ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ ΒΑΣΙΣΜΕΝΕΣ ΣΕ ΣΤΑΤΙΣΤΙΚΟ ΕΛΕΓΧΟ ΤΗΣ ΜΟΝΟΤΡΟΠΙΚΟΤΗΤΑΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Μηχανικών Η/Υ & Πληροφορικής
Εξεταστική Επιτροπή

από τον

ΘΕΟΦΙΛΟ ΧΑΜΑΛΗ

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΟ ΛΟΓΙΣΜΙΚΟ

Οκτώβριος 2015

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά όλους όσους με βοήθησαν και με στήριξαν με οποιοδήποτε τρόπο όλο αυτό το διάστημα ώστε να καταστεί δυνατή η εκπόνησης της μεταπτυχιακής μου εργασίας. Αρχικά, τον επιβλέποντα καθηγητή μου κ. Αριστείδη Λύκα για τη πολύτιμη βοήθεια και την καθοδήγηση που μου παρείχε, καθώς και για τις γνώσεις και την εμπειρία που μου μετέδωσε για τον κόσμο της εξόρυξης γνώσης από δεδομένα.

Επίσης, την οικογένειά μου, τους γονείς μου, την αδερφή μου και την κοπέλα μου για τη στήριξη και την υπομονή τους καθ' όλη τη διάρκεια των σπουδών και της εργασίας μου.

Τέλος, θέλω να ευχαριστήσω τους φίλους μου όπως επίσης και τους Ιωσήφ Πολενάκη, Αργύρη Καλογεράτο, Παναγιώτη Ζαγορίσιο και Γρηγόρη Τζώρτζη για τη βοήθεια και τις συμβουλές τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Θεόφιλος Χαμάλης του Θεοδώρου και της Γαλήνης.
MSc, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Οκτώβριος 2015.
Τίτλος: Μέθοδοι Ομαδοποίησης βασισμένες σε στατιστικό έλεγχο της μονοτροπικότητας των δεδομένων.
Επιβλέπων: Αριστείδης Λύκας.

Η ομαδοποίηση αποτελεί ένα από τα σημαντικότερα αντικείμενα της μηχανικής μάθησης και της εξόρυξης δεδομένων λόγω του πλήθους εφαρμογών της σε προβλήματα ανάλυσης δεδομένων. Ένα βασικό ζήτημα κατά την ομαδοποίηση ενός συνόλου δεδομένων σχετίζεται με την εκτίμηση του αριθμού των ομάδων, ο οποίος συνήθως δεν είναι γνωστός εκ των προτέρων. Μια τεχνική που έχει προταθεί για το πρόβλημα αυτό είναι ο αλγόριθμος dip-means, ο οποίος προτείνει μια μεθοδολογία (κριτήριο) για τον στατιστικό έλεγχο της μονοτροπικότητας ενός συνόλου δεδομένων και χρησιμοποιεί τη μεθοδολογία αυτή για ομαδοποίηση με αυξητικό τρόπο: στην αρχή όλα τα δεδομένα ανήκουν στην ίδια ομάδα και σε κάθε βήμα διασπώνται οι ομάδες που δεν είναι μονοτροπικές σύμφωνα με το κριτήριο.

Στην εργασία αυτή καταρχήν προτείνεται μια παραλλαγή του αλγορίθμου dip-means η οποία ονομάζεται pdip-means (projected dip-means) και η οποία τροποποιεί το κριτήριο ελέγχου μονοτροπικότητας ώστε, αντί να εξετάζονται για μονοτροπικότητα οι γραμμές του πίνακα αποστάσεων μεταξύ των δεδομένων μιας ομάδας, να ελέγχονται για μονοτροπικότητα οι μονοδιάστατες προβολές των δεδομένων της ομάδας σε διάφορες κατευθύνσεις που καθορίζονται ντετερμινιστικά (π.χ. PCA) ή τυχαία (π.χ. Random Projections).

Στη συνέχεια παρουσιάζεται μια συσσωρευτική (agglomerative) μεθοδολογία ομαδοποίησης βασισμένη στον έλεγχο μονοτροπικότητας των δεδομένων μιας ομάδας. Η

μέθοδος αυτή (agglodip) ξεκινά με πολλές αρχικές ομάδες. Σε κάθε βήμα ενώνονται δύο από τις ομάδες σε μια, εάν από τη συνένωσή τους προκύπτει μια μονοτροπική ομάδα σύμφωνα με το κριτήριο ελέγχου της μονοτροπικότητας. Η διαδικασία τερματίζεται όταν δεν υπάρχουν πλέον ομάδες που η συνένωσή τους να δίνει μια νέα μονοτροπική ομάδα. Με βάση αυτή τη βασική ιδέα προτείνονται κάποιες εναλλακτικές προσεγγίσεις. Μια από αυτές επιταχύνει τον έλεγχο μονοτροπικότητας κατά τη συνένωση δύο ομάδων εκμεταλλευόμενη τα κεντροειδή των δύο ομάδων. Μια δεύτερη μετασχηματίζει το πρόβλημα επαναληπτικής συνένωσης των ομάδων του αρχικού συνόλου σε πρόβλημα εύρεσης των συνεκτικών συνιστωσών ενός γραφήματος. Τέλος προτείνεται και μια τρίτη μέθοδος (agglopdip) η οποία χρησιμοποιεί για τον έλεγχο της μονοτροπικότητας τη μέθοδο των προβολών που αναφέρθηκε παραπάνω.

Οι παραπάνω μεθοδολογίες αξιολογήθηκαν πειραματικά σε συνθετικά αλλά και σε πραγματικά σύνολα δεδομένων και οι επιδόσεις τους συγκρίθηκαν τόσο με τον αλγόριθμο dip-means, όσο και με προγενέστερους αυξητικούς αλγορίθμους όπως οι x-means και g-means. Επιπλέον οι προτεινόμενες συσσωρευτικές μεθοδολογίες ομαδοποίησης εφαρμόστηκαν για το πρόβλημα της κατάτμησης εικόνων (image segmentation). Στη μέθοδο κατάτμησης που μελετήθηκε, αρχικά δημιουργείται μια υπερκατάτμηση της εικόνας με μεθόδους δημιουργίας superpixels και στη συνέχεια τα superpixels συνενώνονται σε μεγαλύτερα τμήματα, εάν από τη συνένωσή τους προκύπτουν μονοτροπικές ομάδες.

EXTENDED ABSTRACT IN ENGLISH

Chamalis, Theofilos, C.T.
MSc, Computer Science & Engineering Department, University of Ioannina, Greece.
October 2015.
Thesis Title: Clustering Methods based on statistical testing of the unimodality of the data.
Thesis Supervisor: Aristidis Likas

Clustering is one of the most important fields of machine learning and data mining because of the vast number of applications in data analysis problems. A key issue of clustering a dataset is related with the estimation of the number of clusters, which is usually unknown beforehand. An already proposed technique is the dip-means algorithm, which proposes the use of a methodology (criterion) for the statistical test of the unimodality of a dataset and uses this methodology in an incremental manner: starting with all the data in the same cluster and in each step splitting the multimodal clusters according to the criterion.

Initially, in this thesis we propose a new variant of the dip-means algorithm named pdip-means (projected dip-means) which modifies the unimodality criterion such that the one dimensional projections of the data within a cluster to various directions, that are deterministic defined (e.g. PCA) or random (e.g. Random Projections), are tested for unimodality instead of each line of the distance matrix of the data within a cluster.

