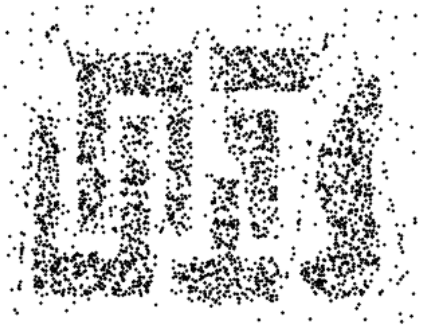
Ομαδοποίηση Δεδομένων Υψηλής Διάστασης

Σ.Κ. Τασουλής
Διατμηματικό Π.Μ.Σ.
Μαθηματικά των Υπολογιστών και των Αποφάσεων
Πανεπιστήμιο Πατρών

Διπλωματική Εργασία
Επιβλέπων: Μ.Ν. Βραχάτης



Τριμελής Επιτροπή: Μ.Ν. Βραχάτης, Π. Αλεβίζος, Β.Π. Πλαγιανάκος

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περίληψη

Η ομαδοποίηση ομαδοποιεί τα δεδομένα βασισμένη μόνο σε πληροφορία που βρίσκεται σε αυτά η οποία περιγράφει τα αντικείμενα και τις σχέσεις τους. Ο στόχος είναι τα αντικείμενα που βρίσκονται σε μια ομάδα να είναι όμοια(ή σχετικά) μεταξύ τους και διαφορετικά από τα αντικείμενα των άλλων ομάδων. Όσο μεγαλύτερη είναι η ομοιότητα(ή η ομοιογένεια) σε μια ομάδα και όσο μεγαλύτερη είναι η διαφορετικότητα ανάμεσα στις ομάδες τόσο καλύτερη είναι η ομαδοποίηση.

Οι μέθοδοι ομαδοποίησης μπορούν να διακριθούν σε τρεις κατηγορίες, ιεραρχικές, διαχωριστικές, και στις βασισμένες στη πυκνότητα. Οι ιεραρχικοί αλγόριθμοι μας δίνουν ιεραρχίες ομάδων σε μία top-down(συγχωνευτική) ή bottom-up(διαχωριστική) μορφή. Η εργασία αυτή επικεντρώνεται στην ιεραρχική διαχωριστική ομαδοποίηση. Ανάμεσα στους ιεραρχικούς διαχωριστικούς αλγορίθμους ξεχωρίζουμε τον αλγόριθμο Principal Direction Divisive Partitioning (PDDP). Ο PDDP χρησιμοποιεί την προβολή των δεδομένων στα κύρια συστατικά της αντίστοιχης μήτρας συνδιασποράς. Αυτό επιτρέπει την εφαρμογή σε δεδομένα υψηλής διάστασης. Στην εργασία αυτή προτείνεται μια βελτίωση του αλγορίθμου Principal Direction Divisive Partitioning. Ο προτεινόμενος αλγόριθμος συνδυάζει στοιχεία από την εκτίμηση πυκνότητας και τις μεθόδους βασισμένες στην προβολή με έναν γρήγορο και αποδοτικό αλγόριθμο, ικανό να αντιμετωπίσει δεδομένα υψηλής διάστασης. Τα πειραματικά αποτελέσματα δείχνουν βελτιωμένη απόδοση ομαδοποίησης σε σύγκριση με άλλες δημοφιλείς μεθόδους. Επίσης ερευνάται το πρόβλημα του αυτόματου καθορισμού του πλήθους των ομάδων που είναι πολύ σημαντικό την ομαδοποίηση.

1 Εισαγωγή

Στην ομαδοποίηση διαιρούμε τα δεδομένα σε ομάδες οι οποίες έχουν κάποιο νόημα, είναι χρήσιμες ή και τα δύο. Αν οι ομάδες με νόημα είναι ο στόχος, τότε οι ομάδες θα πρέπει να έχουν καταλάβει τη φυσική δομή των δεδομένων. Σε κάποιες περιπτώσεις, παρόλα αυτά, η ομαδοποίηση είναι μόνο ένα αρχικό βήμα για άλλους σκοπούς, όπως η σύνοψη των δεδομένων (data summarization). Είτε όταν χρησιμοποιείται για τη κατανόηση είτε για λόγους χρησιμότητας, η ομαδοποίηση μακροπρόθεσμα έχει παίξει σημαντικό ρόλο σε πολλά πεδία: στη ψυχολογία και άλλες κοινωνικές επιστήμες, στη βιολογία, στη στατιστική, στην αναγνώριση προτύπων, στην ανάκτηση πληροφορίας, στην εκμάθηση μηχανών, και στην εξόρυξη δεδομένων.

Υπάρχουν πολλές εφαρμογές της ομαδοποίησης σε πρακτικά προβλήματα. Θα δούμε κάποια συγκεκριμένα παραδείγματα, κατανεμημένα σύμφωνα με το αν ο στόχος της ομαδοποίησης είναι η κατανόηση ή η χρησιμότητα.

Η ομαδοποίηση για την κατανόηση

Οι κλάσεις, ή οι ομάδες αντικειμένων που μοιράζονται κοινά χαρακτηριστικά, παίζουν έναν σημαντικό ρόλο στο πως οι άνθρωποι αναλύουν και περιγράφουν τον κόσμο. Σίγουρα, οι άνθρωποι είναι ικανοί στο να διαιρούν αντικείμενα σε ομάδες και να τοποθετούν συγκεκριμένα αντικείμενα σε αυτές τις ομάδες. Για παράδειγμα, ακόμα και ένα σχετικά μικρό παιδί μπορεί να διαχωρίσει κάποια αντικείμενα σε μια φωτογραφία όπως τα κτίρια, τα οχήματα, οι άνθρωποι, τα ζώα, τα φυτά κ.τ.λ.. Στο πλαίσιο της κατανόησης των δεδομένων, οι ομάδες είναι πιθανές κλάσεις, και η ανάλυση ομάδων είναι η μελέτη των τεχνικών για την αυτόματη εύρεση των κλάσεων. Στη συνέχεια θα δούμε μερικά παραδείγματα.

- Βιολογία. Οι βιολόγοι για πολλά χρόνια έχουν ασχοληθεί με τη δημιουργία μιας ταξονομίας(ιεραρχική κλασικοποίηση) όλων των ζωντανών οργανισμών. Για αυτό το λόγο δεν μας εκπλήσσει ότι αρκετή από την αρχική δουλειά στην ομαδοποίηση επιχειρεί να δημιουργήσει μια μαθηματική ταξινομία που θα μπορούσε αυτόματα να βρει τέτοιες δομές κλασικοποίησης. Πιο πρόσφατα, βιολόγοι έχουν εφαρμόσει την ομαδοποίηση για να αναλύσουν την τεράστια γενετική πληροφορία που τώρα είναι διαθέσιμη. Για παράδειγμα, η ομαδοποίηση έχει χρησιμοποιηθεί για να βρει ομάδες γονιδίων που έχουν παρόμοιες λειτουργίες.

- Ανάκτηση πληροφορίας. Το διαδίκτυο περιέχει εκατομμύρια ιστοσελίδες, και το αποτέλεσμα μιας ερώτησης σε μια μηχανή αναζήτησης μπορεί να δώσει χιλιάδες απαντήσεις. Η ομαδοποίηση μπορεί να χρησιμοποιηθεί για να οργανώσει αυτά τα αποτελέσματα σε έναν μικρό αριθμό ομάδων, καθεμιά εκ των οποίων αναφέρεται σε μία πτυχή της ερώτησης. Για παράδειγμα μια αναζήτηση της λέξης ΄ταινίες΄ μπορεί να μας επιστρέψει σαν αποτέλεσμα ιστοσελίδες ομαδοποιημένες σε κατηγορίες όπως κριτικές, trailer, βαθμολογία, αίθουσες προβολής. Κάθε κατηγορία(ομάδα) μπορεί να διασπαστεί σε υποκατηγορίες(υποομάδες), παράγοντας μια ιεραρχική δομή που βοηθά τον χρήστη στην ανάγνωση των αποτελεσμάτων της αναζήτησης.

- Κλίμα. Η κατανόηση του κλίματος της γης απαιτεί την εύρεση προτύπων στην ατμόσφαιρα και τον ωκεανό. Για αυτό το λόγο, η ομαδοποίηση έχει

εφαρμοστεί στην εύρεση προτύπων στην ατμοσφαιρική πίεση των πολικών περιοχών και περιοχών του ωκεανού που έχουν μεγάλη επιρροή στη διαμόρφωση του κλίματος.

- Ψυχολογία και Φαρμακευτική. Μια αρρώστια ή κάποια συμπτώματα έχουν πλήθος παραλλαγών. Η ομαδοποίηση μπορεί να χρησιμοποιηθεί για την αναγνώριση των διάφορων υποκατηγοριών. Για παράδειγμα, η ομαδοποίηση έχει χρησιμοποιηθεί για την αναγνώριση διαφορετικών ειδών κατάθλιψης.

- Επιχειρήσεις. Οι επιχειρήσεις συλλέγουν πλήθος πληροφοριών για τους υπάρχων ή πιθανούς πελάτες τους. Η ομαδοποίηση μπορεί να χρησιμοποιηθεί στο να χωρίσει τους πελάτες σε μικρό αριθμό ομάδων για επιπλέον ανάλυση και για καθορισμό διαφημιστικών δραστηριοτήτων.

Ομαδοποίηση για λόγους χρησιμότητας

Μερικές τεχνικές ομαδοποίησης χαρακτηρίζουν κάθε ομάδα σύμφωνα με ένα πρότυπο ομάδας, για παράδειγμα ένα αντικείμενο που είναι αντιπροσωπευτικό των άλλων αντικειμένων της ομάδας. Αυτά τα πρότυπα των ομάδων μπορούν να χρησιμοποιηθούν ως βάση για ένα πλήθος τεχνικών ανάλυσης δεδομένων. Έτσι, στα πλαίσια της χρησιμότητας, η ομαδοποίηση είναι η μελέτη τεχνικών που βρίσκουν τα πιο αντιπροσωπευτικά πρότυπα ομάδων.

- Περίληψη. Πολλές τεχνικές ανάλυσης δεδομένων, όπως η οπισθοδρόμηση ή η PCA, έχουν πολυπλοκότητα χρόνου ή χώρου τάξεως $O(m^2)$ ή μεγαλύτερη (όπου m είναι το πλήθος των αντικειμένων), για αυτό το λόγο, δεν είναι πρακτικές για μεγάλα σύνολα δεδομένων. Παρόλα αυτά, αντί να εφαρμόζουμε τον αλγόριθμο σε ολόκληρο το σύνολο δεδομένων, μπορούμε να τον εφαρμόσουμε σε ένα μικρότερο σύνολο δεδομένων που αποτελείται από πρότυπα ομάδων. Ανάλογα με τον τύπο της ανάλυσης, το πλήθος των προτύπων, και την ακρίβεια με την οποία τα πρότυπα αναπαριστούν τα δεδομένα, τα αποτελέσματα μπορεί να είναι συγκρίσιμα με αυτά που θα είχαμε αν είχαν χρησιμοποιηθεί όλα τα δεδομένα.

- Συμπίεση. Τα πρότυπα ομάδων μπορούν επίσης να χρησιμοποιηθούν για συμπίεση δεδομένων. Πρακτικά, δημιουργείτε ένας πίνακας που περιέχει τα πρότυπα της κάθε ομάδας. Για παράδειγμα κάθε πρότυπο αντιστοιχεί σε έναν ακαίρεο αριθμό (ταμπέλα) που είναι η θέση του στον πίνακα. Κάθε αντικείμενο αναπαριστάται από την ταμπέλα του προτύπου που σχετίζεται με την ομάδα που ανήκει. Αυτός ο τύπος συμπίεσης συνήθως εφαρμόζεται σε δεδομένα ήχου, εικόνας, ή video, όπου (1) πολλά αντικείμενα είναι σχετικά όμοια μεταξύ τους, (2) κάποια απώλεια πληροφορίας είναι αποδεκτή και (3) μια ουσιαστική μείωση στο μέγεθος των δεδομένων είναι επιθυμητή.

- Αποδοτικότητα στην εύρεση κοντινότερων γειτόνων. Η εύρεση των κοντινότερων γειτόνων μπορεί να απαιτεί τον υπολογισμό της ανά δύο απόστασης ανάμεσα σε όλα τα σημεία. Συχνά οι ομάδες και τα πρότυπα τους μπορούν να βρεθούν αρκετά πιο αποδοτικά. Αν τα αντικείμενα είναι σχετικά κοντά στο πρότυπο της ομάδας τους, τότε μπορούμε να χρησιμοποιήσουμε τα πρότυπα για να μειώσουμε το πλήθος των υπολογισμών αποστάσεων που χρειάζονται για να βρούμε τους κοντινότερους γείτονες ενός αντικειμένου. Δηλαδή, αν δύο πρότυπα ομάδων είναι μακριά, τότε τα αντικείμενα στην αντίστοιχη ομάδα

δεν μπορούν να είναι κοντινότεροι γείτονες μεταξύ τους. Άρα, για να βρούμε τους κοντινότερους γείτονες ενός αντικειμένου αρκεί να υπολογίσουμε την απόσταση που έχει με τα αντικείμενα των κοντινών ομάδων, όπου το πόσο κοντά είναι δύο ομάδες μετριέται από την απόσταση που έχουν τα πρότυπα τους.

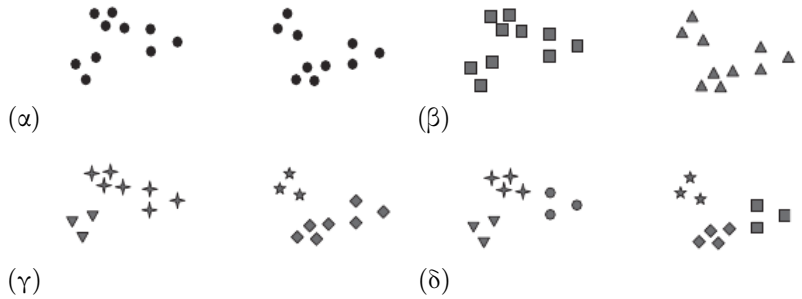
2 Ομαδοποίηση Δεδομένων

Στη συνέχεια θα αναφερθεί κάποιο απαραίτητο υπόβαθρο για τις τεχνικές ομαδοποίησης. Αρχικά αναλύεται περαιτέρω η ομαδοποίηση, ενώ στη συνέχεια θα επικεντρωθούμε σε ένα συγκεκριμένο τύπο ομαδοποίησης

2.1 Τι είναι η ομαδοποίηση

Η ομαδοποίηση ομαδοποιεί τα δεδομένα βασισμένη μόνο σε πληροφορία που βρίσκεται σε αυτά η οποία περιγράφει τα αντικείμενα και τις σχέσεις τους. Ο στόχος είναι τα αντικείμενα που βρίσκονται σε μια ομάδα να είναι όμοια(ή σχετικά) μεταξύ τους και διαφορετικά από τα αντικείμενα των άλλων ομάδων. Όσο μεγαλύτερη είναι η ομοιότητα(ή η ομοιογένεια) σε μια ομάδα και όσο μεγαλύτερη είναι η διαφορετικότητα ανάμεσα στις ομάδες τόσο καλύτερη είναι η ομαδοποίηση.

