεμημένες βάσεις δεδομένων μελετάται στην εργασία [79], ενώ στην [144] σε κάθετα κατανεμημένες βάσεις.

Όσον αφορά την ομαδοποίηση, οι Vaidya και Clifton, πρότειναν μια προσαρμογή του αλγορίθμου *k*-means, χρησιμοποιώντας πολλά στοιχεία από τον ασφαλή υπολογισμό πολλών οντοτήτων. Μια εναλλακτική προσέγγιση, προτείνεται στην εργασία [100], όπου αποφεύγεται το μοίρασμα των δεδομένων, μοιράζοντας ανάμεσα στις οντότητες παραμέτρους μοντέλων παραγωγής των δεδομένων. Τέλος στην εργασία [106], προτείνονται μέθοδοι ιεραρχικής ομαδοποίησης χρησιμοποιώντας διάφορους γεωμετρικούς μετασχηματισμούς.

Σε αυτήν την ενότητα βασιζόμαστε στην υπόθεση ότι, το σύνολο δεδομένων είναι κάθετα κατανεμημένο σε διάφορες οντότητες-τόπους. Κάθε οντότητα, επιθυμεί να εξάγει σφαιρική γνώση για τα δεδομένα, αλλά δεν είναι διατεθειμένη να αποκαλύψει οποιαδήποτε τιμή για αυτά στις υπόλοιπες. Για το σκοπό της ομαδοποίησης κάτω από αυτές τις υποθέσεις αναπτύσσεται ένα αλγοριθμικό σχήμα [129], για τον αλγόριθμο UKW, που:

- Αποτρέπει την αποκάλυψη τιμών για τα δεδομένα από οποιαδήποτε οντότητα.
- Τα αποτελέσματα της ομαδοποίησης δεν αλλάζουν καθόλου από αυτά που θα προέκυπταν αν η βάση δεδομένων ήταν ενιαία.

Πρέπει επίσης να σημειωθεί ότι, θεωρούμε ότι κάποια από τις εμπλεκόμενες οντότητες δεν είναι «μοχθηρή», δηλαδή δεν σκοπεύει στην παραγωγή λανθασμένων αποτελεσμάτων. Αυτή η υπόθεση θεωρείται δικαιολογημένη, καθώς συνήθως οι οντότητες που εμπλέκονται τέτοιες εφαρμογές, έχουν μια εγκαθιδρυμένη συνεργασία, παρά ένα πρόσκαιρο συνεταιρισμό. Η κακή πίστη θεωρείται ότι τιμωρείται σε μη αλγοριθμικό πλαίσιο.

Στις επόμενες παραγράφους θα γίνει μία αναλυτική έκθεση των προτεινόμενων επεκτάσεων στις διαδικασίες της «αρχικοποίησης», «μετακίνησης», «διεύρυνσης» και «ενοποίησης» του αλγορίθμου UKW.

6.3.1 Αρχικοποίηση των Παραθύρων

Όπως έχει ήδη ειπωθεί στο Κεφάλαιο 3, ο αλγόριθμος UKW ξεκινά με την τυχαία αρχικοποίηση παραθύρων, δεδομένου μεγέθους. Στην περίπτωση αυτή, τίποτα δεν είναι διαφορετικό. Κάθε οντότητα μπορεί επιλέξει τις τιμές των κέντρων των αρχικών παραθύρων τυχαία, καθώς και το μέγεθος των παραθύρων, όσον αφορά βέβαια τα ιδιωτικά χαρακτηριστικά της. Σε αυτό το βήμα δεν απαιτείται καμιά είδους επικοινωνία ανάμεσα στις οντότητες.

6.3.2 Μετακινήσεις και Διευρύνσεις

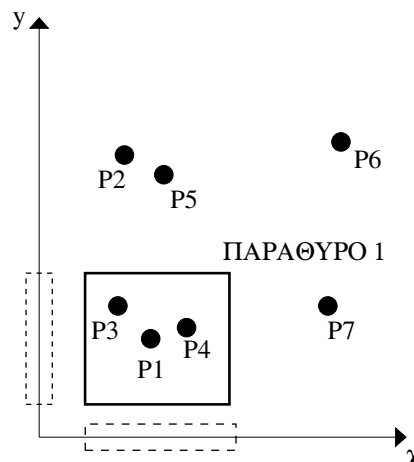
Μετά την ολοκλήρωση του βήματος της αρχικοποίησης, κάθε οντότητα γνωρίζει (α) τις συντεταγμένες των κέντρων των παραθύρων, για τιμές ιδιοτήτων που κατέχει και (β) το μέγεθος των παραθύρων για τις ίδιες συντεταγμένες. Από αυτές τις πληροφορίες μόνο, κάθε οντότητα μπορεί να προσδιορίσει το σύνολο των σημείων που αυτή θεωρεί ότι εσωκλείονται σε ένα συγκεκριμένο παράθυρο. Ωστόσο, το αληθινό σύνολο, των σημείων που εσωκλείονται στο παράθυρο, σύμφωνα με το πλήρες σύνολο

χαρακτηριστικών αποτελείται από την τομή των συνόλων της κάθε οντότητας. Ένα παράδειγμα απεικονίζεται στο Σχήμα. 6.1, όπου το σύνολο δεδομένων απαρτίζεται από 7 διαφορετικά δισδιάστατα σημεία. Επίσης, θεωρούμε ότι το υπάρχουν δύο οντότητες, από τις οποίες η μία έχει τις τιμές της διάστασης x για τα δεδομένα και η άλλη της διάστασης y . Οι πληροφορίες για το κέντρο και το μέγεθος του παραθύρου 1, βρίσκονται κατανεμημένες στις δύο οντότητες. Έτσι η πρώτη οντότητα γνωρίζει το κέντρο του και το μέγεθός του για την συντεταγμένη x , και σύμφωνα με αυτήν το σύνολο των σημείων V_1 που περιλαμβάνονται στο παράθυρο είναι $V_1 = \{P1, P2, P3, P4, P5\}$. Αντίστοιχα για την άλλη οντότητα, που έχει τις πληροφορίες για το κέντρο και το μέγεθος της διάστασης y , το αντίστοιχο σύνολο V_2 , είναι $V_2 = \{P1, P3, P4, P7\}$. Όπως είναι προφανές το αληθινό σύνολο σημείων V που εσωκλείονται στο παράθυρο, για το πλήρες σύνολο συντεταγμένων-χαρακτηριστικών x, y είναι $V = V_1 \cap V_2 = \{P1, P3, P4\}$. Ο υπολογισμός αυτής της τομής γίνεται σύμφωνα με το πρωτόκολλο υπολογισμού τομής συνόλων, για ιδιωτικές βάσεις δεδομένων που προτάθηκε στην εργασία [54], και επιτρέπει τον υπολογισμό της τομής των ξεχωριστών συνόλων χωρίς να αποκαλύπτεται η ταυτότητα των σημείων που ανήκουν στο κάθε ένα επιμέρους σύνολο.

Αυτό το πρωτόκολλο περιλαμβάνει, ομο-μορφικά σχήματα κρυπτογράφησης και μη ενθυμούμενους πολυωνυμικούς υπολογισμούς. Θεωρεί μία από τις οντότητες σε θέση εξυπηρετητή, και τις υπόλοιπες πελάτες αυτής. Έχει επίσης αποδειχθεί ότι με την χρήση του υπολογίζεται σωστά η τομή των συνόλων και διατηρείται η ιδιωτικότητα των δεδομένων για όλες τις οντότητες. Περισσότερες πληροφορίες αναφέρονται στη εργασία [54].

Στη συνέχεια, όπως και στον κανονικό αλγόριθμο UKW, υπολογίζεται ο μέσος όρος των σημείων που βρίσκονται μέσα σε κάθε ένα παράθυρο. Σε αυτή την περίπτωση βέβαια, κάθε οντότητα υπολογίζει ξεχωριστά το μέσο όρο για τα χαρακτηριστικά που κατέχει και μετακινεί αντίστοιχα το παράθυρο. Αυτή η διαδικασία επαναλαμβάνεται, μέχρι να μην εισέρχονται νέα σημεία στο παράθυρο όπως και στην περίπτωση του σειριακού αλγορίθμου UKW. Κάθε φορά βέβαια που πρέπει να υπολογιστεί το σύνολο των σημείων που περιέχονται σε ένα παράθυρο, γίνεται χρήση του πρωτοκόλλου υπολογισμού τομής συνόλων με διατήρηση της ιδιωτικότητας.

Για την διαδικασία της διεύρυνσης, που είναι το επόμενο βήμα του αλγορίθμου



Σχήμα 6.1: Υπολογισμός των σημείων που εσωκλείονται σε ένα παράθυρο όταν τα χαρακτηριστικά (x, y) είναι κατανεμημένα σε 2 διαφορετικές οντότητες.

UKW, τα πράγματα δεν είναι πολύ διαφορετικά. Δεδομένου ότι η διεύρυνση εξετάζεται σε κάθε διάσταση χωριστά, κάθε οντότητα μπορεί να εκτελέσει τη διαδικασία για κάθε ένα από τα χαρακτηριστικά της σταδιακά. Όταν χρειάζεται ο υπολογισμός των σημείων που περιλαμβάνονται σε ένα παράθυρο, χρησιμοποιείται πάλι το προαναφερθέν πρωτόκολλο. Επίσης πρέπει να σημειωθεί ότι σε αυτήν την περίπτωση είναι δυνατόν να χρησιμοποιηθούν και διαφορετικές τιμές για τις παραμέτρους θ_e και θ_c , χωρίς αυτό να επηρεάζει την λειτουργία του συνολικού αλγορίθμου, ικανοποιώντας βέβαια την συνθήκη $\theta_c \geq 1$.

6.3.3 Συγχώνευση των Τελικών Παραθύρων

Για να εκτελεστεί η διαδικασία της συγχώνευσης, πρέπει αν εντοπιστούν τα παράθυρα που επικαλύπτονται, στο πλήρες σύνολο των χαρακτηριστικών. Σε αυτήν την περίπτωση κάτι τέτοιο θα απαιτούσε την αποκάλυψη σημαντικής πληροφορίας ανάμεσα στις οντότητες.

Όμως η διαδικασία της συγχώνευσης, αποφασίζει για τον αν δύο παράθυρα θα συγχωνευτούν ή όχι σύμφωνα με το πλήθος το κοινών τους σημείων. Για οποιαδήποτε ζευγάρι παραθύρων, αυτό μπορεί να υπολογιστεί διατηρώντας την ιδιωτικότητα, με τη χρήση του προαναφερθέντος πρωτοκόλλου τομής συνόλων [54]. Με αυτό τον τρόπο, για κάθε ένα ζευγάρι παραθύρων, ανεξάρτητα αν επικαλύπτονται ή όχι, μπορεί να ανακαλυφθεί για το συνολικό πλήθος των χαρακτηριστικών, το πλήθος των κοινών τους σημείων. Έπειτα σύμφωνα με τη διαδικασία της συγχώνευσης, η αναλογία του πλήθους των κοινών σημείων κάθε ζεύγους παραθύρων, προς τα συνολικό πλήθος σημείων που περιέχουν, και οι τιμές των παραμέτρων θ_s και θ_m , καθορίζουν το τελικό αποτέλεσμα της ομαδοποίησης (Ενότητα 3.3). Σε αυτή την περίπτωση, όπως και στην διαδικασία της διεύρυνσης, κάθε οντότητα μπορεί να επιλέξει ξεχωριστά τιμές για τις παραμέτρους, $0 < \theta_s, \theta_m \leq 1$, και να διαμορφώσει ανάλογα το τελικό αποτέλεσμα, ανάλογα με τις δικές της ανάγκες.

6.3.4 Υπολογιστική Πολυπλοκότητα

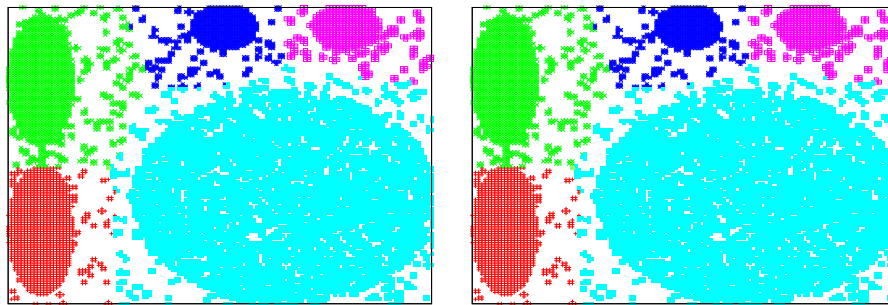
Όπως έχει ήδη αναφερθεί η πολυπλοκότητα του αλγορίθμου UKW, βασίζεται στο υπολογιστικό κόστος που απαιτείται για την *ορθογώνια αναζήτηση διαστήματος* (Ενότητα 3.5). Σε αυτήν την περίπτωση όμως, πρέπει επίσης να συμπεριληφθεί και το κόστος του υπολογισμού της τομής συνόλων με διατήρηση της ιδιωτικότητας. Αυτό το κόστος διαχωρίζεται σε επικοινωνιακό, που σχετίζεται με το πλήθος των μηνυμάτων που πρέπει να ανταλλάγουν ανάμεσα στις οντότητες, και στο υπολογιστικό που αναφέρεται στου επιπλέον υπολογισμούς που χρειάζονται. Το κόστος επικοινωνίας αυτού του πρωτοκόλλου είναι $O(ml_c)$, όπου το m αντιπροσωπεύει τον αριθμό των οντοτήτων, και το l_c το μέγιστο μέγεθος των συνόλων σε κάθε υπολογισμό. Από την άλλη μεριά, το υπολογιστικό κόστος ανέρχεται σε $O(ml_c^2)$, το οποίο όμως, είναι δυνατόν να μειωθεί σε $O(m(l_c + l_c \ln \ln l_c))$, μέσω μιας μεθόδου καταμερισμού σε κουτιά που περιγράφεται στην εργασία [54].

6.4 Πειραματικά Αποτελέσματα

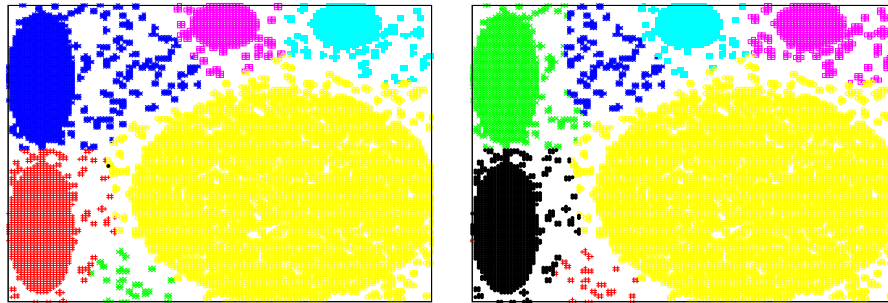
Για να μελετήσουμε την ακρίβεια του αλγορίθμου DUKW, για περιβάλλοντα ελάχιστης επικοινωνίας, που παρουσιάστηκε στην Ενότητα 6.1, θα χρησιμοποιήσουμε πειρά-

ματα που καταλήγουν σε αποτελέσματα που επιτρέπουν την άμεση οπτικοποίηση. Για το σκοπό αυτό χρησιμοποιούμε το δισδιάστατο σύνολο δεδομένων $Dset_d$ που περιέχει 10241 στοιχεία οργανωμένα σε 5 ομάδες. Το σύνολο αυτό, διασπάστηκε τυχαία σε 4, 8, και 16 τόπους. Έπειτα εφαρμόστηκε ο αλγόριθμος DUKW χρησιμοποιώντας 256, 64, 32 και 16 αρχικά παράθυρα ανά τόπο για 1, 4, 8 και 16 τόπους αντίστοιχα ώστε να έχουμε τον ίδιο τελικό αριθμό παραθύρων.

Το σύνολο δεδομένων μαζί με το αποτέλεσμα του αλγορίθμου για 1 και 4 κόμβους φαίνεται στο Σχήμα. 6.2, ενώ στο Σχήμα 6.3 παρουσιάζονται τα αποτελέσματα για 8 και 16 τόπους. Όπως είναι προφανές από τα σχήματα ο αλγόριθμος καταφέρνει να ανακαλύψει την πραγματική δομή των δεδομένων σε όλες τις περιπτώσεις, Επίσης πρέπει να σημειωθεί ότι στην περίπτωση των 8 και 16 τόπων, ανακαλύπτεται μία ακόμη ομάδα που εξαιτίας των λίγων σημείων που περιέχει μπορεί να θεωρηθεί και σαν θόρυβος.



Σχήμα 6.2: Το σύνολο δεδομένων $Dset_d$ με το αποτέλεσμα του αλγορίθμου DUKW, για 1 (αριστερά) και 4 (δεξιά) τόπους.



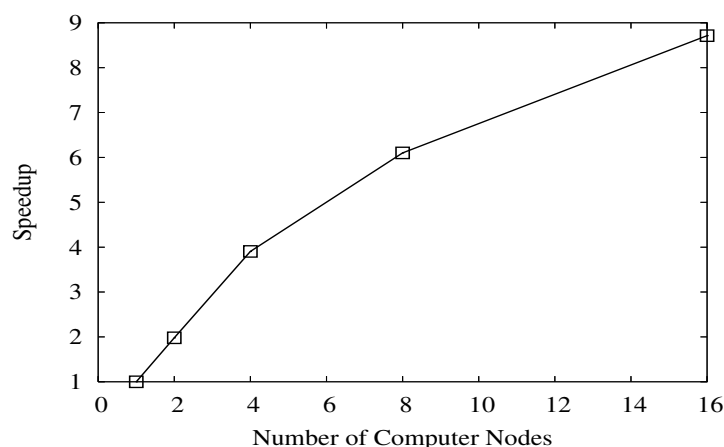
Σχήμα 6.3: Το σύνολο δεδομένων $Dset_d$ με το αποτέλεσμα του αλγορίθμου DUKW, για 16 (αριστερά) και 32 (δεξιά) τόπους.

Τα επόμενα πειραματικά αποτελέσματα αφορούν την μέτρηση της επιτάχυνσης που μπορεί να επιτύχει ο παράλληλος αλγόριθμος PUKW που παρουσιάζεται στην Ενότητα 6.2, ως προς την σειριακή εκτέλεση του UKW.

Για το σκοπό αυτό χρησιμοποιήθηκε η προγραμματιστική βιβλιοθήκη παράλληλου υπολογισμού PVM, σε γλώσσα προγραμματισμού C++, όπου και υλοποιήθηκε ο αλγόριθμος PUKW. Το όλο περιβάλλον βασίστηκε στο λειτουργικό σύστημα Debian GNU, Linux. Το υλικό που χρησιμοποιήθηκε περιλάμβανε 16 Pentium III 900MHz προσωπικούς υπολογιστές, με 32MB μνήμης RAM, ενώ ο κεντρικός τόπος αποτελούν-

ταν από ένα προσωπικό υπολογιστή Pentium 4 1.8GHz με 256MB μνήμης RAM. Όλο το σύστημα ήταν συνδεδεμένο μέσω ενός δικτύου, Fast Ethernet 100MBit/s.

Επίσης κατασκευάστηκε ένα τεχνητό σύνολο δεδομένων $Dset_p$, το οποίο αποτελούνταν από 21000 στοιχεία με 50 χαρακτηριστικά, οργανωμένα σε 10 ομάδες. Όπως απεικονίζεται στο Σχήμα 6.4, ο αλγόριθμος για αυτό το σύνολο δεδομένων κατάφερε να είναι σχεδόν 9 φορές ταχύτερος από τον σειριακό χρησιμοποιώντας και τις 16 μηχανές. Όπως όμως βλέπουμε στο σχήμα, μετά τη χρήση των 8 μηχανών υπάρχει απότομη πτώση της απόδοσης. Αυτό συμβαίνει γιατί από το σημείο αυτό και μετά το πλήθος των μηνυμάτων που μεταδίδονται στο δίκτυο είναι τόσο μεγάλο που περιορίζει την απόδοση του όλου συστήματος.



Σχήμα 6.4: Επιτάχυνση του αλγορίθμου PUKW χρησιμοποιώντας μέχρι και 16 μηχανές.

6.5 Συμπεράσματα - Συνεισφορά

Δοθέντος του γεγονότος ότι η ανάπτυξη αποδοτικών κατανεμημένων και παράλληλων αλγορίθμων, έχει έντονο ενδιαφέρον τα τελευταία χρόνια, παρουσιάστηκαν σε αυτό το κεφάλαιο, δύο εκδοχές του αλγορίθμου UKW, για κατανεμημένα σύστημα με διαφορετικές προϋποθέσεις. Για περιβάλλοντα με πολύ αργά δίκτυα επικοινωνίας ή με έντονα θέματα ασφάλειας παρουσιάστηκε και μελετήθηκε η απόδοση του αλγορίθμου DUKW [134], που στοχεύει σε αποδοτικά αποτελέσματα ομαδοποίησης ελαχιστοποιώντας το επικοινωνιακό κόστος ανάμεσα στους τόπους. Επίσης, παρουσιάστηκε ο παράλληλος αλγόριθμος ομαδοποίησης PUKW [5], και μελετήθηκε η επιτάχυνση που μπορεί να επιτευχθεί σε σχέση με τη σειριακή εκτέλεση. Οι προτεινόμενες τροποποιήσεις του αλγορίθμου μπορεί να δώσουν υψηλής απόδοσης λύσεις στα σύγχρονα κατανεμημένα υπολογιστικά περιβάλλοντα, αξιοποιώντας όσο το δυνατόν καλύτερα τη διαθέσιμη υπολογιστική ισχύ και μειώνοντας το κόστος επικοινωνίας.

Επίσης, ένα φαινόμενο που συναντάται όλο και συχνότερα στις πραγματικές εφαρμογές, είναι διάφοροι οργανισμοί που έχουν διαφορετικές αποθηκευμένες ιδιότητες για τις ίδιες εγγραφές, να θέλουν να εκτελέσουν αλγορίθμους ομαδοποίησης χωρίς να διακινδυνεύσουν την ιδιωτικότητα των δεδομένων. Για αυτό το σκοπό, παρουσιάστηκε μια τροποποίηση του αλγορίθμου UKW, για κάθετα κατανεμημένες βάσεις δεδομένων, με στόχο την διατήρηση της ιδιωτικότητας των δεδομένων [129]. Η προτεινόμενη τροποποίηση, έχει μικρότερο επικοινωνιακό και υπολογιστικό κόστος από άλλες παρόμοιες

μεθόδους [145]. Ακόμα, στα πλεονεκτήματα της συμπεριλαμβάνονται οι δυνατότητες του αυτόματου υπολογισμού του πλήθους των ομάδων και των πολύ αποδοτικών αποτελεσμάτων, που κληρονομούνται άμεσα από τον αρχικό αλγόριθμο UKW.

Ελπιδοφόρες ερευνητικές κατευθύνσεις, φαίνεται να βρίσκονται προς την κατασκευή μεθόδων που ενοποιούν τον υπολογισμό με την επικοινωνία, με ένα τρόπο ώστε να ανιχνεύονται περιοχές του χώρου των δεδομένων, η εξέταση των οποίων πραγματικά απαιτεί την επικοινωνία ανάμεσα στις διαφορετικές οντότητες. Επίσης, η επέκταση των αποτελεσμάτων για την ομαδοποίηση με διατήρηση της ιδιωτικότητας σε ετερογενής κατανεμημένες βάσεις δεδομένων φαίνεται να είναι αρκετά σημαντική για την άμεση χρήση του αλγορίθμου σε πραγματικές εφαρμογές.

Ομαδοποίηση σε Δυναμικές Βάσεις Δεδομένων

Τα άψυχα αντικείμενα μπορούν να κατηγοριοποιηθούν
σε τρεις κατηγορίες· αυτά που δεν δουλεύουν,
αυτά που χαλάνε και αυτά που χάνονται.

—*Russell Baker (1925 μ.Χ. -)*

Οι περισσότεροι αλγόριθμοι ομαδοποίησης στηρίζονται στο γεγονός ότι το σύνολο δεδομένων έχει δημιουργηθεί εκ των προτέρων από μια στατική κατανομή. Όμως, καθώς συλλέγονται τα δεδομένα είναι πιθανόν η διαδικασία που τα παράγει να αλλάξει και μερικές φορές μάλιστα, ραγδαία. Ένα τέτοιο μη στατικό περιβάλλον μπορεί να περιγραφεί από μια δυναμική βάση δεδομένων στην οποία συμβαίνουν εισαγωγές νέων στοιχείων και διαγραφές παλαιών. Αλγόριθμοι ομαδοποίησης που είναι κατασκευασμένοι για ένα τέτοιο περιβάλλον, πρέπει να στοχεύουν στην ανανέωση της ομαδοποίησης σε τακτά χρονικά διαστήματα με το ελάχιστο δυνατό υπολογιστικό κόστος. Μια προσεκτική επισκόπηση των αλγορίθμων ομαδοποίησης [20] ανέδειξε ότι οι περισσότεροι από αυτούς είναι ακατάλληλοι για αυτό το σκοπό. Οι εφαρμογές αυτών των αλγορίθμων είναι πολύ σημαντικές και περιλαμβάνουν την αναδιοργάνωση βάσεων δεδομένων [157], την ανακάλυψη προτύπων χρήσης του Internet [103], και την ομαδοποίηση κειμένων [149].