Next, an agglomerative clustering methodology is presented based on the unimodality test of the data within a cluster. This method (agglodip) starts off with many initial clusters. In each step two clusters are merged into one, only if a unimodal cluster is produced according to the statistical test. This procedure ends when there are no more clusters that their merge

will produce a unimodal cluster. In addition to this concept, some other approaches are proposed as well. The first one, accelerates the unimodality tests while merging two clusters exploiting their centroids. The next approach transforms the iterative cluster merging problem of the cluster of a dataset into the problem of finding the connected components of a graph. Finally, a third method (agglopdip) is proposed which uses the unimodality test criterion in conjunction with the projection method which was referred above.

These methodologies were evaluated experimentally in synthetic as well as real life datasets and we compared their performance with the dip-means algorithm and with prior incremental clustering algorithms such as x-means and g-means. Furthermore, the proposed agglomerative clustering methodologies were applied to the problem of image segmentation. In the segmentation method that was studied, an oversegmentation of the image is created initially using superpixels and then they get merged into bigger segments, if unimodal clusters are produced by their merge.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια, με την ραγδαία ανάπτυξη της τεχνολογίας και την ευρεία χρήση των ηλεκτρονικών υπολογιστών, καθώς και μεγάλου εύρους ηλεκτρονικών συσκευών όπως smartphones, tablets αλλά και "έξυπνων" οικιακών συσκευών όπως τηλεοράσεις με δυνατότητα σύνδεσης στο Internet, τα παραγόμενα δεδομένα που διακινούνται στο Διαδίκτυο και χρήζουν επεξεργασίας από τις συσκευές αυτές έχουν αυξηθεί δραματικά. Το 2014 ήταν η χρονιά κατά την οποία οι συσκευές με πρόσβαση στο Διαδίκτυο ξεπέρασαν τον τρέχοντα πληθυσμό της γης και κατά συνέπεια όλο και περισσότερες συσκευές χρειάζεται να επεξεργαστούν μεγάλο όγκο δεδομένων.

Το πλήθος και το μέγεθος των εγγράφων, εικόνων, βίντεο, μηνυμάτων ηλεκτρονικού ταχυδρομείου και άλλων δεδομένων που διακινούνται ή είναι αποθηκευμένα στο Διαδίκτυο αυξάνεται με εκθετικό ρυθμό καθώς οι ανάγκες των χρηστών αυξάνονται. Εργαζόμενοι επιχειρήσεων αλλά και επιχειρηματικές εφαρμογές καλούνται να ανακτήσουν και να αναζητήσουν γρήγορα και αποτελεσματικά πληροφορία από βάσεις δεδομένων, από εταιρικά έγγραφα, και από άλλες εφαρμογές ώστε να επιτύχουν το μεγαλύτερο δυνατό κέρδος.

Η υπολογιστική ισχύς και η χωρητικότητα των αποθηκευτικών μέσων των σύγχρονων υπολογιστών και των υπόλοιπων συσκευών έχει φτάσει σε πολύ υψηλά επίπεδα όμως ο ρυθμός ανάπτυξής τους (νόμος Moore) υστερεί κατά πολύ σε σχέση με τον πρωτοφανή ρυθμό που παράγονται νέα δεδομένα. Γίνεται επιτακτική λοιπόν η ανάγκη για την εύρεση μεθόδων

ανάλυσης και εξαγωγής μόνο της πληροφορίας που θεωρούμε χρήσιμη για την εκάστοτε περίπτωση.

1.1. Εξόρυξη Δεδομένων (Data Mining)

Λύση στο παραπάνω πρόβλημα έρχεται να δώσει η *εξόρυξη δεδομένων (data mining)* η οποία έχει αναπτυχθεί ιδιαίτερα τα τελευταία χρόνια και χρησιμοποιείται πλέον ευρέως σε διάφορους επιστημονικούς κλάδους αλλά και σε εμπορικές εφαρμογές.

1.1.1. Τι είναι Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων [1] μπορεί να οριστεί με πολλούς τρόπους. Ο κλασσικότερος ορίζει την εξόρυξη δεδομένων ως την διαδικασία της αυτόματης ανακάλυψης μη τετριμμένης χρήσιμης πληροφορίας από μεγάλες αποθήκες δεδομένων. Οι τεχνικές που εφαρμόζονται στην εξόρυξη δεδομένων στοχεύουν στην εξαγωγή χρήσιμων και καινοτόμων προτύπων που με τη χρήση κλασσικών τεχνικών εξαγωγής πληροφορίας μπορεί να παρέμεναν άγνωστα. Επίσης μας δίνει τη δυνατότητα να προβλέψουμε μια μελλοντική συμπεριφορά ή ένα μελλοντικό αποτέλεσμα.

Δεν θεωρούνται όμως όλες οι τεχνικές ανάκτησης πληροφορίας εξόρυξη δεδομένων. Οι τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται για την ενίσχυση των διάφορων παραδοσιακών τεχνικών ανάκτησης πληροφορίας. Για παράδειγμα, δεν μπορεί να θεωρηθεί ως εξόρυξη δεδομένων:

- Η εύρεση ενός αριθμού τηλεφώνου στον τηλεφωνικό κατάλογο.
- Η ερώτηση (query) σε μια βάση δεδομένων για την εύρεση μιας καταχώρησης.
- Η αναζήτηση σε μια μηχανή αναζήτησης για πληροφορίες σχετικά με τα Ιωάννινα.

Η εξόρυξη δεδομένων περιλαμβάνει εργασίες όπως:

- Η εύρεση ονομάτων που εμφανίζονται συχνότερα σε κάποια γεωγραφική περιοχή σε ένα τηλεφωνικό κατάλογο.
- Η ομαδοποίηση άρθρων εφημερίδων που επιστρέφονται από μια μηχανή αναζήτησης σε ομάδες βάσει του περιεχομένου τους ώστε να παρουσιάζονται με πιο οργανωμένο και κατανοητό τρόπο στο χρήστη.
- Η ομαδοποίηση μεγάλου πλήθους καταχωρήσεων σε μια βάση δεδομένων βάσει κάποιου/ων χαρακτηριστικού/ων για γρηγορότερη και ευκολότερη μετέπειτα αναζήτηση.

1.1.2. Εύρεση Γνώσης σε Βάσεις Δεδομένων (KDD)