Σε πολλές εφαρμογές, η έννοια της ομάδας δεν είναι καθορισμένη επαρκώς. Για να κατανοήσουμε καλύτερα την δυσκολία του να αποφασίσουμε τι αποτελεί μια ομάδα, παρατηρούμε την εικόνα (1), όπου βλέπουμε 20 σημεία και τρεις διαφορετικούς τρόπους που μπορούμε να τα χωρίσουμε σε ομάδες. Το σχήμα του κάθε σημείου μας δείχνει σε πια ομάδα ανήκει. Στις εικόνες (1.β) και (1.δ) τα δεδομένα χωρίζονται σε 2 και 6 ομάδες αντίστοιχα. Παρόλα αυτά, ο εμφανής διαχωρισμός των δύο μεγαλύτερων ομάδων σε τρεις υποομάδες μπορεί απλά να είναι κατασκεύασμα της ανθρώπινης όρασης. Επίσης, μπορεί να μην είναι παράλογο να πούμε ότι τα σημεία σχηματίζουν 4 ομάδες, όπως παρατηρούμε στην εικόνα (1.γ). Αυτή η εικόνα μας δείχνει πως ο ορισμός μίας ομάδας είναι αμφιλεγόμενος και ο σωστός ορισμός εξαρτάται από τη φύση των δεδομένων αλλά και από τα επιθυμητά αποτελέσματα.



Σχήμα 1: (α) Αρχικά σημεία. (β) Δύο ομάδες. (γ) Τέσσερις ομάδες. (δ) Έξι ομάδες. [34]

Η ομαδοποίηση σχετίζεται με άλλες τεχνικές που χρησιμοποιούνται για να χωρίζουν τα δεδομένα σε ομάδες. Για παράδειγμα, η ομαδοποίηση μπορεί να θεωρηθεί

ως μία μορφή κλασικοποίησης καθώς δημιουργεί μια ταμπελοποίηση αντικειμένων που το καθένα έχει την ταμπέλα της κλάσης που ανήκει. Παρόλα αυτά, βάζει αυτές τις ταμπέλες βασισμένη μόνο στα δεδομένα. Σε αντίθεση, η κλασικοποίηση είναι ελεγχόμενη. Για παράδειγμα, ένα νέο αντικείμενο που δεν έχει ταμπέλα παίρνει ταμπέλα χρησιμοποιώντας ένα μοντέλο που έχει αναπτυχθεί από αντικείμενα με γνωστές ταμπέλες. Για αυτό το λόγο, η ομαδοποίηση πολλές φορές αναφέρεται ως μη ελεγχόμενη κλασικοποίηση. Όταν ο όρος κλασικοποίηση συνήθως αναφέρεται στην ελεγχόμενη κλασικοποίηση.

Επίσης, ενώ οι όροι κατάτμηση και τμηματοποίηση μερικές φορές χρησιμοποιούνται ως συνώνυμα της ομαδοποίησης, αυτοί οι όροι συνήθως χρησιμοποιούνται για προσεγγίσεις μακριά από τα παραδοσιακά όρια της ομαδοποίησης. Για παράδειγμα, ο όρος τμηματοποίηση συχνά χρησιμοποιείται σε τεχνικές που χωρίζουν γραφήματα σε υπογραφήματα και για αυτό δεν συνδέεται ισχυρά με την ομαδοποίηση. Η κατάτμηση συχνά αναφέρεται στον χωρισμό των δεδομένων σε ομάδες χρησιμοποιώντας απλές τεχνικές. Για παράδειγμα, μια εικόνα μπορεί χωριστεί σε τμήματα (segments) χρησιμοποιώντας μόνο την ευαισθησία των pixel και το χρώμα, ή οι άνθρωποι μπορούν να χωριστούν σε ομάδες βασισμένες στο μισθό τους. Εν τούτοις, κάποια δουλειά στην τμηματοποίηση γραφημάτων και στην κατάτμηση εικόνας και αγοράς σχετίζεται με την ομαδοποίηση.

2.2 Διαφορετικά είδη ομαδοποίησης

Μια ολόκληρη συλλογή από ομάδες συχνά αναφέρεται ως ομαδοποίηση. Σε αυτή την ενότητα διακρίνονται διάφορα είδη ομαδοποίησης. Η ιεραρχική(hierarchical) έναντι της τμηματικής(partitional), η αποκλειστική (exclusive) έναντι της επικαλυπτώμενης (overlapping) έναντι της (fuzzy), και η πλήρης έναντι της μερικής.

Ιεραρχική(hierarchical) έναντι Τμηματικής(partitional)

Η πιο πολυσυζητημένη διάκριση ανάμεσα στους τύπους της ομαδοποίησης είναι για το αν το σύνολο των ομάδων είναι ιεραρχικό ή τμηματικό. Η τμηματική ομαδοποίηση είναι απλά ένας διαχωρισμός του συνόλου δεδομένων σε μη επικαλύπτοντα υποσύνολα (ομάδες), έτσι ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο. Αν τα πάρουμε ξεχωριστά κάθε σύνολο ομάδων στην εικόνα 1 είναι μία τμηματική ομαδοποίηση.

Αν επιτρέψουμε στις ομάδες να έχουν υποομάδες, τότε έχουμε μία ιεραρχική ομαδοποίηση, η οποία αποτελείται ένα σύνολο ομάδων που είναι οργανωμένο στη μορφή ενός δέντρου. Κάθε κόμβος(ομάδα) στο δέντρο (εκτός από τους κόμβους φύλλα) είναι η ένωση των παιδιών του (των υποομάδων του), και η ρίζα του δέντρου είναι η ομάδα που περιέχει όλα τα αντικείμενα. Αν λοιπόν επιτρέψουμε στις ομάδες να είναι εμφωλευμένες τότε μία ερμηνεία της εικόνας $1(a)$ είναι ότι έχει δύο υποομάδες (αυτές που βλέπουμε στην εικόνα $1(b)$) κάθε μια εκ των οποίων έχει έχει τρείς υποομάδες (εικόνα $1(d)$). Οι ομάδες που βλέπουμε στην εικόνα $1(a-d)$ όταν τις παίρνουμε με αυτή τη σειρά, σχηματίζουν επίσης μια ιεραρχική ομαδοποίηση με αντίστοιχα 1,2,4 και 6 ομάδες σε κάθε επίπεδο. Τελικά, ας σημειώσουμε ότι μια ιεραρχική ομαδοποίηση μπορεί να αντιμετωπισθεί ως μια ακολουθία τμηματικών ομαδοποιήσεων και μια τμηματική ομαδοποίηση μπορεί να παραχθεί παίρνοντας οποιοδήποτε τμήμα αυτής της ακολουθίας. Για παράδειγμα κόβοντας το ιεραρχικό δέντρο σε κάποιο επίπεδο.

Αποκλειστική έναντι Επικαλυπτώμενης έναντι Fuzzy

Η ομαδοποίηση που είδαμε στην εικόνα 1 είναι εξ ολοκλήρου αποκλειστική, καθώς κάθε αντικείμενο τοποθετείται σε μία μόνο ομάδα. Υπάρχουν πολλές περιπτώσεις στις οποίες ένα σημείο μπορεί λογικά να τοποθετηθεί σε περισσότερες από μία ομάδες. Αυτές οι περιπτώσεις ανήκουν στην κατηγορία της μη αποκλειστικής ομαδοποίησης. Γενικά, μια επικαλυπτώμενη ή μη αποκλειστική ομαδοποίηση χρησιμοποιείται για να δείξουμε ότι ένα αντικείμενο μπορεί εικονικά να ανήκει σε περισσότερες από μία ομάδες. Για παράδειγμα, ένα άτομο σε ένα πανεπιστήμιο μπορεί να είναι ταυτόχρονα και φοιτητής αλλά και εργαζόμενος στο πανεπιστήμιο. Μία μη αποκλειστική ομαδοποίηση επίσης χρησιμοποιείται συχνά όταν, για παράδειγμα, ένα αντικείμενο βρίσκεται ανάμεσα σε δύο ή περισσότερες ομάδες και μπορεί λογικά να τοποθετηθεί σε κάθε μια από αυτές. Μπορούμε να φανταστούμε ένα σημείο ανάμεσα σε δύο ομάδες της εικόνας 1 να βρίσκεται περίπου στη μέση. Αντί να κάνουμε μια τυχαία επιλογή της ομάδας που θα το τοποθετήσουμε, μπορούμε να επιλέξουμε να το τοποθετήσουμε σε όλες τις ομάδες που είναι πιθανό να ανήκει.

Σε μία fuzzy ομαδοποίηση , κάθε αντικείμενο ανήκει σε κάθε ομάδα με ένα βάρος συμμετοχής μεταξύ 0(δεν ανήκει καθόλου στην ομάδα) και 1(ανήκει απόλυτα στην ομάδα). Με άλλα λόγια οι ομάδες συμπεριφέρονται σαν σύνολο fuzzy. (Μαθηματικά, ένα σύνολο fuzzy είναι αυτό στο οποίο ένα αντικείμενο ανήκει σε οποιοδήποτε σύνολο με ένα βάρος που είναι μεταξύ 0 και 1. Στην fuzzy ομαδοποίηση, συχνά απαιτούμε τον επιπλέον περιορισμό το άθροισμα των βαρών για κάθε αντικείμενο να είναι 1.) Παρόμοια, πιθανοτικές τεχνικές ομαδοποίησης υπολογίζουν την πιθανότητα με την οποία κάθε σημείο ανήκει σε κάθε ομάδα, και το άθροισμα αυτών των πιθανοτήτων πρέπει επίσης να είναι 1. Επειδή, τα βάρη συμμετοχής ή οι πιθανότητες για κάθε αντικείμενο έχουν άθροισμα 1, η fuzzy ή η πιθανοτική ομαδοποίηση δεν επιλύει πραγματικές περιπτώσεις πολλών κλάσεων, όπως την περίπτωση του εργαζόμενου φοιτητή, όπου ένα αντικείμενο ανήκει σε πολλές κλάσεις. Αντίθετα, αυτή η προσέγγιση προτιμάται για την αποφυγή της αυθαιρεσίας του να βάλουμε ένα αντικείμενο σε μόνο μία ομάδα ενώ μπορεί να είναι κοντά σε αρκετές. Πρακτικά η fuzzy ή πιθανοτική ομαδοποίηση συχνά τη μετατρέπουμε σε αποκλειστική ομαδοποίηση τοποθετώντας κάθε αντικείμενο στην ομάδα όπου το βάρος συμμετοχής του ή η πιθανότητα του είναι μεγαλύτερη.

Πλήρης έναντι Μερικής

Η πλήρης ομαδοποίηση τοποθετεί κάθε αντικείμενο σε μία ομάδα, ενώ η μερική ομαδοποίηση όχι. Το κίνητρο για μία μερική ομαδοποίηση είναι το εξής, μερικά αντικείμενα σε ένα σύνολο δεδομένων μπορεί να μην ανήκουν σε καλά ορισμένες ομάδες. Πολλές φορές αντικείμενα στο σύνολο δεδομένων μπορεί να αναπαριστούν θόρυβο, απομακρυσμένα σημεία(outliers) ή «αδιάφορο υπόβαθρο». Για παράδειγμα, μερικά άρθρα εφημερίδας μπορεί να αναφέρονται σε ένα κοινό θέμα, όπως το φαινόμενο του θερμοκηπίου, ενώ άλλα άρθρα μπορεί να είναι πιο γενικά ή να έχουν δικό τους ανεξάρτητο θέμα. Έτσι, για να βρούμε τα σημαντικά θέματα στα άρθρα του τελευταίου μήνα, μπορεί να θέλουμε να ψάξουμε μόνο για ομάδες αρχείων που είναι στενά συνδεμένες με ένα κοινό θέμα. Σε άλλες περιπτώσεις, επιθυμείτε μία πλήρης ομαδοποίηση των αντικειμένων. Για παράδειγμα, μία εφαρμογή που χρησιμοποιεί την ομαδοποίηση για να οργανώσει αρχεία για ανάγκες αναζήτησης όπου εγγυάται ότι όλα τα αρχεία μπορούν να αναζητηθούν.

2.3 Διαφορετικά είδη ομάδων

Σκοπός της ομαδοποίησης είναι η εύρεση χρήσιμων συνόλων αντικειμένων (ομάδες), όπου η χρησιμότητα καθορίζεται από το στόχο της ανάλυσης δεδομένων. Υπάρχουν αρκετές διαφορετικές ερμηνείες για μία ομάδα που είναι χρήσιμη στη πράξη. Για να δούμε μία γραφική απεικόνιση των διαφορών ανάμεσα σε τύπους ομάδων, χρησιμοποιούνται ως δεδομένα δισδιάστατα σημεία, όπως βλέπουμε στην εικόνα 2. Τονίζουμε, εντούτοις, ότι οι τύποι ομάδων που περιγράφονται εδώ ισχύουν εξίσου για άλλα είδη δεδομένων.

Καλά χωρισμένες ομάδες

Μία ομάδα είναι ένα σύνολο αντικειμένων στο οποίο σύνολο κάθε αντικείμενο είναι πιο κοντά (ή πιο όμοιο) με κάθε άλλο αντικείμενο στην ομάδα από ότι με οποιοδήποτε αντικείμενο που δεν βρίσκεται στην ομάδα. Μερικές φορές χρησιμοποιείται ένα κατώφλι για να καθορίσουμε ότι όλα τα αντικείμενα σε μία ομάδα πρέπει να είναι επαρκώς κοντά (ή όμοια) μεταξύ τους. Αυτός ο ουτοπιστικός ορισμός μίας ομάδας ικανοποιείται μόνο όταν τα δεδομένα αποτελούνται από φυσικές ομάδες που είναι αρκετά μακριά μεταξύ τους. Στην εικόνα 2(α) βλέπουμε ένα παράδειγμα καλά χωρισμένων ομάδων που αποτελείται από δύο σύνολα σημείων στον δισδιάστατο χώρο. Η απόσταση ανάμεσα σε οποιαδήποτε δύο σημεία που ανήκουν σε διαφορετικές ομάδες είναι μεγαλύτερη από ότι η απόσταση ανάμεσα σε δύο σημεία που ανήκουν στην ίδια ομάδα. Οι καλά χωρισμένες ομάδες δεν χρειάζεται να είναι σφαιρικές, μπορούν να έχουν οποιοδήποτε σχήμα.

Ομάδες βασισμένες σε πρότυπο

Μία ομάδα είναι ένα σύνολο αντικειμένων στο οποίο κάθε αντικείμενο είναι πιο κοντά ή πιο όμοιο με το πρότυπο που ορίζει την ομάδα από ότι με το πρότυπο οποιασδήποτε άλλης ομάδας. Για δεδομένα με συνεχείς χαρακτηριστικά, το πρότυπο μίας ομάδας συχνά είναι ένα κέντρο, για παράδειγμα, ο μέσος όλων των σημείων στην ομάδα. Όταν ένα κέντρο δεν έχει νόημα, όπως όταν τα δεδομένα έχουν κατηγοριοποιημένες ιδιότητες, το πρότυπο είναι συχνά μία μέση τιμή, για παράδειγμα, το Πίο αντιπροσωπευτικό σημείο μίας ομάδας. Για πολλούς τύπους δεδομένων, το πρότυπο μπορεί να θεωρηθεί ως το πιο κεντρικό σημείο, σε αυτές τις περιπτώσεις, συχνά αναφερόμαστε στις ομάδες βασισμένες σε πρότυπο ως ομάδες βασισμένες σε κέντρο. Τέτοιου τύπου ομάδες συνηθίζουν να είναι σφαιρικές. Στην εικόνα 2(β) βλέπουμε ένα παράδειγμα ομάδων βασισμένων σε κέντρο.