Από το ευρύ πεδίο της εξόρυξης δεδομένων, μια μέθοδος για την ανανέωση κανόνων συσχέτισης παρουσιάστηκε στην εργασία [28]. Ένα αλγόριθμος γενίκευσης για την αυξητική σύνοψη, προτάθηκε στην εργασία [46]. Ένα αυξητικός αλγόριθμος ομαδοποίησης κειμένων, που προσπαθεί να διατηρήσει ομάδες μικρής διαμέτρου, καθώς νέα στοιχεία εισέρχονται στη βάση δεδομένων προτάθηκε στην εργασία [24]. Ένας ακόμη on-line αλγόριθμος αστέρος για την ομαδοποίηση κειμένων έχει προταθεί στην εργασία [9]. Τέλος, μια τροποποίηση του αλγορίθμου GDBSCAN για τέτοιες δυναμικές βάσεις δεδομένων προτάθηκε στην εργασία [45], όπου και παρουσιάστηκαν πειραματικά αποτελέσματα για την επιτάχυνση που επιτυγχάνει σε σχέση με την πλήρη επανεκτέλεση του αλγορίθμου. Με μια συναφή τεχνική παρουσιάστηκε και μια τροποποίηση [85] του αλγορίθμου OPTICS [8], αλλά με πολύ μικρότερες δυνατότητες επιτάχυνσης από αυτές που αναφέρονται στην εργασία [45]. Αυτό αποδίδεται από τους συγγραφείς στη μεγαλύτερη πολυπλοκότητα του αλγορίθμου OPTICS, καθώς αυτός είναι πιο κατάλληλος για ένα ευρύτερο φάσμα εφαρμογών.

Σε αυτό το κεφάλαιο προτείνουμε μια επέκταση του αλγορίθμου UKW [135], για δυναμικές βάσεις δεδομένων στις οποίες συμβαίνουν εισαγωγές νέων στοιχείων και

διαγραφές παλαιών. Η προτεινόμενη επέκταση ενσωματώνει τη δομή δεικτοδότησης Bkd-tree [116]. Αυτή η δομή είναι ικανή να παραμείνει αποδοτική για μεγάλο πλήθος εισαγωγών και διαγραφών, και παρέχει στον αλγόριθμο ένα μηχανισμό για το συγχρονισμό των ανανεώσεων της ομαδοποίησης.

Στο υπόλοιπο κεφάλαιο, αφού παρουσιάσουμε τη δομή Bkd-tree, περιγράφουμε την προτεινόμενη επέκταση και δίνουμε θεωρητικές αναλύσεις της πολυπλοκότητας της. Έπειτα παρουσιάζουμε αναλυτικά πειραματικά αποτελέσματα, και το κεφάλαιο ολοκληρώνεται με συμπεράσματα.

7.1 Η Δομή Bkd-tree

Σε βάσεις δεδομένων, που υπόκεινται σε μεγάλο αριθμό ανανεώσεων γίνεται δύσκολο το θέμα της δεικτοδότησης των δεδομένων. Μία αποδοτική δομή δεικτοδότησης, πρέπει να έχει τα χαρακτηριστικά, του μικρού υπολογιστικού κόστους στη διάρκεια της διαδικασίας ανανέωσης και να χρησιμοποιεί το διαθέσιμο αποθηκευτικό χώρο όσο το δυνατόν καλύτερα. Για αυτό το λόγο χρησιμοποιούμε την δομή Bkd-tree, που προτάθηκε στην εργασία [116], η οποία έχει τη δυνατότητα υψηλής αξιοποίησης του χώρου και άριστης απόδοσης στη διαδικασία ανανέωσης ανεξαρτήτως του αριθμού των ανανεώσεων.

Αυτή η δομή βασίζεται στις επεκτάσεις της πολύ γνωστής kd-tree δομής, που ονομάζεται K-D-B-tree [122], καθώς και στη λογαριθμική μέθοδο για τη μετατροπή μιας στατικής δομής σε δυναμική. Εκτεταμένα πειραματικά αποτελέσματα [116] έχουν δείξει ότι, η δομή Bkd-tree χρησιμοποιεί σχεδόν στο 100% το διαθέσιμο χώρο, και παρέχει τη γρηγόρη επεξεργασία των ερωτημάτων του K-D-B-tree. Ταυτόχρονα είναι ικανή να διατηρήσει αυτές τις ιδιότητες κάτω από ένα μεγάλο πλήθος ανανεώσεων της βάσης δεδομένων.

Η δομή Bkd-tree, αντί να διατηρεί μόνο μια δενδροειδή δομή την οποία να ισορροπεί δυναμικά μετά από κάθε εισαγωγή στοιχείων, διατηρεί ένα σύνολο από $\log_2(n/M)$, στατικές K-D-B-tree δομές, και οι ανανεώσεις εκτελούνται επαναδημιουργώντας ένα προσεκτικά επιλεγμένο σύνολο από αυτές, σε τακτά χρονικά διαστήματα. Στην έκφραση $\log_2(n/M)$ το n αντιπροσωπεύει το πλήθος των στοιχείων και M το μέγεθος της προσωρινής μνήμης σε πλήθος σημείων. Για την απάντηση μιας ερώτησης διαστήματος η δομή Bkd-tree, χρησιμοποιεί όλες τις $\log_2(n/M)$ δομές. Παρόλα αυτά, η πολυπλοκότητα της χειρότερης περίπτωσης είναι:

$$C = O(dn^{1-(1/d)} + s), \quad (7.1)$$

όπου το s αντιπροσωπεύει το πλήθος των ανακτώμενων στοιχείων. Χρησιμοποιώντας ένα βέλτιστο $O(n \log_m(n))$ μαζικής φόρτωσης αλγόριθμο, μια εισαγωγή ενός στοιχείου γίνεται σε χρόνο $O(\log_m(n) \log_2(n/M))$. Μία διαγραφή γίνεται απλώς ανακαλύπτοντας τη δομή που το περιέχει, και έπειτα το στοιχείο διαγράφεται από αυτήν. Εφόσον υπάρχουν το πολύ $\log_2(n/M)$ δομές, ο αριθμός των υπολογισμών που πρέπει να συμβούν για κάθε διαγραφή είναι το πολύ $\log(n) \log_2(n/M)$ [116].

Οι εισαγωγές νέων στοιχείων αντιμετωπίζονται τελείως διαφορετικά. Οι περισσότερες εισαγωγές $((M-1)$ από τις M συνεχόμενες) γίνονται στη δομή T_0 που βρίσκεται στην προσωρινή μνήμη. Όταν αυτή η δομή αποκτήσει το μέγιστο πλήθος σημείων (M) που μπορεί να αποθηκεύσει, τότε υπολογίζεται το μικρότερο j , τέτοιο ώστε η δομή T_j να είναι κενή. Έπειτα, όλα τα στοιχεία από τη δομή T_0 και τις T_i για $0 \leq i < j$,

εισάγονται με τον αλγόριθμο μαζικής φόρτωσης στη δομή T_j . Με άλλα λόγια, τα στοιχεία εισέρχονται στη δομή T_0 , και περιοδικά οργανώνονται σε όλο και μεγαλύτερες K-D-B-tree δομές.

7.2 Ο Αλγόριθμος UKW σε Δυναμικές Βάσεις Δεδομένων

Η προτεινόμενη δυναμική επέκταση του αλγορίθμου UKW, χρησιμοποιεί τη δομή Bkd-tree. Η επιλογή αυτής της δομής, έγινε γιατί επιτρέπει τη γρήγορη επεξεργασία των ερωτημάτων διαστήματος, και ταυτόχρονα παρέχει ένα κριτήριο, για το συγχρονισμό, των διαδικασιών ανανέωσης του αποτελέσματος της ομαδοποίησης. Το παρακάτω σχήμα συνοψίζει τις λειτουργίες του δυναμικού αλγορίθμου:

1. Υποθέτουμε ότι η εκτέλεση του αλγορίθμου στη αρχική βάση δεδομένων έχει γίνει, και έχει καταλήξει σε ένα σύνολο από παράθυρα που αποτελούν το αποτέλεσμα της ομαδοποίησης.
2. Σε συγκεκριμένα χρονικά σημεία εκτέλεσε:
 - (α') Αντιμετώπισε τις εισαγωγές νέων δεδομένων.
 - (β') Αντιμετώπισε τις διαγραφές των παλιών δεδομένων.
3. Μετά το τέλος του κάθε ενός από τα παραπάνω βήματα ανανέωσε το σύνολο των παραθύρων άρα και το αποτέλεσμα της ομαδοποίησης.

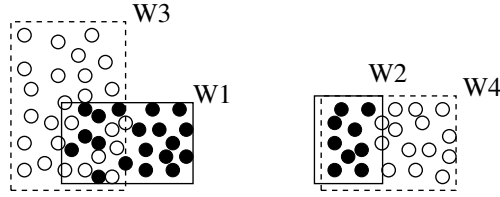
Μέσω αυτού του σχήματος η δυναμική επέκταση που προτείνεται είναι δυνατόν να προσαρμόζει συνεχώς τις ανακαλυφθέντες ομάδες στις αλλαγές των δεδομένων με μικρό κόστος. Στις επόμενες ενότητες θα παρουσιάσουμε κάθε ένα από τα βήματα του σχήματος ξεχωριστά.

7.2.1 Αντιμετώπιση της Εισαγωγής Νέων Στοιχείων

Αρχικά θεωρούμε ότι η εφαρμογή του στατικού UKW αλγορίθμου έχει γίνει στην αρχική βάση δεδομένων, και έχει παράγει ένα σύνολο παραθύρων που περιγράφουν την ομαδοποίηση των δεδομένων.

Καθώς ο αριθμός των εισαγωγών νέων στοιχείων, αναγκάζει την δομή T_0 του Bkd-tree, να αποκτήσει το μέγιστο πλήθος σημείων (M) που μπορεί να αποθηκεύσει, ένα πλήθος παραθύρων αρχικοποιείται πάνω σε αυτά τα σημεία. Έπειτα, τα παράθυρα υπόκεινται στις διαδικασίες μετακίνησης και διεύρυνσης του αλγορίθμου UKW, λαμβάνοντας όμως υπόψιν το πλήρες Bkd-tree, και όχι μόνο τη δομή T_0 . Μετά το τέλος αυτών των διαδικασιών, τα νέα παράθυρα μαζί με τα παλαιά υπόκεινται στη διαδικασία ενοποίησης του αλγορίθμου UKW. Το τελικό σύνολο παραθύρων είναι και αυτό που περιγράφει την ομαδοποίηση των δεδομένων. Ένα παράδειγμα της διαδικασίας απεικονίζεται στο Σχήμα 7.1. Πιο συγκεκριμένα, σε αυτό το σχήμα, οι κύκλοι αναπαριστούν τα αρχικά στοιχεία που υπήρχαν στη βάση, ενώ οι δίσκοι τα εισηγμένα στοιχεία. Τα παράθυρα $W1$ και $W2$, έχουν οριστικοποιηθεί από την αρχική εκτέλεση του αλγορίθμου. Τα παράθυρα $W3$ και $W4$ αρχικοποιούνται στα στοιχεία που έχουν εισαχθεί, και αφού επεξεργαστούν από τις διαδικασίες μετακίνησης και διεύρυνσης, υπόκεινται μαζί με τα αρχικά στη διαδικασία ενοποίησης. Τα παράθυρα $W1$ και $W3$ θεωρούνται

ότι καλύπτουν μέρη της ίδιας ομάδας, ενώ τα παράθυρα $W2$ και $W4$, θεωρούνται ότι καλύπτουν την ίδια ομάδα οπότε διαγράφεται το παράθυρο $W2$.



Σχήμα 7.1: Η αντιμετώπιση των εισαγωγών σημείων από τον αλγόριθμο.

Αν θεωρήσουμε ένα άνω φράγμα L για τον αριθμό των ερωτημάτων διαστημάτων που επιτρέπονται στον αλγόριθμο για την επεξεργασία του κάθε παραθύρου, τότε μπορούμε να διατυπώσουμε το ακόλουθο λήμμα:

Λήμμα 7.1. Έστω D το σύνολο των αρχικών n στοιχείων στη βάση δεδομένων. Έστω επίσης ότι συμβαίνουν m εισαγωγές στοιχείων που σχηματίζουν το σύνολο D' . Έστω ακόμη ότι ο αλγόριθμος UKW, εκτελείται στο σύνολο D , χρησιμοποιώντας $k = c|D|$, $0 < c \leq 1$, αρχικά παράθυρα που καταλήγουν σε λ τελικά παράθυρα. Τότε, η προτεινόμενη δυναμική επέκταση του αλγορίθμου, εφαρμοσμένη στο σύνολο $D \cup D'$, χρησιμοποιώντας $k' = c|D'|$ αρχικά παράθυρα από το σύνολο D' , και λαμβάνοντας υπόψη τα λ παράθυρα στη διαδικασία της ενοποίησης, έχει καλύτερη πολυπλοκότητα χειρότερης περίπτωσης, ως προς τον στατικό αλγόριθμο εφαρμοσμένο στο σύνολο $D \cup D'$ χρησιμοποιώντας $k'' = c|D \cup D'|$ αρχικά παράθυρα.

Απόδειξη. Ακολουθώντας τις παραπάνω υποθέσεις και κάνοντας χρήση της σχέσης (7.1), η πολυπλοκότητα χειρότερης περίπτωσης για τον στατικό UKW αλγόριθμο εφαρμοσμένο στο σύνολο $D \cup D'$ είναι:

$$\begin{aligned} C_s &= O(L k'' (d|D + D'|^{1-(1/d)} + s_1)) = \\ &O(L c|D + D'| (d|D + D'|^{1-(1/d)} + s_1)), \end{aligned} \quad (7.2)$$

όπου s_1 , το άνω όριο των συνολικών ανακτηθέντων στοιχείων, που φράσσεται από το μέγεθος του συνόλου $|D + D'|$. Από την άλλη, η δυναμική έκδοση του αλγορίθμου έχει πολυπλοκότητα χειρότερης περίπτωσης:

$$\begin{aligned} C_d &= O(L k' (d|D'|^{1-(1/d)} + s_2)) = \\ &O(L c|D'| (d|D'|^{1-(1/d)} + s_2)), \end{aligned} \quad (7.3)$$

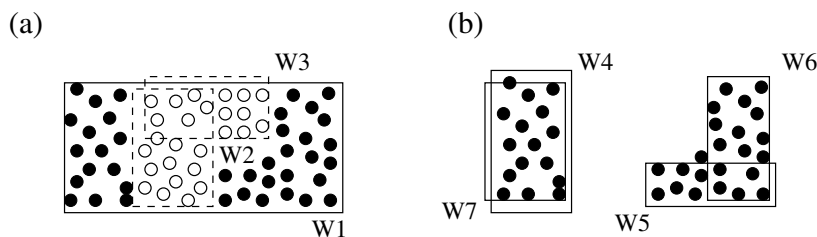
όπου s_2 , το άνω όριο των συνολικών ανακτηθέντων στοιχείων, που φράσσεται από το μέγεθος του συνόλου $|D'|$. Τώρα εφόσον, $|D + D'| < |D'|$ είναι προφανές ότι $C_d < C_s$. $\square$

7.2.2 Αντιμετώπιση της Διαγραφής Στοιχείων

Η προτεινόμενη δυναμική έκδοση του αλγορίθμου, για να αντιμετωπίσει τις διαγραφές στοιχείων, διατηρεί μια δεύτερη δομή Bkd-tree. Κάθε φορά που ένα στοιχείο διαγράφεται από την βάση δεδομένων τότε αυτό εισάγεται στη δεύτερη αυτή δομή. Όταν το μέγεθος της δεύτερης δομής φτάσει το μέγιστο πλήθος σημείων που μπορεί να

αποθηκεύσει, ένα πλήθος παραθύρων αρχικοποιείται πάνω στα διαγραφόμενα στοιχεία. Αυτά τα παράθυρα, επεξεργάζονται από τις διαδικασίες μετακίνησης και διεύρυνσης, με πληροφορίες όμως μόνο από τη δομή των διαγραφόμενων στοιχείων. Έπειτα εξετάζονται για ομοιότητα από τη διαδικασία ενοποίησης, με τα ήδη επεξεργασμένα παράθυρα. Αν ένα επεξεργασμένο παράθυρο βρεθεί να είναι όμοιο με ένα από τα παράθυρα των διαγραφόμενων σημείων, τότε αυτό διαγράφεται. Αν περιείχε ένα αρκετά μεγάλο αριθμό σημείων, τότε νέα παράθυρα αρχικοποιούνται σε αυτά, και επεξεργάζονται κανονικά σαν νέα παράθυρα. Σε αυτό το σημείο, η δομή που κρατά τα διαγραφόμενα στοιχεία αδειάζει.

Στο Σχήμα. 7.2, απεικονίζεται ένα παράδειγμα αυτής της διαδικασίας. Οι κύκλοι αντιπροσωπεύουν, τα αρχικά στοιχεία της βάσης, ενώ οι δίσκοι, τα διαγραφόμενα. Το παράθυρο $W1$, έχει ήδη οριστικοποιηθεί από την αρχική εκτέλεση του αλγορίθμου. Τα παράθυρα $W2$ και $W3$, έχουν αρχικοποιηθεί πάνω από τα διαγραφόμενα σημεία. Αφού επεξεργαστούν από τις διαδικασίες μετακίνησης και διεύρυνσης, εξετάζονται για ομοιότητα με το αρχικό παράθυρο, $W1$. Το παράθυρο $W1$, ικανοποιεί το κριτήριο ομοιότητας με τα παράθυρα $W2$ και $W3$, οπότε διαγράφεται. Εφόσον περιείχε ένα μεγάλο πλήθος σημείων, τέσσερα νέα παράθυρα αρχικοποιούνται σε αυτά, που καταλήγουν στις θέσεις $W4$, $W5$, $W6$ και $W7$. Αυτά τα παράθυρα επεξεργάζονται από την διαδικασία ενοποίησης. Το παράθυρο $W7$ διαγράφεται καθώς ικανοποιεί το κριτήριο ομοιότητας με το παράθυρο $W4$, ενώ τα παράθυρα $W5$ και $W6$ θεωρούνται ότι καλύπτουν μέρη της ίδιας ομάδας.



Σχήμα 7.2: (a) Η εφαρμογή του αλγορίθμου στα διαγραφόμενα σημεία. (b) Τα τελικά παράθυρα που παράγει ο δυναμικός αλγόριθμος.

Ομαδοποιώντας τα διαγραφόμενα στοιχεία, ο δυναμικός αλγόριθμος, είναι ικανός να αναγνωρίσει τα παράθυρα που χρειάζονται επαναεπεξεργασία, και έτσι αναδιοργανώνει το αρχικό του αποτέλεσμα. Με αυτό τον τρόπο, η επιτάχυνση που μπορεί να προσφέρει ο δυναμικός αλγόριθμος δεν εξαρτάται μόνο από το πλήθος των διαγραφών αλλά και από τις αλλαγές που επιβάλουν στο αρχικό αποτέλεσμα της ομαδοποίησης. Το ακόλουθο Λήμμα μπορεί να διατυπωθεί, υπό την υπόθεση ενός μέγιστου αριθμού, L , στα επιτρεπόμενα ερωτήματα διαστήματος για την επεξεργασία του κάθε παραθύρου.

Λήμμα 7.2. Έστω D το αρχικό σύνολο n σημείων στη βάση δεδομένων. Υποθέτουμε ότι συμβαίνουν m διαγραφές στοιχείων που σχηματίζουν το σύνολο $D_0 \subset D$, υπό τον περιορισμό $|D_0| \leq 0.5|D|$. Υποθέτουμε ακόμη ότι ο αλγόριθμος UKW, εφαρμόζεται στο σύνολο D χρησιμοποιώντας $k = c|D|$, $0 < c \leq 1$ αρχικά παράθυρα, που καταλήγουν σε λ τελικά. Έστω ότι, η εφαρμογή του αλγορίθμου UKW, στο σύνολο $D - D_0$ χρησιμοποιώντας $k' = c|D - D_0|$ παράθυρα, καταλήγει σε λ' τελικά παράθυρα, ενώ η εφαρμογή του στο σύνολο D_0 , χρησιμοποιώντας $k_0 = c|D_0|$ καταλήγει σε λ_0 τελικά παράθυρα. Τότε η δυναμική επέκταση του αλγορίθμου έχει καλύτερη πολυπλοκότητα

χειρότερης περίπτωσης από τον στατικό αλγόριθμο, εάν, $|D'| \leq |D| - 2|D_0|$ όπου D' είναι το υπο-σύνολο των σημείων του D , στα οποία αρχικοποιούνται νέα παράθυρα.

Απόδειξη. Από τις υποθέσεις του λήμματος, η πολυπλοκότητα χειρότερης περίπτωσης, του στατικού αλγόριθμου, εφαρμοσμένου στο σύνολο $D - D_0$ είναι:

$$\begin{aligned} C_s &= O(Lk'(d|D - D_0|^{1-(1/d)} + s_1)) = \\ &O(Lc|D - D_0|(d|D - D_0|^{1-(1/d)} + s_1)), \end{aligned} \quad (7.4)$$

όπου όπου s_1 , το άνω όριο των συνολικών ανακτημένων στοιχείων, που φράσσεται από το μέγεθος του συνόλου $|D - D_0|$.

Από την άλλη μεριά, η πολυπλοκότητα χειρότερης περίπτωσης για τον δυναμικό αλγόριθμο είναι:

$$\begin{aligned} C_d &= O(Lk_0(d|D_0|^{1-(1/d)} + s_2)) + O(Lk''(d|D - D_0|^{1-(1/d)} + s_3)) \\ &= O(Lc|D_0|(d|D_0|^{1-(1/d)} + s_2)) + O(Lc|D'|(|D - D_0|^{1-(1/d)} + s_3)), \end{aligned}$$

όπου το k'' αντιπροσωπεύει το πλήθος των νέων παραθύρων που πρέπει να αρχικοποιηθούν, πάνω από το D' , και s_2, s_3 , τα άνω όρια των συνολικών ανακτηθέντων στοιχείων, που φράσσονται από τα μεγέθη των συνόλων $|D_0|$ και $|D - D_0|$, αντίστοιχα.

Εφόσον $|D_0| \leq |D - D_0|$ και $s_2 \leq s_3$, έχουμε το ακόλουθο όριο για το C_d :

$$\begin{aligned} C_d &\leq O(Lc|D_0|(d|D - D_0|^{1-(1/d)} + s_2)) + \\ &O(Lc|D'|(|D - D_0|^{1-(1/d)} + s_3)) \\ &\leq O(Lc(|D_0| + |D'|)(d|D - D_0|^{1-(1/d)} + s_3)). \end{aligned} \quad (7.5)$$

Τώρα εφόσον έχουμε υποθέσει ότι, $|D'| \leq |D| - 2|D_0|$:

$$|D_0| + |D'| \leq |D - D_0|. \quad (7.6)$$

Χρησιμοποιώντας τις σχέσεις (7.4), (7.5) και (7.6), ισχύει ότι:

$$\begin{aligned} O(Lc(|D_0| + |D'|)(d|D - D_0|^{1-(1/d)} + s_3)) &\leq \\ O(Lc|D - D_0|(d|D - D_0|^{1-(1/d)} + s_1)). \end{aligned} \quad (7.7)$$

Από τα προηγούμενα καταλήγουμε στο ότι $C_d \leq C_s$. □

7.2.3 Η Δυναμική Επέκταση του Αλγορίθμου UKW

Βασιζόμενη στις παραπάνω περιγραφές προτείνουμε την ακόλουθη αλγοριθμική διαδικασία για την δυναμική επέκταση του αλγορίθμου UKW.

Δυναμικός Αλγόριθμος UKW

Αρχικοποίηση τις παραμέτρους του αλγορίθμου UKW.

Αρχικοποίηση ένα κενό σύνολο παραθύρων $\mathcal{W}$.

Κάθε φορά που η δομή T_0 γεμίζει:

Αρχικοποίηση ένα σύνολο k παραθύρων $\mathcal{I}$ πάνω στα δεδομένα της δομής T_0 .

Εκτέλεσε τις διαδικασίες μετακίνησης και διεύρυνσης για τα παράθυρα στο $\mathcal{I}$.

Θέσε $\mathcal{W} = \mathcal{W} \cup \mathcal{I}$.

Εκτέλεσε την διαδικασία ενοποίησης για τα παράθυρα στο $\mathcal{W}$.
Εάν συμβεί ένα μεγάλο πλήθος διαγραφών:
Αρχικοποίησε ένα σύνολο k παραθύρων $\mathcal{D}$ πάνω στα διαγραφέντα σημεία.
Εκτέλεσε τον αλγόριθμο UKW για τα παράθυρα στο $\mathcal{D}$.
Εάν παράθυρα από το $\mathcal{D}$ ικανοποιούν το κριτήριο ομοιότητας με παράθυρα στο $\mathcal{W}$:
Τότε διέγραψε από το $\mathcal{W}$ τα παράθυρα αυτά.
Εάν τα διαγραφόμενα παράθυρα περιείχαν μη διαγραφόμενα σημεία, εκτέλεσε τον αλγόριθμο UKW σε αυτά.
Ανέφερε το σύνολο $\mathcal{W}$ που αποτελεί την τελική ομαδοποίηση.

Χρησιμοποιώντας τα προηγούμενα Λήμματα, διατυπώνουμε το ακόλουθο θεώρημα, για τον παραπάνω αλγόριθμο:

Θεώρημα 7.1. Έστω ότι οι υποθέσεις των Λημμάτων 7.1 και 7.2, ισχύουν. Τότε, ο δυναμικός αλγόριθμος UKW επιτυγχάνει μια καλύτερη πολυπλοκότητα χειρότερης περίπτωσης από τον στατικό αλγόριθμο UKW.

Απόδειξη. Η απόδειξη είναι προφανής από τα Λήμματα 7.1 και 7.2. □

Μια σημαντική παράμετρος για την εκτέλεση του ανωτέρου αλγορίθμου είναι η επιλογή, του μεγέθους M , της δομής T_0 του Bkd-tree. Αυτή η παράμετρος έμμεσα καθορίζει τον συγχρονισμό, των ανανεώσεων της βάσης δεδομένων [116], η οποία με τη σειρά της ξεκινά και την ανανέωση των αποτελεσμάτων της ομαδοποίησης. Για αυτό το λόγο κατά την επιλογή της παραμέτρου, πρέπει να λάβουμε υπόψη την διαθέσιμη υπολογιστική ισχύ, τα επιθυμητά διαστήματα ανανέωσης της ομαδοποίησης, καθώς και το μέγεθος της τρέχουσας εφαρμογής.