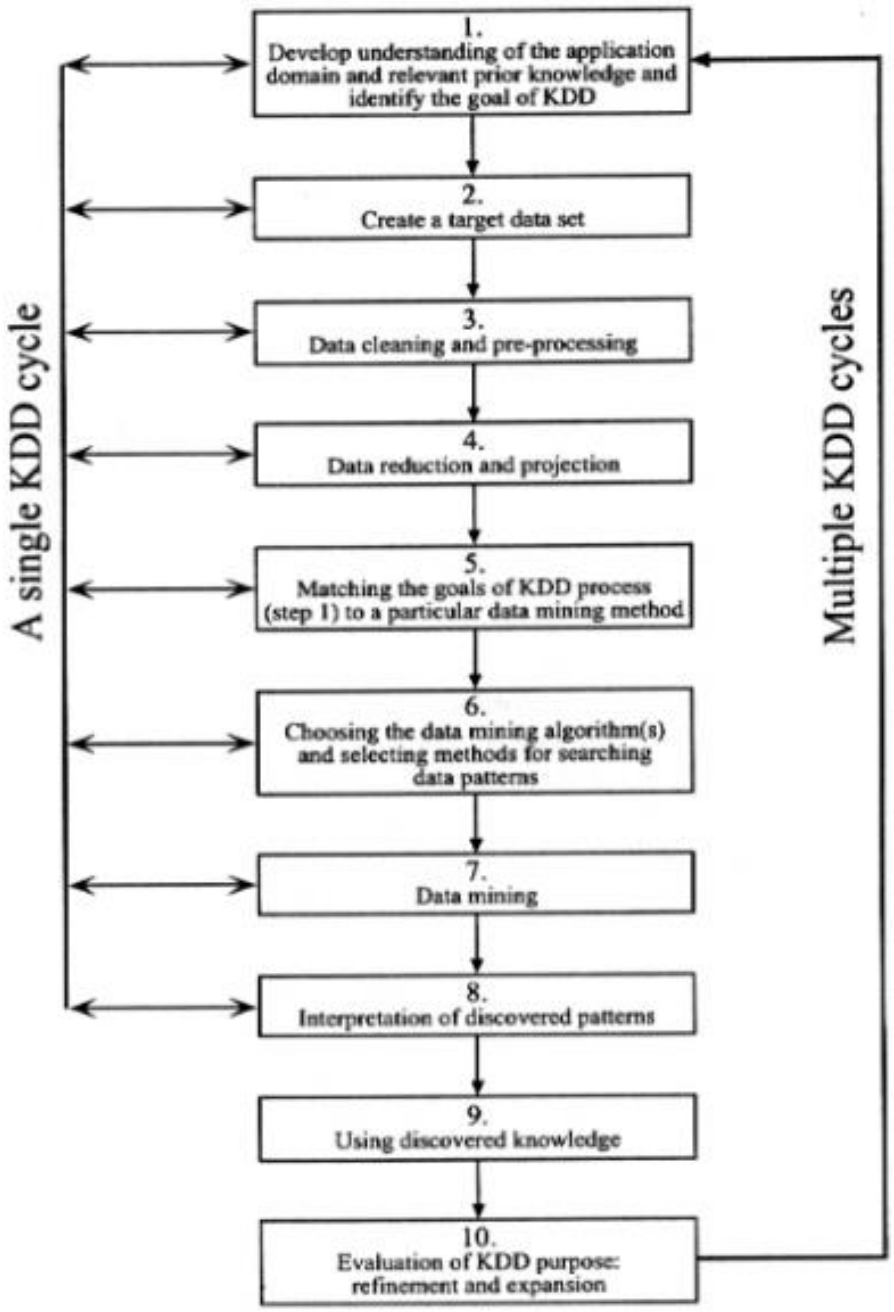
Η εξόρυξη δεδομένων αποτελεί ένα αναπόσπαστο κομμάτι της *εύρεσης γνώσης σε βάσεις δεδομένων (knowledge discovery in databases – KDD),* η οποία αναφέρεται στη διαδικασία μετατροπής των αρχικών (ακατέργαστων) δεδομένων σε χρήσιμη πληροφορία για ανακάλυψη και εξαγωγή νοήματος, συμπερασμάτων, προτύπων ή (και) αλληλοσυσχετίσεων. Το KDD είναι ιδιαίτερα χρήσιμο σε περιπτώσεις όπου τα ακατέργαστα (χαμηλού επιπέδου) δεδομένα είναι δύσκολο να κατανοηθούν ή να μεταφραστούν είτε λόγω του τεράστιου όγκου τους, είτε λόγω της αυξημένης πολυπλοκότητάς τους. Η διαδικασία του KDD αποτελείται από μια σειρά δέκα βημάτων [2].

1. **Κατανόηση του πεδίου της εφαρμογής**. Εδώ είναι απαραίτητη η απόκτηση γνώσης σχετικής με το πεδίο της εφαρμογής, καθώς και τυχόν προηγούμενης γνώσης που μπορεί να έχει ο χρήστης. Καθορίζεται επίσης ο στόχος της διαδικασίας του KDD με βάση το χρήστη.
2. **Δημιουργία ενός συνόλου δεδομένων (data set)**. Αυτό θα αποτελείται από επιλεγμένα δεδομένα και μεταβλητές τα οποία μπορεί να έχουν προέλθει από πηγές εισόδου με δεδομένα διαφορετικής μορφής ή μπορεί να είναι αποθηκευμένα κατανεμημένα σε διαφορετικές τοποθεσίες.

3. **Προ-επεξεργασία (pre-processing) και καθαρισμός δεδομένων**. Αφαιρείται θόρυβος από το σύνολο δεδομένων, διπλότυπες τιμές και αντιμετωπίζονται δεδομένα που έχουν κενά πεδία σε τιμές είτε με τη διαγραφή αυτών είτε με την συμπλήρωσή τους με κάποιο προσεγγιστικό τρόπο.
4. **Μείωση και προβολή δεδομένων**. Τα δεδομένα μετασχηματίζονται κατάλληλα είτε με ελάττωση των διαστάσεων είτε με άλλες μεθόδους μετασχηματισμών με βάση τα πιο χρήσιμα/αντιπροσωπευτικά χαρακτηριστικά προκειμένου να μειωθεί ο αριθμός των μεταβλητών που θα ληφθούν υπόψη από την τεχνική εξόρυξης που θα χρησιμοποιηθεί μετέπειτα.
5. **Επιλογή της κατάλληλης κατηγορίας τεχνικών εξόρυξης**. Ταξινομείται με βάση την απόφαση του χρήστη για το εάν θα χρησιμοποιηθεί αλγόριθμος που εμπίπτει στην κατηγοριοποίηση (classification), την ομαδοποίηση (clustering), τους κανόνες συσχέτισης (association rules) ή κάποια άλλη κατηγορία
6. **Επιλογή των αλγορίθμων που θα εκτελέσουν την εξόρυξη**. Σε αυτό το στάδιο επιλέγεται ο αλγόριθμος ο οποίος θα χρησιμοποιηθεί και εμπίπτει στην κατηγορία που επιλέξαμε παραπάνω. Η επιλογή μπορεί να γίνει με βάση την αναμενόμενη αποτελεσματικότητα του αλγορίθμου, την ταχύτητα εκτέλεσης ή με άλλα κριτήρια.
7. **Εξόρυξη της γνώσης**. Αυτό το βήμα αποτελεί ίσως την κρισιμότερη λειτουργία του KDD και είναι αυτό κατά το οποίο εφαρμόζεται ο αλγόριθμος που επιλέχθηκε στα δεδομένα για την εξαγωγή χρήσιμων προτύπων, αλληλοσυσχετίσεων ή άλλης μορφής γνώσης.
8. **Ερμηνεία των αποτελεσμάτων**. Ερμηνεύεται με τη χρήση κριτηρίων και μετρικών η γνώση που προήλθε ως έξοδος από την εξόρυξη και οπτικοποιείται για καλύτερη κατανόηση. Τα αποτελέσματα μπορεί να εκφράζονται σε πολλές μορφές όπως κανόνες κατηγοριοποίησης, ομάδες, δέντρα κτλ. Επίσης αν κριθεί απαραίτητο μπορούμε να επιστρέψουμε σε οποιοδήποτε από τα προηγούμενα βήματα από αυτό το στάδιο και να τα επαναλάβουμε.
9. **Χρήση της γνώσης**. Η γνώση που εξήχθη από τα αποτελέσματα των αλγορίθμων εξόρυξης μπορεί να αποτελέσει είσοδο σε κάποιο άλλο σύστημα για περαιτέρω επεξεργασία ή συμπληρωματικά με άλλα δεδομένα εισόδου.
10. **Αξιολόγηση του σκοπού της χρήσης του KDD**. Τα συμπεράσματα που προέκυψαν από την ερμηνεία των αποτελεσμάτων μπορούν να βοηθήσουν στη θεμελίωση υποθέσεων πολλών θεμάτων καθώς και να δώσουν τροφή για νέες ερωτήσεις στο

εκάστοτε αντικείμενο. Έτσι μπορεί να επαναληφθεί η όλη διαδικασία με διαφορετική παραμετροποίηση για τη βελτίωση των αποτελεσμάτων αν αυτό κριθεί αναγκαίο.

Στο Σχήμα 1.1.1 παρουσιάζονται οπτικά τα παραπάνω βήματα [2].