Ομάδες βασισμένες σε γραφήματα

Αν τα δεδομένα αναπαριστώνται σαν ένα γράφημα, όπου κόμβοι είναι τα αντικείμενα και οι ακμές αναπαριστούν συνδέσεις ανάμεσα σε αντικείμενα, τότε μία ομάδα μπορεί να οριστεί ως ένα συνδεμένο συστατικό. Για παράδειγμα, ένα σύνολο αντικειμένων που είναι συνδεμένα μεταξύ τους αλλά δεν έχουν καμία σύνδεση με αντικείμενα εκτός τις ομάδας. Ένα σημαντικό παράδειγμα ομάδων βασισμένων σε γραφήματα είναι οι ομάδες βασισμένες στη γειτνίαση, όπου δύο αντικείμενα συνδέονται μόνο αν βρίσκονται εντός μίας ορισμένης μεταξύ τους απόστασης. Αυτό υπονοεί ότι κάθε αντικείμενο στην βασισμένη στην γειτνίαση ομάδα είναι πιο κοντά σε κάποιο άλλο αντικείμενο της ομάδας από ότι σε οποιοδήποτε άλλο σημείο διαφορετικής ομάδας. Αυτός ο ορισμός μίας ομάδας είναι χρήσιμος όταν οι ομάδες

είναι ακανόνιστες, αλλά μπορεί να έχει προβλήματα όταν υπάρχει θόρυβος, όπως βλέπουμε από τις δύο σφαιρικές ομάδες στην εικόνα 2(γ), μία μικρή γέφυρα από σημεία μπορεί να ενώσει δύο διαφορετικές ομάδες.

Ομάδες βασισμένες στη πυκνότητα

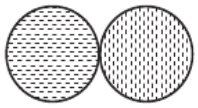
Μία ομάδα είναι μία πυκνή περιοχή αντικειμένων που περιβάλλεται από μία περιοχή χαμηλής πυκνότητας. Στην εικόνα 2(δ) βλέπουμε μερικές ομάδες βασισμένες στη πυκνότητα για δεδομένα που δημιουργήθηκαν προσθέτοντας θόρυβο στα δεδομένα της εικόνας 2(γ). Οι δύο κυκλικές ομάδες δεν ενώνονται όπως στην εικόνα 2(γ), γιατί η μεταξύ τους γέφυρα χάνεται μέσα στο θόρυβο. Παρόμοια, η καμπύλη που υπάρχει στην εικόνα 2(γ) επίσης χάνεται μέσα στο θόρυβο και δεν σχηματίζει μία ομάδα στην εικόνα 2(δ). Ένας ορισμός βασισμένος στην πυκνότητα συχνά χρειάζεται όταν οι ομάδες είναι ακανόνιστες, και όταν υπάρχουν θόρυβος και απομακρυσμένα σημεία. Αντίθετα, ένας ορισμός βασισμένος στη γειτνίαση δεν θα δούλευε καλά για τα δεδομένα της εικόνας 2(δ) αφού ο θόρυβος θα σχημάτιζε γέφυρες ανάμεσα στις ομάδες.

Μοιρασμένη ιδιότητα (εννοιολογικές ομάδες)

Γενικότερα, μπορούμε να ορίσουμε μία ομάδα ως ένα σύνολο αντικειμένων που μοιράζονται κάποια ιδιότητα. Αυτός ο ορισμός εμπεριέχει όλους τους προηγούμενους ορισμούς μίας ομάδας. Για παράδειγμα, αντικείμενα σε μία ομάδα βασισμένη στο κέντρο μοιράζονται την ιδιότητα ότι είναι όλα πιο κοντά στο ίδιο κέντρο ή μέσο. Παρόλα αυτά, η προσέγγιση της μοιρασμένης ιδιότητας επίσης περιέχει νέους τύπους ομάδων. Ας σκεφτούμε τις ομάδες της εικόνας 2(ε). Μία τριγωνική περιοχή (ομάδα) είναι παρακείμενη σε μία ορθογώνια, και βλέπουμε δύο τέμνοντες κύκλους (ομάδες). Και στις δύο περιπτώσεις, ένας αλγόριθμος ομαδοποίησης θα χρειαζόταν μία πολύ συγκεκριμένη έννοια μίας ομάδας για να ανιχνεύσει επιτυχώς αυτές τις ομάδες. Η διαδικασία εύρεσης τέτοιου τύπου ομάδες ονομάζεται εννοιολογική ομαδοποίηση (conceptual clustering).

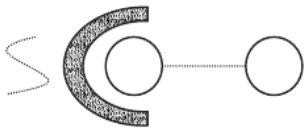
(α)

Καλά χωρισμένες ομάδες. Κάθε σημείο είναι πιο κοντά σε όλα τα σημεία της ομάδας του από ότι σε οποιοδήποτε σημείο σε άλλη ομάδα.



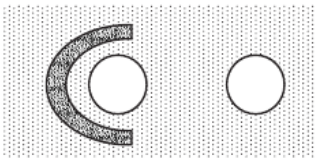
(β)

Ομάδες βασισμένες στο κέντρο. Κάθε σημείο είναι πιο κοντά στο κέντρο της ομάδας του από ότι στο κέντρο οποιασδήποτε άλλης ομάδας



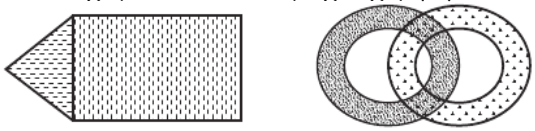
(γ)

Ομάδες βασισμένες στη γειτνίαση. Κάθε σημείο είναι πιο κοντά σε ένα τουλάχιστον σημείο της ομάδας του από ότι σε οποιοδήποτε άλλο σημείο σε άλλη ομάδα



(δ)

Ομάδες βασισμένες στη πυκνότητα. Οι ομάδες είναι περιοχές μεγάλης πυκνότητας που χωρίζονται από περιοχές χαμηλής πυκνότητας



(ε)

Εννοιολογικές ομάδες. Τα σημεία μίας ομάδας μοιράζονται κάποια γενική ιδιότητα που προκύπτει από ολόκληρο το σύνολο των σημείων.(Τα σημεία στην τομή των κύκλων ανήκουν και στους δύο).

Σχήμα 2: Διαφορετικά είδη ομάδων που παρουσιάζονται από σύνολα δισδιάστατων σημείων [34]

2.4 Αλγόριθμοι ομαδοποίησης

Στη συνέχεια θα παρουσιαστούν περιγραφικά μερικοί από τους πιο γνωστούς και απλούς αλγορίθμους ομαδοποίησης

2.4.1 K-means

Οι τεχνικές ομαδοποίησης που είναι βασισμένες σε πρότυπο δημιουργούν μία ενός επιπέδου τμηματοποίηση των δεδομένων. Υπάρχει ένα πλήθος τέτοιων τεχνικών, αλλά μία από τις πιο προεξέχων είναι η K-means. Η τεχνική K-means ορίζει ένα πρωτότυπο σύμφωνα με ένα κέντρο, που συνήθως είναι η μέση τιμή ενός συνόλου στοιχείων, και τυπικά εφαρμόζεται σε αντικείμενα ενός συνεχή n-διάστατου χώρου.

Ο βασικός K-means αλγόριθμος

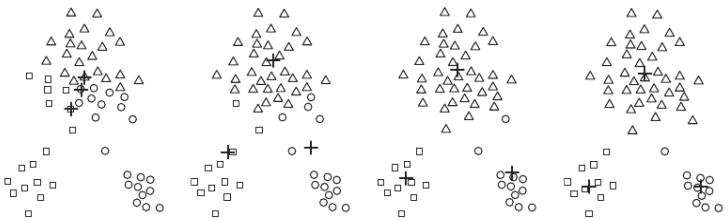
Η τεχνική ομαδοποίησης K-means είναι απλή. Επιλέγουμε K αρχικά κέντρα, όπου K είναι μία παράμετρος καθορισμένη από τον χρήστη, ονομαζόμενη ως το πλήθος των ομάδων που επιθυμούμε. Έπειτα κάθε στοιχείο ανατίθεται στο κοντινότερο κέντρο, και κάθε συλλογή στοιχείων που έχουν ανατεθεί σε ένα κέντρο σχηματίζουν μία ομάδα. Τότε επαναπροσδιορίζουμε το κέντρο βασιζόμενοι στα στοιχεία που του έχουν ανατεθεί. Τέλος επαναλαμβάνουμε την διαδικασία μέχρι κανένα στοιχείο να μην αλλάζει ομάδα, ή όμοια, μέχρι να παραμείνουν ίδια τα κέντρα.

Ο αλγόριθμος K-means περιγράφεται τυπικά στον πίνακα 1. Η λειτουργία του αλγορίθμου παρουσιάζεται στην εικόνα 3, στην οποία βλέπουμε πως, ξεκινώντας από τρία κέντρα, οι τελικές ομάδες βρίσκονται μετά από τέσσερις επαναλήψεις της διαδικασίας. Σε κάθε βήμα μπορούμε να δούμε τα κέντρα και τα σημεία που τους ανατίθενται Τα κέντρα παρουσιάζονται με το σύμβολο + και όλα τα σημεία που ανήκουν στην ίδια ομάδα συμβολίζονται με το ίδιο σχήμα.

Function K-means {
1. Διάλεξε K στοιχεία ως αρχικά κέντρα
2. επανέλαβε
3. Δημιούργησε K ομάδες τοποθετώντας το κάθε στοιχείο στην ομάδα του κοντινότερου κέντρου
4. Επαναπροσδιόρισε το κέντρο της κάθε ομάδας
5. Μέχρι τα κέντρα να μην αλλάζουν
}

Πίνακας 1: Ο βασικός K-means αλγόριθμος.

Στο πρώτο βήμα που βλέπουμε στην εικόνα 3, τα στοιχεία ανατίθενται στα αρχικά κέντρα, τα οποία είναι όλα στη μεγάλη ομάδα των στοιχείων. Για αυτό το παράδειγμα χρησιμοποιήθηκε ο μέσος ως κέντρο. Αφού τα στοιχεία ανατεθούν στα κέντρα, τα κέντρα ανανεώνονται. Στο δεύτερο βήμα τα στοιχεία ανατίθενται στα ανανεωμένα κέντρα, και τα κέντρα ανανεώνονται και πάλι. Στα επόμενα βήματα που βλέπουμε στην εικόνα τα δύο κέντρα μετακινούνται προς τις μικρότερες ομάδες στο κάτω μέρος. Όταν ο αλγόριθμος τερματίζεται στο τελευταίο βήμα καθώς τα κέντρα του δεν αλλάζουν ξανά, έχει αναγνωρίσει τις φυσικές ομάδες των στοιχείων.



Σχήμα 3: Τα τέσσερα βήματα του αλγορίθμου K-means από αριστερά προς τα δεξιά για την εύρεση τριών ομάδων [34]

Πολυπλοκότητα Χρόνου και Χώρου

Οι απαιτήσεις σε χώρο του αλγορίθμου K-means είναι μέτριες καθώς μόνο τα δεδομένα και κέντρα αποθηκεύονται. Συγκεκριμένα ο χώρος που χρειάζεται είναι $O((m + K)n)$, όπου m είναι το πλήθος των στοιχείων και n είναι το πλήθος των χαρακτηριστικών. Η απαιτήσεις του χρόνου του αλγορίθμου είναι επίσης μέτριες. Συγκεκριμένα ο χρόνος που χρειάζεται είναι $O(I * K * m * n)$, όπου I είναι το πλήθος των επαναλήψεων που χρειάζονται για να συγκλίνει ο αλγόριθμος.

Πλεονεκτήματα και Μειονεκτήματα

Ο αλγόριθμος K-means είναι απλός και μπορεί να χρησιμοποιηθεί σε μεγάλη ποικιλία τύπων δεδομένων. Είναι επίσης ιδιαίτερα αποδοτικός, παρότι συχνά εκτελούνται πολλαπλά τρεξίματα. Μερικές τροποποιήσεις του αλγορίθμου όπως ο bisecting K-means, είναι ακόμα πιο αποδοτικές, και λιγότερο ευαίσθητες σε προβλήματα αρχικοποίησης. Παρόλα αυτά ο K-means δεν είναι κατάλληλος για όλους τους τύπους δεδομένων. Δεν μπορεί να χειριστεί μη σφαιρικού τύπου ομάδες ή ομάδες διαφορετικού μεγέθους και πυκνότητας, παρότι μπορεί τυπικά να βρει καθαρές υποομάδες αν ένα αρκετά μεγάλο πλήθος ομάδων έχει καθορισθεί. Ο αλγόριθμος K-means επίσης αντιμετωπίζει προβλήματα με σύνολα δεδομένων που περιέχουν απομακρυσμένα σημεία (outliers). Η εκ των προτέρων εύρεση των απομακρυσμένων σημείων και η αφαίρεση τους βοηθάει ιδιαίτερα σε αυτές τις περιπτώσεις. Τέλος, ο αλγόριθμος K-means είναι περιορισμένος στην εφαρμογή του σε δεδομένα που υπάρχει η έννοια του κέντρου. Μία παρόμοια τεχνική, η ομαδοποίηση K-medoid, δεν έχει αυτόν τον περιορισμό , αλλά είναι αρκετά πιο ακριβή μέθοδος.

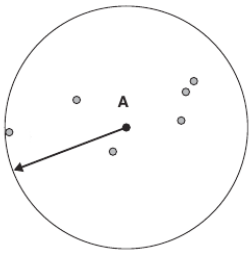
2.4.2 DBSCAN

Η ομαδοποίηση που είναι βασισμένη στην πυκνότητα εντοπίζει περιοχές υψηλής πυκνότητας οι οποίες χωρίζονται από περιοχές χαμηλής πυκνότητας. Ο DBSCAN είναι ένας απλός και αποδοτικός αλγόριθμος βασισμένος στην πυκνότητα ο οποίος παρουσιάζει πλήθος σημαντικών χαρακτηριστικών που είναι σημαντικά για την βασισμένη στην πυκνότητα ομαδοποίηση. Πριν περιγράψουμε τον αλγόριθμο DBSCAN θα αναφέρουμε έννοιες κλειδί για την πυκνότητα.

Παραδοσιακή Πυκνότητα: Βασισμένη στο Κέντρο

Παρότι δεν υπάρχουν τόσες προσεγγίσεις για τον ορισμό της πυκνότητα όπως υπάρχουν για τον ορισμό της ομοιότητας, υπάρχουν αρκετές διαφορετικές μέθοδοι. Σε αυτή τη παράγραφο αναλύουμε την βασισμένη στο κέντρο εκδοχή στην οποία βασίζεται ο αλγόριθμος DBSCAN.

Στην βασισμένη στο κέντρο εκδοχή, η πυκνότητα εκτιμάται για ένα συγκεκριμένο στοιχείο στο σύνολο δεδομένων μετρώντας το πλήθος των στοιχείων που βρίσκονται σε μία καθορισμένη ακτίνα γύρο από αυτό. Εκεί συμπεριλαμβάνεται και το ίδιο το στοιχείο. Αυτή τη τεχνική την βλέπουμε στην εικόνα 4. Το πλήθος των στοιχείων μέσα στην ακτίνα του στοιχείου A είναι 7, συμπεριλαμβανομένου του A.