7.3 Πειραματικά Αποτελέσματα

Για την μέτρηση της απόδοσης του προτεινόμενου αλγορίθμου χρησιμοποιήθηκαν δύο τεχνητά σύνολα δεδομένων (Dset1 και Dset2) και ένα πραγματικό Dset3. Τα δύο τεχνητά είναι δισδιάστατα για να επιτρέψουν την οπτικοποίηση των αποτελεσμάτων. Τέτοιου είδους σύνολα δεδομένων έχουν χρησιμοποιηθεί για την αξιολόγηση και άλλων αλγορίθμων [65, 82, 123].

Το σύνολο δεδομένων Dset1, περιέχει 1600 στοιχεία οργανωμένα σε 4 ομάδες διαφορετικών μεγεθών. Το σύνολο αυτό χωρίστηκε σε 4 μέρη των 400 σημείων το κάθε ένα τυχαία, τα οποία παρουσιάστηκαν σταδιακά στον αλγόριθμο. Κάθε φορά που παρουσιάζεται ένα μέρος σαν εισαγωγή νέων στοιχείων στη βάση δεδομένων, 32 παράθυρα αρχικοποιόντουσαν σε αυτό. Τα αποτελέσματα της ανανέωσης της ομαδοποίησης του προτεινόμενου δυναμικού αλγορίθμου, απεικονίζονται στο Σχήμα 7.3. Όπως είναι φανερό στα τμήματα (a) και (b) ο αλγόριθμος αναγνωρίζει 7 ομάδες που όπως φαίνεται οπτικά είναι και το σωστό αποτέλεσμα. Στο τμήμα (c), οι πάνω δεξιά και κάτω αριστερά ομάδες ενοποιούνται, οπότε και τελικά φτάνουμε στο τμήμα (d), όπου έχουν αναγνωριστεί σωστά οι 4 ομάδες του συνόλου.

Το δεύτερο τεχνητό σύνολο δεδομένων Dset2, περιέχει 2760 στοιχεία οργανωμένα σε τρεις κυρτές μορφολογικά και μία μη-κυρτή ομάδα. Αυτό το σύνολο χωρίστηκε σε 5 μέρη. Το πρώτο μέρος περιείχε 920, στοιχεία, ενώ τα υπόλοιπα 4 μέρη 460. Σταδιακά

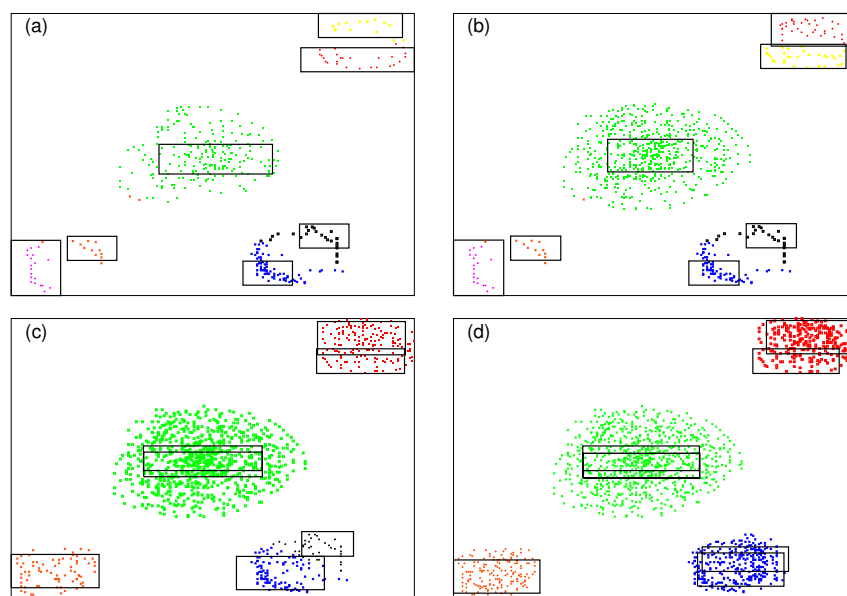
το κάθε μέρος παρουσιάζονταν σαν εισαγωγή νέων σημείων, στον αλγόριθμο, ο οποίος αρχικοποιούσε 64 παράθυρα στα νέα σημεία κάθε φορά. Στο Σχήμα 7.4, απεικονίζονται τα αποτελέσματα του δυναμικού αλγόριθμου. Όπως φαίνεται στα τμήματα (a) και (b), ο αλγόριθμος ανακαλύπτει τις δύο ομάδες. Στη συνέχεια στα τμήματα (c), (d) και (e), αναγνωρίζονται σωστά οι 3 κυρτές ομάδες καθώς και η μία μη-κυρτή. Σε αυτό το σημείο, αφαιρούνται 460 σημεία από τη βάση δεδομένων, και εκτελείται η διαδικασία του δυναμικού αλγορίθμου, που αντιμετωπίζει τις διαγραφές σημείων, αρχικοποιώντας 64 παράθυρα στα διαγραφόμενα σημεία. Μετά το τέλος, της διαδικασίας τα αποτελέσματα του αλγορίθμου απεικονίζονται στο τμήμα (f), του Σχήματος 7.4

Το σύνολο δεδομένων Dset3 αποτελεί ένα μέρος του διαγωνισμού KDD 1999 [83], και έχει κατασκευαστεί το 1998 από το DARPA Intrusion Detection Evaluation Program στο MIT Lincoln Labs. Ο σκοπός εκείνου του προγράμματος ήταν η επισκόπηση της έρευνας στην ανίχνευση διαρρήξεων υπολογιστικών συστημάτων. Από αυτό το σύνολο χρησιμοποιήθηκαν τα πρώτα 100000 στοιχεία. Τα 77888 από αυτά έχουν χαρακτηριστεί σαν normal ενώ τα υπόλοιπα 22112, σαν επιθέσεις στα συστήματα τύπου Denial of Service (DoS). Από τα 42 χαρακτηριστικά χρησιμοποιήθηκαν τα 37 αριθμητικά. Αφού τα δεδομένα ανακατεύτηκαν, χωρίστηκαν σε 10 μέρη με 10000 στοιχεία το καθένα. Το κάθε μέρος παρουσιάστηκε σταδιακά στον αλγόριθμο, σαν εισαγωγές νέων στοιχείων στη βάση δεδομένων. Για να αξιολογήσουμε τα αποτελέσματα του δυναμικού αλγορίθμου θα τα συγκρίνουμε με αυτά του στατικού UKW, όταν εφαρμόζεται σε ολόκληρο το σύνολο δεδομένων. Εφαρμόζοντας τον αλγόριθμο UKW σε ολόκληρο το σύνολο δεδομένων με 16 αρχικά παράθυρα, αυτός καταλήγει σε 7 ομάδες, από τις οποίες μία περιέχει 22087 στοιχεία τύπου DoS. Οι υπόλοιπες έξι περιέχουν αποκλειστικά στοιχεία τύπου normal, εκτός από μία που περιέχει και 24 DoS, στοιχεία. Η εφαρμογή του δυναμικού αλγορίθμου, αρχικοποιώντας 8 παράθυρα κάθε φορά που επεξεργάζεται ένα νέο τμήμα, καταλήγει σε 22 ομάδες. Όμοια πάλι υπάρχει μια ομάδα με 22087 στοιχεία τύπου DoS, ενώ οι υπόλοιπες ομάδες περιέχουν αποκλειστικά στοιχεία τύπου normal, εκτός 24 στοιχεία τύπου DoS που είναι διασκορπισμένα σε 4 ομάδες. Όπως είναι προφανές η δυναμική έκδοση του αλγορίθμου καταλήγει σε μία ομαδοποίηση με την ίδια φυσική αναπαράσταση που επιτυγχάνει η ομαδοποίηση του στατικού UKW.

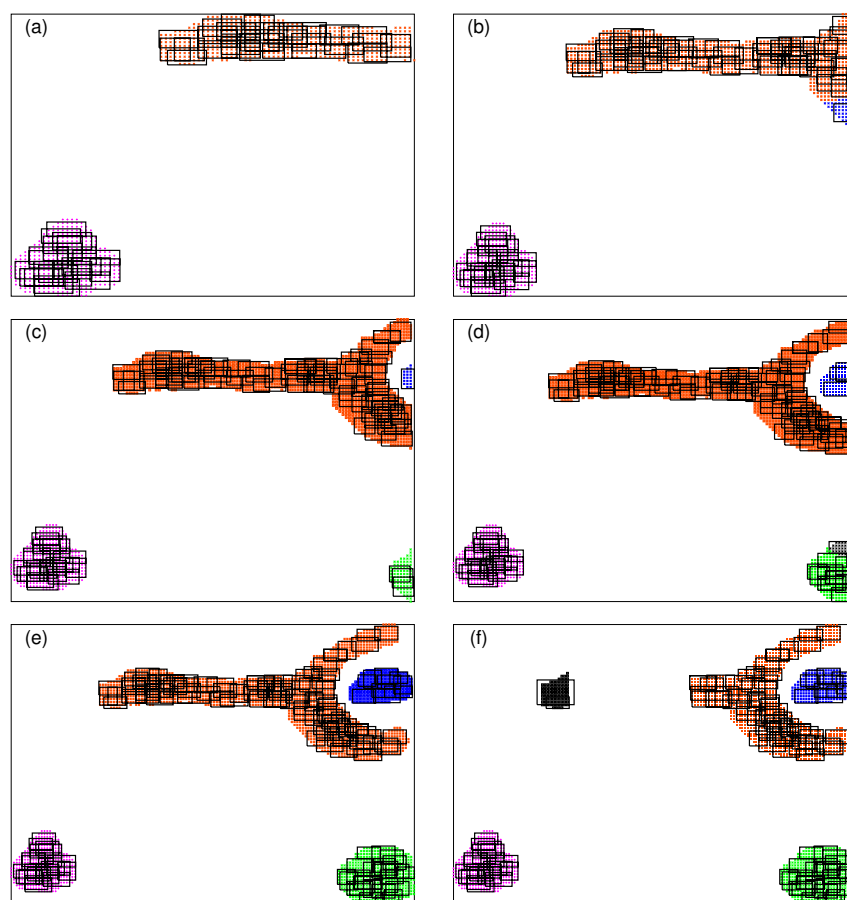
Τέλος, για συγκριτικούς σκοπούς με τις εργασίες [85,123], εκτελέστηκαν και πειραματικές μετρήσεις της επιτάχυνσης (Speedup), που επιτυγχάνει ο δυναμικός αλγόριθμος. Για το σκοπό αυτό κατασκευάστηκε, το σύνολο δεδομένων Dset4, που περιέχει 100 τυχαία διασκορπισμένες ομάδες στο διάστημα $[10, 200]$ ¹⁰. Κάθε ομάδα αποτελείται από 1000 στοιχεία που έχουν ληφθεί δειγματολογικά από μια κανονική κατανομή με διασπορά σε κάθε διάσταση έναν τυχαίο αριθμό στο διάστημα $[1, 3]$. Για να υπολογίσουμε το (Speedup), μετρήσαμε τον υπολογιστικό χρόνο που χρειάζεται ο στατικός αλγόριθμος για να επανεκτελεστεί στην ανανεωμένη βάση δεδομένων και το συγκρίναμε με το χρόνο που χρειάζεται ο δυναμικός αλγόριθμος,

Τα αποτελέσματα απεικονίζονται στα Σχήματα 7.5, και 7.6. Για την περίπτωση των εισαγωγών νέων στοιχείων (Σχήμα 7.5), ο δυναμικός αλγόριθμος καταφέρνει να είναι μέχρι και 906.96 φορές γρηγορότερος από τον στατικό όταν συμβαίνουν 100 εισαγωγές στοιχείων σε μία βάση δεδομένων μεγέθους 90000 στοιχείων. Για μεγαλύτερο αριθμό εισαγωγών 1000, η επιτάχυνση παρόλο που είναι μικρότερη 92.414, δείχνει να είναι ανάλογη του λόγου του αριθμού των εισαγωγών προς το μέγεθος της βάσης δεδομένων.

Στην περίπτωση των διαγραφών στοιχείων (Σχήμα 7.6) οι επιταχύνσεις είναι ακόμη μεγαλύτερες. Για παράδειγμα, όταν το μέγεθος της βάσης είναι 900100 η επιτάχυνση

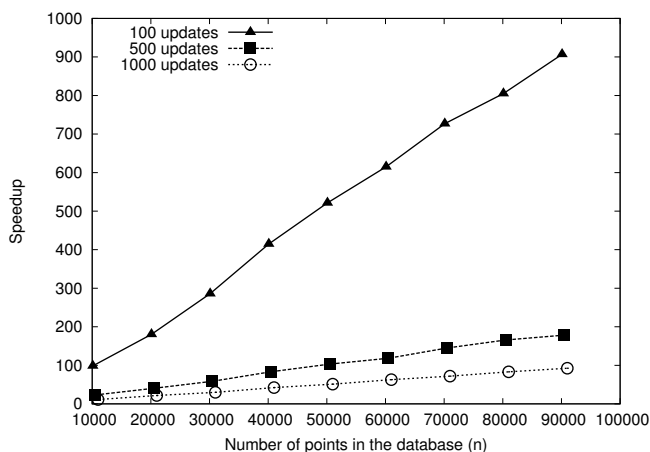


Σχήμα 7.3: Εφαρμόζοντας τον δυναμικό αλγόριθμο στις 4 διαφορετικές καταστάσεις του συνόλου Dset1.

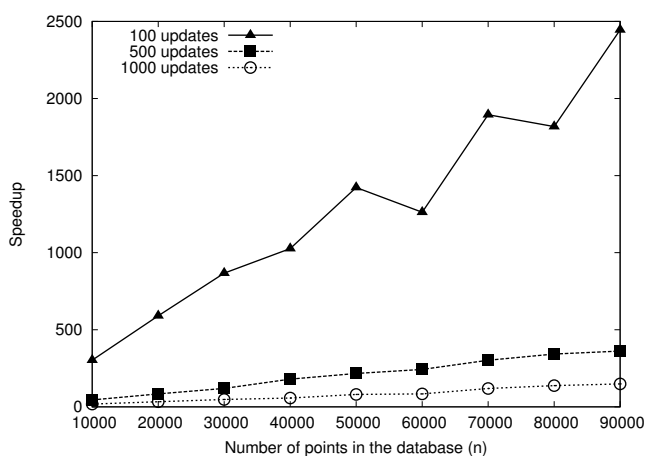


Σχήμα 7.4: Εφαρμόζοντας τον δυναμικό αλγόριθμο στα 6 διαφορετικές καταστάσεις του συνόλου Dset2.

φτάνει τις τιμές 2445.23 και 148.934 για 100 και 1000 διαγραφές αντίστοιχα. Όπως επίσης φαίνεται και στο σχήμα στην περίπτωση αυτή η επιτάχυνση δεν είναι μονότονη ως προς το μέγεθος της βάσης αφού εξαρτάται και από την μεταβολή που επιφέρουν οι αλλαγές στην εκάστοτε ομαδοποίηση,



Σχήμα 7.5: Η επιτάχυνση (Speedup) του δυναμικού αλγορίθμου, για διάφορους αριθμούς εισαγωγής στοιχείων, και για διάφορα μεγέθη της βάσης δεδομένων.



Σχήμα 7.6: Η επιτάχυνση (Speedup) του δυναμικού αλγορίθμου, για διάφορους αριθμούς διαγραφής στοιχείων, και για διάφορα μεγέθη της βάσης δεδομένων.

7.4 Συμπεράσματα - Συνεισφορά

Οι σύγχρονες βάσεις δεδομένων, συγκεντρώνουν δεδομένα με εκπληκτικά γρήγορους ρυθμούς, για πολύ εκτενείς χρονικές περιόδους. Αυτό το γεγονός, έχει δημιουργήσει νέες προκλήσεις για την εξόρυξη πληροφορίας. Μια τέτοια πρόκληση είναι η ανάπτυξη αλγορίθμων που έχουν την δυνατότητα, να παρακολουθήσουν την αλλαγή στην ομαδοποίηση των δεδομένων καθώς αυτή συμβαίνει σταδιακά. Σε αυτό το κεφάλαιο παρουσιάσαμε μια επέκταση του αλγορίθμου UKW, για δυναμικές βάσεις δεδομένων, που ενσωματώνει τη δομή δεικτοδότησης Bkd-tree. Επιπλέον παρουσιάσαμε

θεωρητικά αποτελέσματα για την πολυπλοκότητα χειρότερης περίπτωσης του αλγορίθμου. Τα πειραματικά αποτελέσματα, ανέδειξαν την ικανότητα του αλγορίθμου να ανανεώνει το αποτέλεσμα της ομαδοποίησης με μεγάλη ακρίβεια, επιτυγχάνοντας ταυτόχρονα επιτάχυνση σε σχέση με τη στατική έκδοση, που είναι συγκρίσιμη με αυτές άλλων προσεγγίσεων [45, 85].

Μέρος III

Εφαρμογές

Ομαδοποίηση σε Εκφράσεις mRNA

Τα κύτταρα μας επιτρέπουν να περπατάμε,
να μιλάμε, να σκεπτόμαστε, να κάνουμε έρωτα
και να καταλαβαίνουμε ότι το νερό του μπάνιου είναι κρύο.

—Lorraine Lee Cudmore

Σε κάθε ζωντανό οργανισμό που υπόκειται μία βιολογική διαδικασία, εκφράζονται διαφορετικά υποσύνολα των γονιδίων του. Η κανονική λειτουργία ενός κυττάρου, επηρεάζεται σημαντικά, από την γονιδιακή του έκφραση σε ένα δεδομένο στάδιο. Για να γίνουν κατανοητές οι βιολογικές διαδικασίες, θα πρέπει να μετρηθούν, τα επίπεδα έκφρασης των γονιδίων σε διαφορετικές φάσεις ανάπτυξης, σε διαφορετικούς ιστούς, διαφορετικές κλινικές συνθήκες και διαφορετικούς οργανισμούς. Αυτό το είδος της πληροφορίας μπορεί να βοηθήσει στον χαρακτηρισμό της γονιδιακής λειτουργίας, και την κατανόηση μοριακών βιολογικών διαδικασιών [76].

Συγκρινόμενοι με τις παραδοσιακές προσεγγίσεις στη γονιδιακή έρευνα, που βασίζονται στη συλλογή και εξέταση δεδομένων για ένα μόνο γονίδιο τοπικά, οι τεχνολογίες μικρο-συστοιχιών DNA έχουν δώσει τη δυνατότητα της παρακολούθησης των εκφράσεων χιλιάδων γονιδίων ταυτόχρονα. Δυστυχώς, η πραγματική γονιδιακή έκφρασή λαμβάνεται μαζί με θόρυβο, τιμές που λείπουν, και συστηματικές διαφοροποιήσεις εξαιτίας της πειραματικής διαδικασίας. Πολλές μεθοδολογίες έχουν χρησιμοποιηθεί για να αντιμετωπίσουν αυτά τα προβλήματα, όπως οι τεχνικές Singular Value Decomposition, weighted k -nearest neighbors, μέσοι όροι γραμμών, επανάληψη των πειραμάτων για την μοντελοποίηση του θορύβου, και κανονικοποίηση, που είναι η διαδικασία της αναγνώρισης και διαγραφής των συστηματικών πηγών διαφοροποίησης. Αφού τα επίπεδα των γονιδιακών εκφράσεων μετρηθούν, τα δεδομένα αναπαρίστανται από ένα μητρώο X πραγματικών τιμών. Οι γραμμές του μητρώου αντιστοιχούν σε διανύσματα που αντιπροσωπεύουν τα διανύσματα έκφρασης γονιδίων, ενώ οι στήλες του αντιστοιχούν στα δείγματα διαφορετικών συνθηκών. Κάθε στοιχείο, x_{ij} , του μητρώου αντιστοιχεί στα επίπεδα έκφρασης του γονιδίου i στο δείγμα j .

Η ανακάλυψη κρυμμένων προτύπων σε δεδομένα γονιδιακών εκφράσεων μικρο-συστοιχιών, είναι μια πολύ μεγάλη πρόκληση για τη Γονιδιακή και Πρωτεϊνική έρευνα [76]. Μία ελπιδοφόρα τεχνική φαίνεται να είναι η χρήση μεθόδων εξόρυξης πληροφορίας. Ακόμη περισσότερο η ομαδοποίηση μοιάζει να είναι το βήμα κλειδί για την κατανόηση, του πώς η δραστηριότητα των γονιδίων μεταβάλλεται κατά τη διάρκεια των γονιδιακών λειτουργιών και πως επηρεάζεται από τις καταστάσεις των ασθενειών και του κυτταρικού περιβάλλοντος. Πιο συγκεκριμένα, η ομαδοποίηση μπορεί να χρησιμοποιηθεί, είτε για να αναγνωρίσει ομάδες γονιδίων σύμφωνα με την έκφραση τους σε ένα

σύνολο δειγμάτων [44,148], είτε για να ομαδοποιήσει δείγματα σε ομοιόμορφες ομάδες που αντιστοιχούν σε συγκεκριμένους μακροσκοπικούς φαινότυπους [61]. Ο δεύτερος ρόλος της ομαδοποίησης φαίνεται να είναι και πιο δύσκολος εξαιτίας της «κατάρας της διάστασης» (Curse of Dimensionality) [11] (μικρός αριθμός δειγμάτων με πολύ μεγάλη διάσταση).

Στο κεφάλαιο αυτό μελετάμε την αποτελεσματικότητα των αλγορίθμων ομαδοποίησης σε δεδομένα γονιδιακών εκφράσεων και παρουσιάζουμε μια υβριδική μέθοδο [132], που προσπαθεί ταυτόχρονα, να μειώσει την διάσταση των δεδομένων, και τελικά να κατηγοριοποιήσει τα δεδομένα. Στο υπόλοιπο του κεφάλαιου αρχικά κάνουμε μια σύντομη βιβλιογραφική επισκόπηση διαφόρων συναφών μεθόδων. Στη συνέχεια παρουσιάζουμε την απόδοση διαφόρων αλγορίθμων ομαδοποίησης σε δεδομένα γονιδιακών εκφράσεων. Έπειτα, παρουσιάζουμε τα αποτελέσματά της υβριδικής τεχνικής και την αξιολόγηση τους σε διάφορα σύνολα δεδομένων, και τέλος το κεφάλαιο ολοκληρώνεται με μια συζήτηση των συμπερασμάτων.

8.1 Σύντομη Βιβλιογραφική Επισκόπηση

Παρά την ύπαρξη πολλών αλγορίθμων ομαδοποίησης [75], κυρίως έχουν χρησιμοποιηθεί ιεραρχικοί αλγόριθμοι σε δεδομένα μικρο-συστοιχιών. Μια από τις πιο αντιπροσωπευτικές εργασίες είναι η [44]. Σε αυτή την εργασία, οι συγγραφείς, χρησιμοποιούν ένα συγχωνευτικό ιεραρχικό αλγόριθμο, και προσαρμόζουν μια μέθοδο για την οπτική αναπαράσταση των ομαδοποιημένων συνόλων. Αυτή η μέθοδος, έχει χρησιμοποιηθεί ευρέως από πολλούς βιολόγους [6,47,111]. Παρόλα αυτά η μεγάλη ευαισθησία των συγχωνευτικών μεθόδων, στις εισόδους και το υψηλό υπολογιστικό κόστος, δυσκολεύει σημαντικά τη χρήση της σε πραγματικές εφαρμογές.

Οι Xing και Karp [151], ανέπτυξαν μια δειγματοληπτική μέθοδο που ονομάζεται CLIFF (CLustering via Iterative Feature Filtering), που χρησιμοποιεί, δοκιμαστικές ομαδοποιήσεις για να φιλτράρει γονίδια. Η επιλογή των γονιδίων βασίζεται στο αποτέλεσμα ενός αλγορίθμου NCut, ο οποίος δεν είναι σθεναρός στο θόρυβο.

Μια ακόμη γραφο-θεωρητική προσέγγισή είναι ο αλγόριθμος CLICK (CLuster Identification via Connectivity Kernels) [126], που επιχειρεί να αναγνωρίσει πολύ συνδεδεμένα μέρη του γραφήματος ομοιότητας σαν ομάδες. Οι συγγραφείς στην εργασία αυτή ανέδειξαν την υπεροχή του CLICK ως προς την προσέγγισή της εργασίας [44], και των χαρτών αυτο-οργάνωσης που προτείνονται στην εργασία [127]. Βέβαια όπως αναφέρεται στην εργασία [76], ο αλγόριθμος CLICK, δεν εγγυάται ότι δεν παράγει υψηλά μη-ισορροπημένες διαμερίσεις. Επίσης, σε δεδομένα γονιδιακών εκφράσεων, δύο ομάδες συν-εκφραζόμενων γονιδίων, $C1$ και $C2$, μπορεί να έχουν ένα ποσοστό επικάλυψης. Σε αυτήν την περίπτωση ο αλγόριθμος είναι αδύνατο να τις χωρίσει και θα τις αναφέρει σαν μία ομάδα.

Τέλος, στην εργασία [7], εξετάζοντας την προβολή των δεδομένων σε ένα μικρό αριθμό από principal components, επιχειρείται να καλυφθεί η πλειοψηφία των διαφοροποιήσεων των γονιδίων. Παρόλα αυτά ο υψηλός αριθμός των μη σχετικών γονιδίων, δεν εγγυάται, ότι η σημαντική πληροφορία θα είναι προφανής στα προβαλλόμενα δεδομένα. Μια εκτενής επισκόπηση της βιβλιογραφίας παρουσιάζεται στην εργασία [76].