Σχήμα 1.1.1 Εύρεση γνώσης σε βάσεις δεδομένων

1.1.3. Ιστορική αναδρομή της Εξόρυξης Δεδομένων

Η εξαγωγή γνώσης και προτύπων από δεδομένα με μη αυτόματο τρόπο χρησιμοποιείται αρκετούς αιώνες τώρα. Τεχνικές για τον προσδιορισμό προτύπων εμφανίστηκαν μέσω του τομέα της στατιστικής με την ανάπτυξη της *θεωρίας Bayes* (*Bayes theorem* - 1700) αλλά και με την *ανάλυση παλινδρόμησης* (*regression analysis* - 1800).

Ο όρος εξόρυξη δεδομένων [3] εμφανίστηκε λίγο πριν το 1990 αρχικά μεταξύ μελών της κοινότητας βάσεων δεδομένων. Προέκυψε από την ανάγκη ανάλυσης και επεξεργασίας δεδομένων μεγάλου όγκου και πολυπλοκότητας με αυτόματο τρόπο ώστε τα συγκεντρωμένα δεδομένα να είναι οργανωμένα και εύκολα διαχειρίσιμα κρατώντας μόνο τα χρήσιμα κάθε φορά.

Παράλληλα ερευνητικοί τομείς της επιστήμης της πληροφορικής όπως η *τεχνητή νοημοσύνη (artificial intelligence)*, η *μηχανική μάθηση (machine learning)* και η *αναγνώριση προτύπων (pattern recognition)* εμφανίστηκαν λίγο πριν το 1950 και γνώρισαν ραγδαία ανάπτυξη. Ο τομέας λοιπόν της εξόρυξης δεδομένων δανείστηκε ιδέες και αρχές από όλους τους παραπάνω τομείς καθώς και από πιο τεχνικούς τομείς της επιστήμης των υπολογιστών όπως η *τεχνολογία βάσεων δεδομένων (database technology)*, η *παράλληλη επεξεργασία* και τα *κατανεμημένα συστήματα (parallel/distributed computing)* κυρίως για θέματα απόδοσης. Για παράδειγμα, ιδέες όπως η δειγματοληψία και τα στατιστικά τεστ προέρχονται από τον τομέα της στατιστικής, θεωρίες μάθησης, ασαφής λογική από την τεχνητή νοημοσύνη και τη μηχανική μάθηση και τεχνικές για ανάπτυξη παράλληλων αλγορίθμων για πολυεπεξεργαστικά συστήματα από τον παράλληλο υπολογισμό. Το Σχήμα 1.1.2 δείχνει οπτικά την αλληλοεπικάλυψη των περιοχών αυτών για τη δημιουργία του τομέα της εξόρυξης δεδομένων.

Σχήμα 1.1.2 Προέλευση της εξόρυξης δεδομένων

1.2. Τεχνικές Εξόρυξης Δεδομένων

Τα βασικά προβλήματα εξόρυξης γνώσης από δεδομένα είναι: η *κατηγοριοποίηση (classification)*, η *παλινδρόμηση (regression)*, η *ανίχνευση ανωμαλιών (anomaly detection)*, η *ομαδοποίηση (clustering)*, η *εύρεση κανόνων συσχέτισης (association rule discovery)* και η *εύρεση προτύπων ακολουθιών (sequential pattern discovery)*. Τα συστήματα μπορούν να οργανωθούν σε δύο ομάδες, τα *συστήματα πρόβλεψης (prediction systems)* και τα *περιγραφικά μοντέλα (descriptive models)* [1]. Στη συνέχεια αναλύονται οι παραπάνω τεχνικές και ο τρόπος οργάνωσής τους:

Τεχνικές πρόβλεψης. Η ομάδα αυτή των τεχνικών εξόρυξης έχει ως στόχο την πρόβλεψη άγνωστων ή μελλοντικών τιμών ενός χαρακτηριστικού βασιζόμενη σε τιμές άλλων χαρακτηριστικών που έχουν παρατηρηθεί. Τα χαρακτηριστικά που είναι προς πρόβλεψη ονομάζονται και εξαρτημένες μεταβλητές, ενώ αυτά που χρησιμοποιούνται για να γίνει η πρόβλεψη των άγνωστων ονομάζονται ανεξάρτητες μεταβλητές. Επειδή στις τεχνικές πρόβλεψης υπάρχουν οι τιμές-στόχοι που θέλουμε να προβλέψουμε και στην ουσία ένα σύστημα πρόβλεψης μαθαίνει με βάση υπάρχουσες τιμές, μπορούμε χρησιμοποιώντας όρους μηχανικής μάθησης να πούμε πως η διαδικασία μάθησης που συντελείται είναι *μάθηση επίβλεψη (supervised learning)*. Οι τεχνικές που ανήκουν σε αυτή την ομάδα είναι:

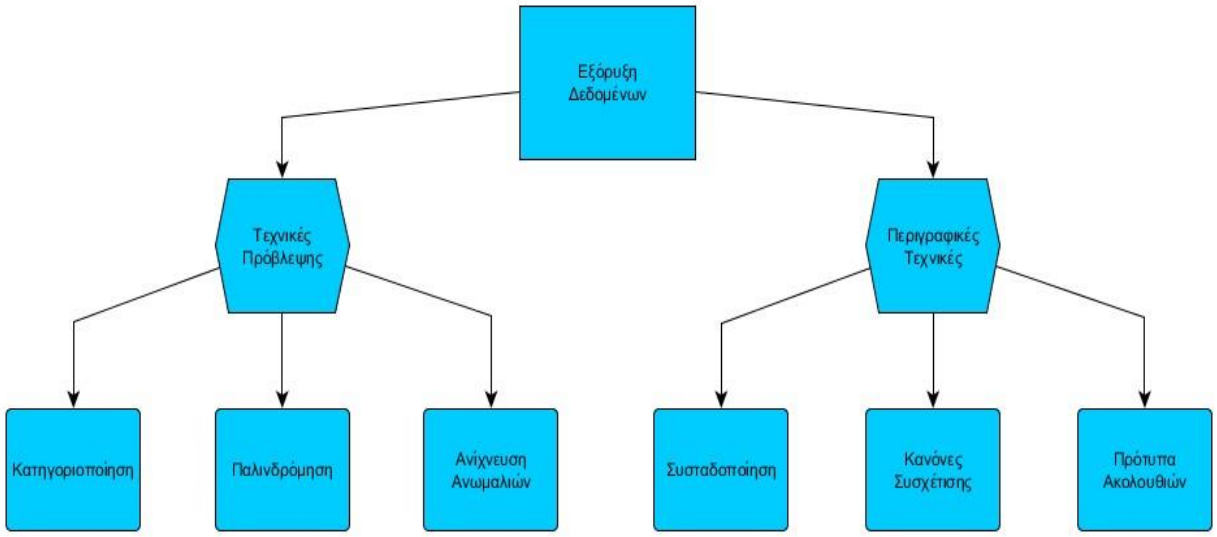
- **Κατηγοριοποίηση (Classification)**. Η κατηγοριοποίηση είναι μία από τις βασικότερες τεχνικές στην εξόρυξη δεδομένων. Η κατηγοριοποίηση βασίζεται στον προκαθορισμένο ορισμό των κατηγοριών ενός συνόλου παρατηρήσεων/παραδειγμάτων το οποίο θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου που θα δημιουργηθεί. Αυτό το μοντέλο αργότερα θα έχει τη δυνατότητα να παίρνει ως είσοδο νέες παρατηρήσεις των οποίων δεν ξέρουμε την κατηγορία και να τις αναθέτει στην κατάλληλη χρησιμοποιώντας γνώση που εξήγαγε από τις προηγούμενες παρατηρήσεις. Ένα παράδειγμα θα μπορούσε να είναι η κατηγοριοποίηση ενός ατόμου αν είναι θετικός σε κάποια ασθένεια με βάση κάποια χαρακτηριστικά/μεταβλητές που έχει όπως το φύλο, την ηλικία και την αρτηριακή πίεση, χρησιμοποιώντας κατηγοριοποιήσεις άλλων ανθρώπων, ασθενών και μη.
- **Παλινδρόμηση (Regression)**. Η παλινδρόμηση είναι μια ευρέως χρησιμοποιούμενη τεχνική που προήλθε από τον τομέα της στατιστικής και χρησιμοποιείται κυρίως για τη μοντελοποίηση και την εύρεση συσχέτισης μεταξύ μιας εξαρτημένης και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Σκοπός της στην εξόρυξη δεδομένων είναι να προβλέψει την τιμή μιας μεταβλητής χρησιμοποιώντας τις τιμές που είχε στο παρελθόν. Για παράδειγμα μπορεί να υπολογιστεί η πιθανότητα με την οποία ένας ασθενής θα αναρρώσει με βάση τα αποτελέσματα της διάγνωσης.
- **Ανίχνευση ανωμαλιών (Anomaly detection)**. Η τεχνική αυτή αναφέρεται στην αναγνώριση αντικειμένων ή προτύπων από ένα σύνολο δεδομένων που δεν ακολουθούν την αναμενόμενη συμπεριφορά αλλά παρεκκλίνουν. Αυτά τα αντικείμενα μπορεί να είναι κάποιου είδους θόρυβος, πρόβλημα ή λάθος των μηχανημάτων που κάνουν τις μετρήσεις ή ακόμη και ανθρώπινο λάθος κατά την εισαγωγή των δεδομένων. Ο κύριος σκοπός της όμως είναι ο εντοπισμός ύποπτων ή επιβλαβών συμπεριφορών και γεγονότων με όσο το δυνατό υψηλότερο επίπεδο πιθανών ανωμαλιών και όσο το δυνατό μικρότερο ποσοστό λάθος ανίχνευσης σε εφαρμογές όπως η ανίχνευση εισβολής σε τραπεζικά συστήματα ή δίκτυα υπολογιστών.