Σχήμα 4: Πυκνότητα βασισμένη στο κέντρο [34].

Αυτή η μέθοδος είναι απλή να εφαρμοστεί, αλλά η πυκνότητα οποιουδήποτε στοιχείο θα εξαρτάται από τη συγκεκριμένη ακτίνα. Για παράδειγμα, αν η ακτίνα είναι αρκετά μεγάλη, τότε όλα τα στοιχεία θα έχουν πυκνότητα m, το πλήθος των στοιχείων στο σύνολο δεδομένων. Επιπλέον, αν η ακτίνα είναι πολύ μικρή, τότε όλα τα στοιχεία θα έχουν πυκνότητα 1.

Κλασικοποίηση Στοιχείων Σύμφωνα με τη Βασισμένη στο Κέντρο Πυκνότητα

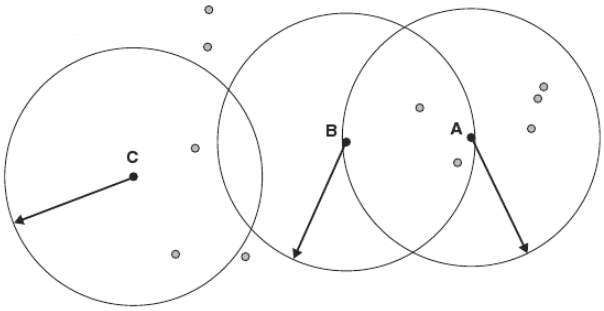
Η βασισμένη στο κέντρο εκδοχή της πυκνότητας μας επιτρέπει να κλασικοποιήσουμε ένα στοιχείο ανάλογα όταν είναι (1) στο εσωτερικό μιας πυκνής περιοχής (ένα στοιχείο πυρήνα), (2) στο σύνορο μιας πυκνής περιοχής (ένα συνοριακό στοιχείο), ή (3) σε μία πιο αραιή περιοχή (ένα στοιχείο θόρυβος). Στην εικόνα 5 βλέπουμε μία γραφική απεικόνιση των περιπτώσεων των στοιχείων πυρήνα, συνοριακών και θορύβου χρησιμοποιώντας μία συλλογή δισδιάστατων σημείων. Στη συνέχεια βλέπουμε μια πιο ακριβής περιγραφή.

- **Στοιχεία Πυρήνα:** Αυτά τα στοιχεία είναι στο εσωτερικό μίας βασισμένης στην πυκνότητα ομάδας. Ένα στοιχείο είναι ένα στοιχείο πυρήνα όταν το πλήθος των στοιχείων μέσα σε μία δεδομένη γειτονιά γύρο από το στοιχείο όπως ορίζεται από τη συνάρτηση απόστασης και από την καθορισμένη από το χρήστη παράμετρο απόστασης, ξεπερνάει ένα συγκεκριμένο όριο, MinPts, που είναι επίσης μία καθορισμένη από το χρήστη παράμετρος. Στην εικόνα 5, το στοιχείο A είναι ένα στοιχείο πυρήνα, για τη δεδομένη ακτίνα αν θέσουμε $MinPts \leq 7$.

- **Συνοριακά Στοιχεία:** Ένα συνοριακό στοιχείο δεν είναι στοιχείο πυρήνα, αλλά σχηματίζει γειτονιά με ένα στοιχείο πυρήνα. Στην εικόνα 5, το

στοιχείο B είναι ένα συνοριακό στοιχείο. Ένα συνοριακό στοιχείο μπορεί να σχηματίζει γειτονιά με αρκετά στοιχεία πυρήνα.

- **Στοιχεία Θορύβου:** Ένα στοιχείο θορύβου είναι κάθε στοιχείο που δεν είναι ούτε στοιχείο πυρήνα ούτε συνοριακό. Στην εικόνα 5, το στοιχείο C είναι στοιχείο θορύβου.



Σχήμα 5: Στοιχεία πυρήνα, συνοριακά και θορύβου [34].

Ο Αλγόριθμος DBSCAN

Δεδομένων των παραπάνω ορισμών ο αλγόριθμος DBSCAN μπορεί να ορισθεί ως ακολούθως Κάθε δύο στοιχεία πυρήνα που είναι αρκετά κοντά μεταξύ τους, δηλαδή το ένα εντός της ακτίνας του άλλου ανήκουν στην ίδια ομάδα. Παρόμοιο, κάθε συνοριακό στοιχείο που είναι κοντά σε ένα στοιχείο πυρήνα τοποθετείται στην ίδια ομάδα με το στοιχείο πυρήνα. (Ίσως χρειαστεί να επιλυθούν περιπτώσεις όπου ένα συνοριακό στοιχείο είναι κοντά σε δύο στοιχεία πυρήνα που ανήκουν σε διαφορετικές ομάδες.) Τα στοιχεία θορύβου απορρίπτονται Οι τυπικές λεπτομέρειες δίνονται στον πίνακα 2.

Function DBSCAN {
1. Ταυτοποίησε όλα τα στοιχεία ανάλογα με το αν είναι πυρήνα, συνοριακά ή θορύβου.
2. Διέγραψε τα στοιχεία θορύβου
3. Βάλε μία ακμή ανάμεσα σε όλα τα στοιχεία πυρήνα που είναι εντός της μεταξύ τους ακτίνας
4. Φτιάξε κάθε σύνολο συνδεμένων στοιχείων πυρήνα σε μία ξεχωριστή ομάδα
5. Τοποθέτησε κάθε συνοριακό στοιχείο σε μία από τις ομάδες των γειτονικών του στοιχείων πυρήνα
}

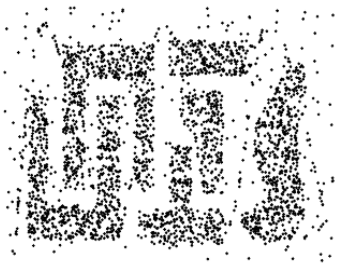
Πίνακας 2: Ο βασικός DBSCAN αλγόριθμος.

Πολυπλοκότητα Χρόνου και Χώρου

Η βασική πολυπλοκότητα χρόνου του DBSCAN αλγορίθμου είναι O($m \times$ ο χρόνος για την εύρεση της γειτονιάς), όπου m είναι το πλήθος των στοιχείων. Στην χειρότερη περίπτωση, αυτή η πολυπλοκότητα είναι O(m^2). Παρόλα αυτά, σε χώρους χαμηλής διάστασης, υπάρχουν δομές δεδομένων που επιτρέπουν την αποδοτική εύρεση όλων των στοιχείων που βρίσκονται σε μία δεδομένη απόσταση από ένα συγκεκριμένο στοιχείο, και η πολυπλοκότητα χρόνου μπορεί να είναι μέχρι και O($m \log m$). Ο χώρος που χρειάζεται ο αλγόριθμος DBSCAN ακόμα και για δεδομένα υψηλής διάστασης είναι O(m) γιατί το μόνα που χρειάζεται είναι να κρατάμε ένα μικρό ποσό δεδομένων για κάθε στοιχείο, για παράδειγμα, την ταμπέλα της ομάδας και την αναγνώριση του κάθε στοιχείου ως στοιχείο πυρήνα, συνοριακό ή θορύβου.

Πλεονεκτήματα και Μειονεκτήματα

Επειδή ο αλγόριθμος DBSCAN χρησιμοποιεί έναν βασισμένο στην πυκνότητα ορισμό μιας ομάδας, είναι σχετικά ανθεκτικός σε θόρυβο και μπορεί να αντιμετωπίσει ομάδες διαφόρων μορφών και μεγεθών. Για αυτό το λόγο ο αλγόριθμος DBSCAN μπορεί να εντοπίσει ομάδες που δεν μπορούν να βρεθούν από τον αλγόριθμο K-means όπως αυτές στην εικόνα 6. Παρόλα αυτά, ο DBSCAN έχει προβλήματα όταν οι ομάδες έχουν πολύ διαφορετικές πυκνότητες. Και επίσης έχει προβλήματα με δεδομένα υψηλής διάστασης καθώς η πυκνότητα είναι πολύ δύσκολο να ορισθεί για τέτοιου τύπου δεδομένα. Τέλος ο DBSCAN μπορεί αν είναι ιδιαίτερα ακριβός σε δεδομένα υψηλής διάστασης.



Σχήμα 6: Δείγμα δεδομένων [34].

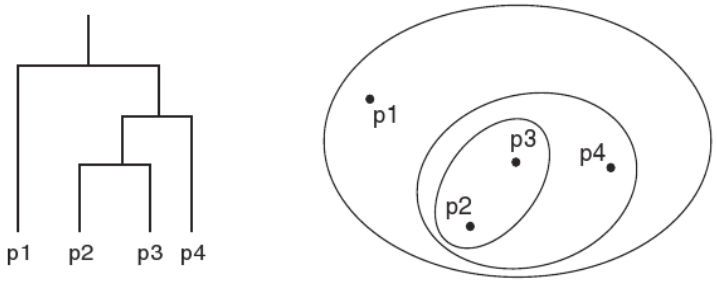
3 Ιεραρχική Ομαδοποίηση

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 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 μέθοδοι

χωρίζουν το σύνολο των n αντικειμένων σε πιο εκλεπτυσμένες διαμερίσεις και τελικά καταλήγουν στον εντοπισμό n ομάδων που η καθεμία περιέχει ένα αντικείμενο.

Το αποτέλεσμα των συναθροιστικών και των διαιρετικών μεθόδων παριστάνονται σε ένα δισδιάστατο διάγραμμα γνωστό ως δεντρόγραμμα. Οι παραγόμενες ομάδες είναι φωλιασμένες και καθεμιά μπορεί να θεωρηθεί ως ένα μέρος μίας ευρύτερης και πιο περιεκτικής ομάδας που χαρακτηρίζεται από ένα υψηλότερο επίπεδο ομοιότητας [3,36,50,125].



Σχήμα 7: Ιεραρχική ομαδοποίηση ως δεντρόγραμμα και ως φωλιασμένες ομάδες

3.1 Ιεραρχικές Μέθοδοι Διαίρεσης

Μία ιεραρχική μέθοδος διαίρεσης στην εκκίνηση της θεωρεί ότι όλες οι οντότητες ανήκουν σε μία ομάδα και στην 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 υποδιαιρέσεις των ήδη σχηματιζόμενων ομάδων [36,50,125]. Υπάρχουν δύο στρατηγικές διαίρεσης[3,50,125]:

- Μονοθετικές (Monothetic): Μονοθετική χαρακτηρίζεται μία ομάδα στην οποία όλες οι οντότητες έχουν προσεγγιστικά την ίδια τιμή για μία συγκεκριμένη μεταβλητή. Δηλαδή, οι μονοθετικές ομάδες καθορίζονται από συγκεκριμένες μεταβλητές στις οποίες συγκεκριμένες τιμές είναι απαραίτητες για να γίνουν οι οντότητες μέλη 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 δεντρόγραμμα που αναπαριστά και το τελικό αποτέλεσμα του αλγορίθμου.

Ανάμεσα στην κατηγορία των διαιρετικών ιεραρχικών αλγορίθμων, ο αλγόριθμος PDDP(Principal Direction Divisive Partitioning)[5] είναι ιδιαίτερης αξίας. Ο αλγόριθμος αυτός, βασίζεται στην τεχνική Ανάλυσης σε Πρωτεύουσες Συνιστώσες (Principal Compenent Analysis(PCA))[3], και αξιοποιεί την σποραδικότητα των δεδομένων. Αυτή η τεχνική επιτρέπει την εφαρμογή του αλγορίθμου σε δεδομένα υψηλής διάστασης, που για άλλους αλγόριθμους είναι απαγορευτική. Συγκρινόμενος με άλλες παρόμοιες τεχνικές (όπως Latent SemanticIndexing και Linear Least Square Fit), ο PDDP έχει το πλεονέκτημα της πολύ χαμηλής υπολογιστικής πολυπλοκότητας. Αυτό επιτυγχάνεται λαμβάνοντας πληροφορία μόνο από το πρώτο ιδιάζων διάνυσμα, και όχι από μία πλήρης ανάλυση της μήτρας των δεδομένων. Βέβαια σαν ιεραρχικός αλγόριθμος ομαδοποίησης το αποτέλεσμα του είναι ένα ιεραρχικό δενδρόγραμμα των ομάδων.

Πριν συνεχίσουμε στην εκτενή ανάλυση του αλγορίθμου PDDP θα κάνουμε μία αναφορά στην τεχνική Ανάλυσης σε Πρωτεύουσες Συνιστώσες.

3.2 Ανάλυση Πρωτευουσών Συνιστωσών

Η ανάλυση πρωτευουσών συνιστωσών Principal Components Analysis, PCA είναι ένας τρόπος αναγνώρισης προτύπων στα δεδομένα και εμφάνισης των δεδομένων αυτών με τέτοιο τρόπο ώστε να δίνεται έμφαση στις ομοιότητες και τις διαφορές τους. Δεδομένου ότι τα πρότυπα των δεδομένων είναι δύσκολο να βρεθούν για δεδομένα μεγάλης διάστασης, όπου η πολυτέλεια του να έχουμε γραφική απεικόνιση δεν υπάρχει, η τεχνική PCA είναι ένα δυνατό εργαλείο για την ανάλυση τους. Το κύριο χαρακτηριστικό του PCA είναι ότι είναι μια τεχνική που χρησιμοποιείται για να μειώσει τη διάσταση πολυδιάστατων συνόλων δεδομένων για την ανάλυση. Από τεχνικής άποψης η τεχνική PCA είναι ένας ορθογώνιος γραμμικός μετασχηματισμός που μετασχηματίζει τα δεδομένα σε ένα νέο σύστημα συντεταγμένων έτσι ώστε η μέγιστη διαφορά από οποιαδήποτε προβολή των στοιχείων έρχεται να βρεθεί στην πρώτη συντεταγμένη(πρώτη πρωτεύουσα συνιστώσα), η δεύτερη μέγιστη διαφορά στη δεύτερη συντεταγμένη και ούτω καθ' εξής.