8.2 Επιλογή των Χαρακτηριστικών

Όπως έχει ήδη αναφερθεί σε κάθε πρόβλημα ομαδοποίησης η επιλογή των χαρακτηριστικών από τα διαθέσιμα είναι μια πολύ σημαντική διαδικασία καθώς μπορεί να επηρεάσει σημαντικά το αποτέλεσμα. Στα δεδομένα μικρο-συστοιχιών υπάρχει ένας πολύ μεγάλος αριθμός γονιδίων της τάξης των χιλιάδων από τα οποία μόνο μερικά είναι σημαντικά. Για παράδειγμα στην εργασία [61], επιλέχθηκαν από τους συγγραφείς 50 από τα 7129 διαθέσιμα σε 72 ασθενείς λευχαιμίας, χρησιμοποιώντας δικές τους ειδικές γνώσεις. Όμως η επιλογή των γονιδίων από το χρήστη, είναι μια χρονοβόρα και επιρρεπής στα λάθη διαδικασία, ειδικά όταν δεν υπάρχει κάποια ειδική πληροφορία.

Ένα άλλος τρόπος αντιμετώπισης της μεγάλης διάστασης είναι η χρήση της τεχνικής Principal Component Analysis (PCA) [77], που μετατρέπει τα δεδομένα με βέλτιστο τρόπο σε χαμηλότερης διάστασης προβολές, όπου κάθε συντεταγμένη του νέου χώρου είναι γραμμικός συνδυασμός των αρχικών. Παρόλο που η τεχνική PCA, είναι ικανή να μειώσει την διάσταση, τα νέα χαρακτηριστικά δεν έχουν φυσική σημασία. Επιπλέον, για να υπολογιστούν τα νέα χαρακτηριστικά απαιτείται πληροφορία από όλα τα υπόλοιπα.

Όμως και τεχνικές ομαδοποίησης μπορούν να εφαρμοστούν για την επιλογή των χαρακτηριστικών [91]. Σε αυτό το κεφάλαιο έχουμε εφαρμόσει τον αλγόριθμο UKW, για να αναγνωρίσουμε το υποσύνολο των χαρακτηριστικών που θα χρησιμοποιηθεί στην τελική ομαδοποίηση των δεδομένων. Ο αλγόριθμος ομαδοποίησης εφαρμόστηκε, για να αναγνωρίσει ομάδες γονιδίων με παρόμοια έκφραση σε όλα τα δείγματα. Έπειτα από κάθε ομάδα επιλέχτηκε ένα αντιπροσωπευτικό γονίδιο, αυτό με την μικρότερη απόσταση από το κέντρο της ομάδας.

8.3 Πειραματικά Αποτελέσματα

Για να αναλύσουμε την απόδοση του αλγορίθμου UKW, σε δεδομένα γονιδιακών εκφράσεων μικρο-συστοιχιών, αρχικά χρησιμοποιήσαμε δεδομένα από μια προηγούμενη έρευνα, που εξέταζε τα προφίλ εκφράσεων από 72 ασθενείς λευχαιμίας, για να σχεδιάσει μια μέθοδο κατηγοριοποίησης για την οξεία λευχαιμία [61]. Εκτελέστηκαν δύο σύνολα πειραμάτων. Στο πρώτο σύνολο, επιλέχθηκαν δύο σύνολα γονιδίων από δύο προηγούμενες εργασίες αλλά χρησιμοποιήθηκε και η ένωση τους. Από τους αλγόριθμους ομαδοποίησης που δοκιμάστηκαν ο UKW παρουσίασε τα καλύτερα αποτελέσματα.

Το δεύτερο σύνολο πειραμάτων περιλάμβανε, τεχνικές που δεν λαμβάνουν υπόψη εκ των προτέρων πληροφορία για να επιλέξουν γονίδια. Για το σκοπό αυτό χρησιμοποιήθηκε η τεχνική PCA, καθώς και ο αλγόριθμος UKW για την επιλογή των γονιδίων. Έπειτα εφαρμόστηκε ξανά ο αλγόριθμος UKW στο σύνολο με τα επιλεγμένα γονίδια, για ομαδοποιήσει αυτή τη φορά τα δείγματα. Αυτό το σύνολο πειραμάτων ταιριάζει πολύ σε ένα πραγματικό πρόβλημα, όπου δεν υπάρχει πληροφορία για την κλάση των δειγμάτων. Τα αποτελέσματα παρουσίασαν ότι ο αλγόριθμος μπορεί να καταλήξει σε σθεναρές ομαδοποιήσεις. Επιπλέον, χρησιμοποιήθηκε η υβριδοποίηση των τεχνικών UKW και PCA, καταλήγοντας σε αποτελέσματα που ήταν ισοδύναμα με αυτά της χειροκίνητης επιλογής γονιδίων. Αυτό το γεγονός, σε συνδυασμό με το μεγάλο αυτοματισμό της τεχνικής την αναδεικνύει σε πολλά υποσχόμενη για τη χρήση σε πραγματικές συνθήκες. Για να ισχυροποιήσουμε αυτό το γεγονός εκτελέστηκε και ένα τρίτο σύνολο πειραμάτων σε τρία άλλα σύνολα δεδομένων μικρο-συστοιχιών.

8.3.1 Ομαδοποίηση Βασισμένη σε Επιβλεπόμενη Επιλογή Γονιδίων

Στο σύνολο δεδομένων που χρησιμοποιήθηκε, κάθε δείγμα χαρακτηρίζονταν από 7129 γονίδια. Τα πρώτα 38 δείγματα χρησιμοποιήθηκαν για την διαδικασία της ομαδοποίησης (σύνολο εκπαίδευσης), ενώ τα υπόλοιπα 34 χρησιμοποιήθηκαν για την αξιολόγηση του αποτελέσματος (σύνολο ελέγχου). Τα αρχικά 38 δείγματα περιέχουν 27 δείγματα χαρακτηρισμένα ως acute myeloid leukemia (ALL), και 11 ως acute lymphoblastic leukemia (AML). Το σύνολο ελέγχου περιέχει 20 δείγματα χαρακτηρισμένα ως ALL και 14 χαρακτηρισμένα ως AML. Στην εργασία [61], εφαρμόστηκαν χάρτες αυτοοργάνωσης για την ομαδοποίηση [84], χρησιμοποιώντας τα 50 πιο συσχετισμένα γονίδια ως προς τη διάκριση των χαρακτηρισμών ALL-AML. Τα 38 δείγματα ομαδοποιήθηκαν από τους χάρτες αυτοοργάνωσης σε δύο ομάδες, όπου η μία περιείχε 24 από τα 25 ALL δείγματα, και η άλλη 10 από τα 13 AML.

Εφαρμόζοντας τον αλγόριθμο UKW, απευθείας χρησιμοποιώντας και τα 7129 γονίδια, ο αλγόριθμος κατέληξε σε ομάδες που περιείχαν τόσο δείγματα ALL όσο και AML. Για να βελτιώσουμε το αποτέλεσμα της ομαδοποίησης έγινε προφανές ότι πρέπει να επιλεχθούν προσεχτικά τα κατάλληλα γονίδια. Με αυτό τον τρόπο, κατασκευάσαμε τα σύνολα $GeneSet_1$, με την επιλογή των 50 γονιδίων που προτάθηκε στην εργασία [61], $GeneSet_2$ με την επιλογή των 50 γονιδίων που προτάθηκε στην εργασία [141], και τέλος το σύνολο $GeneSet_3$ που αποτελεί την ένωση των δύο προηγούμενων αποτελούμενη από 72 γονίδια.

Αποτελέσματα για το σύνολο εκπαίδευσης με γονίδια από το $GeneSet_1$ ALL απόδοση αντιστοίχισης: 87.5% — AML απόδοση αντιστοίχισης: 100%						
Τύπος ασθένειας	ALL Ομάδες				AML Ομάδες	
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 1	Ομάδα 2
ALL	4	4	12	4	3	0
AML	0	0	0	0	4	7
Αποτελέσματα για το σύνολο εκπαίδευσης με γονίδια από το $GeneSet_2$ ALL απόδοση αντιστοίχισης: 100.0% — AML απόδοση αντιστοίχισης: 100%						
Τύπος ασθένειας	ALL Ομάδες				AML Ομάδες	
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 1	Ομάδα 2
ALL	10	3	10	4	0	0
AML	0	0	0	0	8	3
Αποτελέσματα για το σύνολο εκπαίδευσης με γονίδια από το $GeneSet_3$ ALL απόδοση αντιστοίχισης: 95.83% — AML απόδοση αντιστοίχισης: 100%						
Τύπος ασθένειας	ALL Ομάδες				AML Ομάδες	
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 1	Ομάδα 2
ALL	8	9	5	4	0	1
AML	0	0	0	0	7	4

Πίνακας 8.1: Η απόδοση του αλγορίθμου UKW για το σύνολο εκπαίδευσης με διαφορετικά σύνολα γονιδίων.

Η εφαρμογή του αλγορίθμου UKW, στα 3 αυτά σύνολα, παρήγαγε 6 ομάδες που περιείχαν ALL ή AML στοιχεία. Στον Πίνακα 8.1, παρουσιάζονται τα αποτελέσματα αναλυτικά. Στο σύνολο $GeneSet_1$ ο αλγόριθμος, υπολογίζει 4 ALL ομάδες και 2 AML, με 3 από τα στοιχεία να θεωρούνται σαν λάθος κατηγοριοποιήσεις. Στο σύνολο $GeneSet_2$, τα αποτελέσματα ήταν παρόμοια αλλά δεν υπήρχαν καθόλου λάθος

κατηγοριοποιήσεις, και τέλος στο σύνολο *GeneSet*₃, η μόνη διαφορά ήταν ότι υπήρχε 1 λάθος κατηγοριοποίησης. Προφανώς στο σύνολο εκπαίδευσης η επιλογή γονιδίων *GeneSet*₂, φαίνεται να είναι η καλύτερη ώστε η ομαδοποίηση να διαχωρίζει πλήρως τις 2 κατηγορίες ALL–AML.

Για την εκτίμηση των παραπάνω ομαδοποιήσεων, τα δείγματα από το σύνολο ελέγχου ανατέθηκαν στις ομάδες, ανάλογα με την απόσταση τους από το κέντρο της ομάδας. Εάν ένα ALL (AML) δείγμα αντιστοιχίζοταν σε μια AML (AML αντίστοιχα) ομάδα τότε αυτό θεωρούνταν λάθος στην αντιστοίχιση. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 8.2. Η χρήση της ομαδοποίησης με γονίδια από το *GeneSet*₁, καταλήγει σε 73.5% απόδοση αντιστοίχισης. Η χρήση γονιδίων από το σύνολο *GeneSet*₂, καταλήγει σε 3 δείγματα λάθος αντιστοίχισμένα (91.2% απόδοση), ενώ με γονίδια από το *GeneSet*₃, η ομαδοποίηση έχει την καλύτερη απόδοση αντιστοίχισης 94.1%, με μόνο 2 λάθη.

Αποτελέσματα για το σύνολο ελέγχου με γονίδια από το <i>GeneSet</i> ₁ ALL απόδοση αντιστοίχισης: 60.00% — AML απόδοση αντιστοίχισης: 92.85%						
Τύπος ασθένειας	ALL Ομάδες				AML Ομάδες	
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 1	Ομάδα 2
ALL	2	0	7	3	8	0
AML	1	0	0	0	8	5
Αποτελέσματα για το σύνολο ελέγχου με γονίδια από το <i>GeneSet</i> ₂ ALL απόδοση αντιστοίχισης: 100% — AML απόδοση αντιστοίχισης: 78.57%						
Τύπος ασθένειας	ALL Ομάδες				AML Ομάδες	
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 1	Ομάδα 2
ALL	8	0	9	3	0	0
AML	0	0	3	0	8	3
Αποτελέσματα για το σύνολο ελέγχου με γονίδια από το <i>GeneSet</i> ₃ ALL απόδοση αντιστοίχισης: 90% — AML απόδοση αντιστοίχισης: 100%						
Τύπος ασθένειας	ALL Ομάδες				AML Ομάδες	
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 1	Ομάδα 2
ALL	10	4	3	1	0	2
AML	0	0	0	0	5	9

Πίνακας 8.2: Η απόδοση του αλγορίθμου UKW για το σύνολο ελέγχου με διαφορετικά σύνολα γονιδίων.

Στον Πίνακα 8.3, παρουσιάζονται συγκριτικά αποτελέσματα με τη χρήση του συνόλου γονιδίων *GeneSet*₃, καθώς με αυτό όλοι οι αλγόριθμοι πέτυχαν την καλύτερη απόδοση. Από τους συγκρινόμενους αλγόριθμους οι UKW, και FCM, είχαν την καλύτερη απόδοση, ακολουθούμενοι από τους DBSCAN και GNG. Βέβαια ο αλγόριθμος FCM, απαιτεί από τον χρήστη τον ορισμό του πλήθους των ομάδων, και ο αλγόριθμος DBSCAN, δεν αντιστόιχε 7 δείγματα από το σύνολο εκπαίδευσης και 5 από το σύνολο ελέγχου (όλα από την AML κατηγορία), καθώς τα χαρακτήρισε σαν θόρυβο.

Παρόλο που ο αλγόριθμος PDDP, είχε την χειρότερη απόδοση, ήταν ο μόνος που είχε την ικανότητα να επεξεργαστεί και τα 7129 γονίδια. Αυτός ο αλγόριθμος, χρησιμοποιώντας το πλήρες σύνολο γονιδίων αντιστόιχε λάθος 2 γονίδια από το σύνολο εκπαίδευσής και 8 από το σύνολο ελέγχου.

	Λάθος Αντιστοιχίσεις		Πλήθος των Ομάδων		Απόδοση (%)	
	Σύνολο		Σύνολο		AML	ALL
	Εκπαίδευσης	Ελέγχου	Εκπαίδευσης	Ελέγχου	AML	ALL
DBSCAN	1	3	4	4	78.5	100
FCM	1	2	4	4	85.7	100
GNG	1	3	3	3	78.5	100
PDDP	2	4	6	6	71.4	100
UKW	1	2	6	6	100.0	90.0

Πίνακας 8.3: Συγκριτικά αποτελέσματα για διάφορους αλγορίθμους με γονίδια από το σύνολο *GeneSet₃*

8.3.2 Ομαδοποίηση Βασισμένη σε Μη-Επιβλεπόμενη Επιλογή Γονιδίων

Στην ενότητα αυτή, εξετάζουμε την απόδοση του αλγορίθμου UKW σε σύνολα γονιδίων που έχουν κατασκευαστεί από μη-επιβλεπόμενες μεθόδους (δεν χρησιμοποιήθηκε η πληροφορία της κατηγορίας για την κατασκευή τους). Χρησιμοποιούμε μόνο τον αλγόριθμο UKW, καθώς όπως έδειξαν τα προηγούμενα πειράματα έχει καλύτερη ή όμοια απόδοση με του υπόλοιπους.

Αρχικά υπολογίζουμε το σύνολο των γονιδίων, ομαδοποιώντας τις εκφράσεις τους χαρακτηρισμένες σε όλα τα δείγματα. Από κάθε ομάδα επιλέγουμε το γονίδιο που βρίσκεται πλησιέστερα στο κέντρο της, για την κατασκευή του συνόλου γονιδίων *GeneSet₄*. Η εφαρμογή του αλγορίθμου κατέληξε σε 293 ομάδες, που από την κάθε μία επιλέχθηκε ένα γονίδιο. Οπότε το σύνολο *GeneSet₄* περιείχε 293 γονίδια.

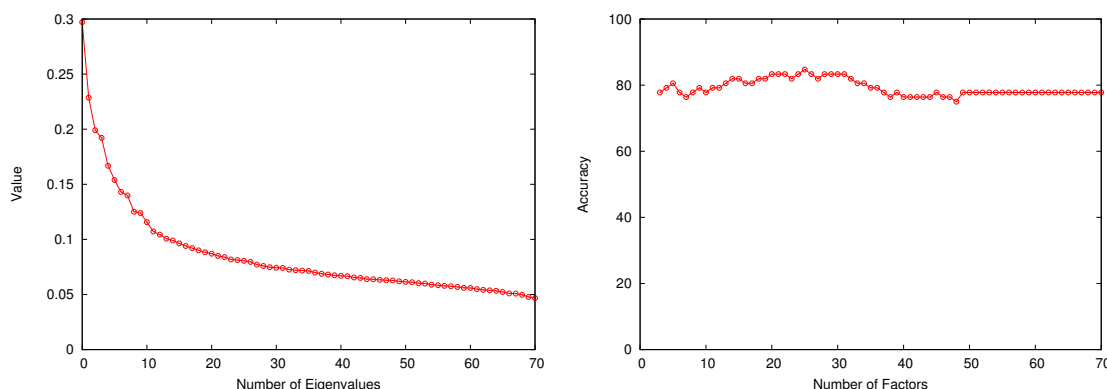
Έπειτα ο αλγόριθμος, εφαρμόστηκε στο πλήρες σύνολο των 72 δειγμάτων χαρακτηρισμένα στα 293 γονίδια του συνόλου *GeneSet₄*, και τα αποτελέσματα απεικονίζονται στον Πίνακα 8.4. Από τον πίνακα αυτό είναι προφανές ότι ακόμη και αν δεν χρησιμοποιηθεί η πληροφορία της κλάσης είναι δυνατόν να καταλήξουμε σε μια ομαδοποίηση με υψηλό ποσοστό αντιστοίχισης: 93.6% και 76% τις κατηγορίες ALL και AML αντίστοιχα.

Αποτελέσματα ομαδοποίησης με τη χρήση γονιδίων από το σύνολο <i>GeneSet₄</i>						
ALL απόδοση αντιστοίχισης: 93.61% — AML απόδοση αντιστοίχισης: 76%						
Τύπος Ασθένειας	ALL Ομάδες					AML Ομάδα
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 5	Ομάδα 1
ALL	12	5	8	16	3	3
AML	2	0	3	0	1	19

Πίνακας 8.4: Η απόδοση του αλγορίθμου UKW όταν χρησιμοποιείται και για την επιλογή των γονιδίων (*GeneSet₄*) αλλά και για την ομαδοποίηση των δειγμάτων.

Τα επόμενα πειράματα αφορούν την χρήση της τεχνικής PCA, για τη μείωση της διάστασης. Το συνηθισμένο πρόβλημα της τεχνικής, PCA, είναι η επιλογή του αριθμού των εξαγόμενων χαρακτηριστικών. Ένα απλός κανόνας, είναι η οπτικοποίηση του γραφήματος των ιδιοτίμων (*scree plot*) σε φθίνουσα σειρά. Το γράφημα αυτό συνήθως μοιάζει με την πλαγιά ενός λόφου, και ο κανόνας υπαγορεύει την επιλογή του πλήθους των ιδιοτιμών μέχρι το σημείο που η πλαγιά να εξομαλύνεται σε πεδιάδα. Για το συγκεκριμένο πρόβλημα ωστόσο αυτό ο κανόνας δεν μπορεί να εφαρμοστεί ξεκάθα-

ρα όπως φαίνεται από το αριστερό μέρος του Σχήματος 8.1, που απεικονίζονται οι 70 πρώτες ιδιοτιμές. Γενικά θα μπορούσαμε να πούμε ότι η χρήση περισσότερων από 10 χαρακτηριστικών δεν προσφέρει περισσότερη πληροφορία. Παρόλα αυτά δοκιμάστηκαν πειραματικά όλα τα παραγόμενα χαρακτηριστικά από 2 ως 70. Η απόδοση αντιστοίχησης απεικονίζεται στο δεξί μέρος του Σχήματος 8.1. Η καλύτερη απόδοση (84.72%) επιτεύχθηκε για 25 χαρακτηριστικά.



Σχήμα 8.1: Το γράφημα των ιδιοτιμών σε φθίνουσα σειρά (αριστερά), και η απόδοση αντιστοίχησης με τη χρήση 2 ως 70 χαρακτηριστικών (δεξιά).

Παρόλο που η τεχνική PCA μετασχηματίζει με το βέλτιστο τρόπο τα δεδομένα σε ένα χώρο μικρότερης διάστασης, η απόδοση αντιστοίχησης της ομαδοποίησης δεν είναι τόσο υψηλή όσο των ομαδοποιήσεων που προήλθαν από επιβεβλημένη επιλογή των γονιδίων (βλ. Ενότητα 8.3.1). Στη συνέχεια προτείνουμε μια υβριδική τεχνική που αναμειγνύει τον αλγόριθμο UKW, και την τεχνική PCA για την εξαγωγή χαρακτηριστικών μειωμένης διάστασης.

Αρχικά, χρησιμοποιείται ο αλγόριθμος UKW για την ομαδοποίηση των γονιδίων σε ομάδες με ομοιόμορφες εκφράσεις σε όλα τα δείγματα. Έπειτα σε κάθε ομάδα γονιδίων εφαρμόζουμε ξεχωριστά την τεχνική PCA, η οποία μετασχηματίζει το εκάστοτε σύνολο των γονιδίων ξεχωριστά. Από κάθε ομάδα διαλέγουμε τα δύο πρώτα εξαγόμενα χαρακτηριστικά. Με αυτό τον τρόπο κατασκευάζεται τελικά το σύνολο *GeneSet₅*. Ο αριθμός των εξαγόμενων χαρακτηριστικών σε αυτή την περίπτωση δεν είναι τόσο κρίσιμος καθώς το αρχικό πλήθος των ομάδων είναι αυτό που καθορίζει τη διάσταση των τελικών χαρακτηριστικών. Με αυτό τον τρόπο το σύνολο *GeneSet₅*, έχει 586 χαρακτηριστικά. Η απόδοση του αλγορίθμου φαίνεται στον Πίνακα 8.5, που φτάνει τα ποσοστά αντιστοίχησης: 97.87% και 88% τις κατηγορίες ALL και AML αντίστοιχα.

Το αποτελέσματα της ομαδοποίησης για το σύνολο χαρακτηριστικών <i>GeneSet₅</i>						
ALL απόδοση αντιστοίχησης: 97.87% — AML απόδοση αντιστοίχησης: 88%						
Τύπος ασθένειας	ALL Ομάδες				AML Ομάδες	
	Ομάδα 1	Ομάδα 2	Ομάδα 3	Ομάδα 4	Ομάδα 1	Ομάδα 2
ALL	7	14	14	11	1	0
AML	0	0	3	0	13	9

Πίνακας 8.5: Η απόδοση του υβριδικού σχήματος εξαγωγής χαρακτηριστικών.

Γενικά, τα πειραματικά αποτελέσματα, δείχνουν ότι με τη χρήση της ένωσης των

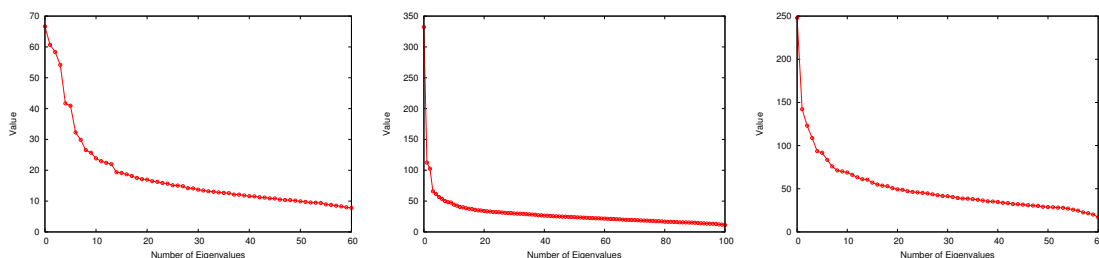
συνόλων γονιδίων $GeneSet_1$, και $GeneSet_2$, μπορούμε να καταλήξουμε σε ομαδοποιήσεις με τα καλύτερα αποτελέσματα. Όμως η πληροφορία που μας δίνει τη δυνατότητα να επιλέξουμε γονίδια με αυτόν τον τρόπο τις περισσότερες φορές δεν υπάρχει. Για το σκοπό αυτό αναγκάζομαστε να εξάγουμε χαρακτηριστικά με μη-επιβλεπόμενες μεθόδους. Το προτεινόμενο υβριδικό σχήμα δείχνει ικανό να εξάγει χαρακτηριστικά που μπορούν να οδηγήσουν σε ομαδοποιήσεις με ποσοστά αντιστοίχησης παρόμοια με αυτά της επιβλεπόμενης επιλογής γονιδίων.

8.3.3 Αξιολόγηση του Υβριδικού Σχήματος

Η περαιτέρω αξιολόγηση του υβριδικού σχήματος γίνεται με τη χρήση τριών συνόλων δεδομένων διαθέσιμα προς κοινή χρήση:

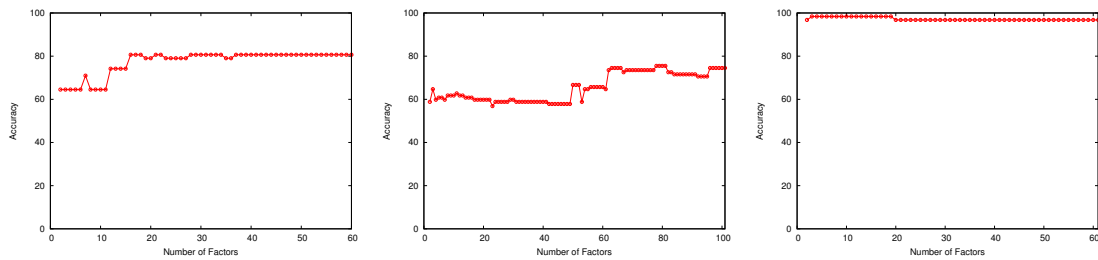
- Το σύνολο COLON [6] που αποτελείται από 40 ιστούς όγκων του παχέως εντέρου και 22 κανονικούς. Κάθε ένα δείγμα περιγράφεται από 2000 γονιδιακές εκφράσεις. Το σύνολο αυτό είναι δημόσια διαθέσιμο από τη διεύθυνση <http://microarray.princeton.edu/oncology>
- Το σύνολο PROSTATE [48] που περιέχει 52 δείγματα από καρκίνο του προστάτη και 50 κανονικά δείγματα. Για κάθε δείγμα, έχουν μετρηθεί τα επίπεδα έκφρασης 6033 γονιδίων. Το σύνολο αυτό είναι δημόσια διαθέσιμο από τη διεύθυνση <http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>.
- Το σύνολο LYMPHOMA [47] που περιέχει 62 δείγματα από 3 τύπους lymphoid malignancies, μετρημένα στα επίπεδα έκφρασης 4026 γονιδίων. Το σύνολο αυτό είναι δημόσια διαθέσιμο από τη διεύθυνση <http://genome-www.stanford.edu/>.