Περιγραφικές τεχνικές. Οι τεχνικές που ανήκουν σε αυτή την ομάδα έχουν ως στόχο την εύρεση προτύπων διαφόρων μορφών (συσχετίσεις, ομάδες, ανωμαλίες) οι οποίες είναι κατανοητές από τον άνθρωπο και περιγράφουν τις σχέσεις που αναπτύσσονται ανάμεσα στα δεδομένα. Εδώ, σε αντίθεση με τις τεχνικές πρόβλεψης, δεν υπάρχουν τιμές-στόχοι και η διαδικασία μάθησης γίνεται χωρίς κάποια εξωτερική συμβολή που μπορεί να διορθώνει τυχόν

λάθη που προκύπτουν κατά τη διάρκειά της. Σε αυτή την περίπτωση έχουμε *μάθηση χωρίς επίβλεψη (unsupervised learning)* και οι τεχνικές που ανήκουν σε αυτή την ομάδα είναι οι εξής:

- **Ομαδοποίηση (Clustering)**. Η ομαδοποίηση είναι η διαδικασία κατά την οποία ένα σύνολο παρατηρήσεων χωρίζεται σε ένα σύνολο από ομάδες/συστάδες (clusters). Η ανάθεση των παρατηρήσεων στις ομάδες που ανήκουν γίνεται με βάση δύο κριτήρια: την μεγιστοποίηση της ομοιότητας των παρατηρήσεων στο εσωτερικό κάθε ομάδας και ταυτόχρονα την ελαχιστοποίηση της ομοιότητας των παρατηρήσεων μιας ομάδας με τις παρατηρήσεις κάθε άλλης. Αυτό έχει ως αποτέλεσμα να δημιουργούνται "λογικές" ομάδες με τις παρατηρήσεις που μοιάζουν περισσότερο να βρίσκονται μαζί και να δημιουργείται έτσι μια "περίληψη των δεδομένων". Ένα παράδειγμα εφαρμογής είναι στη συμπίεση δεδομένων όπου αρχικά μπορούμε να χωρίσουμε σε ομάδες ένα μεγάλο πλήθος δεδομένων και όταν θέλουμε στη συνέχεια να επεξεργαστούμε τα δεδομένα με κάποιο τρόπο, να χρησιμοποιούμε μόνο τον αντιπρόσωπο από κάθε ομάδα. Η παρούσα διατριβή θα ασχοληθεί με αλγορίθμους ομαδοποίησης. Στο κεφάλαιο 2 θα αναλυθεί εκτενέστερα η τεχνική αυτή.
- **Εύρεση κανόνων συσχέτισης (Association rule discovery)**. Η εύρεση κανόνων συσχέτισης θεωρείται μια από τις πιο ενδιαφέρουσες τεχνικές της εξόρυξης δεδομένων. Αυτό που ξεχωρίζει την τεχνική αυτή από τις υπόλοιπες είναι ότι έχει την ικανότητα να ανακαλύπτει κρυμμένες συσχετίσεις μεταξύ των χαρακτηριστικών του συνόλου δεδομένων εισόδου και να τα παρουσιάζει σε μορφή ευκολονόητη από τον άνθρωπο. Η μορφή αυτή ενός κανόνα συσχέτισης $A \Rightarrow B$ δηλώνει μια σχέση εξάρτησης ανάμεσα σε δύο ξένα σύνολα στοιχείων Α και Β. Για παράδειγμα, ένας κανόνας συσχέτισης που μπορεί να εκφράζει τους καταναλωτές ενός πολυκαταστήματος είναι ότι εάν αγοράσουν ντομάτες και κρεμμύδια είναι πολύ πιθανό να αγοράσουν και λάδι δηλαδή $\{\text{ντομάτες, κρεμμύδια}\} \Rightarrow \{\text{λάδι}\}$.
- **Εύρεση προτύπων ακολουθιών (Sequential pattern discovery)**. Η εύρεση προτύπων ακολουθιών είναι η ανακάλυψη των προτύπων που εμφανίζονται συχνότερα από ένα σύνολο δεδομένων. Τα πρότυπα αυτά συνήθως σχετίζονται με τον χρόνο όμως μπορούν να σχετίζονται και με άλλου είδους ακολουθίες όπως ακολουθίες που εμφανίζονται στο DNA ή ακολουθίες αριθμών που εμφανίζονται σε κάποιο τυχερό παιχνίδι.

Μια οπτική αναπαράσταση της οργάνωσης των τεχνικών εξόρυξης δεδομένων παρουσιάζεται στο Σχήμα 1.2.1



Σχήμα 1.2.1 Οργάνωση τεχνικών εξόρυξης δεδομένων

1.3. Δομή της Διατριβής

Η διατριβή ξεκινά με μια εισαγωγή στην τομέα της εξόρυξης δεδομένων. Παρατίθεται μια σύντομη ιστορική αναδρομή που σκοπό έχει να καθορίσει τις ρίζες της εξόρυξης δεδομένων και τον τρόπο δημιουργίας τους από τομείς άλλων επιστημονικών πεδίων. Γίνεται αναφορά στους κυριότερους κλάδους που την αποτελούν και παρουσιάζονται τα πεδία εφαρμογής της σε ένα πλήθος προβλημάτων που εμφανίζονται και τονίζεται η σημαντικότητά της.