Η Μέθοδος

Στη συνέχεια θα δούμε τα βήματα εφαρμογής της τεχνικής PCA. Αρχικά πρέπει να αφαιρέσουμε τη μέση τιμή κάθε διάστασης από κάθε κάθε στοιχείο της αντίστοιχης διάστασης του δείγματος μας. Η μέση τιμή που αφαιρούμε είναι ο δειγματικός μέσος της κάθε διάστασης. Για παράδειγμα αφαιρούμε όλες τις τιμές τις διάστασης x με τον δειγματικό μέσο $\hat{x}$. Έτσι θα πάρουμε ένα νέο σύνολο δεδομένων που η μέση τιμή του είναι μηδέν. Στη συνέχεια υπολογίζουμε τον πίνακα συνδιασποράς του μητρώου που περιέχει τα κεντροποιημένα στοιχεία. Αφού ο πίνακας συνδιασποράς είναι τετραγωνικός μπορούμε να υπολογίσουμε τα ιδιοδιανύσματα και τις ιδιοτιμές του. Αυτά είναι πολύ σημαντικά καθώς μας δίνουν χρήσιμες πληροφορίες για τα δεδομένα μας. Από τα ιδιοδιανύσματα μπορούμε να εξάγουμε γραμμές οι οποίες χαρακτηρίζουν τα δεδομένα μας. Τώρα παρατηρώντας τα ιδιοδιανύσματα και τις ιδιοτιμές βλέπουμε ότι οι ιδιοτιμές έχουν διαφορετικές τιμές μεταξύ τους. Τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές είναι οι πρωτεύουσες συνιστώσες του συνόλου δεδομένων μας. Αυτό που κάνουμε είναι να διατάξουμε τα ιδιοδιανύσματα σύμφωνα με τις ιδιοτιμές τους από τη μεγαλύτερη στη μικρότερη. Έτσι παίρνουμε τις συνιστώσες με τη σειρά σημαντικότητας τους. Τώρα μπορούμε να αποφασίσουμε αν θα αγνοήσουμε τις λιγότερο σημαντικές συνιστώσες. Αν

αγνοήσουμε θα χάσουμε κάποιες πληροφορίες αλλά αφού οι ιδιοτιμές τους είναι μικρές δε χάνουμε και τόσα πολλά. Αν αφήσουμε εκτός κάποιες συνιστώσες το τελικό σύνολο δεδομένων θα έχει λιγότερες διαστάσεις από το αρχικό. Στη συνέχεια κατασκευάζουμε ένα μητρώο με τα ιδιοδιανύσματα που θέλουμε να κρατήσουμε στις στήλες του, έτσι μπορούμε να πούμε ότι έχουμε ένα διάνυσμα, έστω το A του οποίου τα στοιχεία είναι τα ιδιοδιανύσματα $a_1, a_2, a_3, ..., a_n$, που ονομάζεται διάνυσμα χαρακτηριστικών γνωρισμάτων.

$$\text{A}=(a_1,\, a_2,\, a_3\, \ldots\, a_n)$$

όπου a_i ,$i = 1..n$ τα n ιδιοδιανύσματα που έχουμε κρατήσει.

Αφού έχουμε διαλέξει τις συνιστώσες(ιδιοδιανύσματα) που θέλουμε να κρατήσουμε και έχουμε κατασκευάσει το διάνυσμα χαρακτηριστικών γνωρισμάτων, πολλαπλασιάζουμε το ανάστροφο του διανύσματος αυτού με το ανάστροφο διάνυσμα των δεδομένων μας. Έτσι έχουμε

$$F= A' \times D$$

Όπου A' ο πίνακας των ιδιοδιανυσμάτων ανεστραμμένος έτσι ώστε τα ιδιοδιανύσματα τώρα να είναι στις γραμμές, ταξινομημένα από πάνω προς τα κάτω με σειρά σημαντικότητας, και D είναι ο ανάστροφος πίνακας των δεδομένων που έχουμε αφαιρέσει τη μέση τιμή. Τώρα η κάθε σειρά είναι μια διάσταση και τα δεδομένα εμφανίζονται σε κάθε στήλη. F λοιπόν θα είναι το τελικό σύνολο δεδομένων όπου τα δεδομένα εμφανίζονται σε κάθε στήλη και οι γραμμές είναι οι διαστάσεις. Τώρα έχουμε τα δεδομένα μας εκφραζόμενα με τα ιδιοδιανύσματα που έχουμε διαλέξει. Αν δεν έχουμε αφήσει κανένα ιδιοδιάνυσμα εκτός τότε απλώς έχουμε εκφράσει τα δεδομένα μας σε ένα άλλο σύστημα συντεταγμένων, αυτό που καθορίζουν τα ιδιοδιανύσματα.

4 Ο Αλγόριθμος PDDP

Όπως έχουμε ήδη αναφέρει ο αλγόριθμος PDDP, είναι μία ιεραρχική μέθοδος διαίρεσης, που χρησιμοποιεί τις προβολές των δεδομένων στις Πρωτεύουσες Συνιστώσες. Γενικά, κάθε διαιρετικός αλγόριθμος επαναληπτικά και ιεραρχικά χωρίζει το σύνολο δεδομένων σε ομάδες. Για να το πετύχει αυτό πρέπει να αντιμετωπίσει τρία ερωτήματα:

Q_1: Ποια ομάδα να χωρίσει στη συνέχεια;

Q_2: Πως να χωρίσει την επιλεγμένη ομάδα;

Q_3: Πότε πρέπει να τερματιστεί η επανάληψη;

Για να περιγράψουμε αναλυτικότερα τον αλγόριθμο ας υποθέσουμε ότι τα δεδομένα αναπαριστώνται από ένα $n \times a$ μητρώο D του οποίου οι γραμμές αναπαριστούν ένα δείγμα των δεδομένων d_i, για $i = 1, \ldots, n$. Επίσης ορίζουμε το διάνυσμα b, και το μητρώο Σ, που αναπαριστούν το διάνυσμα των μέσων και τη συνδιασπορά των δεδομένων αντίστοιχα:

$$b = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad \Sigma = \frac{1}{n}(D - be)^{\top}(D - be),$$

Function PDDP (D, c_{max}) {

1. Όσο το πλήθος των φύλλων του δέντρου*pddp* είναι μικρότερο από c_{max}

2. Διάλεξε το φύλλο P_k του δέντρου*pddp* με τη μεγαλύτερη τιμή διασποράς: $k = \arg\max_i\{scat(P_i)\}$

3. Χώρισε το P_k σύμφωνα με το πρόσημο την αντίστοιχης προβολής του $d_j \in P_k$, $j = 1, \ldots, |P_i|$

4. Κατασκεύασε τις υποομάδες του P_k και πρόσθεσε τες στο δέντρο *pddp*

}

Πίνακας 3: Ο αλγόριθμος PDDP.

όπου e είναι ένα διάνυσμα στήλη με μοναδιαία στοιχεία. Η μήτρα συνδιασποράς Σ είναι συμμετρική και θετικά ημιορισμένη, έτσι όλες οι ιδιοτιμές της είναι πραγματικές και μη αρνητικές. Τα ιδιοδιανύσματα u_j $j = 1, \ldots, k$ που αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές ονομάζονται οι πρωτεύουσες συνιστώσες ή πρωτεύουσες κατευθύνσεις. Οι προβολές p_i:

$$p_i = u_1(d_i - b), \quad i = 1, \ldots, n$$

στην πρώτη πρωτεύουσα συνιστώσα u_1, είναι η πληροφορία που χρησιμοποιεί ο αλγόριθμος PDDP για να χωρίσει αρχικά το σύνολο δεδομένων σε δύο υποσύνολα P_1,P_2, με τον παρακάτω κανόνα:

- $\forall p_i \in D$, Αν $p_i \geq 0$ τότε το i-οστό στοιχείο ανήκει στο πρώτο υποσύνολο $P_1 = P_1 \cup d_i$,

  αλλιώς ανήκει στο δεύτερο υποσύνολο $P_2 = P_2 \cup d_i$.

Διαλέγουμε το πρώτο ιδιοδιάνυσμα γιατί είναι η κατεύθυνση με τη μέγιστη διασπορά, και ως εκ τούτου η κατεύθυνση στην οποία τα δεδομένα απλώνονται περισσότερο. Σε αυτό το σημείο, ο αλγόριθμος έχει χωρίσει το αρχικό σύνολο δεδομένων σε δυο ομάδες, και είναι έτοιμος να επαναλάβει αυτή τη διαδικασία για τη μία εκ των δύο. Η επιλογή για το πια ομάδα θα χωρίσει βασίζεται στον παρακάτω κανόνα:

- Διάλεξε το P_1 Αν $\|(P_1 - b_1 e\| \geq \|(P_2 - b_2 e\|$,

  διαφορετικά διάλεξε το P_2.

Τα διανύσματα b_1 και b_2 είναι τα διανύσματα των μέσων των P_1 και P_2 αντίστοιχα, και το $\|(P_1 - b_1 e\|$ μπορεί επίσης να περιγραφεί ως η τιμή διασποράς $scat(P_1)$ του τμήματος P_1. Αυτό είναι ένα μέτρο συνεκτικότητας μιας ομάδας.

Αυτή η στρατηγική τμηματοποίησης δημιουργεί ένα δυαδικό δέντρο, που ονομάζεται *pddp δέντρο*, του οποίου τα φύλλα αποτελούν το τελικό αποτέλεσμα της ομαδοποίησης. Ο αλγόριθμος τερματίζει όταν βρεθεί το προκαθορισμένο πλήθος ομάδων c_{max}. Μία αναλυτική περιγραφή του αλγορίθμου παρουσιάζεται στον πίνακα 3. Όπως βλέπουμε τα κριτήρια που απαντούν στα ερωτήματα Q_1, Q_2 και Q_3 είναι τα παρακάτω:

5 Βελτιώνοντας τον αλγόριθμο PDDP (Improving PDDP)

Σε αυτή την ενότητα αναλύεται και απαντάται κάθε ένα από τα τρία βασικά ερωτήματα Q_1, Q_2 και Q_3 χωριστά, ως καθοδήγηση στο σχεδιασμό μίας νέας μεθόδου.

5.1 Πως διασπάται η επιλεγμένη ομάδα;

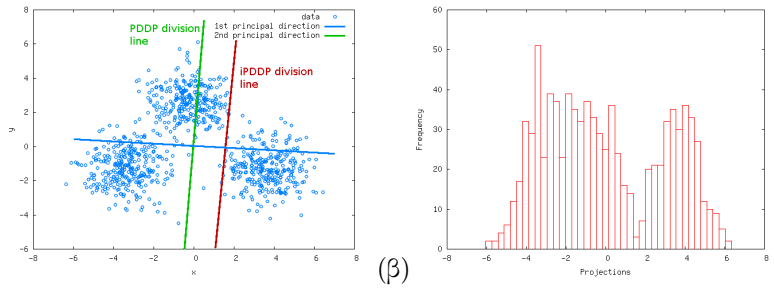
Όπως έχουμε αναφέρει σε προηγούμενη ενότητα ο αλγόριθμος PDDP χρησιμοποιεί το πρόσημο της προβολής του κάθε στοιχείου των δεδομένων ως κριτήριο διάσπασης ομάδας (σημείο διάσπασης 0). Αυτό δείχνει και τις text mining ρίζες του αλγορίθμου, τα στοιχεία (documents) που έχουν θετικές προβολές αναμένεται να είναι περισσότερο όμοια μεταξύ τους από ότι με στοιχεία (documents) που έχουν αρνητικές προβολές.

Στο nilsson2002hcu, προτάθηκε να χρησιμοποιηθεί αντί για την κύρια πρωτεύουσα συνιστώσα, η δεύτερη, η τρίτη κτλ. Με αυτόν το τρόπο, οι συγγραφείς προσπάθησαν να αποφύγουν ανεπιθύμητες διασπάσεις πραγματικών ομάδων στο σύνολο δεδομένων. Η απόφαση για το ποία προβολή θα χρησιμοποιηθεί, βασιζόταν στην τιμή διασποράς (scat value) των παραγόμενων ομάδων. Αυτό το κριτήριο αυξάνει υπερβολικά την υπολογιστική πολυπλοκότητα του αλγορίθμου. Παρόλα αυτά οι συγγραφείς δείχνουν ότι χρησιμοποιώντας μόνο τις δύο πρώτες πρωτεύουσες συνιστώσες βελτιώνουν σημαντικά την ποιότητα των ομάδων.

Σε αυτή την εργασία προτείνεται ένας εναλλακτικός κανόνας για αυτή την απόφαση εμπνευσμένος από στοιχεία βασισμένα στην πυκνότητα. Για να δούμε αναλυτικότερα την αρχή της προτεινόμενης μεθόδου θα χρησιμοποιήσουμε ένα απλό παράδειγμα όπως στο [14]. Χρησιμοποιούμε ένα δισδιάστατο σύνολο δεδομένων όπως βλέπουμε στην εικόνα 8(α). Αυτό το σύνολο δεδομένων κατασκευάστηκε τεχνητά λαμβάνοντας τυχαία στοιχεία από μία πεπερασμένη μίξη τριών Γκαουσιανών κατανομών με διαφορετικές μέσες τιμές και μοναδιαία διασπορά.

Εφαρμόζοντας την ανάλυση πρωτευουσών συνιστωσών, και προβάλλοντας τις δύο πρωτεύουσες συνιστώσες παίρνουμε τις γραμμές που βλέπουμε στην εικόνα 8(α). Αφού υπολογίζουμε τις προβολές των δεδομένων στην κύρια πρωτεύουσα συνιστώσα, ο αλγόριθμος PDDP κανονικά θα χώριζε τα δεδομένα σύμφωνα με το πρόσημο της αντίστοιχης προβολής για κάθε στοιχείο. Αυτό θα είχε ως αποτέλεσμα η ομάδα που βρίσκεται πιο πάνω από τις άλλες να χωριστεί στη μέση όπως βλέπουμε από την γραμμή ορισμένη ως γραμμή διάσπασης PDDP. Επίσης σημειώνουμε ότι ακόμα και αν χρησιμοποιήσουμε τη δεύτερη πρωτεύουσα συνιστώσα και πάλι δεν έχουμε ικανοποιητικό αποτέλεσμα, παρότι καλύτερο από την αρχική περίπτωση. Αυτό είναι το πρόβλημα που επιλύεται σε αυτή την εργασία.

Για να βρούμε ένα νέο καλύτερο υπερεπίπεδο διάσπασης, εξετάζουμε το ιστόγραμμα των προβολών των δεδομένων. Την γραφική απεικόνιση του ιστογράμματος την βλέπουμε στην εικόνα 8(β). Όπως παρατηρούμε, γύρο από το σημείο 0 που είναι το σημείο διάσπασης του αλγορίθμου PDDP, υπάρχει αρκετά μεγάλη συγκέντρωση προβολών. Ενώ, γύρο από το σημείο 1.6 τα δεδομένα δείχνουν να είναι πολύ λιγότερο συγκεντρωμένα. Αν χωρίσουμε την ομάδα βασισμένοι σε αυτόν τον αριθμό φαίνεται να μειώνουμε την πιθανότητα να διασπάσουμε μία ομάδα. Τη γραμμή διάσπασης που προκύπτει από αυτό το σημείο τη βλέπουμε στην εικόνα 8(α) ως την γραμμή διάσπασης iPDDP.



(α) (β)

Σχήμα 8: (α) Ένα σύνολο δεδομένων με τις πρωτεύουσες συνιστώσες του. (β) Το ιστόγραμμα των προβολών των δεδομένων στην κύρια πρωτεύουσα συνιστώσα.

Συνάρτηση FindCutoff(D) {

1.Για κάθε $d_i \in D$ υπολόγισε τις προβολές p_i,
στην κύρια πρωτεύουσα συνιστώσα u_1

2.Για κάθε p_i, $i = 1, \ldots, n$, Βρες $j = argmin_j\{\|p_i - p_j\|\}$
και $p_j \leq p_i$, και υπολόγισε το $pc_i = \|p_i - p_j\|$

3.Υπολόγισε το $c = \arg\max_i\{pc_i\}$ και $m = \max\{pc_i\}$

4.Επέστρεψε $\{p_c, m\}$

}

Πίνακας 4: Η συνάρτηση FindCutoff(D) για ένα $n \times a$ μητρώο D.