Για κάθε ένα σύνολο συγκρίνουμε την υβριδική προσέγγιση με την τεχνική PCA. Καθώς για την επιλογή των διαστάσεων της τεχνικής PCA, τα γραφήματα των ιδιοτιμών δεν είναι ξεκάθαρα ως προς την επιλογή των τιμών (Σχήμα 8.2). Για αυτό το λόγο δοκιμάστηκαν όλες οι διαθέσιμες και για τα 3 σύνολα δεδομένων και οι αποδόσεις αντιστοίχησης, απεικονίζονται στο Σχήμα 8.3. Για το σύνολο COLON η καλύτερη απόδοση που επιτεύχθηκε ήταν 80.64% χρησιμοποιώντας 16 χαρακτηριστικά, για το σύνολο PROSTATE 82.35% με 71 χαρακτηριστικά, ενώ για το σύνολο LYMPHOMA 98.38%, χρησιμοποιώντας μόνο 3 χαρακτηριστικά.



Σχήμα 8.2: Γραφήματα των ιδιοτιμών σε φθίνουσα σειρά για τα σύνολα δεδομένων COLON (αριστερά), PROSTATE (μέση) και LYMPHOMA (δεξιά).

Τα αποτελέσματα της υβριδικής προσέγγισης, για τα τρία σύνολα, παρουσιάζονται στον Πίνακα 8.6. Όπως είναι προφανές, η απόδοση αντιστοίχησης είναι μεγαλύτερη



Σχήμα 8.3: Η απόδοση αντιστοίχισης με τη χρήση όλων των διαθέσιμων χαρακτηριστικών για τα σύνολα COLON (αριστερά), PROSTATE (μέση) και LYMPHOMA (δεξιά).

από αυτήν της τεχνικής PCA, και στις 3 περιπτώσεις. Ο μεγάλος αριθμός χαρακτηριστικών δεν είναι πρόβλημα πλέον, καθώς τα χαρακτηριστικά προέρχονται από διαφορετικές ομάδες και δεν είναι συσχετισμένα μεταξύ τους. Επίσης η αυτόματη αναγνώριση του πλήθους των χαρακτηριστικών προσφέρει μια σθεναρή αντιμετώπιση που δεν είναι δυνατή μέσω της τεχνικής PCA. Αυτά τα ποσοστά αντιστοίχισης, είναι πολύ ψηλά, ακόμα και σε σύγκριση με επιβλεπόμενες τεχνικές [73].

Σύνολο Δεδομένων	Πλήθος Χαρακτηριστικών	Ποσοστό Αντιστοίχισης (%)
COLON	229	82.25
PROSTATE	84	83.3
LYMPHOMA	103	99.01

Πίνακας 8.6: Η απόδοση του υβριδικού σχήματος στα 3 σύνολα δεδομένων.

8.4 Συμπεράσματα - Συνεισφορά

Οι τεχνολογίες μικρο-συστοιχιών DNA, μπορούν να παρέχουν μετρήσεις των επιπέδων έκφρασης χιλιάδων γονιδίων ταυτόχρονα. Παρόλα αυτά, ο αριθμός των γονιδίων είναι πολύ μεγάλος σε σχέση με το πλήθος των δειγμάτων σε αυτές τις εφαρμογές. Για να αντιμετωπιστεί η δυσκολία της μεγάλης διάστασης και του θορύβου, τα δεδομένα συνήθως μετατρέπονται με όσο το δυνατόν μικρότερη απώλεια πληροφορίας, σε ένα χώρο πολύ μικρότερης διάστασης.

Η ανάλυση ομαδοποίησης που παρουσιάζεται σε αυτό το κεφάλαιο, στοχεύει στην ομαδοποίηση δειγμάτων βασισμένη στην ομοιόμορφη έκφραση τους σε ένα σύνολο γονιδίων. Για το σκοπό αυτό χρησιμοποιήσαμε τον αλγόριθμο UKW, καθώς όπως φάνηκε από τα συγκριτικά αποτελέσματα αλλά και από παρόμοιες εργασίες [112,113, 130,131], καταφέρνει να προσφέρει υψηλής ποιότητας ομαδοποιήσεις, σε τέτοιου είδους δεδομένα.

Όμως η επιλογή των γονιδίων με βάση μια προκαθορισμένη πληροφορία, για κάθε ένα δείγμα δεν είναι εφικτή σε πρακτικές συνθήκες. Για το σκοπό αυτό προτείνουμε και αξιολογήσαμε ένα υβριδικό σχήμα που καταφέρνει να αυτοματοποιήσει την όλη διαδικασία. Τα αποτελέσματα της τελικής ομαδοποίησης σε δεδομένα που προήλθαν από το υβριδικό σχήμα, ήταν πολύ ενθαρρυντικά συγκρινόμενα ακόμη και με επιβλεπόμενες μεθόδους [73].

Υβριδικές Εφαρμογές Υπολογιστικής Νοημοσύνης και Ομαδοποίησης

Έρευνα είναι αυτό που κάνω όταν δεν ξέρω τι κάνω.

—Wernher von Braun (1912–1977)

Όλο και πιο δημοφιλής είναι η χρήση «Υβριδικών» τεχνικών για την αντιμετώπιση πραγματικών προβλημάτων στις μέρες μας. Η «Υβριδική Υπολογιστική Νοημοσύνη» ορίζεται σαν οποιαδήποτε αποτελεσματικός συνδυασμός ευφυών τεχνικών που παρουσιάζει καλύτερα ή ανταγωνιστικά αποτελέσματα σε σχέση με τις μεμονωμένες τεχνικές.

Σε αυτό το κεφάλαιο, προτείνουμε την υβριδική συνέργεια μεταξύ μεθόδων ομαδοποίησης και Τεχνητών Νευρωνικών Δικτύων. Πιο συγκεκριμένα, η προτεινόμενη μεθοδολογία αποτελείται από δύο βήματα. Αρχικά χρησιμοποιούνται μέθοδοι ομαδοποίησης, για να διαχωρίσουν τον χώρο των δεδομένων, σε ομάδες με κοινά χαρακτηριστικά. Έπειτα για κάθε μια ομάδα, εκπαιδεύουμε νευρωνικά δίκτυα με στόχο την κατηγοριοποίηση των δεδομένων που ανήκουν μόνο σε αυτήν την ομάδα. Η προτεινόμενη αυτή υβριδική τεχνική, εφαρμόζεται σε δύο δύσκολα προβλήματα, και όπως φαίνεται στα αποτελέσματα που αναφέρονται στις επόμενες ενότητες, καταφέρνει να υπερκεράσει τα αποτελέσματα άλλων απλών τεχνικών.

9.1 Αυτόματος Εντοπισμός Ισχαιμικών Επεισοδίων

Η μυοκαρδιακή Ισχαιμία είναι μια από τις πιο συχνές αιτίες θανάτου στις βιομηχανικές χώρες, οπότε η έγκαιρη διάγνωση και θεραπεία της έχει μεγάλη σημασία. Η διάγνωση της μυοκαρδιακής Ισχαιμίας χρησιμοποιώντας, μεγάλες εγγραφές ηλεκτροκαρδιογραφήματων, είναι μία απλή μέθοδος, που χρειάζεται περαιτέρω ανάπτυξη πριν τη χρήση της στην κλινική Ιατρική. Η ικανότητα ακριβών και πρόωρων εντοπισμών των οξείων ισχαιμικών επεισοδίων είναι κρίσιμη για την πειστικότητα της κατάλληλης θεραπείας. Το Ηλεκτρο-καρδιογράφημα (EEG), καταγράφει τις αλλαγές που συμβαίνουν στο ηλεκτρικό δυναμικό μεταξύ διαφορετικών σημείων του δέρματος, εξαιτίας της καρδιακής λειτουργίας. Εφόσον το EEG καταγράφεται εύκολα, είναι πολύ σημαντικό να υπάρχουν μέσα για τον αξιόπιστο εντοπισμό της ισχαιμίας από την ανάλυση του.

Υπάρχουν μερικά υποχρεωτικά βήματα για τον αυτόματο εντοπισμό ισχαιμικών επεισοδίων. Μετά από την πρωταρχική αφαίρεση του θορύβου, στο δεύτερο βήμα εξάγονται όλα τα σημαντικά χαρακτηριστικά του EEG (σημείο J, ισο-ηλεκτρική γραμμή, και κορυφή κύματος T). Χρησιμοποιώντας τα παραπάνω χαρακτηριστικά, στο τρίτο στάδιο ο καρδιακός παλμός κατηγοριοποιείται σαν κανονικός ή ισχαιμικός. Τέλος, στο τελευταίο στάδιο συνεχόμενοι ισχαιμικοί παλμοί ομαδοποιούνται κατάλληλα ώστε να αναγνωρισθούν τα ισχαιμικά επεισόδια.

Το ST-T Complex του EEG, αντιπροσωπεύει, τη χρονική περίοδο, ανάμεσα στο τέλος της εκ-πόλωσης της κοιλιακής χώρας ως το τέλος της αντίστοιχης επαναπόλωσης στο ηλεκτρικό καρδιακό κύκλο. Οι ισχαιμικές αλλαγές συχνά επηρεάζουν ολόκληρη τη μορφή του κύματος επαναπόλωσης, και έτσι δεν περιγράφονται επαρκώς από απομονωμένα χαρακτηριστικά, ακόμα και εάν αυτά λαμβάνονται σαν το μέσο από πολλά δείγματα σήματος. [92]. Η προτεινόμενη αντιμετώπιση δεν χρησιμοποιεί τοπικά απομονωμένα χαρακτηριστικά, αλλά βασίζεται στη χρήση της τεχνικής PCA, για την εξαγωγή των χαρακτηριστικών, που περιγράφουν το ολικό περιεχόμενο του ST-T Complex.

9.1.1 Προ-Επεξεργασία Δεδομένων

Μια περιγραφή της Ευρωπαϊκής Βάσης Δεδομένων ST-T, παρουσιάζεται στην εργασία [146], όπου εξηγούνται με ακρίβεια οι κανόνες για τον εντοπισμό των επεισοδίων ST και T. Μετά από αυτό το βήμα χρησιμοποιήθηκαν 5 χαρακτηριστικά εξαγόμενα από την τεχνική PCA. Αυτά επεξεργάστηκαν περαιτέρω μέσω μιας τεχνικής soft thresholding [39], από την οποία επιλέχτηκαν πέντε επίπεδα του μετασχηματισμού κυματομορφής Daubechies, που σχηματίζουν και την τελική μορφή των δεδομένων.

9.1.2 Η Μεθοδολογία Κατηγοριοποίησης

Σε αυτήν την ενότητα παρουσιάζουμε τη μεθοδολογία που χρησιμοποιήθηκε για την αυτόματη κατηγοριοποίηση των επεξεργασμένων δεδομένων EEG, σε ισχαιμικά ή μη [133]. Αντί να χρησιμοποιηθεί ένα ολικό μοντέλο για την κατηγοριοποίηση των δεδομένων, κατασκευάζονται πολλά τοπικά μοντέλα, για γειτονιές του χώρου των δεδομένων. Για να επιτευχθεί αυτό αρχικά ομαδοποιούμε τα δεδομένα μέσω του αλγορίθμου UKW, και έπειτα για κάθε μια ανιχνευμένη ομάδα, ένα ξεχωριστό Νευρωνικό Δίκτυο Εμπρόσθιας Τροφοδότησης (FNN) χρησιμοποιείται σαν τοπικός κατηγοροποιήτης. Πιο συγκεκριμένα η μεθοδολογία συνοψίζεται ως εξής:

1. Αναγνώριση των ομάδων του συνόλου δεδομένων εκπαίδευσής.
2. Για κάθε μια ομάδα, εκπαιδεύεται ένα ξεχωριστό FNN με δεδομένα μόνο από τη συγκεκριμένη ομάδα.
3. Τα στοιχεία του συνόλου ελέγχου ανατίθενται σε ομάδες ανάλογα με την απόστασή τους από το κέντρο των ομάδων.
4. Τα εκπαιδευμένα δίκτυα χρησιμοποιούνται για να ληφθούν οι αποφάσεις κατηγοριοποίησης για το σύνολο ελέγχου.

9.1.3 Αριθμητικά Αποτελέσματα

Η αποδοτική επιβλεπόμενη εκπαίδευση FNNs, είναι ένα θέμα συνεχούς έρευνας. Για το σκοπό αυτό έχουν προταθεί διάφοροι αλγόριθμοι. Εμείς χρησιμοποιούμε τις

ακόλουθες μεθόδους εκπαίδευσης:

- Resilient Back Propagation (RPROP) [120],
- Improved Resilient Back Propagation (iRPROP) [72],
- Scaled Conjugate Gradient (SCG) [102],
- Adaptive On-Line Back Propagation (AOBP) [94],
- Back Propagation with Variable Stepsize(BPVS) [97],

Μετά από εκτεταμένα πειράματα, η καλύτερη τοπολογία δικτύων περιείχε δύο κρυφά επίπεδα, από τα οποία το πρώτο είχε 8 και το δεύτερο 7 νευρώνες. Δύο κόμβοι χρησιμοποιήθηκαν στο επίπεδο εξόδου, από τους οποίους αυτός με τη μεγαλύτερη έξοδο καθόριζε την απόφαση του δικτύου. Η τελική τοπολογία των δικτύων ήταν (5–8–7–2). Όλα τα νευρωνικά δίκτυα εκπαιδεύτηκαν για 300 περιόδους, και έπειτα αξιολογήθηκε η απόδοση τους στο σύνολο ελέγχου. Αυτή η διαδικασία, επαναλήφθηκε 100 φορές για όλους τους αλγόριθμους εκπαίδευσής. Η απόδοση της ικανότητας κατηγοριοποίησης των FNNs στο σύνολο ελέγχου παρουσιάζεται στον Πίνακα 9.1.

Σύνολο ελέγχου	Απόδοση Κατηγοριοποίησης					
	E103	E104	E106	E107	E108	E111
RPROP						
mean	75.60	86.63	59.78	79.07	73.95	92.19
std	0.39	0.96	2.20	2.48	1.13	2.28
max	76.21	88.87	66.04	85.76	75.85	94.55
min	74.76	84.34	57.58	73.67	71.57	87.51
iRPROP						
mean	75.64	87.07	59.87	78.48	73.91	91.84
std	0.54	0.73	2.88	1.62	1.10	4.00
max	76.53	88.29	69.17	81.71	75.75	94.92
min	74.36	84.57	57.48	75.24	71.97	80.41
SCG						
mean	77.46	85.77	62.51	79.07	73.57	92.45
std	0.77	0.80	4.74	2.13	1.61	1.49
max	79.05	87.09	71.15	84.38	76.16	94.77
min	75.76	83.64	58.54	72.61	69.62	89.82
BPVS						
mean	72.50	84.29	67.94	83.28	70.00	93.91
std	0.02	0.45	0.17	0.07	0.007	0.03
max	72.57	85.27	68.08	83.34	70.01	93.99
min	72.49	84.06	67.72	83.17	69.99	93.86
AOBP						
mean	73.69	83.83	71.68	83.10	69.67	93.21
std	0.59	0.09	0.00	0.01	0.005	0.30
max	73.88	83.91	71.68	83.14	69.67	93.86
min	71.93	83.72	71.68	83.10	69.64	93.05

Πίνακας 9.1: Ποσοστά σωστής κατηγοριοποίησης των FNNs εκπαιδευμένων με διαφορετικές μεθόδους για όλα τα 100 πειράματα.

Στα σύνολα δεδομένων E103, E104, και E108, τα δίκτυα που εκπαιδεύτηκαν με τις μεθόδους RPROP και iRPROP, είχαν την καλύτερη απόδοση αλλά με μεγάλη τυπική απόκλιση στα διαφορετικά πειράματα. Τα δίκτυα εκπαιδευμένα με τη μέθοδο SCG,

είχαν το ίδιο μειονέκτημα. Για τα υπόλοιπα σύνολα δεδομένων, δίκτυα εκπαιδευμένα με τις μεθόδους BPVS και AOBP, είχαν τα καλύτερα αποτελέσματα. Ένα σημαντικό πλεονέκτημα της μεθόδου AOBP και σε ένα βαθμό και της BPVS, ήταν η πολύ σθεναρή συμπεριφορά καθώς είχαν ελάχιστη διαφοροποίηση των αποτελεσμάτων στα 100 πειράματα. Γενικά, για τα σύνολα δεδομένων E104, E107 και E111, τα ποσοστά σωστής κατηγοριοποίησης της προτεινόμενης μεθοδολογίας ήταν πολύ υψηλά.

9.2 Αυτόματη Αναγνώριση Ανωμαλιών σε Κολονοσκοπήσεις

Ο καρκίνος του παχέος εντέρου μπορεί να προληφθεί και να θεραπευτεί αν ανιχνευτεί νωρίς. Ένα ευφύες περιβάλλον που θα βοηθούσε τους παθολόγους στην πρόωρη διάγνωση της ασθένειας αυτής, σίγουρα θα αύξανε την αποδοτικότητα των χαρακτηριστικών μεθοδολογιών απεικόνισης. Με πιο τεχνικούς όρους, το πρόβλημα στην αυτόματη ερμηνεία εικόνων είναι να συνδέσει σύνολα εικονοκυττάρων (pixels) (δομές) που εμφανίζονται σε μια εικόνα με άγνωστα αντικείμενα που είναι παρόντα στη σκηνή από την οποία η εικόνα λήφθηκε.

Κατά τη διάρκεια της κολονοσκόπησης ο παθολόγος ερμηνεύει τις φυσικές ιδιότητες επιφάνειας του ιστού, όπως τραχύτητα, ομαλότητα, τακτικότητα και μορφή, για να ανιχνεύσει τις ανωμαλίες. Αυτός ο στόχος είναι ιδιαίτερα δύσκολος. Ακόμη και μέσα στο ίδιο δείγμα, ένα τμήμα μπορεί να έχει πολύ διαφορετικά χαρακτηριστικά από ένα άλλο. Οι παρακαείμενες επιφάνειες της επένδυσης του εντέρου που παρουσιάζουν διαφορετικές ιδιότητες διακρίνονται βάση των διαφοροποιήσεων της υφής του ιστού τους. Αυτές οι διαφοροποιήσεις της υφής της βλεννογόνου επιφάνειας του παχέος εντέρου αναδεικνύουν, ότι αυτή η ιδιότητα θα μπορούσε επίσης να χρησιμοποιηθεί για την αυτόματη ανίχνευση όγκων [81, 95, 96].

Η αυτόματη ανίχνευση των όγκων στην κολονοσκόπηση υπόκειται αβεβαιότητες λόγω των ανακριβών μετρήσεων και την έλλειψη μοντέλων με μεγάλη ακρίβεια των χαρακτηριστικών εικόνων των όγκων. Λαμβάνοντας υπόψη μια εικόνα κολονοσκόπησης, τα «αληθινά» χαρακτηριστικά γνωρίσματα που συνδέονται με τη φυσική επιφάνεια του ιστού δεν είναι γνωστά στον σχεδιαστή του αυτόματου συστήματος [96]. Συνήθως, ένα ή περισσότερα μοντέλα εξαγωγής χαρακτηριστικών χρησιμοποιούνται για να παρέχουν τις τιμές για κάθε ένα στοιχείο [90]. Σε αυτόν τον ίδιο στόχο της ερμηνείας των τοπικών αλλαγών στις ιδιότητες του ιστού υπό εξέταση, η απόδοση της ανθρωπίνης αντίληψης θεωρείται εξαιρετική. Επιπλέον, οι ιατρικοί εμπειρογνώμονες έχουν τη δυνατότητα σε καθεμία να προσθέτουν ή να αφαιρούν συστατικά από μια εικόνα για να προσδώσουν νόημα σε αυτά που βλέπουν. Οι παθολόγοι μπορούν επίσης να προσαρμόσουν στις αλλαγές μέχρι το σημείο που μπορεί να αναγνωριστεί ακόμη και μια διαστρεβλωμένη εικόνα.

Στην ενότητα αυτή προτείνουμε μια υβριδική μέθοδο για την αυτόματη ανίχνευση όγκων σε εικόνες από κολονοσκόπησης που συνδυάζει, την κατάτμηση των εικόνων, την ομαδοποίηση και τα τεχνητά νευρωνικά δίκτυα [93].

9.2.1 Προτεινόμενη Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε για την αντιμετώπιση αυτού του προβλήματος είναι ίδια με αυτήν που παρουσιάστηκε στη προηγούμενη ενότητα. Όμοια, κατασκευάζον-

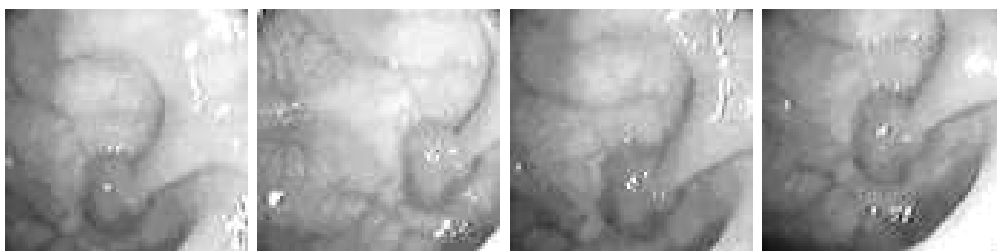
ται πολλά τοπικά μοντέλα, για γειτονιές του χώρου των δεδομένων, χρησιμοποιώντας το αποτέλεσμα της ομαδοποίησης στο σύνολο εκπαίδευσης. Έπειτα για κάθε μια ανιχνευμένη ομάδα, ένα ξεχωριστό Νευρωνικό Δίκτυο Εμπρόσθιας Τροφοδότησης (FNN) χρησιμοποιείται σαν τοπικός κατηγοροποιητής. Ο αλγόριθμος UKW, χρησιμοποιήθηκε για την διαδικασία της ομαδοποίησης. Πιο συγκεκριμένα η μεθοδολογία συνοψίζεται ως εξής:

1. Παραγωγή των χαρακτηριστικών από τις εικόνες.
2. Αναγνώριση των ομάδων του συνόλου δεδομένων εκπαίδευσης.
3. Για κάθε μια ομάδα, ένα ξεχωριστό FNN εκπαιδεύεται με δεδομένα μόνο από τη συγκεκριμένη ομάδα.
4. Τα στοιχεία του συνόλου ελέγχου ανατίθενται σε ομάδες ανάλογα με την απόστασή τους από το κέντρο των ομάδων.
5. Τα εκπαιδευμένα δίκτυα χρησιμοποιούνται για να ληφθούν οι αποφάσεις κατηγοριοποίησης για το σύνολο ελέγχου.

9.2.2 Πειραματικά Αποτελέσματα

Οι κύριες δυσκολίες στην αναγνώριση των περιοχών ενδιαφέροντος για τις εικόνες κολονοσκόπησης οφείλονται στην αλλαγή της ανάλυσης και τις μεταβλητές συνθήκες αντίληψης (σκιές, φωτισμός, και αντανάκλασεις) που έχουν σαν αποτέλεσμα θορυβώδη στοιχεία. Στα συγκεκριμένα πειράματα, υφές από κανονικά και ανώμαλα δείγματα ιστού επιλέχτηκαν τυχαία από τέσσερα πλαίσια (Σχήμα 9.1) της ίδιας τηλεοπτικής ακολουθίας, που περιείχαν αλλαγές ανάλυσης, διαφορετική αντιληπτική κατεύθυνση του παθολόγου, και διαφορετικές συνθήκες φωτισμού, που χρησιμοποιήθηκαν για την εκπαίδευση και τον έλεγχο της απόδοσης.

Δεν εφαρμόστηκε κανενός είδους φιλτράρισμα ή προ-επεξεργασία των στοιχείων. Χρησιμοποιήθηκε Η μέθοδος μητρώων co-occurrence [66], για την εξαγωγή χαρακτηριστικών από τις εικόνες, που αντιπροσωπεύει τη χωρική κατανομή και την εξάρτηση των επιπέδων του γκρι σε μια τοπική περιοχή. Πιο συγκεκριμένα, κάθε εικόνα κολονοσκόπησης, χωρίστηκε σε κομμάτια μεγέθους 16×16 εικονοκυττάρων. Κατόπιν χρησιμοποιήθηκε ο αλγόριθμος μητρώων co-occurrence, για να συγκεντρώσει τις πληροφορίες σχετικά με κάθε εικονοκύτταρο σε κάθε κομμάτι της εικόνας [81, 96]. Στα πειράματα, κάθε στοιχείο αντιπροσωπεύονταν από 16 χαρακτηριστικά. Οπότε και χρησιμοποιήθηκαν Νευρωνικά Δίκτυα Εμπρόσθιας Τροφοδότησης (FNN) με 16 νευρώνες



Σχήμα 9.1: Διαδοχικές εικόνες από μια κολονοσκόπηση όπου απεικονίζεται ένας πολυποειδής όγκος του παχέος εντέρου.