Στο δεύτερο κεφάλαιο παρουσιάζεται αναλυτικότερα το αντικείμενο της ομαδοποίησης και οι εφαρμογές του. Αναπτύσσονται οι κυριότεροι τρόποι ομαδοποίησης και δίνεται μεγαλύτερη έμφαση στον αλγόριθμο ομαδοποίησης Κ-Μέσων. Παρουσιάζεται αναλυτικά ο τρόπος λειτουργίας του, ο οποίος χαρακτηρίζεται για την απλότητα και την αμεσότητα που παρέχει στην αντιμετώπιση προβλημάτων ομαδοποίησης, τα προτερήματά του καθώς και τις αδυναμίες τις οποίες παρουσιάζει. Επιπλέον παρουσιάζεται η τεχνική της αυξητικής ομαδοποίησης η οποία αντιμετωπίζει σε μεγάλο βαθμό ένα από τα βασικά προβλήματά του που είναι η αρχικοποίηση.

Το κεφάλαιο 3 περιέχει μεθόδους ομαδοποίησης που έχουν αναπτυχθεί παλαιότερα και στοχεύουν στην αντιμετώπιση των αδυναμιών του K-Μέσων και πιο συγκεκριμένα στην αυτόματη εύρεση του αριθμού των ομάδων του υποκείμενου συνόλου δεδομένων. Περιγράφεται ο τρόπος λειτουργίας τους και αναφέρονται τα προτερήματα και οι αδυναμίες του καθενός.

Στο τέταρτο κεφάλαιο παρουσιάζονται οι μέθοδοι ομαδοποίησης που αναπτύχθηκαν στο πλαίσιο της παρούσας εργασίας. Αναλύεται η δομή τους, ο τρόπος εφαρμογής ενός στατιστικού κριτηρίου για τον έλεγχο της μονοτροπικότητας των δεδομένων καθώς και τεχνικές που στοχεύουν στην επιτάχυνση του εκάστοτε αλγορίθμου.

Στο κεφάλαιο 5 παρουσιάζεται η πειραματική μελέτη και η σύγκριση όλων των παραπάνω αλγορίθμων καθώς και των τεχνικών επιτάχυνσης. Η σύγκριση αυτή γίνεται με βάση ενός συνόλου δεικτών εκτίμησης της ποιότητας της ομαδοποίησης και με οπτική απεικόνιση σε σύνολα δεδομένων 2 διαστάσεων, όπου ήταν αυτό εφικτό. Στο έκτο κεφάλαιο παρατίθεται μια εφαρμογή ενός από τους αλγορίθμους που αναπτύχθηκαν για την κατάτμηση εικόνων και παρουσιάζονται στη συνέχεια αποτελέσματα από την κατάτμηση διάφορων εικόνων που έγιναν με τη χρήση του αλγορίθμου αυτού. Τέλος, το κεφάλαιο 7 αποτελεί τον επίλογο, στον οποίο παρουσιάζονται συνοπτικά τα συμπεράσματα της διατριβής όπως επίσης και κατευθύνσεις για μελλοντική εργασία.

ΚΕΦΑΛΑΙΟ 2. ΟΜΑΔΟΠΟΙΗΣΗ

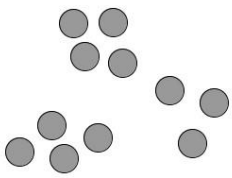
2.1. Εισαγωγή στην Ομαδοποίηση (Clustering)

Η Ομαδοποίηση (clustering) ή όπως αλλιώς λέγεται ‘ανάλυση ομάδων’ (cluster analysis) είναι πρακτικά η διαδικασία της οργάνωσης των στοιχείων ενός συνόλου δεδομένων σε ομάδες με βάση κάποιο μέτρο ομοιότητας. Η βασική ιδέα στην οποία στηρίζεται η ομαδοποίηση καθορίζει τους στόχους όλων των αλγορίθμων που ανήκουν σε αυτή την κατηγορία, οι οποίοι είναι:

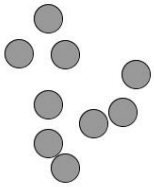
- Κάθε ομάδα αντικειμένων από το σύνολο των ομάδων που έχουν δημιουργηθεί να περιέχει αντικείμενα τα οποία είναι “όμοια” μεταξύ τους, δηλαδή να είναι “ομοιογενής”.
- Κάθε ομάδα αντικειμένων να περιέχει αντικείμενα τα οποία είναι “ανόμοια” με τα αντικείμενα κάθε άλλης ομάδας. Δηλαδή κάθε ομάδα να είναι “ανόμοια” με κάθε άλλη.

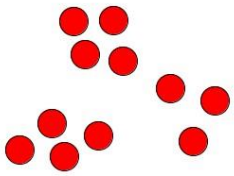
Η ομαδοποίηση όμως είναι γενικά ένα δύσκολο πρόβλημα. Αυτό σε μεγάλο βαθμό προκύπτει από μια θεωρητική δυσκολία που εμφανίζεται στην παραπάνω ιδέα. Τι ορίζουμε ως ομάδα; Ποιο αποτέλεσμα θεωρούμε σωστό και ποιο όχι; Δεν υπάρχει κάποιος ξεκάθαρος

ορισμός για το πώς ορίζεται μια ομάδα, ούτε κάποια “συνταγή” για το ποια είναι η σωστή λύση. Το θεωρητικό αυτό κενό της ομαδοποίησης εν μέρει αντιμετωπίζεται σε σύνολα δεδομένων όπου υπάρχει προηγούμενη γνώση από τον χρήστη για τη φύση των αντικειμένων που τα απαρτίζουν ώστε να παραχθούν “ορθότερες λύσεις” και να είναι ευκολότερη η ερμηνεία τους. Στα παρακάτω σχήματα (Σχήματα 2.1.1 – 2.1.2) γίνεται εμφανής η παραπάνω δυσκολία. Στο παράδειγμα αυτό [1] έχουμε 20 σημεία τα οποία αποτελούν το σύνολο δεδομένων το οποίο απεικονίζεται στο Σχήμα 2.1.1 και αποτελείται από τα αντικείμενα πριν χωριστούν σε ομάδες. Στα επόμενα σχήματα (Σχήματα 2.1.2 – 2.1.4) μπορούμε να δούμε πως διαμορφώνεται κάθε λύση, καταλήγοντας σε διαφορετικό πλήθος ομάδων κάθε φορά ανάλογα με τον τρόπο που ορίζουμε την ομάδα, με τα μέλη διαφορετικών ομάδων αναπαρίστανται με διαφορετικό χρώμα. Όλες οι λύσεις μπορούν να θεωρηθούν σωστές και για τον λόγο αυτό, το κομμάτι της ερμηνείας των λύσεων παίζει εξίσου σημαντικό ρόλο με τη διαδικασία της ομαδοποίησης.

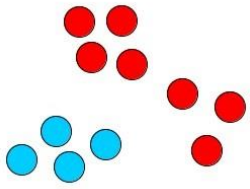


Σχήμα 2.1.1 Αρχικό σύνολο παραδείγματος

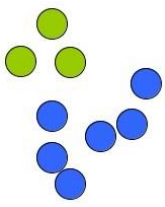


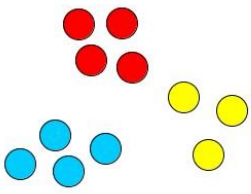


Σχήμα 2.1.2 Ομαδοποίηση σε 2 ομάδες

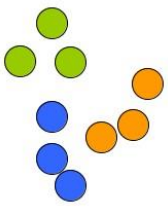


Σχήμα 2.1.3 Ομαδοποίηση σε 4 ομάδες





Σχήμα 2.1.4 Ομαδοποίηση σε 6 ομάδες



Οι παρατηρήσεις τις περισσότερες φορές αναπαρίστανται ως διανύσματα συνεχών ή τιμών. Επίσης διευκολύνει η αναπαράσταση ως σημείων στο δισδιάστατο ή τρισδιάστατο χώρο για καλύτερη εποπτεία και ερμηνεία των αποτελεσμάτων από τον άνθρωπο. Σε πολλά προβλήματα όμως κυρίως λόγω της πολυπλοκότητας που παρουσιάζουν ή της μεγάλης διάστασής τους δεν είναι εφικτό να χρησιμοποιούνται τα δεδομένα στην αρχική τους μορφή, οπότε απαιτείται η εφαρμογή κάποιων μετασχηματισμών ή άλλων διαδικασιών όπως είναι η

αφαίρεση θορύβου κατά τη διαδικασία της προεπεξεργασίας. Τέτοιες περιπτώσεις είναι για παράδειγμα το σύνολο των pixels μιας εικόνας και οι συμβολοσειρές που απαρτίζουν ένα κείμενο.