Για να ανακαλύψουμε ένα τέτοιο σημείο στην γενική περίπτωση του μητρώου δεδομένων D χρησιμοποιούμε την συνάρτηση FindCutoff(), που βλέπουμε ως ψευδοκώδικα στον πίνακα 4. Με αυτό τον τρόπο στην πραγματικότητα υπολογίζουμε την πιο αραιή περιοχή των προβολών των δεδομένων, διατάσσοντας αρχικά και στη συνέχεια υπολογίζοντας την μέγιστη απόσταση ανάμεσα σε δύο διαδοχικές προβολές. Έτσι η διάσπαση των δεδομένων πραγματοποιείται βασισμένη στο παρακάτω κριτήριο:

$C_{2,1}$: Υπολόγισε το $\{p_c, m\}$ =**FindCutoff**(D),

$\forall p_i \in D$, Αν $(p_i - p_c) \geq 0$ τότε το i-οστό στοιχείο ανήκει στο πρώτο τμήμα $P_1 = P_1 \cup d_i$,

διαφορετικά ανήκει στο δεύτερο τμήμα $P_2 = P_2 \cup d_i$.

Η μεθοδολογία που έχει περιγραφεί μέχρι στιγμής έχει επίσης ένα μειονέκτημα. Στην περίπτωση όπου το σύνολο δεδομένων έχει πολλά απομακρυσμένα σημεία (σημεία που δεν ανήκουν σε καμία ομάδα), είναι πιθανόν η διαδικασία να αποφασίσει να χωρίσει ομάδες στις εξωτερικές τους περιοχές αφού θα είναι αραιές από την άποψη πυκνότητας. Για παράδειγμα Στην περίπτωση που παρουσιάζεται στην εικόνα 8(β), αυτές οι περιοχές θα είναι οι περιοχές $(-5, -6)$ και $(5, 6)$. Για να διευθετήσουμε αυτό το πρόβλημα είμαστε υποχρεωμένοι να εισάγουμε μία ελεύθερη παράμετρο $MinPts$, στη διαδικασία που καθορίζει το ελάχιστο πλήθος στοιχείων που απαιτείται για να σχηματίζεται μία κανονική ομάδα. Αυτή είναι μια συνηθισμένη διαδικασία για αλγόριθμους που είναι σχεδιασμένοι να αντιμετωπίζουν σύνολα δεδομένων με αρκετό θόρυβο [15].

5.2 Κριτήριο Τερματισμού

Ο αλγόριθμος PDDP καθώς και οι παραλλαγές του που προτείνονται στο [14, 24], τερματίζουν την επαναληπτική διαδικασία διάσπασης των ομάδων όταν το πλήθος των ομάδων που έχουν βρεθεί έχει φτάσει το μέγιστο πλήθος ομάδων που έχει καθοριστεί από τον χρήστη. Στις περισσότερες περιπτώσεις, αυτή είναι μία αποδεκτή τεχνική αφού χρησιμοποιείται ευρέος στην ομαδοποίηση. Παρόλα αυτά, είναι πιθανό να σχεδιαστούν αυτόματα κριτήρια τερματισμού, που θα επιτρέπουν την αυτόματη εύρεση του πλήθους των ομάδων που βρίσκονται στο σύνολο δεδομένων. Αυτό είναι ένα θεμελιώδες πρόβλημα στην ομαδοποίηση, ανεξάρτητα από τη συγκεκριμένη τεχνική ομαδοποίησης που εφαρμόζεται, αυτό το πρόβλημα παραμένει. Οι περισσότερες δημοφιλής τεχνικές όπως των k-μέσων, χρειάζονται μία τεχνική επιλογής μοντέλου, βασισμένη σε μετρήσεις όπως το AIC ή το BIC σε ένα σύνολο διαφορετικών αποτελεσμάτων για να εκτιμήσουν το πλήθος των ομάδων [10]. Προσεγγίσεις συσσωρευτικής ιεραρχικής ομαδοποίησης όπως οι BIRCH [25], CHAMELEON [12] και CURE [8], που παρέχουν εκτίμηση του πλήθους των ομάδων αντιμετωπίζουν εμπόδια όπως η μη γραμμική πολυπλοκότητα χρόνου, και υψηλή συμμετοχή του χρήστη. Οι προσεγγίσεις που είναι βασισμένες στη πυκνότητα παρότι είναι υπολογιστικά αποδοτικές εξαρτώνται και αυτές πολύ από παραμέτρους που καθορίζονται από το χρήστη, για να παρέχουν αποδοτικές εκτιμήσεις.

Τα κριτήρια που χρησιμοποιούνται σε άλλους αλγορίθμους θα μπορούσαν επίσης να χρησιμοποιηθούν στον αλγόριθμο PDDP. Για παράδειγμα στο [13] προτείνεται να χρησιμοποιηθεί το BIC για να καθοριστεί αν η επόμενη διάσπαση θα βελτιώσει το αποτέλεσμα της ομαδοποίησης ή όχι. Αυτό μπορούμε να το επιτύχουμε με τον παρακάτω κανόνα:

$C_{3,1}$: Αν για όλες τις ομάδες P_i, ισχύει ότι $\mathrm{BIC}(P_{i_1}, P_{i_2}) \leq \mathrm{BIC}(P_i)$

τότε τερμάτισε τη διαδικασία.

Όπου P_{i_1}, P_{i_2}, είναι οι παραγόμενες υποομάδες από τη διάσπαση της ομάδας P_i. Για τον υπολογισμό του BIC αναφερόμαστε στο [13]. Επίσης μπορούμε να χρησιμοποιήσουμε στατιστικές του κοντινότερου γείτονα όπως αυτές που χρησιμοποιούνται στο [19, 23].

Για μία ακόμα πιο απλή περίπτωση μπορούμε να ορίσουμε έναν πολύ μεγάλο αριθμό ομάδων ως μέγιστο ή τουλάχιστον μεγαλύτερο από τον πραγματικό. Διαισθητικά, αυτό θα επέτρεπε στον αλγόριθμο αρχικά να βρει τις ομάδες που είναι εύκολα διαχωρίσιμες. Στη συνέχεια, αναμένεται οι υπόλοιπες διασπάσεις να γίνουν σε εξωτερικά στοιχεία στα σύνορα των ομάδων. Με αυτόν τον τρόπο κάθε διάσπαση δεν θα επηρεάζει τον πυρήνα των ομάδων. Στον καθορισμό του τελικού αποτελέσματος της ομαδοποίησης μπορούμε να αγνοήσουμε τις ομάδες που περιέχουν λιγότερα από $MinPts$ στοιχεία. Έτσι έχουμε και το πλεονέκτημα ταυτόχρονα να ανακαλύπτουμε τα εξωτερικά στοιχεία:

$C_{3,2}$: Βρες c_M ομάδες, όπου $C_M > c_{\mathrm{real}}$,

Ανέφερε ως ομάδες αυτές που περιέχουν περισσότερα από $MinPts$ στοιχεία

Όρισε τα στοιχεία που περιέχουν οι υπόλοιπες ομάδες ως εξωτερικά στοιχεία

Σημειώνουμε ότι το c_{real} είναι το πλήθος των πραγματικών ομάδων στο σύνολο δεδομένων.

5.3 iPDDP

Σε αυτό το σημείο στον πίνακα 5 παρουσιάζεται ο ολοκληρωμένος *βελτιωμένος PDDP* (iPDDP) αλγόριθμος, ο οποίος έχει τρία ορίσματα, το μητρώο δεδομένων D, το πλήθος των επιθυμητών ομάδων c_{max} και τη παράμετρο $MinPts$. Ο αλγόριθμος αυτός είναι βασισμένος στα κριτήρια $C_{1,1}$, $C_{2,1}$ και C_3. Αλλάζοντας τα διάφορα διαθέσιμα κριτήρια μπορούμε να σχεδιάσουμε διαφορετικές εκδοχές αυτού του αλγορίθμου. Παρόλα αυτά, τα συγκεκριμένα κριτήρια επιλέχθηκαν γιατί εισάγουν την ελάχιστη παραλλαγή που είναι απαραίτητη για να βελτιωθεί η απόδοση του αρχικού αλγορίθμου, όπως θα δούμε παρακάτω στα πειραματικά αποτελέσματα. Επίσης η υπολογιστική πολυπλοκότητα δεν αυξάνεται σημαντικά.

Η υπολογιστική πολυπλοκότητα του PDDP αλγορίθμου επηρεάζεται περισσότερο από τον υπολογισμό των πρωτεύων διανυσμάτων. Για να τα υπολογίσουμε χρησιμοποιούμε τη διάσπαση ιδιαζουσών τιμών (Singular Value Decomposition) του μητρώου δεδομένων D. Αυτό μας δίνει στην χειρότερη περίπτωση πολυπλοκότητα $\mathrm{O}(c_{max}(2 + k_{SVD})s_{nz}\ n\ a)$, όπου k_{SVD} είναι οι επαναλήψεις που χρειάζονται από τον υπολογιστικό αλγόριθμο Lanczos SVD και s_{nz} είναι το ποσοστό των μη μηδενικών στοιχείων στο D. Για περισσότερες πληροφορίες αναφερόμαστε στο [5]. Στον iPDDP αλγόριθμο τα επιπλέον υπολογιστικά βήματα που χρειάζονται για τη συνάρτηση FindCutoff() αλλάζουν την πολυπλοκότητα σε $\mathrm{O}(c_{max}(2 + k_{SVD})(s_{nz}n\ a + n\log(n)))$, που παρότι αυξάνεται είναι και πάλι συγκρίσιμη με την πολυπλοκότητα των περισσότερων αλγορίθμων ομαδοποίησης. Επίσης επισημαίνουμε ότι το επιπλέον κόστος δεν επηρεάζεται από την διάσταση των δεδομένων. Έτσι, η ικανότητα των αλγορίθμων να αντιμετωπίζουν δεδομένα πολύ υψηλής διάστασης διατηρείται

Function iPDDP $(D, c_{max}, MinPts)$ {
1. Όσο το πλήθος των φύλλων του δέντρου *pddp*
με περισσότερα από $MinPts$ στοιχεία είναι μικρότερο από c_{max}
2. Για κάθε φύλλο P_iτου δέντρου *pddp*
υπολόγισε το $\{p_{c,i}, m_i\}$ =**FindCutoff**(P_i)
3. Διάλεξε το φύλλο P_k του δέντρου *pddp*
με το μεγαλύτερο m_k: $k = \arg\max_i\{m_i)\}$
υπότον περιορισμό $|P_k| > MinPts$
3. Χώρισε το P_k σύμφωνα με το πρόσημο του $(p_j - p_{k,i})$
4. Δημιούργησε τις υποομάδες του P_k και πρόσθεσε τες
στο δέντρο *pddp*
}

Πίνακας 5: Ο τελικός αλγόριθμος iPDDP.

6 Πειραματικά αποτελέσματα

Για να εξετάσουμε την αποδοτικότητα του νέου iPDDP αλγορίθμου χρησιμοποιούμε μία τεχνητή μέθοδος κατασκευής ομάδων που συνήθως χρησιμοποιείται σε εμπειρικές αναλογίες [2, 13, 14]. Έτσι κατασκευάζουμε σύνολα δεδομένων δημιουργώντας στοιχεία από μία πεπερασμένη μίξη από k Γκαουσιανές κατανομές που τοποθετούνται τυχαία στο διάστημα $[100, 200]^d$. Με αυτό τον τρόπο το k αναπαριστά το πραγματικό πλήθος ομάδων, αλλά επίσης είναι πιθανό περισσότερες

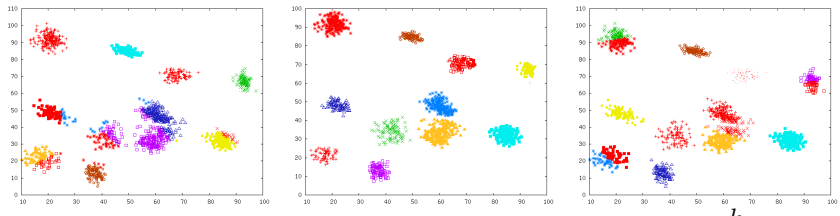
από μία κατανομές να καταλαμβάνουν τον ίδιο χώρο, έτσι ώστε να μην μπορούν να διακριθούν στην διαδικασία της ομαδοποίησης. Το μητρώο συνδιασποράς κάθε κατανομής επίσης δημιουργείται τυχαία με μία κατάλληλη διαδικασία έτσι ώστε να διασφαλιστεί ότι είναι συμμετρικό και θετικά ορισμένο. Στα παρακάτω πειράματα δημιουργήθηκαν 200 στοιχεία από κάθε κατανομή και έτσι το συνολικό πλήθος στοιχείων είναι $k \times 200$.

Οι αλγόριθμοι με τους οποίους συγκρίνεται η νέα μέθοδος είναι ο PDDP ως ο αρχικός αλγόριθμος που βελτιώνεται με τη νέα τεχνική και ο k-means ως ένας πολύ δημοφιλής αλγόριθμος. Για να μπορέσουμε να μετρήσουμε την αποδοτικότητα των αλγορίθμων χρησιμοποιούμε ένα μέτρο που αντιπροσωπεύει την μέση διασπορά των k ομάδων που ορίζεται ως:

$$S = \frac{1}{k} \sum_{i=1}^{k} \|P_i\|$$

όπου $\|P_i\|$ αναπαριστά τη Frobenius νόρμα της ομάδας P_i, για $i = 1, \ldots, k$. Για διάφορες τιμές όσο αφορά το πλήθος των πραγματικών ομάδων αλλά και τη διάσταση του συνόλου δεδομένων, στο πίνακα 6 βλέπουμε τα αποτελέσματα των τριών αλγορίθμων σύμφωνα με τη μέση τιμή του S σε 100 διαφορετικά πειράματα. Επίσης σημειώνουμε ότι και στους τρεις αλγορίθμους δόθηκε ως είσοδος το πραγματικό πλήθος των ομάδων και η παράμετρος $MinPts$ του iPDDP είχε την τιμή 2. Παρόμοια αποτελέσματα έχουμε όταν οι τιμές της παραμέτρου είναι στο διάστημα $[1, 10]$.

Όπως βλέπουμε ο αλγόριθμος iPDDP πάντα έχει καλύτερα αποτελέσματα από τον PDDP. Όπως αναμενόταν ακόμα και για μικρό πλήθος ομάδων και σε χαμηλή διαστατικότητα, η μεθοδολογία διάσπασης του iPDDP αλγορίθμου βελτιώνει τα αποτελέσματα. Σε σύγκριση με τον αλγόριθμο k-means, παρατηρούμε ότι σε πειράματα μέτριας δυσκολίας (μικρό πλήθος ομάδων, λίγες διαστάσεις) η αποδοτικότητα είναι περίπου ίδια. Αυτό οφείλεται στο γεγονός ότι και οι δύο αλγόριθμοι δημιουργούν όμοιες ομάδες. Όταν το πλήθος των ομάδων που περιέχει το σύνολο δεδομένων αυξάνεται και αυξάνεται και η δυσκολία του προβλήματος, ο αλγόριθμος iPDDP δείχνει τη δύναμη του πετυχαίνοντας μικρότερες τιμές του S έναντι και του k-means και του PDDP.



PDDP iPDDP k-means

Σχήμα 9: (α) Παραδειγματικό σύνολο δεδομένων με τα αποτελέσματα των αλγορίθμων PDDP, iPDDP και k-means αντίστοιχα.