εισόδου, δύο κρυφών επιπέδων από 8 και 7 νευρώνες το καθένα. Δύο κόμβοι χρησιμοποιήθηκαν στο επίπεδο εξόδου, από τους οποίους αυτός με τη μεγαλύτερη έξοδο καθόριζε την απόφαση του δικτύου. Η τελική τοπολογία των δικτύων ήταν (16–8–7–2). Για το σύνολο εκπαίδευσης 1200 τυχαία επιλεγμένα στοιχεία χρησιμοποιήθηκαν, και η εκπαίδευσή τελείωνε μετά από 300 επαναλήψεις. Οι αλγόριθμοι εκπαίδευσης που χρησιμοποιήθηκαν ήταν:

- Resilient Back Propagation (RPROP) [120].
- Improved Resilient Back Propagation (iRPROP) [72].
- Scaled Conjugate Gradient (SCG) [102].
- Adaptive On-Line Back Propagation (AOBP) [94].

Για κάθε ένα αλγόριθμο έγιναν 100 διαφορετικές διαδικασίες εκπαίδευσης για να αποκτηθεί ένα στατιστικά σημαντικό δείγμα της απόδοσης.

Η απόδοση των εκπαιδευμένων FNNs, όσον αφορά την ακρίβεια σωστής ταξινόμησης στα σύνολα δοκιμής αναφέρεται στον Πίνακα 9.2. Όλοι οι αλγόριθμοι που χρησιμοποιούνται για να εκπαιδεύσουν FNNs ήταν ικανοί να κατασκευάσουν δίκτυα με υψηλά ποσοστά σωστής ταξινόμησης.

Αλγόριθμος Εκπαίδευσης	Σωστή Ταξινόμηση στα 100 Πειράματα(%)			
	Μέση Τιμή	Διασπορά	Μέγιστη Τιμή	Ελάχιστη Τιμή
Εικόνα 1				
RPROP	94.89	0.48	95.76	94.30
iRPROP	93.11	3.48	95.43	77.52
SCG	91.57	6.20	94.55	59.58
AOBP	91.84	3.07	93.44	76.06
Εικόνα 2				
RPROP	95.09	0.49	95.96	94.12
iRPROP	93.85	1.16	95.71	91.35
SCG	90.56	8.26	94.48	59.73
AOBP	90.10	8.53	93.65	58.75
Εικόνα 3				
RPROP	73.26	1.83	76.97	67.67
iRPROP	71.53	2.60	75.58	63.36
SCG	70.95	2.15	76.71	67.27
AOBP	70.83	1.52	75.03	68.07
Εικόνα 4				
RPROP	74.99	2.74	78.93	65.55
iRPROP	70.97	3.75	78.35	63.29
SCG	68.88	5.37	86.39	57.79
AOBP	74.95	1.48	79.66	71.88

Πίνακας 9.2: Αποτελέσματα σωστής ταξινόμησης του συνόλου ελέγχου για κάθε αλγόριθμο.

9.3 Συμπεράσματα - Συνεισφορά

Στο κεφάλαιο αυτό παρουσιάστηκε μια υβριδική μεθοδολογία, που βασίζεται στις αρχές της ομαδοποίησης και των τεχνητών νευρωνικών δικτύων. Η μεθοδολογία αυτή αρχικά εφαρμόστηκε για την αυτόματη αναγνώριση ισχαιμικών επεισοδίων. Συνολικά αποτελείται από τέσσερα στάδια. Το πρώτο στάδιο περιλαμβάνει, την ομαδοποίηση των δεδομένων εκπαίδευσης. Έπειτα, για κάθε εντοπισμένη ομάδα εκπαιδεύεται ένα ξεχωριστό δίκτυο το οποίο είναι υπεύθυνο για την λήψη αποφάσεων για τα δεδομένα από το σύνολο ελέγχου, που αντιστοιχούν μόνο στη συγκεκριμένη ομάδα. Αυτή η πολυ-μοντελική προσέγγιση, δίνει τη δυνατότητα πολύ υψηλών ποσοστών κατηγοριοποίησης όπως αποδείχτηκε από τα πειραματικά αποτελέσματα.

Στη συνέχεια εφαρμόστηκε για την αυτόματη ανίχνευση όγκων σε εικόνες κολοноσκόπησης. Όμοια με την προηγούμενη εφαρμογή, χρησιμοποιώντας τον αλγόριθμο UKW, ο χώρος δεδομένων διαμερίστηκε, και έπειτα κατασκευάστηκαν τοπικά μοντέλα κατηγοριοποίησης για κάθε ένα κομμάτι ξεχωριστά. Η κατηγοριοποίηση έγινε μέσω Νευρωνικών Δικτύων Εμπρόσθιας Τροφοδότησης, τα οποία κατάφεραν να πετύχουν πολύ ελπιδοφόρα συνολικά ποσοστά σωστής κατηγοριοποίησης. Περαιτέρω έρευνα είναι αναγκαία, σε μια μεγαλύτερη κλίμακα ώστε να ερευνηθούν πλήρως τα οφέλη και να προσδιοριστούν οι περιορισμοί της προτεινόμενης πολυ-μοντελικής προσέγγισης.

Μέρος IV

Συμπεράσματα – Βιβλιογραφία –
Ευρετήριο

Συμπεράσματα Διατριβής

Οσοδήποτε μεγάλο πλήθος πειραμάτων
δεν μπορεί να αποδείξει ότι έχω δίκιο,
αλλά μόνο ένα μπορεί να αποδείξει ότι έχω άδικο.

—Albert Einstein (1879–1955)

Η συνεισφορά της διατριβής αυτής βρίσκεται στην ανάπτυξη και μελέτη, νέων μεθόδων υπολογιστικής νοημοσύνης για την ομαδοποίηση δεδομένων. Ο όρος «ομαδοποίηση δεδομένων», αναφέρεται στη διαδικασία εντοπισμού ξεχωριστών ομάδων σε ένα σύνολο από δεδομένα, ώστε τα στοιχεία που βρίσκονται στην ίδια ομάδα να είναι περισσότερο «όμοια» μεταξύ τους από στοιχεία που βρίσκονται σε διαφορετικές ομάδες.

Η ευρύτητα των εφαρμογών της ομαδοποίησης και των επιστημονικών περιοχών που χρησιμοποιείται, αντικατοπτρίζει την μεγάλη σημασία της και χρησιμότητά της στην ανάλυση δεδομένων. Παρόλο που έχουν προταθεί διάφοροι αλγόριθμοι ομαδοποίησης, ο συνεχώς αυξανόμενος όγκος των βάσεων δεδομένων και οι νέες εξαιρετικά πολύπλοκες εφαρμογές που γεννιούνται καθημερινά, δημιουργούν την ανάγκη για κατασκευή νέων αλγορίθμων. Οι νέοι αλγόριθμοι θα πρέπει να έχουν την ικανότητα να αξιοποιούν τις δυνατότητες των βάσεων δεδομένων, και να μπορούν να ενσωματώνουν γνώση για τα διαφορετικά πεδία εφαρμογής εύκολα.

Το πρωτότυπο έργο της διατριβής ξεκινά στο Κεφάλαιο 3, παρουσιάζοντας τον αλγόριθμο UKW, που κατά την εκτέλεσή του έχει την ικανότητα να προσεγγίζει το πλήθος των ομάδων σε ένα σύνολο δεδομένων. Βασισμένο στον αλγόριθμο k -windows, το σχήμα αυτό καταφέρει με μια πολύ μικρή χρονική πολυπλοκότητα σε σχέση με άλλους αντίστοιχους αλγόριθμους, να καταλήξει σε ακριβείς ομαδοποιήσεις, όπως αναδεικνύουν τα παρουσιαζόμενα πειραματικά αποτελέσματα. Ο αλγόριθμος συνολικά χρησιμοποιεί 6 παραμέτρους, που είναι όμως άμεσα συσχετισμένες με την εφαρμογή, οι οποίες μπορούν εύκολα να προσαρμοστούν στην κάθε περίπτωση.

Για την αντιμετώπιση δεδομένων που αναπαρίστανται με διακριτές τιμές χωρίς καμία διάταξη, στο Κεφάλαιο 4, προτείνεται μια επέκταση του αλγορίθμου UKW, σε μετρικούς χώρους. Η προτεινόμενη επέκταση καταφέρει να διατηρήσει όλα τα πλεονεκτήματα του αλγορίθμου UKW. Η υπολογιστική πολυπλοκότητα παραμένει υπογραμμική σε σχέση με το πλήθος των δεδομένων. Ταυτόχρονα, ο γενικευμένος αλγόριθμος έχει την δυνατότητα να προσεγγίζει αυτόματα το πλήθος των ομάδων στο σύνολο δεδομένων. Όπως δείχνουν τα παρουσιαζόμενα πειραματικά αποτελέσματα, ο γενικευμένος αλγόριθμος, καταλήγει σε ομαδοποιήσεις που ανταγωνίζονται επάξια άλλους δημοφιλείς αλγόριθμους.

Το ερευνητικό έργο της διατριβής συνεχίζεται στο Κεφάλαιο 5, παρουσιάζοντας τροποποιήσεις του αλγόριθμου UKW, με στόχο την βελτίωση των αποτελεσμάτων του. Οι προτεινόμενες τροποποιήσεις αξιοποιούν πληροφορία από τα τοπικά χαρακτηριστικά των δεδομένων, ώστε να κατευθύνουν όσο το δυνατόν καλύτερα την αλγοριθμική διαδικασία. Η πρώτη τροποποίηση χρησιμοποιεί την Fractal διάσταση, των σημείων για να κατευθύνει την όλη διαδικασία, και στοχεύει στην ανακάλυψη περιοχών με διαφορετική Fractal διάσταση, ακόμα και μέσα στην ίδια ομάδα. Αλγόριθμοι που μπορούν να εξάγουν τέτοιου είδους πληροφορία, είναι πολύτιμοι σε διάφορες εφαρμογές όπως ανάλυση χρονοσειρών, ανάλυση εικόνας, και ιατρικές εφαρμογές. Έπειτα προτείνεται η χρήση της τεχνικής PCA, για να εξαχθεί πληροφορία για τον προσανατολισμό των σημείων. Με αυτό τον τρόπο η διαδικασία της ομαδοποίησης τροποποιείται ώστε να λαμβάνει υπόψιν τον προσανατολισμό αυτό, καταφέροντας να εντοπίζει καλύτερα τις υποκείμενες ομάδες. Τα αποτελέσματα αυτής της τροποποίησης φαίνεται να βελτιώνουν σημαντικά τον αλγόριθμο UKW.

Τις τελευταίες δεκαετίες η ικανότητα συγκέντρωσης, αποθήκευσης και ανάκτησης των δεδομένων αυξάνεται συνεχώς. Αυτό το γεγονός έχει δημιουργήσει βάσεις δεδομένων που είναι καταναμημένες σε διαφορετικές θέσεις και συνδέονται με ένα δίκτυο επικοινωνίας. Παρόλο που έχουν προταθεί αρκετές προσεγγίσεις για την παράλληλη και καταναμημένη εξόρυξη πληροφορίας, δεν υπάρχουν τόσο πολλές προσεγγίσεις στον τομέα της ομαδοποίησης. Για αυτό ακριβώς το λόγο στο Κεφάλαιο 6, προσαρμόζουμε τον αλγόριθμο UKW, ώστε να είναι εφαρμόσιμος σε καταναμημένες βάσεις δεδομένων. Εξαιτίας των διαφορετικών υποθέσεων που μπορούν να γίνουν για τη φύση του περιβάλλοντος επικοινωνίας, εξετάζουμε διάφορες περιπτώσεις. Αρχικά θεωρούμε ένα περιβάλλον ελάχιστης επικοινωνίας. Αυτό μπορεί να οφείλεται είτε στο πολύ υψηλό κόστος της επικοινωνίας είτε σε θέματα ασφάλειας και προσωπικών δεδομένων. Για αυτήν την περίπτωση προτείνουμε ένα αλγόριθμο που όπως δείχνουν τα πειραματικά αποτελέσματα καταλήγει σε ομαδοποιήσεις με υψηλή απόδοση, ελαχιστοποιώντας το επικοινωνιακό κόστος ανάμεσα στους τόπους.

Έπειτα βασιζόμαστε στην υπόθεση ότι, οι τόποι κατανομής της βάσης δεδομένων συνδέονται με ένα υψηλής ταχύτητας δίκτυο, και ότι επιτρέπεται η αποκάλυψη των δεδομένων του καθενός τόπου. Για αυτήν την περίπτωση, παρουσιάζουμε μια εκδοχή του αλγόριθμου που καταφέρει να λειτουργεί σημαντικά ταχύτερα από μια υποθετική σειριακή εκτέλεση, αξιοποιώντας την υπολογιστική ισχύ του κάθε τόπου ξεχωριστά. Επίσης παρουσιάζουμε τις μαθηματικές συνθήκες που πρέπει να ισχύουν για το δίκτυο επικοινωνίας ώστε η παράλληλη εκτέλεση να καταλήγει σε επιτάχυνση.

Επίσης, ένα φαινόμενο που συναντάται όλο και συχνότερα στις πραγματικές εφαρμογές με καταναμημένες βάσεις δεδομένων, είναι διάφοροι οργανισμοί με διαφορετικές αποθηκευμένες ιδιότητες για τις ίδιες εγγραφές, να θέλουν να εκτελέσουν αλγορίθμους ομαδοποίησης χωρίς να διακινδυνεύσουν την ιδιωτικότητα των δεδομένων. Για αυτό το σκοπό, παρουσιάστηκε μια τροποποίηση του αλγόριθμου UKW, για τέτοιες κάθετα καταναμημένες βάσεις δεδομένων, με στόχο την διατήρηση της ιδιωτικότητας των αποθηκευμένων στοιχείων. Η προτεινόμενη τροποποίηση, έχει μικρότερο επικοινωνιακό και υπολογιστικό κόστος από άλλες παρόμοιες μεθόδους. Επιπρόσθετα, στα πλεονεκτήματα της συμπεριλαμβάνονται οι δυνατότητες του αυτόματου υπολογισμού του πλήθους των ομάδων και των πολύ αποδοτικών αποτελεσμάτων, που κληρονομούνται άμεσα από τον αρχικό αλγόριθμο UKW.

Η διατριβή συνεχίζεται στο Κεφάλαιο 7, εξετάζοντας την περίπτωση δυναμικών βάσεων δεδομένων. Οι περισσότεροι αλγόριθμοι ομαδοποίησης στηρίζονται στην υπό-

θεση ότι το σύνολο δεδομένων έχει δημιουργηθεί εκ των προτέρων από μια στατική κατανομή. Όμως, καθώς συλλέγονται τα δεδομένα είναι πιθανόν η διαδικασία που τα παράγει να αλλάξει και μερικές φορές μάλιστα, ραγδαία. Για ένα τέτοιο μη στατικό περιβάλλον αναπτύσσεται μια επέκταση του αλγορίθμου UKW, που ενσωματώνει τη δυναμική δομή δεικτοδότησης Bkd-tree. Επιπλέον παρουσιάζονται θεωρητικά αποτελέσματα για την πολυπλοκότητα χειρότερης περίπτωσης του αλγορίθμου. Ακόμα, τα πειραματικά αποτελέσματα, ανέδειξαν την ικανότητα του αλγορίθμου να ανανεώνει το αποτέλεσμα της ομαδοποίησης με μεγάλη ακρίβεια, επιτυγχάνοντας ταυτόχρονα επιτάχυνση σε σχέση με τη στατική έκδοση, που είναι συγκρίσιμη με αυτές άλλων προσεγγίσεων.

Το υπόλοιπο μέρος της διατριβής ασχολείται με εφαρμογές ομαδοποίησης σε πραγματικά προβλήματα. Το Κεφάλαιο 8, μελετά την εφαρμογή αλγορίθμων ομαδοποίησης σε δεδομένα γονιδιακών εκφράσεων μικρο-συστοιχιών. Η ανακάλυψη πληροφορίας σε τέτοια δεδομένα αποτελεί μια πολύ μεγάλη πρόκληση για τα Γονιδιακή και Πρωτεϊνική έρευνα. Ακόμη περισσότερο η ομαδοποίηση μοιάζει να είναι το βήμα κλειδί για την κατανόηση της μεταβολής της δραστηριότητας των γονιδίων κατά τη διάρκεια των γονιδιακών λειτουργιών και πως επηρεάζεται από τις καταστάσεις των ασθενειών και του κυτταρικού περιβάλλοντος. Πιο συγκεκριμένα, η ομαδοποίηση μπορεί να χρησιμοποιηθεί, είτε για να αναγνωρίσει ομάδες γονιδίων σύμφωνα με την έκφραση τους σε ένα σύνολο δειγμάτων, είτε για να ομαδοποιήσει δείγματα σε ομοιόμορφες ομάδες που αντιστοιχούν σε συγκεκριμένους μακροσκοπικούς φαινότυπους. Σε αυτήν την κατεύθυνση, αρχικά μελετώνται πόσο καλά διάφοροι αλγόριθμοι ομαδοποίησης, μπορούν να διαχωρίσουν ένα σύνολο δειγμάτων, βασιζόμενοι στην ομοιόμορφη έκφραση τους σε ένα σύνολο γονιδίων. Για το σκοπό αυτό χρησιμοποιήσαμε τον αλγόριθμο UKW, καθώς όπως φάνηκε από τα συγκριτικά αποτελέσματα καταφέρει να προσφέρει υψηλής ποιότητας ομαδοποιήσεις, σε τέτοιου είδους δεδομένα.

Όμως, σε πρακτικές συνθήκες η επιλογή των γονιδίων με βάση μια προκαθορισμένη πληροφορία, δεν είναι πάντα εφικτή. Για το σκοπό αυτό προτείνουμε και αξιολογήσαμε ένα υβριδικό σχήμα που καταφέρει να αυτοματοποιήσει την όλη διαδικασία. Τα αποτελέσματα της τελικής ομαδοποίησης σε δεδομένα που προήλθαν από το υβριδικό σχήμα, ήταν πολύ ενθαρρυντικά συγκρινόμενα ακόμη και με επιβλεπόμενες μεθόδους.

Το τελευταίο κεφάλαιο της διατριβής, ασχολείται με την ανάπτυξη υβριδικών τεχνικών που συνδυάζουν την ομαδοποίηση και τα Τεχνητά Νευρωνικά Δίκτυα, και αναδεικνύει τις δυνατότητες τους σε δύο πραγματικά προβλήματα. Πιο συγκεκριμένα, η προτεινόμενη μεθοδολογία αποτελείται από δύο βήματα. Αρχικά χρησιμοποιούνται μέθοδοι ομαδοποίησης, για να διαχωρίσουν τον χώρο των δεδομένων, σε ομάδες με κοινά χαρακτηριστικά. Έπειτα για κάθε μια ομάδα, εκπαιδεύονται νευρωνικά δίκτυα με στόχο την κατηγοριοποίηση των δεδομένων που ανήκουν μόνο σε αυτήν την ομάδα. Η προτεινόμενη αυτή υβριδική τεχνική, όπως φαίνεται στα πειραματικά αποτελέσματα, καταφέρει να υπερκεράσει τα αποτελέσματα άλλων μεμονωμένων τεχνικών.

Βιβλιογραφία

- [1] D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms”, In the proceedings of the *Symposium on Principles of Database Systems*, 247–255, 2001.
- [2] R. Agrawal and R. Srikant, “Privacy-preserving data mining”, In the proceedings of the *Proc. of the ACM SIGMOD Conference on Management of Data*, ACM Press, 439–450, 2000.
- [3] M.S. Aldenderfer and R.K. Blashfield, *Cluster Analysis*, Quantitative Applications in the Social Sciences, SAGE Publications, London, 1984.
- [4] P. Alevizos, “An algorithm for orthogonal range search in $d \geq 3$ dimensions”, In the proceedings of the *14th European Workshop on Computational Geometry*, Barcelona, 1998.
- [5] P. Alevizos, D.K. Tasoulis, and M.N. Vrahatis, “Parallelizing the unsupervised k -windows clustering algorithm”, In *Lecture Notes in Computer Science*, edited by R. Wyrzykowski, vol. 3019, Springer-Verlag, 225–232, 2004.
- [6] U. Alon, N. Barkai, D.A. Notterman, K.Gish, S. Ybarra, D. Mack, and A.J. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array”, *Proc. Natl. Acad. Sci. USA*, vol. 96, No. 12, 6745–6750, 1999.
- [7] O. Alter, P.O. Brown, and D. Bostein, “Singular value decomposition for genome-wide expression data processing and modeling”, *Proc. Natl. Acad. Sci. USA*, vol. 97, No. 18, 10101–10106, 2000.
- [8] M. Ankerst, M. Breunig, H. P. Kriegel, and J. Sander, “OPTICS: Ordering points to identify the clustering structure”, In the proceedings of the *ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 99)*, 49–60, 1999.
- [9] J. Aslam, K. Pelekhev, and D. Rus, “A practical clustering algorithm for static and dynamic information organization”, In the proceedings of the *ACM-SIAM Symposium on Discrete Algorithms*, 51–60, 1999.
- [10] D. Barbarä and P. Chen, “Using the fractal dimension to cluster datasets”, In the proceedings of the *KDD*, ACM Press, 260–264, 2000.
- [11] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.
- [12] J. L. Bentley and H. A. Maurer, “Efficient worst-case data structures for range searching”, *Acta Informatica*, vol. 13, 1551–68, 1980.
- [13] P. Berkhin, “Survey of clustering data mininnng techniques”, Technical report, Accrue Software, 2002.
- [14] M.W. Berry, “Large scale singular value decomposition”, *Int. J. Supercomp. Appl.*, vol. 6, 13–49, 1992.

- [15] C.L. Blake and C.J. Merz, “UCI repository of machine learning databases”, 1998.
- [16] C. Bohm, K. Kailing, P. Kroger, and A. Zimek, “Computing clusters of correlation connected objects”, In the proceedings of the *ACM SIGMOD international conference on Management of data*, ACM Press, 455–466, 2004.
- [17] D. Boley, “Principal direction divisive partitioning”, *Data Mining and Knowledge Discovery*, vol. 2, No. 4, 325–344, 1998.
- [18] B. Boutsinas, D.K. Tasoulis, and M.N. Vrahatis, “Estimating the number of clusters using a windowing technique”, *Pattern Recognition and Image Analysis (to appear)*, vol. 16, No. 2, 2006.
- [19] T. Bozkaya and M. Ozsoyoglu, “Indexing large metric spaces for similarity search queries”, *ACM Transactions on Database Systems*, vol. 24, No. 3, 1999.
- [20] F. Can, “Incremental clustering for dynamic information processing”, *ACM Trans. Inf. Syst.*, vol. 11, No. 2, 143–164, 1993.
- [21] Earthquake Catalogue, “<http://www.gein.noa.gr/services/cat.html>, Institute of Geodynamics, National Observatory of Athens”, 2005.
- [22] K. Chakrabarti and S. Mehrotra, “Local dimensionality reduction: A new approach to indexing high dimensional spaces”, In the proceedings of the *VLDB*, Morgan Kaufmann Publishers Inc., 89–100, 2000.
- [23] P.K. Chan and S.J. Stolfo, “Sharing learned models among remote database partitions by local meta-learning”, In the proceedings of the *Knowledge Discovery and Data Mining*, 2–7, 1996.
- [24] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, “Incremental clustering and dynamic information retrieval”, *SIAM Journal on Computing*, vol. 33, No. 6, 1417–1440, 2004.
- [25] B. Chazelle, “Filtering search: A new approach to query-answering”, *SIAM J. Comput.*, vol. 15, No. 3, 703–724, 1986.
- [26] B. Chazelle and L. J. Guibas, “Fractional cascading: II applications”, *Algorithmica*, vol. 1, 163–191, 1986.
- [27] C.H. Cheng, “A branch-and-bound clustering algorithm”, *IEEE Trans. Syst. Man Cybern.*, vol. 25, 895–898, 1995.
- [28] D. W.L. Cheung, S.D. Lee, and B. Kao, “A general incremental technique for maintaining discovered association rules”, In the proceedings of the *Database Systems for Advanced Applications*, 185–194, 1997.
- [29] D. Comaniciu, *Nonparametric Robust Methods for Computer Vision*, Ph.D. thesis, Rutgers, The State University of New Jersey, 2000.
- [30] D. Comaniciu, V. Ramesh, and P. Meer, “The variable bandwidth mean shift and data-driven scale selection”, In the proceedings of the *8th Int. Conference on Computer Vision*, 438–445, 2001.
- [31] B. Cui, B. C. Ooi, J. Su, and K. L. Tan, “Indexing high-dimensional data for efficient in-memory similarity search”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, No. 3, 333–353, 2005.