Εξετάζοντας τη σχέση μεταξύ των τεχνικών της ομαδοποίησης και της κατηγοριοποίησης μπορούμε να βρούμε αρκετές ομοιότητες. Η ομαδοποίηση μπορεί να θεωρηθεί σαν μια μορφή κατηγοριοποίησης με την ιδιότητα να δημιουργεί τις κλάσεις /κατηγορίες αυτόματα. Εν αντιθέσει, η κατηγοριοποίηση προϋποθέτει ότι τα αντικείμενα που θα έχει ως είσοδο αρχικά θα πρέπει να έχουν ήδη κατηγοριοποιηθεί στις υπάρχουσες (γνωστές) κατηγορίες, ώστε στη συνέχεια να ταξινομηθούν σωστά νέα άγνωστα αντικείμενα στις υπάρχουσες κλάσεις. Έτσι η ομαδοποίηση μπορεί να λογιστεί και ως μια διαδικασία κατηγοριοποίησης χωρίς επίβλεψη (unsupervised classification).

2.2. Προσεγγίσεις Ομαδοποίησης και Τύποι Ομάδων

Κάθε αλγόριθμος ομαδοποίησης έχει δημιουργηθεί βασιζόμενος σε διαφορετική φιλοσοφία είτε προς την διαδικασία με την οποία συντελείται η ομαδοποίηση είτε προς τη μορφή των δεδομένων που χρησιμοποιεί ως είσοδο είτε με τον τρόπο τον οποίο ορίζει τι είναι ομάδα και τι μορφή θα έχει η παραγόμενη λύση ομαδοποίησης. Έτσι, υπάρχουν διάφοροι τύποι ομαδοποίησης: *ιεραρχική (hierarchical)* ή *διαιρετική (partitional), αποκλειστική (exclusive)* ή *επικαλυπτόμενη (overlapping)* ή *ασαφής (fuzzy)* και *πλήρης (complete)* ή *μερική (partial).*

- **Ιεραρχική (Hierarchical)**. Όταν ένας αλγόριθμος συντελεί ιεραρχική ομαδοποίηση εννοούμε πως η οργάνωση των ομάδων γίνεται σαν μια δενδροειδή δομή. Αυτό πρακτικά σημαίνει πως η κάθε ομάδα αποτελείται από την ένωση τυχόν υποομάδων που έχει και πως υπάρχει μια ομάδα που περιέχει όλες τις υπόλοιπες, δηλαδή είναι η ρίζα του δέντρου των ομάδων. Τα φύλα αυτού του δέντρου, δηλαδή οι ομάδες του χαμηλότερου επιπέδου που δεν έχουν υποομάδες, ενδέχεται πολλές φορές να αποτελούνται από μία μόνο παρατήρηση του συνόλου δεδομένων.

- **Διαιρετική (Partitional)**. Αντίθετα, όταν συντελείται διαιρετική ομαδοποίηση το σύνολο των ομάδων που αποτελούν τη λύση είναι μια διαμέριση των αρχικών στοιχείων του συνόλου δεδομένων χωρίς την ύπαρξη υποομάδων.
- **Αποκλειστική (Exclusive)**. Αποκλειστική ομαδοποίηση έχουμε όταν ένα αντικείμενο του συνόλου δεδομένων ανήκει αυστηρά σε μία και μόνο ομάδα. Αυτή είναι από τις πιο συνήθεις προσεγγίσεις λόγω της απλότητας της και της εύρεσης "καλών" λύσεων στα περισσότερα προβλήματα.
- **Επικαλυπτόμενη (Overlapping)**. Εδώ το κύριο χαρακτηριστικό αυτής της προσέγγισης είναι πως κάθε στοιχείο έχει τη δυνατότητα να ανήκει σε δυο ή περισσότερες ομάδες κατά τη διαδικασία της ομαδοποίησης. Για παράδειγμα μπορεί ένας άνθρωπος να είναι ταυτόχρονα ασθενής αλλά και γιατρός.
- **Ασαφής (Fuzzy)**. Η προσέγγιση αυτή δίνει τη δυνατότητα σε κάθε παρατήρηση του συνόλου δεδομένων να ανήκει σε δύο ή περισσότερες ομάδες ταυτόχρονα. Αυτό που τη διαφοροποιεί όμως με την επικαλυπτόμενη, είναι πως μπορεί να καθοριστεί βάσει ενός βάρους κατά πόσο μια παρατήρηση είναι μέλος μιας ομάδας. Για παράδειγμα μπορεί μια παρατήρηση να ανήκει σε όλες τις ομάδες όμως σε κάποια να ανήκει με μεγαλύτερο βάρος, δηλαδή περισσότερο. Αν μια παρατήρηση δεν ανήκει σε μια ομάδα τότε έχει βάρος ίσο με 0 προς αυτή. Επίσης όλα τα βάρη πρέπει αθροιζόμενα να έχουν αποτέλεσμα 1.
- **Πλήρης (Complete)**. Η πλήρης βασίζεται στην ιδέα πως κάθε αντικείμενο του συνόλου δεδομένων πρέπει να ανήκει σε μία ή περισσότερες ομάδες και πως θα αποτελεί μέρος της λύσης.
- **Μερική (Partial)**. Σε αντίθεση με την παραπάνω ιδέα, όταν η ομαδοποίηση που συντελείται είναι μερική, σημαίνει πως ενδέχεται κάποια δεδομένα να μην αποτελούν μέρος της λύσης. Αυτά μπορεί να ταυτοποιηθούν είτε ως θόρυβος, είτε ως outliers (ακραίες τιμές) είτε ως μη ενδιαφέροντα για την εύρεση λύσης ομαδοποίησης.

Μια εξίσου σημαντική κατηγοριοποίηση βασίζεται στην έννοια της ομάδας όπως ορίζεται από κάθε αλγόριθμο ομαδοποίησης. Οι βασικές κατηγορίες ομάδων είναι οι:

- **Καλά Διαχωρισμένες (Well Separated)**. Ο ορισμός των καλά διαχωρισμένων ομάδων μιας λύσης ομαδοποίησης συνίσταται στο ότι η απόσταση (ή αντίστροφα η

ομοιότητα) κάθε ζεύγους σημείων που βρίσκονται εσωτερικά των ομάδων να είναι μικρότερη από την απόσταση κάθε ζεύγους σημείων από δύο διαφορετικές ομάδες. Αυτό είναι πολλές φορές δύσκολο και προϋποθέτει την ύπαρξη φυσικών ομάδων μεταξύ των δεδομένων.