Για να καταλάβουμε καλύτερα την απόδοση των αλγορίθμων παραθέτουμε μία γραφική απεικόνιση μίας δισδιάστατης περίπτωσης όπου περιέχονται 15 ομάδες στην εικόνα 9. Βλέπουμε τα αποτέλεσμα της ομαδοποίησης για του αλγορίθμους PDDP, iPDDP και k-means από αριστερά προς τα δεξιά αντίστοιχα, όπου οι διαφορετικές ομάδες που έχουν βρει οι αλγόριθμοι είναι σχηματισμένες με διαφορετικό

Διάσταση 2			
Πλήθος Ομάδων	PDDP	k-means	iPDDP
5	161.896	64.977	63.681
10	206.147	135.534	132.620
15	348.557	197.929	200.923
25	634.576	292.460	284.642
Διάσταση 5			
Πλήθος Ομάδων	PDDP	k-means	iPDDP
5	435.195	139.731	139.731
10	755.875	363.112	269.942
15	1566.961	449.348	392.206
25	2456.778	747.340	683.968
Διάσταση 10			
Πλήθος Ομάδων	PDDP	k-means	iPDDP
5	587.177	285.957	241.830
10	2020.161	585.123	486.649
15	2777.603	796.393	732.911
25	6241.165	1468.147	1211.124

Πίνακας 6: Αποτελέσματα σύμφωνα με τη διασπορά S των τελικών ομαδοποιήσεων για διάφορες μεθόδους.

χρώμα και τύπο σημείων. Παρότι συνολικά υπάρχουν 15 Γκαουσιανές συνιστώσες στο σύνολο των δεδομένων το πραγματικό αποτέλεσμα της ομαδοποίησης δεν θα περιέχει περισσότερες από 11 ομάδες. Αυτό οφείλεται στο γεγονός ότι μερικές συνιστώσες βρίσκονται στο ίδιο σημείο έτσι δεν διακρίνονται η μία από την άλλη. Παρατηρώντας το αποτέλεσμα της ομαδοποίησης βλέπουμε ότι ο αλγόριθμος PDDP αναγνωρίζει μερικές ομάδες πολύ καλά, αλλά αποτυχαίνει ιδιαίτερα στη κεντρική περιοχή της εικόνας. Από την άλλη μεριά ο αλγόριθμος k-means έχει αρκετά καλύτερο αποτέλεσμα από τον PDDP αλλά μη κατάλληλο πλήθος ομάδων σε αυτή τη περίπτωση ωθεί τον αλγόριθμο να χωρίσει κανονικές ομάδες. Σε διαφορετική περίπτωση πιθανόν θα ένωνε ομάδες. Παρόλα αυτά, ο iPDDP αλγόριθμος αφού εντοπίσει τις 11 ομάδες, ξεκινάει χωρίζει ομάδες που έχουν λιγότερα από $MinPts$ στοιχεία και δεν τις βλέπουμε. Αυτή η παρατήρηση μας οδηγεί στο συμπέρασμα ότι ο αλγόριθμος θα μπορούσε να χρησιμοποιηθεί στον αυτόματο καθορισμό του πλήθους των ομάδων.

Τα παραπάνω πειράματα μας παρέχουν μία ανάλυση για την απόδοση των αλγορίθμων σε τυχαίες περιπτώσεις. Παρόλα αυτά για να ερευνήσουμε τα αποτελέσματα σε σχέση με τη βιβλιογραφία της ομαδοποίησης δεδομένων, και άλλες παραλλαγές του PDDP αλγορίθμου [13, 14] χρησιμοποιούμε το πολύ δημοφιλές σύνολο δεδομένων Iris, $\mathrm{DS}_{\mathrm{IRIS}}$, από το UCI αρχείο εκμάθησης μηχανής [4]. Το σύνολο δεδομένων αποτελείται από 150 αντικείμενα τεσσάρων χαρακτηριστικών, κανονικοποιημένα στο διάστημα $[10, 100]$, και κατηγοριοποιημένα σε τρεις ομάδες που ονομάζονται Setosa, Versicolour και Virginica. Επίσης χρησιμοποιούμε τους

	κλάση Iris		
Ομάδες	Setosa	Versicolour	Virginica
Ομάδα 1	- - 50 - -	- 50 - - 47	43 50 - 36 4
Ομάδα 2	- - - - 48	40 - 46 48 -	7 - 4 14 -
Ομάδα 3	3 - - 50 -	10 - 4 - 2	0 - 46 - 46
Ομάδα 4	47 - - - -	- - - - -	0 - - - -
	Outliers - - - - 2		

Πίνακας 7: Μητρώα συχέτισης για DS_{IRIS}: Το πρώτο, δεύτερο, τρίτο, τέταρτο και πέμπτο στοιχείο κάθε κελιού αντιστοιχεί στους PDDP, DBSCAN, UKW, k-means, iPDDP αντίστοιχα.

αλγορίθμους DBSCAN [15], και UKW [17] ως αντιπρωσοπευτικούς από τη βασισμένη στην πυκνότητα πλευρά της ομαδοποίησης. Η αποδοτικότητα των αλγορίθμων αναφέρεται σύμφωνα με τα μητρώα συσχέτισης, όπως βλέπουμε στον πίνακα 7. Σε αυτή την περίπτωση παρότι ενδιαφερόμαστε για ένα αποτέλεσμα με 3 ομάδες, επιδιώκουμε ένα αποτέλεσμα με 4 ομάδες από τον αλγόριθμο PDDP καθώς το αποτέλεσμα του διαφορετικά δεν είναι καθόλου ικανοποιητικό. Στην περίπτωση των 4 ομάδων ο PDDP αναγνωρίζει τις 3 βασικές κλάσεις ως ομάδες αλλά κατηγοριοποιεί λάθος συνολικά 13 στοιχεία, έχοντας επίσης μία μικρή ομάδα με 10 αντικείμενα από την κλάση Versicolour και 3 από την Setosa. Ο αλγόριθμος DBSCAN αποτυχαίνει τελείως να χωρίσει τις κλάσεις Versicolour και Virginica αλλά αυτό είναι κατανοητό καθώς υποφέρει από το φαινόμενο της αλυσίδας [17]. Από την άλλη μεριά ο UKW παρέχει πολύ καλό αποτέλεσμα με τις τρεις κλάσεις καλά χωρισμένες και συνολικά 8 λάθος κατηγοριοποιημένα στοιχεία. Τα αποτελέσματα των αλγορίθμων DBSCAN και UKW μπορούν να βελτιωθούν αν χρησιμοποιηθούν εκδοχές τους που βοηθιούνται από την τεχνική PCA [18]. Εν τέλει ο αλγόριθμος iPDDP κατηγοριοποιεί λάθος συνολικά 6 στοιχεία και 2 επιπλέον στοιχεία αναφέρονται ως απομακρυσμένα, τα οποία ακόμα και αν μετρηθούν ως λάθος, έχουμε ένα πολύ καλο αποτέλεσμα.

6.1 Πραγματική Περίπτωση Ι: Δεδομένα έκφρασης γονιδίων

Σε κάθε ζωντανό οργανισμό που υπόκειται μία βιολογική διαδικασία, εκφράζονται διαφορετικά υποσύνολα των γονιδίων του. Η κανονική λειτουργία ενός κυττάρου, επηρεάζεται σημαντικά, από τη γονιδιακή του έκφραση σε ένα δεδομένο στάδιο. Για να γίνουν κατανοητές οι βιολογικές διαδικασίες, θα πρέπει να μετρηθούν, τα επίπεδα έκφρασης 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 [26]

Συγκρινόμενοι με τις παραδοσιακές προσεγγίσεις στη γονιδιακή έρευνα, που βασίζονται στη συλλογή και εξέταση δεδομένων για ένα μόνο γονίδιο τοπικά, οι τεχνολογίες μικροστοιχειών DNA έχουν δώσει τη δυνατότητα της παρακολούθησης των εκφράσεων χιλιάδων γονιδίων ταυτόχρονα. Δυστυχώς, η πραγματική γονιδια-

κή έκφραση 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 neghbors, μέσοι όροι γραμμών, επανάληψη των πειραμάτων για τη μοντελοποίηση του θορύβου, την κανονικοποίησης, 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 μικροστοιχειών, είναι μία πολύ μεγάλη πρόκληση για τη Γονιδιακή και Πρωτεϊνική έρευνα [26]. 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 [27][28], είτε για να ομαδοποιήσει δείγματα σε ομοιόμορφες ομάδες που αντιστοιχούν σε συγκεκριμένους μακροσκοπικούς φαινότυπους [29]. Ο δεύτερος ρόλος της ομαδοποίησης φαίνεται να είναι και ο πιό δύσκολος εξαιτίας της *κατάρας της διάστασης*(Curse of Dimensionality) [30](μικρός αριθμός δειγμάτων με πολύ μεγάλη διάσταση).

Η περίπτωση των δεδομένων γονιδιακής έκφρασης, είναι ένα τυπικό παράδειγμα δεδομένων υψηλής διάστασης. Πρόσφατες εξελίξεις στις τεχνολογίες μικροστοιχειών επιτρέπουν στους επιστήμονες να ανακαλύψουν, να φωτογραφίσουν και να μετρήσουν γονιδιακές εκφράσεις πολλών χιλιάδων γονιδίων σε ένα απλό πείραμα. Επίσης, σε ένα τυπικό βιολογικό σύστημα, συχνά δεν είναι γνωστό πόσα γονίδια είναι αρκετά ώστε να χαρακτηριστεί ένα μακροσκοπικό φαινότυπο. Πρακτικά, μία μηχανιστική υπόθεση που λειτουργεί και μπορεί να δοκιμαστεί και βρίσκει σε μεγάλο βαθμό τη βιολογική αλήθεια σπάνια περιλαμβάνει περισσότερα από μερικές δεκάδες γονίδια, και η γνώση της ταυτότητας αυτών των σχετικών γονιδίων είναι πολύ σημαντική [21]. Τεχνικές ομαδοποίησης έχουν εφαρμοστεί σε δεδομένα έκφρασης γονιδίων [16, 21] και έχουν αποδειχτεί χρήσιμες για την αναγνώριση βιολογικά σχετικών ομάδων γονιδίων και δειγμάτων. Με αυτό τον τρόπο οι τεχνικές ομαδοποίησης έχουν βοηθήσει περαιτέρω στο να διευθετηθούν ερωτήματα όπως η λειτουργικότητα των γονιδίων, η κανονικότητα των γονιδίων και η διαφοροποίηση των γονιδιακών εκφράσεων υπό διάφορες συνθήκες.

Σε αυτό το σημείο για να εξετάσουμε την απόδοση του iPDDP αλγορίθμου σε τέτοιου τύπου δεδομένα χρησιμοποιούμε το σύνολο δεδομένων COLON [1] που περιέχει 40 καρκινικούς και 22 φυσιολογικούς ιστούς colon. Για κάθε δείγμα ιστού υπάρχουν 2000 μετρήσεις γονιδιακών εκφράσεων. Το σύνολο δεδομένων είναι διαθέσιμο στη διεύθυνση http://microarray.princeton.edu/oncology. Οι αλγόριθμοι PDDP και iPDDP μπορούν να εφαρμοστούν απευθείας σε αυτό το σύνολο δεδομένων χωρίς να είναι απαραίτητη κάποια προεργασία που συνήθως χρειάζεται για άλλους αλγορίθμους [16]. Τα πειράματα έγιναν για πολλές τιμές τελικού πλήθους ομάδων, και παρουσιάζονται τα καλύτερα αποτελέσματα που βρέθηκαν στον πίνακα 8. Όπως παρατηρούμε, αν αναθέσουμε ταμπέλες σε κάθε ομάδα, σύμφωνα με

Ομάδες	PDDP		iPDDP	
	τύπος ιστού		τύπος ιστού	
	Κανονικός	Καρκινικός	Κανονικός	Καρκινικός
Ομάδα 1	0	1	6	27
Ομάδα 2	2	0	6	0
Ομάδα 3	2	3	5	2
Ομάδα 4	6	2	4	2
Ομάδα 5	5	13	1	4
Ομάδα 6	7	21	0	5

Πίνακας 8: Τα αποτελέσματα των αλγορίθμων PDDP και iPDDP για το σύνολο δεδομένων COLON.

την κλάση που έχει τα περισσότερα στοιχεία στην κάθε ομάδα μπορούμε να δούμε ότι ο αλγόριθμος PDDP συνολικά τοποθετεί λάθος 16 στοιχεία, 2 από την κλάση των καρκινικών και 14 από την κλάση των κανονικών. Αν λάβουμε υπόψη ότι συνολικά έχουμε 22 στοιχεία κανονικού τύπου τότε περισσότερα από τα μισά στοιχεία αυτής της κλάσης τοποθετούνται λάθος. Από την άλλη μεριά ο αλγόριθμος iPDDP συνολικά τοποθετεί λάθος 11 στοιχεία, 4 από τη καρκινική κλάση και 7 από την κανονική. Αυτό το αποτέλεσμα είναι αρκετά πιο ζυγισμένο όσο αφορά το σφάλμα για την κάθε κλάση και δείχνει την δυνατότητα του iPDDP αλγορίθμου να αντιμετωπίζει ανομοιόμορφα σύνολα δεδομένων.

6.2 Πραγματική Περίπτωση II: Ομαδοποίηση εγγράφων

Η ομαδοποίηση εγγράφων είναι ένα θεμελιώδης και εύχρηστο εργαλείο για την αποδοτική οργάνωση, αναζήτηση και ανάκτηση των εγγράφων. Η ιεραρχική ομαδοποίηση εγγράφων οργανώνει τις ομάδες σε ένα δέντρο ή μία ιεραρχία ώστε να εξυπηρετείται η αναζήτηση. Τη σχέση γονέα παιδιού ανάμεσα στους κόμβους του δέντρου μπορούμε να τη δούμε ως τη σχέση κεφελίδα-υποκεφαλίδα σε μία θεματολογική ιεραρχία όπως το αρχείο Yahoo. Η ομαδοποίηση εγγράφων είναι ιδιαίτερα εφαρμόσιμη σε περιοχές όπως οι μηχανές αναζήτησης, η εξόρυξη δικτύου, η ανάκτηση πληροφορίας και η τοπολογική ανάλυση. Οι περισσότερες μέθοδοι ομαδοποίησης εγγράφων διενεργούν πολλά βήματα προεργασίας. Κάθε έγγραφο αναπαριστάται από ένα διάνυσμα συχνοτήτων εμφάνισης της κάθε λέξης μέσα στο έγγραφο. Παρότι πολλοί αλγόριθμοι ομαδοποίησης εγγράφων έχουν προταθεί στη βιβλιογραφία, οι περισσότεροι δεν ικανοποιούν τις ειδικές απαιτήσεις της ομαδοποίησης εγγράφων:

- Υψηλή διαστατικότητα. Το πλήθος των όρων σε ένα σύνολο εγγράφων είναι συνήθως μερικές χιλιάδες αν όχι δεκάδες χιλιάδες. Κάθε ένας από αυτούς του όρους δημιουργεί μία διάσταση στο διάνυσμα του εγγράφου. Οι φυσικές ομάδες συνήθως δεν υπάρχουν στο πλήρους διάστασης χώρο, αλλά σε ένα υπόχωρο που σχηματίζεται από ένα σύνολο συσχετισμένων διαστάσεων. Το να βρούμε ομάδες σε αυτό τον υπόχωρο είναι μία πρόκληση.
- Εξελιξιμότητα. Τα πραγματικά σύνολα δεδομένων μπορεί να περιέχουν εκατοντάδες χιλιάδες έγγραφα. Πολλοί αλγόριθμοι δουλεύουν καλά σε μικρά σύνολα δεδομένων αλλά αποτυγχάνουν να χειριστούν μεγάλα σύνολα δεδομένων αποδοτικά

- Ακρίβεια. Έγγραφα που βρίσκονται στην ίδια ομάδα θα πρέπει να είναι όμοια μεταξύ τους αλλά ανόμοια με τα έγγραφα των άλλων ομάδων
- Εύκολη αναζήτηση με κατανοητή περιγραφή ομάδων. Η ιεραρχία των επικεφαλίδων που παίρνουμε ως αποτέλεσμα θα πρέπει να παρέχει μία λογική δομή, μαζί με κατανοητή περιγραφή ομάδων, για να υποστηρίξουν την αναζήτηση.
- Προγενέστερη γνώση. Πολλοί αλγόριθμοι ομαδοποίησης απαιτούν από το χρήστη να καθορίσει κάποιες παραμέτρους εισόδου, για παράδειγμα, το πλήθος των ομάδων. Παρόλα αυτά ο χρήστης συχνά δεν μπορεί να έχει αυτή τη προγενέστερη γνώση. Η ακρίβεια της ομαδοποίησης μπορεί να μειώνεται δραματικά αν ο αλγόριθμος είναι πολύ ευαίσθητος σε τέτοιου τύπου παραμέτρους.