- [32] G. Cybenko, “Approximations by superpositions of sigmoidal functions”, *Mathematics of Control, Signals, and Systems*, vol. 2, 303–314, 1989.
- [33] M. de Berg, M. Van Kreveld, M. Overmars, and O. Schwarzkopf, *Computational Geometry: Algorithms and Applications*, Springer, 2000.
- [34] I.S. Dhillon and D.S. Modha, “A data-clustering algorithm on distributed memory multiprocessors”, In the proceedings of the *Large-Scale Parallel Data Mining, Lecture Notes in Artificial Intelligence*, 245–260, 2000.
- [35] E. Didai and J.C. Simon, “Clustering analysis”, *Digital Pattern Recognition*, 47–94, 1976.
- [36] W.R. Dillon and M. Goldstein, *Multivariate Analysis Methods and Applications*, John Wiley & Sons, Inc., 1984.
- [37] V. Dohnal, C. Gennaro, P. Savino, and P. Zezula, “D-index: Distance searching index for metric data sets”, *Multimedia Tools Appl.*, vol. 21, No. 1, 9–33, 2003.
- [38] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma, “Subspace clustering of high dimensional data.” In the proceedings of the *SIAM Data Mining 2004*, 2004.
- [39] D.L. Donoho and I.M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika*, vol. 81, 425–455, 1994.
- [40] E. R. Dougherty and M. Bruna, “A probabilistic theory of clustering”, *Pattern Recognition*, vol. 37, 917–925, 2004.
- [41] R.C. Dubes, “How many clusters are best?—an experiment”, *Pattern Recogn.*, vol. 20, 645–663, 1987.
- [42] R.C. Dubes, “Cluster analysis and related issues”, In *Handbook of Pattern Recognition & Computer Vision*, edited by C. Chen, L. Pau, and P. Wang, World Scientific Publishing Co., Inc., 3–32, 1993.
- [43] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, Inc, 1973.
- [44] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns”, *Proc. Natl. Acad. Sci. USA*, vol. 95, 14863–14868, 1998.
- [45] M. Ester, H.P. Kriegel, J. Sander, M. Wimmer, and X. Xu, “Incremental clustering for mining in a data warehousing environment”, In the proceedings of the *24rd International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., 323–333, 1998.
- [46] M. Ester and R. Wittmann, “Incremental generalization for mining in a data warehousing environment”, In the proceedings of the *6th International Conference on Extending Database Technology*, Springer-Verlag, 135–149, 1998.
- [47] A.A. Alizadeh et al., “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling”, *Nature*, vol. 403, 503–511, 2000.
- [48] D. Singh et al., “Gene expression correlates of clinical prostate cancer behavior”, *Cancer Cell*, vol. 1, 203–209, 2002.
- [49] E. Anderson et al., *LAPACK Users’ Guide 3d ed.*, SIAM, Philadelphia, 1999.

- [50] B. Everitt, *Cluster Analysis*, Social Science Research Council, 1974.
- [51] A. Evfimievski, R. Srikant, R. Agarwal, and J. Gehrke, “Privacy preserving mining of association rules”, *Inf. Syst.*, vol. 29, No. 4, 343–364, 2004.
- [52] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.
- [53] D. H. Fisher, “Knowledge acquisition via incremental conceptual clustering”, *Mach. Learn.*, vol. 2, No. 2, 139–172, 1987.
- [54] M. Freedman, K. Nissim, and B. Pinkas, “Efficient private matching and set intersection”, In the proceedings of the *Advances in Cryptology – EUROCRYPT*, 2004.
- [55] B. Fritzke, “Growing cell structures: self-organizing network for unsupervised and supervised learning”, *Neural Netw.*, vol. 7, No. 9, 1441–1460, 1994.
- [56] B. Fritzke, “A growing neural gas network learns topologies”, In *Advances in Neural Information Processing Systems*, edited by G. Tesauro, D. S. Touretzky, and T. K. Leen, MIT Press, Cambridge MA, 625–632, 1995.
- [57] V. Gaede and O. Günther, “Multidimensional access methods”, *ACM Computing Surveys*, vol. 30, No. 2, 170–231, 1998.
- [58] B. Georgescu, I. Shimshoni, and P. Meer, “Mean shift based clustering in high dimensions: a texture classification example”, In the proceedings of the *Ninth IEEE International Conference on Computer Vision*, 456–463, 2003.
- [59] O. Goldreich, S. Micali, and A. Wigderson, “How to play any mental game - a completeness theorem for protocols with honest majority”, In the proceedings of the *19th ACM Symposium on the Theory of Computing*, 218–229, 1987.
- [60] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1996.
- [61] T.R. Golub, D.K. Slomin, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Collier, M.L. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring”, *Science*, vol. 286, 531–537, 1999.
- [62] K. C. Gowda and E. Diday, “Symbolic clustering using a new dissimilarity measure”, *Pattern Recogn.*, vol. 24, No. 6, 567–578, 1991.
- [63] J. C. Gower, “A general coefficient of similarity and some of its properties”, *Biometrics*, vol. 27, 857–871, 1971.
- [64] P. Grassberg and I. Procaccia, “Characterization of strange attractors”, *Physical Review Letters*, vol. 50, No. 5, 1983.
- [65] S. Guha, R. Rastogi, and K. Shim, “CURE: An efficient algorithm for clustering large databases”, In the proceedings of the *ACM-SIGMOD International Conference on Management of Data*, Seattle, 73–84, 1998.
- [66] R.M. Haralick, “Statistical and structural approaches to texture”, *IEEE Proc.*, vol. 67, 786–804, 1979.
- [67] J.A. Hartigan and M.A. Wong, “A k -means clustering algorithm”, *Applied Statistics*, vol. 28, 100–108, 1979.

- [68] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, 1995.
- [69] S. Haykin, *Neural Networks: A Comprehensive Foundation*, New York: Macmillan College Publishing Company, 1999.
- [70] G. R. Hjaltason and H. Samet, “Index-driven similarity search in metric spaces”, *ACM Trans. Database Syst.*, vol. 28, No. 4, 517–580, 2003.
- [71] Z. Huang, “Extensions to the k -means algorithm for clustering large data sets with categorical values”, *Data Min. Knowl. Discov.*, vol. 2, No. 3, 283–304, 1998.
- [72] C. Igel and M. Hüsken, “Improving the Rprop learning algorithm”, In the proceedings of the *Second International ICSC Symposium on Neural Computation (NC 2000)*, edited by H. Bothe and R. Rojas, ICSC Academic Press, 115–121, 2000.
- [73] T. Xiong J. Ye, T. Li and R. Janardan, “Using uncorrelated discriminant analysis for tissue classification with gene expression data”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, No. 4, 181–190, 2004.
- [74] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Advanced Reference Series, Prentice-Hall, Inc, 1988.
- [75] A.K. Jain, M.N. Murty, and P.J. Flynn, “Data clustering: a review”, *ACM Computing Surveys*, vol. 31, No. 3, 264–323, 1999.
- [76] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: A survey”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, No. 11, 1370–1386.
- [77] G.H. John, R. Kohavi, and K. Pfleger, “Irrelevant features and the subset selection problem”, In the proceedings of the *International Conference on Machine Learning*, 121–129, 1994.
- [78] I. T. Jolliffe, *Principal Component Analysis*, New York: Springer Verlag, 2002.
- [79] M. Kantarcioglu and J. Vaidya, “An architecture for privacy-preserving mining of client information”, In the proceedings of the *IEEE International Conference on Data Mining, Workshop on Privacy, Security, and Data Mining*, edited by C. Clifton and V. Estivill-Castro, vol. 14, Maebashi City, Japan, Australian Computer Society, 37–42, 2002.
- [80] H. Kargupta, W. Huang, K. Sivakumar, and E.L. Johnson, “Distributed clustering using collective principal component analysis”, *Knowledge and Information Systems*, vol. 3, No. 4, 422–448, 2001.
- [81] S.A. Karkanis, G.D. Magoulas, and N. Theofanous, “Image recognition and neuronal networks: Intelligent systems for the improvement of imaging information”, *Minimal Invasive Therapy & Allied Technologies*, vol. 9, 225–230, 2000.
- [82] G. Karyapis, E.H. Han, and V. Kumar, “CHAMELEON: A hierarchical clustering algorithm using dynamic modeling”, *IEEE Computer*, vol. 32, No. 8, 68–75, 1999.
- [83] KDD, “Cup data”, 1999.

- [84] T. Kohonen, *Self-Organized Maps*, Springer Verlag, New York, Berlin, 1997.
- [85] H P. Kriegel, P. Kroger, and I. Gotlibovich, “Incremental OPTICS: Efficient computation of updates in a hierarchical cluster ordering”, In the proceedings of the *5th Int. Conf. on Data Warehousing and Knowledge Discovery*, 2003.
- [86] P. Kumar, J.S.B. Mitchell, and E.A. Yildirim, “Approximate minimum enclosing balls in high dimensions using core-sets”, *J. Exp. Algorithmics*, vol. 8, 2003.
- [87] W. Lam and A.M. Segre, “Distributed data mining of probabilistic knowledge”, In the proceedings of the *17th International Conference on Distributed Computing Systems, Washington*, IEEE Computer Society Press, 178–185, 1997.
- [88] L.S. Liebovitch and T. Toth, “A fast algorithm to determine fractal dimensions by box counting”, *Physics Letters*, vol. 141A, No. 8, 1989.
- [89] Y. Lindell and B. Pinkas, “Privacy preserving data mining”, In the proceedings of the *Advances in Cryptology, CRYPTO 2000*, Springer-Verlag, 36–54, 2000.
- [90] C.G. Looney, *Pattern recognition using neural networks*, Oxford University Press, Oxford, UK, 1997.
- [91] S.C. Madeira and A.L. Oliveira, “Biclustering algorithms for biological data analysis: A survey”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, 24–45, 2004.
- [92] N. Maglaveras, T. Stamkopoulos, C. Pappas, and M.G. Strintzis, “An adaptive backpropagation neural network for real-time ischemia episodes detection”, *IEEE Transactions on Biomedical Engineering*, vol. 45, 805–813, 1998.
- [93] G.D. Magoulas, V.P. Plagianakos, D.K. Tasoulis, and M.N. Vrahatis, “Tumor detection in colonoscopy using the unsupervised k -windows clustering algorithm and neural networks”, In the proceedings of the *Fourth European Symposium on “Biomedical Engineering”*, 2004.
- [94] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, “Adaptive stepsize algorithms for on-line training of neural networks”, *Nonlinear Analysis, T.M.A.*, vol. 47, No. 5, 3425–3430, 2001.
- [95] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, “Improved neural network-based interpretation of colonoscopy images through on-line learning and evolution”, In the proceedings of the *EUNITE Conference*, Tenerife, Spain, 2001.
- [96] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, “Neural network-based colonoscopic diagnosis using on-line learning and differential evolution”, *Applied Soft Computing*, vol. 4, No. 4, 369–379, 2004.
- [97] G.D. Magoulas, M.N. Vrahatis, and G.S. Androulakis, “Effective backpropagation training with variable stepsize”, *Neural Networks*, vol. 10, No. 1, 69–82, 1997.
- [98] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, 1983.
- [99] J. Mao and A.K. Jain, “A self-organizing network for hyperellipsoidal clustering (HEC)”, *IEEE Trans. Neural Netw.*, vol. 7, 16–29, 1996.

- [100] S. Merugu and J. Ghosh, “Privacy-preserving distributed clustering using generative models”, In the proceedings of the *Third IEEE International Conference on Data Mining*, 2003.
- [101] R.S. Michalski and R.E. Stepp, “Automated construction of classifications: Conceptual clustering versus numerical taxonomy”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, No. 4, 396 – 409, 1983.
- [102] M. F. Möller, “A scaled conjugate gradient algorithm for fast supervised learning”, *Neural Networks*, vol. 6, 525–533, 1993.
- [103] O. Nasraoui and C. Rojas, “From static to dynamic web usage mining: Towards scalable profiling and personalization with evolutionary computation”, In the proceedings of the *Workshop on Information Technology Rabat, Morocco*, 2003.
- [104] G. Navarro, “Searching in metric spaces by spatial approximation”, *The VLDB Journal*, vol. 11, No. 1, 28–46, 2002.
- [105] R.T. Ng and J. Han, “CLARANS: A method for clustering objects for spatial data mining”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, No. 5, 1003–1016, 2002.
- [106] S. R. M. Oliveira and O. R. Zaoane, “Privacy preserving clustering by data transformation”, In the proceedings of the *18th Brazilian Symposium on Databases (SBBD), Manaus, Brazil*, 304–318, 2003.
- [107] N.G. Pavlidis, D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis, “Computational intelligence methods for financial time series modeling”, *International Journal of Bifurcation and Chaos (to appear)*, 2006.
- [108] N.G. Pavlidis, D.K. Tasoulis, and M.N. Vrahatis, “Financial forecasting through unsupervised clustering and evolutionary trained neural networks”, In the proceedings of the *Congress on Evolutionary Computation*, Canberra Australia, 2003.
- [109] N.G. Pavlidis and D.K. Tasoulis M.N. Vrahatis, “Time series forecasting methodology for multiple-step-ahead prediction”, In the proceedings of the *Fourth IASTED International Conference on Computational Intelligence*, 456–461, 2005.
- [110] A.P. Pentland, “Fractal-based description of natural scenes”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, No. 6, 661–674, 1984.
- [111] C.M. Perou, S.S. Jeffrey, M. Vande Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams, S.X. Zhu, D. Lashkari, J.C. Lee, D. Shalon, P.O. Brown, and D. Botstein, “Distinctive gene expression patterns in human mammary epithelial cells and breast cancers”, *Proc. Natl. Acad. Sci. USA*, vol. 96, 9212–9217, 1999.
- [112] V.P. Plagianakos, D.K. Tasoulis, and M.N. Vrahatis, “Computational intelligence techniques for acute leukemia gene expression data classification”, In the proceedings of the *International Joint Conference on Neural Networks*, vol. 4, 2469–2474, 2005.

- [113] V.P. Plagianakos, D.K. Tasoulis, and M.N. Vrahatis, “Hybrid dimension reduction approach for gene expression data classification”, In the proceedings of the *International Joint Conference on Neural Networks 2005, Post-Conference Workshop on Computational Intelligence Approaches for the Analysis of Bioinformatics Data*, 2005.
- [114] M.G.P. Prasad, S. Dube, and K. Sridharan, “An efficient fractals-based algorithm for clustering”, In the proceedings of the *IEEE Region 10 Conference on Convergent Technologies For The Asia-Pacific*, 2003.
- [115] F. Preparata and M. Shamos, *Computational Geometry*, Springer Verlag, New York, Berlin, 1985.
- [116] O. Procopiuc, P.K. Agarwal, L. Arge, and J.S. Vitter, “Bkd-tree: A dynamic scalable kd-tree”, In the proceedings of the *Advances in Spatial and Temporal Databases, SSTD*, edited by T. Hadzilacos, Y. Manolopoulos, and J. Roddick, vol. 2750 of *Lecture Notes in Computer Science*, Springer, 46–65, 2003.
- [117] G. Qian, Q. Zhu, Q. Xue, and S. Pramanik, “The ND-tree: A dynamic indexing technique for multidimensional non-ordered discrete data spaces”, In the proceedings of the *29th International Conference on Very Large Data Bases*, Berlin, Germany, 620–631, 2003.
- [118] H. Ralambondrainy, “A conceptual version of the k -means algorithm”, *Pattern Recogn. Lett.*, vol. 16, No. 11, 1147–1157, 1995.
- [119] V. Ramasubramanian and K. Paliwal, “Fast k -dimensional tree algorithms for nearest neighbor search with application to vector quantization encoding”, *IEEE Transactions on Signal Processing*, vol. 40, No. 3, 518–531, 1992.
- [120] M. Riedmiller and H. Braun, “A direct adaptive method for faster backpropagation learning: The RPROP algorithm”, In the proceedings of the *IEEE International Conference on Neural Networks, San Francisco, CA*, 586–591, 1993.
- [121] M. Rigou, S. Sirmakessis, and A. Tsakalidis, “A computational geometry approach to web personalization”, In the proceedings of the *IEEE International Conference on E-Commerce Technology (CEC’04)*, San Diego, California, 377–380, 2004.
- [122] J.T. Robinson, “The K-D-B-tree: A search structure for large multidimensional dynamic indexes”, In the proceedings of the *SIGMOD Int. Conf. on Management of Data*, 10–18, 1981.
- [123] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications”, *Data Mining and Knowledge Discovery*, vol. 2, 169–194, 1998.
- [124] N. Sarkar and B.B. Chaudhuri, “An efficient approach to estimate fractal dimension of textural images”, *Pattern Recognition*, vol. 25, No. 9, 1035–1041, 1992.
- [125] G.A.F. Seber, *Multivariate Observations*, John Wiley & Sons, Inc., 1984.
- [126] R. Shamir and R. Sharan, “CLICK: A clustering algorithm for gene expression analysis”, In the proceedings of the *8th International Conference on Intelligent Systems for Molecular Biology (ISMB 00)*, AAAIPress, 2000.

- [127] P. Tamayo, D. Slonim, Q. Mesirov, J. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, , and T.R. Golub, “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation”, *Proc. Natl. Acad. Sci. USA*, vol. 96, 2907–2912, 1999.
- [128] D.K. Tasoulis, P. Alevizos, B. Boutsinas, and M.N. Vrahatis, “Parallel unsupervised k -windows: an efficient parallel clustering algorithm”, In *Lecture Notes in Computer Science*, edited by V. Malyskin, vol. 2763, Springer-Verlag, 336–344, 2003.
- [129] D.K. Tasoulis, E.C. Laskari, G.C. Meletiou, and M.N. Vrahatis, “Privacy preserving unsupervised clustering over vertically partitioned data”, In the proceedings of the *International Conference on Computational Science and Its Applications - ICCSA 2006*, Lecture Notes in Computer Science series, 2006.
- [130] D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis, “Unsupervised cluster analysis in bioinformatics”, In the proceedings of the *Fourth European Symposium on “Biomedical Engineering”*, 2004.
- [131] D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis, “Unsupervised clustering of bioinformatics data”, In the proceedings of the *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Eunite*, 47–53, 2004.
- [132] D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis, “Unsupervised clustering in mRNA expression profiles”, *Computers in Biology and Medicine (to appear)*, 2006.
- [133] D.K. Tasoulis, L. Vladutu, V.P. Plagianakos, A. Bezerianos, and M.N. Vrahatis, “On-line neural network training for automatic ischemia episode detection”, In *Lecture Notes in Computer Science*, edited by L. Rutkowski, J. H. Siekmann, R. Tadeusiewicz, and L. A. Zadeh, vol. 2070, Springer-Verlag, 1062–1068, 2003.
- [134] D.K. Tasoulis and M.N. Vrahatis, “Unsupervised distributed clustering”, In the proceedings of the *IASTED International Conference on Parallel and Distributed Computing and Networks*, Innsbruck, Austria, 347–351, 2004.
- [135] D.K. Tasoulis and M.N. Vrahatis, “Unsupervised clustering on dynamic databases”, *Pattern Recognition Letters*, vol. 26, No. 13, 2116–2127, 2005.
- [136] D.K. Tasoulis and M.N. Vrahatis, “Generalizing the k -windows clustering algorithm in metric spaces”, *Mathematical and Computer Modelling (to appear)*, 2006.
- [137] D.K. Tasoulis and M.N. Vrahatis, “Unsupervised clustering using fractal dimension”, *International Journal of Bifurcation and Chaos (to appear)*, 2006.
- [138] D.K. Tasoulis, D. Zeimpekis, E. Gallopoulos, and M.N. Vrahatis, “Unsupervised clustering using principal directions guidance”, In the proceedings of the *3rd world conference on Computational Statistics & Data Analysis*, 2005.
- [139] D.K. Tasoulis, D. Zeimpekis, E. Gallopoulos, and M.N. Vrahatis, “Oriented k -windows: A PCA driven clustering method”, In the proceedings of the *2nd Workshop on Algorithmic Techniques for Data Mining*, 2006.

- [140] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 1999.
- [141] J.G. Thomas, J.M. Olson, S.J. Tapscott, and L.P. Zhao, “An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles”, *Genome Research*, vol. 11, 1227–1236, 2001.
- [142] A. Törn and A. Žilinskas, *Global Optimization*, Springer-Verlag, Berlin, 1989.
- [143] C. Tryon, *Cluster Analysis*, Ann Arbor, MI: Edward Brothers, 1939.
- [144] J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, In the proceedings of the *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 639–644, 2002.
- [145] J. Vaidya and C. Clifton, “Privacy preserving k -means clustering over vertically partitioned data”, In the proceedings of the *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.
- [146] L. Vladutu, A. Bezerianos, and S. Papadimitriou, “Hierarchical state space partitioning with a network self-organizing map for the recognition of ST-T segment changes”, *Medical & Biological Engineering & Computing*, vol. 38, 406–415, 2000.
- [147] M.N. Vrahatis, B. Boutsinas, P. Alevizos, and G. Pavlides, “The new k -windows algorithm for improving the k -means clustering algorithm”, *Journal of Complexity*, vol. 18, 375–391, 2002.
- [148] X. Wen, S. Fuhrman, G. Michaels, D. Carr, S. Smith, J. Barker, and R. Somogyi, “Large-scale temporal gene expression mapping of cns development”, *Proceedings of the National Academy of Science USA*, vol. 95, 334–339, 1998.
- [149] P. Willett, “Recent trends in hierarchic document clustering: A critical review”, *Inf. Process. Manage.*, vol. 24, No. 5, 577–597, 1988.
- [150] D.R. Wilson and T.R. Martinez, “Improved heterogeneous distance functions”, *Journal of Artificial Intelligence Research*, vol. 6, 1–34, 1997.
- [151] E.P. Xing and R.M. Karp, “CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts”, *Bioinformatics Discovery Note*, vol. 1, 1–9, 2001.
- [152] X. Xu, J. Idgerand, and H.P. Kriegel, “A fast parallel clustering algorithm for large spatial databases”, *Data Mining and Knowledge Discovery*, vol. 3, 263–290, 1999.
- [153] A. C. Yao, “How to generate and exchange secrets”, In the proceedings of the *27th IEEE Symposium on Foundations of Computer Science*, 162–167, 1986.
- [154] N.P. Yianilos, “Data structures and algorithms for nearest neighbor search in general metric spaces”, In the proceedings of the *ACM-SIAM Symposium on Discrete Algorithms*, ACM Press, 311–321, 1993.
- [155] C.T. Zahn, “Graph-theoretical methods for detecting and describing gestalt clusters”, *IEEE Transactions on Computing*, vol. C-20, 68–86, 1971.

-
- [156] D. Zeimpekis and E. Gallopoulos, “PDDP(l): Towards a Flexing Principal Direction Divisive Partitioning Clustering Algorithms”, In the proceedings of the *IEEE ICDM Workshop on Clustering Large Data Sets*, edited by D. Boley, I. Dhillon, J. Ghosh, and J. Kogan, Melbourne, Florida, 26–35, 2003.
 - [157] C. Zou, B. Salzberg, and R. Ladin, “Back to the future: Dynamic hierarchical clustering”, In the proceedings of the *International Conference on Data Engineering*, IEEE Computer Society, 578–587, 1998.

Κατάλογος Δημοσιεύσεων Υποψηφίου

Διεθνή Περιοδικά με Κριτές:

- J1. B. Boutsinas, D.K. Tasoulis, and M.N. Vrahatis. Estimating the number of clusters using a windowing technique. *Pattern Recognition and Image Analysis*, 16(2), 2006.
- J2. E.C. Laskari, G.C. Meletiou, D.K. Tasoulis, and M.N. Vrahatis. Studying the performance of artificial neural networks on problems related to cryptography. *Nonlinear Analysis Series B: Real World Applications*, 2006, in press.
- J3. E.C. Laskari, G.C. Meletiou, Y.C. Stamatiou, D.K. Tasoulis, and M.N. Vrahatis. Assessing the Effectiveness of Artificial Neural Networks on Problems Related to Elliptic Curve Cryptography *Mathematical and Computer Modelling*, 2006, accepted for publication.
- J4. B. Mourrain, N.G. Pavlidis, D.K. Tasoulis and M.N. Vrahatis. Determining the Number of Real Roots of Polynomials Through Neural Networks *Computers and Mathematics with Applications*, 2006, to appear.
- J5. N.G. Pavlidis, D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis. Computational Intelligence Methods for Financial Time Series Modeling *International Journal of Bifurcation and Chaos*, 2006, accepted for publication.
- J6. D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis. Unsupervised clustering in mRNA expression profiles. *Computers in Biology and Medicine*, 2006, in press.
- J7. D.K. Tasoulis and M.N. Vrahatis. Unsupervised clustering Using Fractal Dimension. *International Journal of Bifurcation and Chaos*, 2006, in press.
- J8. D.K. Tasoulis and M.N. Vrahatis. Unsupervised clustering on dynamic databases. *Pattern Recognition Letters*, 26(13): 2116–2127, 2005.
- J9. D.K. Tasoulis and M.N. Vrahatis Generalizing the k -windows Clustering Algorithm in Metric Spaces. *Mathematical and Computer Modelling*, 2005, in press.
- J10. T. Ganchev, D.K. Tasoulis, M.N. Vrahatis, and N. Fakotakis. Locally recurrent probabilistic neural networks with application to speaker verification. *GESTS International Transaction on Speech Science and Engineering*, 1(2):1–13, 2004.
- J11. E. C. Laskari, G. C. Meletiou, D. K. Tasoulis, and M. N. Vrahatis. Privacy preserving electronic data gathering. *Mathematical and Computer Modelling*, 2004. accepted for publication.

- J12. G.C. Meletiou, D.K. Tasoulis, and M.N. Vrahatis. Cryptography through interpolation approximation and computational intelligence methods. *Bulletin of the Greek Mathematical Society*, 48:61–75, 2003.