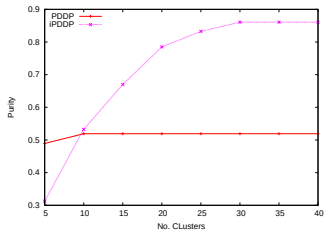
Το πιό σοβαρό πρόβλημα της ομαδοποίησης αρχείων είναι η υψηλή διαστατικότητα του κειμένου της φυσικής γλώσσας, η οποία κάνει τους αλγόριθμους ομαδοποίησης που χρησιμοποιούν προβολές όπως ο PDDP ιδανικούς για αυτή τη περίπτωση [5]. Σε αυτή την ενότητα θα δούμε τις ικανότητες του αλγορίθμου iPDDP σε σχέση με τον αρχικό αλγόριθμο και όχι μία πλήρη ανάλυση της έρευνας του text mining. Σε αυτό το σημείο χρησιμοποιούμε ένα μέρος της συλλογής Reuters-21578 η οποία συνολικά αποτελείται από περισσότερα από 10,000 έγγραφα κατηγοριοποιημένα σε διάφορα θέματα. Μόνο ένα υποσύνολο αυτής της συλλογής χρησιμοποιήθηκε που περιέχει 1200 αρχεία με 3562 συνολικά όρους (λέξεις), οργανωμένα σε 30 διαφορετικά θέματα.

Για να μετρήσουμε την ποιότητα της ομαδοποίησης, επιλέγουμε να χρησιμοποιήσουμε την αγνότητα(Purity) όπως ορίζεται στο [11, 22], $P = \frac{\sum_{i=1}^{k} |D_i|}{|N|}$, όπου $|D_i|$ είναι το πλήθος των στοιχείων με την ταμπέλα(label) της κύριας κλάσης στην ομάδα i, $|N|$ το συνολικό πλήθος των στοιχείων, και k, το πλήθος των ομάδων. Διαισθητικά, έτσι μετράμε την αγνότητα των ομάδων σύμφωνα με τις ταμπέλες των πραγματικών ομάδων που είναι γνωστές για αυτό το σύνολο δεδομένων, έτσι μπορεί να θεωρηθεί ως μία μορφή της εκδοχής της ομαδοποίησης για την ακρίβεια της τμηματοποίησης. Στην εικόνα 10, μπορούμε να δούμε την μέτρηση (Purity) του αποτελέσματος της ομαδοποίησης που πήραμε για διάφορες τιμές του τελικού πλήθους των ομάδων για τους αλγορίθμους PDDP και iPDDP. Όπως παρατηρούμε ο αλγόριθμος PDDP ξεκινάει με καλά αποτελέσματα στην περίπτωση των 5 ομάδων, καλύτερα ακόμα και από τον iPDDP. Αυτό συμβαίνει γιατί μέχρι εκείνο το σημείο ο αλγόριθμος iPDDP βρίσκει ομάδες με λιγότερα από MinPts στοιχεία και αφήνει το κύριο όγκο των αρχείων ανέπαφο, με αποτέλεσμα να έχουμε μία πολύ μεγάλη ομάδα που περιέχει την πλειονότητα των στοιχείων. Παρόλα αυτά καθώς το πλήθος των τελικών ομάδων αυξάνεται, ο αλγόριθμος iPDDP δείχνει τις δυνατότητες του χωρίζοντας αποδοτικά την μεγάλη ομάδα με αποτέλεσμα να έχουμε μία ραγδαία άνοδο του purity.

6.3 Αυτόματος καθορισμός του πλήθους των ομάδων

Όπως αναφέρθηκε στην ενότητα 5.2, μπορούμε να σχεδιάσουμε κριτήρια που προσπαθούν να προσεγγίσουν το πλήθος των ομάδων στο σύνολο δεδομένων. Ενώ μπορούμε να χρησιμοποιήσουμε διάφορα τέτοια κριτήρια επιλέγουμε το πιο απλό που είναι το $C_{3,1}$. Σχεδιάζουμε αυτόν τον αλγόριθμο ως iPDDP*.

Για να μετρήσουμε την αποδοτικότητα του iPDDP*, θα χρησιμοποιήσουμε μία παρόμοια πειραματική διαδικασία με αυτή που περιγράφτηκε προηγουμένως.



Σχήμα 10: Τα αποτελέσματα των αλγορίθμων PDDP και iPDDP για την ομαδοποίηση εγγράφων.

Τα σύνολα δεδομένων κατασκευάζονται τεχνητά λαμβάνοντας στοιχεία από μία πεπερασμένη μίξη από k Γκαουσιανές κατανομές στο διάστημα $[100, 200]^d$. Πάλι, το μητρώο συνδιασποράς κάθε κατανομής επίσης δημιουργείται τυχαία από μία κατάλληλη διαδικασία, έτσι ώστε να εξασφαλίζεται ότι είναι συμμετρικό και θετικά ορισμένο, και συνολικά κάθε σύνολο δεδομένων περιέχει $k \times 200$ στοιχεία αφού 200 στοιχεία λαμβάνονται από κάθε κατανομή. Για να μετρήσουμε την αποδοτικότητα της ομαδοποίησης και πάλι θα χρησιμοποιήσουμε το μέτρο Purity P, που ορίσαμε παραπάνω. Σημειώνουμε ότι αυτό το μέτρο σύγκρισης ευνοεί τον αλγόριθμο UKW αφού δεν λαμβάνει υπόψη τη διασπορά των ομάδων.

Στον πίνακα 9, αναφέρονται τα αποτελέσματα για τους αλγορίθμους UKW και iPDDP*, για τη μέση τιμή της μέτρησης P σε 100 τρεξίματα, και την αντίστοιχη μέση τιμή του πλήθους των ομάδων που βρέθηκαν. Αρχικά θέτουμε την παράμετρο $MinPts$ του αλγορίθμου iPDDP* με τιμή 6 και το κριτήριο τερματισμού τέσσερις φορές το πραγματικό πλήθος των ομάδων. Για τον αλγόριθμο UKW θέτουμε το αρχικό μέγεθος του παραθύρου να είναι 6, αφού με αυτή την τιμή έχουμε τα καλύτερα αποτελέσματα. Για την περίπτωση των 5 διαστάσεων και οι δύο αλγόριθμοι είναι ικανοί να αναγνωρίσουν με ακρίβεια το πραγματικό πλήθος των ομάδων πετυχαίνοντας ομαδοποίηση υψηλού Purity P, ανεξάρτητα από το πλήθος των ομάδων στο σύνολο δεδομένων. Αν παρόλα αυτά αλλάξουμε την διάσταση σε 20 τα αποτελέσματα αλλάζουν μόνο για τον αλγόριθμο UKW. Ο αλγόριθμος iPDDP* διατηρεί ακριβώς την ίδια απόδοση, κάτι που μας δείχνει την ικανότητα του αλγορίθμου να λειτουργεί αποδοτικά ανεξάρτητα της διάστασης των δεδομένων.

Από την άλλη, παρότι ο αλγόριθμος UKW παράγει καλά αποτελέσματα, βρίσκει συνέχεια περισσότερες από τις πραγματικές ομάδες. Παρόλα αυτά, αν αλλάξουμε το αρχικό μέγεθος του παραθύρου σε 8 παίρνουμε πιό ακριβή αποτελέσματα, όπως μπορούμε να δούμε από τους τιμές στις παρενθέσεις στην τελευταία στήλη του πίνακα 9. Τ ίδιο συμβαίνει και στην περίπτωση του αλγορίθμου DBSCAN και για αυτό το λόγο αυτά τα αποτελέσματα δεν περιλαμβάνονται.

7 Συμπεράσματα

Πολλές μέθοδοι έχουν προταθεί για την ομαδοποίηση δεδομένων και ανάμεσα τους ο αλγόριθμος PDDP [5] είναι μία επιτυχημένη τεχνική, ειδικά για εφαρμογές text mining.

Σε αυτή την εργασία ερευνώνται τα μειονεκτήματα του αλγορίθμου PDDP και

Διάσταση 5				
Πλήθος Ομάδων	iPDDP*		UKW	
	P	Ομάδες	P	Ομάδες
10	0.9791	10.2	0.9999	12
20	0.9769	20.3	0.9949	21
30	0.9753	30	0.9966	30.6
Διάσταση 20				
10	0.9808	10.1	1.0	27.4 (14)
20	0.9774	20.1	1.0	34.9 (22.7)
30	0.9777	30	1.0	44.6 (33.5)

Πίνακας 9: Αποτελέσματα στον αυτόματο καθορισμό του πλήθους των ομάδων.

προτείνεται μία βελτιωμένη μέθοδος που είναι ιδιαίτερα αποδοτική, όπως είδαμε και από τα πειραματικά αποτελέσματα, και συγκρίνεται καλά έναντι σε άλλες δημοφιλής μεθόδους. Επίσης η νέα μέθοδος δείχνει να έχει τη δυνατότητα να αντιμετωπίσει το πολύ σημαντικό και ανοιχτό ζήτημα στην ομαδοποίηση που είναι ο προσδιορισμός του πλήθους των ομάδων. Επιπλέον πειράματα μας δείχνουν την ικανότητα του αλγορίθμου να μας παρέχει εκτιμήσεις του πλήθους των ομάδων για διάφορες διαστάσεις και μεγέθη συνόλου δεδομένων χωρίς την παρέμβαση του χρήστη.

Πιθανές μελλοντικές κατευθύνσεις περιέχουν την χρήση διαφορετικών μηχανισμών προβολής των δεδομένων, και η ενσωμάτωση μηχανισμών επιλογής μοντέλου ως βοήθεια στον αυτόματο καθορισμό του πλήθους των ομάδων. Επίσης περισσότερη μελέτη χρειάζεται για την περίπτωση των ομάδων με μη κυρτό σχήμα που δεν περιλαμβάνεται σε αυτή την εργασία.

Αναφορές

[1] 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*, 96(12):6745–6750, 1999.

[2] P. Berkhin. A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data: Recent Advances in Clustering*, pages 25–72. Springer, Berlin, 2006.

[3] I. T. Jolliffe Principal Component Analysis New York: Spriger Verlag, 2002.

[4] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[5] D. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344, 1998.

[6] C.G. Chute and Y. Yang. An overview of statistical methods for the classification and retrieval of patient events. *Methods Inf Med*, 34(1-2):104–10, 1995.

[7] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[8] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient algorithm for clustering large databases. In *Proceedings of ACM-SIGMOD 1998 International Conference on Management of Data*, pages 73–84. Seattle, 1998.

[9] J.A. Hartigan and M.A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[10] T. Ishioka. Extended K-means with an Efficient Estimation of the Number of Clusters. In *Second International Conference on Intelligent Data Engineering and Automated Learning-IDEAL 2000*. Springer, 2000.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[12] G. Karypis, E.H. Han, and V. Kumar. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.

[13] C. Kruengkrai, V. Sornlertlamvanich, and H. Isahara. Refining A Divisive Partitioning Algorithm for Unsupervised Clustering. *Proceedings of the 3rd International Conference on Hybrid Intelligent Systems*, pages 535–542, 2003.

[14] M. Nilsson. Hierarchical Clustering Using Non-Greedy Principal Direction Divisive Partitioning. *Information Retrieval*, 5(4):311–321, 2002.

[15] 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*, 2(2):169–194, 1998.

[16] D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis. Unsupervised clustering in mRNA expression profiles. *Computers in Biology and Medicine*, 36:1126–1142, 2006.

[17] D.K. Tasoulis and M.N. Vrahatis. Novel approaches to unsupervised clustering through the k-windows algorithm. In S. Sirmakessis, editor, *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, pages 51–78. Springer-Verlag, 2005.

[18] D.K. Tasoulis, D. Zeimpekis, E. Gallopoulos, and M.N. Vrahatis. Oriented k-windows: A pca driven clustering method. In *Advances in Web Intelligence and Data Mining*, Studies in Computational Intelligence, pages 319–329. 2006.

[19] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 2006.

[20] C. Tryon. *Cluster Analysis*. Ann Arbor, MI: Edward Brothers, 1939.

[21] E.P. Xing and R.M. Karp. CLIFF: Clustering of high–dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics Discovery Note*, 1:1–9, 2001.

[22] Zhao Y. and Karypis G. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.

[23] D.L. Young. The linear nearest neighbour statistic. *Biometrica*, 69(2):477–480, 1982.

[24] D. Zeimpekis and E. Gallopoulos. PDDP(l): Towards a Flexing Principal Direction Divisive Partitioning Clustering Algorithms. In D. Boley, I. Dhillon, J. Ghosh, and J. Kogan, editors, *Proc. IEEE ICDM '03 Workshop on Clustering Large Data Sets*, pages 26–35, Melbourne, Florida, 2003.

[25] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. pages 103–114, 1996.

[26] D. Jiang, C. Tang, and A. Zhangi. Clusters analysis for gene expression data: A survey IEEE transactions on Knowledge and data Engineering, vol.16, No.11, 1370-1376

[27] M.B. Esien, P.T. Spellman, P.O. Brown, and D. Bostein Clusters analysis and display of genome-wide expression patterns Proc.Natl.Acad.Sci.USA, vol. 95, 14863-14868, 1998

[28] 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

[29] T.R. Golub, D.K Slomin, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, 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

[30] R. Bellman Adaptive Control Processes: A Guided Tour Princeton University Press, 1961

[31] Juyang Weng, Member, IEEE, Yilu Zhang, Student Member, IEEE, and Wey-Shiuan Hwang, Member, IEEE. Candid Covariance-Free Incremental Principal Component Analysis. IEEE transactions on pattern analysis and machine intelligence, VOL. 25, NO. 8, AUGUST 2003.

[32] Lindsay I Smith. A tutorial on Principal Components Analysis.

[33] Benjamin C. M. Fung, Ke Wang, and Martin Ester, Simon Fraser University, Canada. Hierarchical Document Clustering.

[34] Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota Vipin Kumar, University of Minnesota. Introduction to Data Mining. Chapter 8. Cluster Analysis: Basic Concepts and Algorithms