Τόμοι με Κριτές:

- E1. D.K. Tasoulis, D. Zeimpekis, E. Gallopoulos, and M.N. Vrahatis. Oriented k-windows: A PCA driven clustering method. *2nd Workshop on Algorithmic Techniques for Data Mining*, 2006. *Lecture Notes in Computer Science*, in press, 2006.
- E2. D.K. Tasoulis, E.C. Laskari, G.C. Meletiou, and M.N. Vrahatis. Privacy preserving unsupervised clustering over vertically partitioned data. *2006 International Conference on Computational Science and Its Applications - ICCSA 2006 Lecture Notes in Computer Science series*, in press 2006.
- E3. D.K. Tasoulis and M.N. Vrahatis. Novel Approaches to Unsupervised Clustering Through the k-Windows Algorithm. *Knowledge Mining: Series Studies in Fuzziness and Soft Computing*. Series: Studies in Fuzziness and Soft Computing, Vol. 185, pages 51-78, Sirmakessis, Spiros (Ed.) 2005.
- E4. P. Alevizos, D.K. Tasoulis, and M.N. Vrahatis. Parallelizing the unsupervised k-windows clustering algorithm. In R. Wyrzykowski, editor, *Lecture Notes in Computer Science*, 3019:225–232 Springer-Verlag, 2004.
- E5. Y.G. Petalas, D.K. Tasoulis, and M.N. Vrahatis. Dynamic search trajectory methods for neural network training. In *Lecture Notes in Computer Science*, 3070:241–246. Springer-Verlag, 2004.
- E6. N.G. Pavlidis, D.K. Tasoulis, G.S. Androulakis and M.N. Vrahatis. Exchange Rate Forecasting through Distributed Time-Lagged Feedforward Neural Networks. *Supply Chain and Finance Series on Computers and Operations Research*, 2, pp. 283–298, 2002.
- E7. D.K. Tasoulis, P. Alevizos, B. Boutsinas, and M.N. Vrahatis. Parallel unsupervised k-windows: an efficient parallel clustering algorithm. In V. Malyskin, editor, *Lecture Notes in Computer Science*, volume 2763, pages 336–344. Springer-Verlag, 2003.
- E8. D.K. Tasoulis, P. Spyridonos, N.G. Pavlidis, D. Cavouras, P. Ravazoula, G. Nikiiforidis, and M.N. Vrahatis. A neural network approach to the comparative evaluation of physician’s subjective assessment versus quantitative nuclear features in grading urine bladder tumors. In *Lecture Notes in Computer Science*, 2774:199–206. Springer-Verlag, 2003.
- E9. D.K. Tasoulis, L. Vladutu, V.P. Plagianakos, A. Bezerianos, and M.N. Vrahatis. On-line neural network training for automatic ischemia episode detection. In Leszek Rutkowski, Jörg H. Siekmann, Ryszard Tadeusiewicz, and Lotfi A. Zadeh, editors, *Lecture Notes in Computer Science*, 2070:1062–1068. Springer-Verlag, 2003.

- E10. E.C. Laskari, G.C. Meletiou, D.K. Tasoulis and M.N. Vrahatis. Data Mining and Cryptology. *International Conference of Computational Methods in Sciences and Engineering 2003 (ICCMSE 2003)*, pp. 346-349, 2003.

Πρακτικά Συνεδρίων με Κριτές:

- C1. N.G. Pavlidis, V.P. Plagianakos, D.K. Tasoulis, and M.N. Vrahatis. Human Designed Vs. Genetically Programmed Differential Evolution Operators. *2006 IEEE Congress on Evolutionary Computation*, IEEE CEC 2006, in press.
- C2. D.K. Tasoulis, D. Zeimpekis, E. Gallopoulos, and M.N. Vrahatis, Unsupervised clustering using principal directions guidance. *3rd world conference on Computational Statistics & Data Analysis*, 2005.
- C3. N.G. Pavlidis, D.K. Tasoulis and M.N. Vrahatis. Time Series Forecasting Methodology for Multiple-Step-Ahead Prediction. *Fourth IASTED International Conference on Computational Intelligence*, 2005.
- C4. N.G. Pavlidis, D.K. Tasoulis, V.P. Plagianakos, G. Nikiforidis, and M.N. Vrahatis. Spiking neural network training using evolutionary algorithms. In *International Joint Conference on Neural Networks*, 2005.
- C5. N.G. Pavlidis, D.K. Tasoulis, V.P. Plagianakos, C. Siriopoulos and M.N. Vrahatis. Computational Intelligence Methods for Financial Forecasting. In *In T.E. Simos, editor, Lecture Series on Computer and Computational Sciences*, Proceedings of the International Conference of Computational Methods in Sciences and Engineering (ICCMSE 2005), volume 4, pages 1416-1419, 2005.
- C6. V.P. Plagianakos and D.K. Tasoulis and M.N. Vrahatis Hybrid Dimension Reduction Approach for Gene Expression Data Classification. *International Joint Conference on Neural Networks 2005. Post-Conference Workshop on Computational Intelligence Approaches for the Analysis of Bioinformatics Data*, 2005.
- C7. V.P. Plagianakos, D.K. Tasoulis, and M.N. Vrahatis. Recent computational intelligence techniques for gene expression data classification. In *International Joint Conference on Neural Networks*, 2005.
- C8. D.K. Tasoulis and M.N. Vrahatis The new window density function for efficient evolutionary unsupervised clustering. *2005 IEEE Congress on Evolutionary Computation*, IEEE CEC 2005, Volume 3, Pages 2388-2394.
- C9. D.K. Tasoulis, V.P. Plagianakos and M.N. Vrahatis Clustering in Evolutionary Algorithms to Efficiently Compute Simultaneously Local and Global Minima. *Congress on Evolutionary Computation (CEC 2005)*, 2005.
- C10. T. Ganchev, N. Fakotakis, D.K. Tasoulis, and M.N. Vrahatis. Generalized locally recurrent probabilistic neural networks for text-independent speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 2004.

- C11. G.D. Magoulas, V.P. Plagianakos, D.K. Tasoulis, and M.N. Vrahatis. Tumor detection in colonoscopy using the unsupervised k -windows clustering algorithm and neural networks. In *Fourth European Symposium on "Biomedical Engineering"*, 2004.
- C12. B. Mourrain, N.G. Pavlidis, D.K. Tasoulis, and M.N. Vrahatis. Computing the number of real roots of polynomials through neural networks. In T.E. Simos and Ch. Tsitouras, editors, *International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2004)*, pages 276–279. Wiley-VCH Verlag GmbH & Co., Weinheim, 2004.
- C13. K. E. Parsopoulos, D. K. Tasoulis, and M. N. Vrahatis. Multiobjective optimization using parallel vector evaluated particle swarm optimization. In M.H. Hamza, editor, *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004)*, pages 823–828, Innsbruck, Austria, 2004. IASTED/ACTA Press.
- C14. K.E Parsopoulos, D.K. Tasoulis, N.G. Pavlidis, V.P. Plagianakos, and M.N. Vrahatis. Vector evaluated differential evolution for multiobjective optimization. In *Congress on Evolutionary Computation (CEC 2004)*, Portland, Oregon, 2004.
- C15. Y.G. Petalas, D.K. Tasoulis, and M.N. Vrahatis. Trajectory methods for neural network training. In M.H. Hamza, editor, *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004)*, pages 823–828, Innsbruck, Austria, 2004. IASTED/ACTA Press.
- C16. D.K. Tasoulis, N.G. Pavlidis, V.P. Plagianakos, and M.N. Vrahatis. Parallel differential evolution. In *Congress on Evolutionary Computation (CEC 2004)*, Portland, Oregon, 2004.
- C17. D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis. Unsupervised cluster analysis in bioinformatics. In *Fourth European Symposium on "Biomedical Engineering"*, 2004.
- C18. D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis. Unsupervised clustering of bioinformatics data. In *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Eunite*, pages 47–53, 2004.
- C19. D.K. Tasoulis and M.N. Vrahatis. Novel approaches in unsupervised clustering. In *NEMIS 2004 Final Conference in Knowledge Mining*, 2004.
- C20. D.K. Tasoulis and M.N. Vrahatis. Unsupervised clustering using semi-algebraic data structures. In *Lecture Series on Computer and Computational Sciences, Proceedings of the "International Conference of Computational Methods in Sciences and Engineering"*, (ICCMSE 2004), volume 1, pages 884–887, 2004.
- C21. D.K. Tasoulis and M.N. Vrahatis. Unsupervised distributed clustering. In *IASTED International Conference on Parallel and Distributed Computing and Networks*, pages 347–351. Innsbruck, Austria, 2004.

- C22. T. Ganchev, D.K Tasoulis, M.N. Vrahatis, and N. Fakotakis. Locally recurrent probabilistic neural network for text-independent speech recognition. In *8th European Conference on Speech Communication and Technology Geneva, Switzerland*, 2003.
- C23. N.G. Pavlidis, D.K. Tasoulis, and M.N. Vrahatis. Exchange rate forecasting through unsupervised clustering and artificial neural networks. In *Computational Management Science Conference, Chania Greece*, 2003.
- C24. N.G. Pavlidis, D.K. Tasoulis, and M.N. Vrahatis. Financial forecasting through unsupervised clustering and evolutionary trained neural networks. In *Congress on Evolutionary Computation*, (IEEE CEC), pages 2314-2321, Canberra Australia, 2003.
- C25. P. Alevizos, B. Boutsinas, D.K. Tasoulis, and M.N. Vrahatis. Improving the orthogonal range search k-windows clustering algorithm. In *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence*, pages 239–245. Washington, D.C., 2002.
- C26. G.C. Meletiou, D.K Tasoulis, and M.N. Vrahatis. A first study of the neural network approach to the RSA cryptosystem. In *Proceedings of the IASTED 2002 Conference on Artificial Intelligence*, pages 483–488, 2002.
- C27. G.C. Meletiou, D.K Tasoulis, and M.N. Vrahatis. A first study of the neural network approach in problems related to cryptography. In P.M Pardalos, A. Migdalas, and G Baourakis, editors, *FEES conference on Financial Engineering, E-commerce & Supply Chain, and Strategies of Development*, 2002.

Αναφορές από άλλους ερευνητές:

1. M. Akira. Application to the PNN hardware and the image recognition which use the re-announcement integrated circuit, Master thesis, System Information Engineering Postgraduate Course, Tsukuba University, 2004. Αναφέρει την Εργασία C22.
2. B. Jansen and K. Nakayama. Neural Networks Following a Binary Approach Applied to the Integer Prime-Factorization Problem, International Joint Conference on Neural Networks, pp. 2577–2582, 2005. Αναφέρει την Εργασία: C26.
3. S.Y. Chen, C.Y. Yao, G. Xiao, Y.S. Ying, and W.L. Wang. Fault detection and prediction of clocks and timers based on computer audition and probabilistic neural networks. Lecture Notes in Computer Science, 3512:952–959, 2005. Αναφέρει την Εργασία C10.
4. G. Demir and B. Yener. Automated cancer diagnosis based on histopathological images: a systematic survey. Technical Report TR-05-09, Computer Science Department at Rensselaer Polytechnic Institute, Troy NY, U.S., 2005. It cites Paper E8.

5. R. Farjam, H. Soltanian-Zadeh, R.A. Zoroofi, and K. Jafari-Khouzani. Tree-structured grading of pathological images of prostate. In *Progress in Biomedical Optics and Imaging, Proceedings of SPIE*, volume 5747, pages 840–851, 2005. Αναφέρει την Εργασία Ε8.
6. D.H. Mantzaris, G.C. Anastassopoulos, A.D. Tsalkidis, and A.V. Adamopoulos. Intelligent prediction of vesicoureteral reflux disease. *WSEAS Transactions on Systems*, 4(9):1440–1449, 2005. Αναφέρει την Εργασία Ε8.
7. M. Ben. *Approches robustes pour la vérification automatique du locuteur par normalisation et adaptation hiérarchique*. PhD thesis, Université de Rennes 1, IRISA, Rennes, 2004. Αναφέρει την Εργασία C22.
8. Maiga Chang and Ko-Kang Chu. A hybrid training mechanism for applying neural networks to web-based applications. *IEEE International Conference on Systems, Man and Cybernetics, (SMC 2004)*, pages 3543–3547, The Hague, The Netherlands, 2004. Αναφέρει την Εργασία Ε8.
9. C. Demity and B. Yener. Automated cancer diagnosis based on histopathological images: a systematic survey. Technical Report TR-05-09, Rensselaer Polytechnic Institute, Department of Computer Science, 2004. Αναφέρει την Εργασία Ε8.
10. G. Garai and B.B. Chaudhuri. A novel genetic algorithm for automatic clustering. *Pattern Recognition Letters*, 25:173–187, 2004. Αναφέρει την Εργασία C25.
11. C. Gunduz, B. Yener, and S. H. Gultekin. The cell graphs of cancer. *Bioinformatics*, 20:1145–1151, 2004. Αναφέρει την Εργασία Ε8.
12. P. Hore. *Distributed Clustering for Scaling Classic Algorithms*. PhD thesis, Department of Computer Science and Engineering College of Engineering University of South, 2004. Αναφέρει την Εργασία C21.
13. S. Merugu and J. Ghosh. A privacy-sensitive approach to distributed clustering. *Pattern Recognition Letters (in press)*, 2004. Αναφέρει την Εργασία C21.
14. J. Nenortaitė and R. Simutis. Application of particle swarm optimization algorithm in stock markets. In *Second IASTED International Conference on Financial Engineering and Applications (FEA 2004)*, pages 185–190, 2004. Αναφέρει την Εργασία C24.
15. J. Nenortaitė and R. Simutis. Stocks’ trading system based on the particle swarm optimization algorithm. In *Lecture Notes in Computer Science*, volume 3039, pages 843–850. Springer-Verlag, 2004. Αναφέρει την Εργασία C24.
16. A.H. Pilevar and M. Sukumar. GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. *Pattern Recognition Letters (in press)*, 2004. Αναφέρει την Εργασία C25.

17. M.S. Reyes and C.A. Coello. A new multi-objective particle swarm optimizer with improved selection and diversity mechanisms. Technical Report EVOCINV-05-2004, Seccion de Computacisn, Departamento de Ingenierna Electrica, Mexico, 2004. Αναφέρει την Εργασία C13.
18. M. Rigou, S. Sirmakessis, and A. Tsakalidis. A computational geometry approach to web personalization. In IEEE International Conference on E-Commerce Technology (CEC'04), San Diego, California, 2004. Αναφέρει την Εργασία C25.
19. G. Demir and S. H. Gultekin. Learning the topological properties of brain tumors. Technical Report TR-04-14, Computer Science Department at Rensselaer Polytechnic Institute, Troy NY, U.S., 2003. Αναφέρει την Εργασία E8.
20. S. Kukkonen and J. Lampinen. GDE3: The third evolution step of generalized differential evolution, In: Proceedings of the *IEEE Congress on "Evolutionary Computation"*, (CEC 2005), Vol. 1, 2005, pp.443–450. Αναφέρει την Εργασία C13.
21. J.G. Vlachogiannis and K.Y. Lee. Determining Generator Contributions to Transmission System Using Parallel Vector Evaluated Particle Swarm Optimization. *IEEE Transactions on Power Systems*, 20(4):1765–1774, 2005. Αναφέρει την Εργασία C13.
22. K.-J. Lee, J.Joo, J. Yang and S.-Y. Park. Unit volume based distributed clustering using probabilistic mixture model. In: Proceedings of the *Discovery Science: 8th International Conference*, Singapore, October 8 - 11, 2005. A.Hoffmann, H.Motoda, and T.Scheffer (Eds.), LNAI 3735, pp.338–345, 2005. Αναφέρει την Εργασία C25.
23. G. Demir and S. H. Gultekin. Learning the topological properties of brain tumors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3), pp.262-269, 2005. Αναφέρει την Εργασία E8.
24. D. Palmer-Brown and M. Kang. ADFUNN: An adaptive function neural network In: *Adaptive and Natural Computing Algorithms: Proceedings of the International Conference 2005*, Ribeiro, B.; Albrecht, R.F.; Dobnikar, A.; Pearson, D.W.; Steele, N.C. (Eds.) Springer-Verlag New York, LLC, May 3, 2005, pp.1-4 Αναφέρει την Εργασία C4.
25. L.-B. Xu, Y.-D. Wang, X. Li and D.Y. Yang. A gene mining method on disease subclass by gene expression. *China Journal of Bioinformatics*, 2005 Vol.3 No.2 pp.69-72. Αναφέρει την Εργασία C17.
26. B.S. Suryavanshi, N. Shiri and S.P. Mudur. Incremental Relational Fuzzy Subtractive Clustering for Dynamic Web Usage Profiling. *Workshop on Knowledge Discovery in the Web (WebKDD 2005)*, 2005 pp.44-55. Αναφέρει την Εργασία J8.
27. M.A. Villalobos-Arias, G. Toscano Pulido and C.A. Coello Coello. A Proposal to Use Stripes to Maintain Diversity in a Multi-Objective Particle Swarm

- Optimizer, *2005 IEEE Swarm Intelligence Symposium (SIS'05)*, 2005, pp. 22–29, IEEE Press. Αναφέρει την Εργασία C13.
28. J. Nenortaite. Computation improvement of stock market decision making model through the Application of GRID. *Information Technology and Control*, 2005, Vol.34, No. 3, pp.269-275. Αναφέρει την Εργασία C24.
 29. K. Zielinski, D. Peters and R. Laur. Constrained Multi-Objective Optimization Using Differential Evolution, *Proceedings of the Third International Conference on Computational Intelligence, Robotics and Autonomous Systems, (CIRAS 2005)*, 13 - 16 December 2005, Singapore. Αναφέρει την Εργασία C13.
 30. E. Alba, E-G. Talbi, G. Luque and N. Melab. Metaheuristics and Parallelism. *Parallel Metaheuristics: A New Class of Algorithms*, 2005 E. Alba (ed.), Chapter 4, pp.79-104, Wiley Series on Parallel and Distributed Computing, John Wiley & Sons Inc, Hoboken, New Jersey, USA, Αναφέρει την Εργασία C13.
 31. A.J. Nebro, F. Luna, E-G. Talbi and E. Alba. Parallel Multiobjective Optimization. *Parallel Metaheuristics: A New Class of Algorithms*, 2005 E. Alba (ed.), Chapter 16, pp.371-394, Wiley Series on Parallel and Distributed Computing, John Wiley & Sons Inc, Hoboken, New Jersey, USA, Αναφέρει την Εργασία C13.
 32. S.-W. Ke, M.P. Oakes, C. Bowerman. Mining Personal Data Collections to Discover Categories and Category Labels. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2005)*. Text Mining Workshop, September 21-23, 2005, Borovets, Bulgaria, pp.17-22. Αναφέρει την Εργασία J8.
 33. R. Tir. DATA MINING EN ASSURANCE: Quelques Utilisations In: *IV Forum des Assurances d'Alger, Re'formes dans les assurances: plus de rigueur au service de la socie'te'*, November 28-29, 2005, Alger. Αναφέρει την Εργασία C24.
 34. R. Tir. A propos de l'WUtilisation de l'WIntelligence Artificielle en Finance: Apercu de quelques techniques, In *Proceedings of the Le Système National d'WInformation Economique: Etat et perspectives*, Division Recherche en Sciences de l'Information, Equipe de recherche PNR économie, Centre de Recherche sur L'information Scientifique et Technique (CERIST), January 31-February 1, 2005, Ben Aknoun, Alger. Αναφέρει την Εργασία C24.
 35. J. Nenortaite and R. Simutis, Adapting Particle Swarm Optimization to Stock Markets, In: *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications (ISDA '05)*, pp. 520–525, 2005. Αναφέρει την Εργασία C24.
 36. R. Christu Paul, P. Asokan and V.I. Prabhakar A solution to the facility layout problem having passages and inner structure walls using particle swarm optimization. *Int. J. Adv. Manuf. Technol.* 2006. Αναφέρει την Εργασία C13.

37. H. Yapicioglu, A.E. Smith and Gerry Dozier. Solving the semi-desirable facility location problem using bi-objective particle swarm. *European Journal of Operational Research*, in press 2006. Αναφέρει την Εργασία C13.
38. R. Angira and B.V. Babu, Optimization of non-linear chemical processes using modified differential evolution. 2nd Indian International Conference on Artificial Intelligence (IICAI-05), Bhanu Prasad (Ed.): IICAI-05, pp.911-923, 2005. Αναφέρει την Εργασία C16.
39. N. Zong, F. Gui, and M. Adjouadi. A new clustering algorithm of large datasets with $O(N)$ computational complexity. In: *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications, ISDA '05*, pp. 79- 82, 8-10 Sept. 2005. Αναφέρει την Εργασία C18.
40. J.G. Vlachogiannis and K.Y. Lee. Reactive Power Control based on Particle Swarm Multi-Objective Optimization. In: *International Conference on Intelligent Systems Application to Power Systems*, Page(s):494 - 498, Nov. 6-10, 2005. Αναφέρει την Εργασία C13.
41. K.P. Wong and Z.Y. Dong. Differential evolution, an alternative approach to evolutionary algorithm. In: *roceedings of the hirteenth International Conference on Intelligent Systems Application to Power Systems, (ISAP 2005)*, ovember 6-10, 2005, Arlington, U.S.A., pp.73-83, 2005. Αναφέρει την Εργασία C16.
42. A.P. Engelbrecht. Fundamentals of Computational Swarm Intelligence - Particle Swarm Optimisations. John Wiley and Sons Ltd, Chichester, England, November 4, 2005 [ISBN: 0-470-09191-6] Αναφέρει τις Εργασίες C13,C14.
43. B.V. Babu and A. Rakesh. Modified Differential Evolution (MDE) for Optimization of Non-Linear Chemical Processes. *Computers & Chemical Engineering*, (In Press), 2006. Αναφέρει την Εργασία C16.
44. L.T. Bui, H.A. Abbass, and D. Essam. Local models: an approach to distributed multiobjective optimization. Technical Report Series, No. TR-ALAR-200601002, The Artificial Life and Adaptive Robotics Laboratory (ALAR), School of Information Technology and Electrical Engineering, University of New South Wales, Canberra, Australia, February 20, 2006. Αναφέρει την Εργασία C16.

Ευρετήριο

Symbols			
DBSCAN	25	Γενικευμένη Διαδικασία Μετακίνησης	46
Density Τεχνικές	25	Γενικευμένη Διαδικασία Συγχώνευσης	47
GDBSCAN	25		
Mode-Seeking Τεχνικές	25	Δ	
hierarchical agglomerative	15	Δενδρόγραμμα	15
hierarchical divisive	15	Διάσταση fractal	55
overfitting	6	Διαδικασία Διεύρυνσης	38
k -means	21	Διαδικασία Μετακίνησης	37
k -windows	37	Διακριτικοί Μέθοδοι	14
Adaptive On-Line Back Propagation	101	Διαχωριστικοί αλγόριθμοι	20
Back Propagation with Variable Stepsize	101	Δομή Bkd-tree	76
Bandwith	24	Δομή K-D-B-tree	76
Cluster Tendency	11	Δυναμικές Βάσεις Δεδομένων	75
Fuzzy c -means	22		
Growing Neural Gas	31	E	
Improved Resilient Back Propagation	101	Εύρος	24
Mean Shift	23	Επίπεδο Διέγερσης	4
Multidim. Binary Tree	42	Επιβλεπόμενη Επιλογή Γονιδίων	92
PCA	91	Επιλογή Χαρακτηριστικών	10
Principal Component Analysis	13	Εσωτερικής Εκτίμησης	11
Resilient Back Propagation	101	Ευκλείδεια Απόσταση	13
Scaled Conjugate Gradient	101	Εξαγωγή Χαρακτηριστικών	10
Scree Plot	95	Εξελικτικές Στρατηγικές	6
fuzzy c -means	22	Εξελικτικός Προγραμματισμός	6
mode	47	Εξελικτικοί Αλγόριθμοι	6
		Εξωτερικής Εκτίμησης	11
A			
Αφαίρεση Δεδομένων	11	I	
Ανάλυση Εγκυρότητας	11	Ιεραρχικές Μέθοδοι Διάρθρωσης	18
Ανάλυση Κυρίων Παραγόντων	13	Ιεραρχικές Συναθροιστικές Μέθοδοι	15
Ανάλυση Ομάδων	9	Ιεραρχικοί Αλγόριθμοι	15
Αναπαράσταση Δεδομένων	10		
Απλή σύνδεση	16	K	
Αριθμητική Ταξινόμηση	9	κάθετα κατανεμημένη βάση δεδομένων	66
Ασαφής Συστήματα	3	Κέντρο Βάρους	37
Ασαφής Μέθοδοι	14	Κέντρο Παραθύρου	37
Αυστηροί Μέθοδοι	14	Κατηγορικά Χαρακτηριστικά	13
Αυτόματη Κατηγοριοποίηση	9	κατηγοριοποίηση με επίβλεψη	9
		κατηγοριοποίηση χωρίς επίβλεψη	9
Γ		Κολονοσκόπηση	102
Γενετικός προγραμματισμός	6		
Γενετικοί Αλγόριθμοι	6	M	
Γενικευμένη Διαδικασία Διεύρυνσης	47	Μέθοδος box-counting	55
		Μέθοδος του Ward	16
		Μέτρα Ομοιότητας	13
		Μέτρο Ομοιότητας	12

Μετακίνηση	38
Μετρική Minkowski	13
Μηχανική Μάθηση	4
Μικρο-Συστοιχίες	89
Μονοθετικές Μέθοδοι Διαιρέσης	18

N

Νευρωνικά Δίκτυα Εμπρόσθιας Τροφοδό- τησης	4
Ντετερμινιστικοί Μέθοδοι	14

O

Ομαδοποίηση Δεδομένων	9
Ορθογώνια Αναζήτηση Διαστήματος	42

Π

Παράθυρο	37, 46
Παράμετρος Ενοποίησης	40
Παράμετρος Ομοιότητας	40
Πλήρης σύνδεση	16
Ποιοτικά Χαρακτηριστικά	13
Πολυθετικές Μέθοδοι Διαιρέσης	18
Ποσοτικά Χαρακτηριστικά	13

Σ

Σύνδεση διαμέσου	16
Σύνδεση κεντροειδούς	16
Σύνδεση μέσου όρου	16
Σύνολο fractal	55
Σύνολο Δεδομένων	12
Σύνολο Μορφοκλασματικής Δομής	55
Σχετικού Ελέγχου	11
Στοιχείο	11
Στοχαστικοί Μέθοδοι	14
Συγχωνευτικοί Μέθοδοι	14
Συνάρτηση Απόστασης	13
Συνάρτηση Ενεργοποίησης	4
Συνάρτηση Πυρήνα	23
Συνάψεις	4

T

Τάση Ομαδοποίησης	11
Τεχνητά Νευρωνικά Δίκτυα	4
Τεχνητός Νευρώνας	4
Τυπολογική Ανάλυση	9

Υ

Υβριδική Υπολογιστική Νοημοσύνη	99
Υπερπροσαρμογή	6
Υπολογιστική Νοημοσύνη	3

X

Χάρτες Αυτο-οργάνωσης	29
-----------------------	----