- **Με Πρωτότυπο (Prototype-Based)**. Αυτός ο ορισμός υποθέτει πως κάθε ομάδα αποτελείται από ένα σύνολο στοιχείων τα οποία είναι πιο κοντά (ή πιο όμοια) στο πρωτότυπο της ομάδας αυτής από οποιασδήποτε άλλης. Αυτό το πρωτότυπο μπορεί να είναι το *κεντροειδές (centroid)*, δηλαδή ο μέσος όρος όλων των στοιχείων που απαρτίζουν την ομάδα και το οποίο μπορεί να μην είναι κάποιο υπάρχων σημείο της, είτε ο ενδιάμεσος *(medoid)*, δηλαδή το κεντρικό στοιχείο της ομάδας το οποίο είναι απαραίτητα μέλος της.
- **Με Γράφημα (Graph-Based)**. Κάποιες φορές το σύνολο δεδομένων έχει τη μορφή γραφήματος, με τα αντικείμενα να συμβολίζουν τους κόμβους και τις σχέσεις μεταξύ τους τις ακμές. Τότε σαν ομάδα μπορεί να οριστεί μια συνεκτική συνιστώσα, δηλαδή ένα υποσύνολο του συνόλου δεδομένων στο οποίο κάθε κόμβος έχει μια ακμή προς κάθε άλλο εντός του υποσυνόλου και δεν έχουν ακμή προς αντικείμενα εκτός αυτού.
- **Με Πυκνότητα (Density-Based)**. Σε αντίθεση με τους παραπάνω ορισμούς είναι δυνατό μια ομάδα να οριστεί με βάση την πυκνότητα των αντικειμένων ενός συνόλου δεδομένων ως μια πυκνή περιοχή που διακρίνεται από τη γύρω της περιοχή αραιής πυκνότητας. Αυτός ο τρόπος ορισμού ομάδων είναι ιδιαίτερα χρήσιμος όταν στα δεδομένα παρουσιάζεται θόρυβος ή ακραίες τιμές, καθώς και σε περιπτώσεις που μια ομάδα επικαλύπτει μέρος μιας άλλης και μέσω της πυκνότητας των στοιχείων τους μπορεί να εντοπιστεί η διαφοροποίηση.
- **Συνδυασμοί των παραπάνω κατηγοριών**. Σε πολύπλοκα και πολλές φορές τεχνητά σύνολα δεδομένων μπορούν να εμφανιστούν και συνδυασμοί των παραπάνω τύπων ομάδων στο ίδιο σύνολο δεδομένων.

Στη συνέχεια θα παρουσιαστεί ο αλγόριθμος ομαδοποίησης K-Μέσων (k-means) πάνω στον οποίο βασίζονται οι αλγόριθμοι της παρούσας διατριβής, οι οποίοι θα παρουσιαστούν στο κεφάλαιο 4 [1].

2.3. Αλγόριθμος K-Μέσων (K-Means)

Ο δημοφιλέστερος ίσως αλγόριθμος ομαδοποίησης είναι ο αλγόριθμος K-Μέσων. Ανήκει στις διαιρετικές τεχνικές με ομάδες ορισμένες με τη χρήση πρωτοτύπων. Υπάρχει μια οικογένεια αλγορίθμων που βασίζονται στις ιδέες και στον τρόπο λειτουργίας του, με τους πιο γνωστούς να είναι ο K-ενδιάμεσων (k-medoids) και ο διαιρετικός K-Μέσων (bisecting k-means). Ο παρακάτω πίνακας (Πίνακας 2.3) περιέχει τους συμβολισμούς οι οποίοι θα χρησιμοποιηθούν στη συνέχεια.

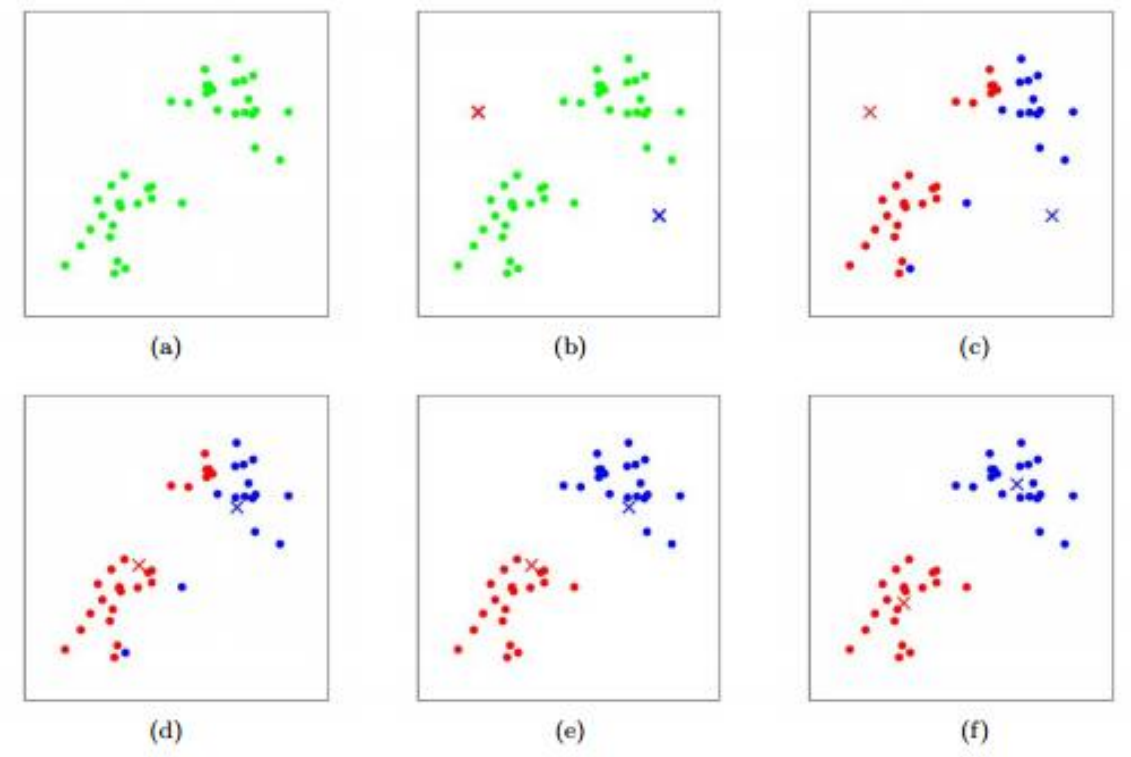
Συμβολισμός	Περιγραφή
X	Το σύνολο δεδομένων
x	Ένα σημείο/αντικείμενο του συνόλου δεδομένων
C_i	Η i-οστή ομάδα
c_i	Το κεντροειδές της i-οστής ομάδας
c	Το κεντροειδές όλων των σημείων
N_i	Το πλήθος των στοιχείων της i-οστής ομάδας
N	Το πλήθος των στοιχείων όλου του συνόλου δεδομένων
K	Ο αριθμός των ομάδων

Πίνακας 2.3.1 Συμβολισμοί αλγορίθμου K-Μέσων

2.3.1. Βασικός Αλγόριθμος K-Μέσων

Ο βασικός αλγόριθμος K-Μέσων είναι από τους πιο ευκολονόητους ως προς τον τρόπο λειτουργίας του καθώς ακολουθεί μια απλή διαδικασία βημάτων για ομαδοποίηση. Αρχικά, ορίζονται *K* κεντροειδή από το χρήστη (ή τυχαία), τα οποία αποτελούν το επιθυμητό πλήθος των ομάδων. Αυτό παραμένει σταθερό καθ' όλη τη διαδικασία της ομαδοποίησης. Έπειτα κάθε στοιχείο ανατίθεται στο κοντινότερο προς αυτό κεντροειδές και κάθε τέτοια συλλογή σημείων που προκύπτει σε αυτό το βήμα είναι μια ομάδα. Έτσι έχει δημιουργηθεί μια αρχική διαμέριση *K* ομάδων την οποία προσπαθεί να βελτιώσει ο αλγόριθμος επαναληπτικά. Στο επόμενο βήμα, το κεντροειδές κάθε ομάδας ανανεώνεται ώστε να είναι ο μέσος όρος των τιμών των αντικειμένων της. Έπειτα κάθε στοιχείο ανατίθεται εκ νέου στο κοντινότερο προς

αυτό κεντροειδές και επαναλαμβάνεται επαναληπτικά η όλη διαδικασία μέχρι τα κεντροειδή να παραμείνουν σταθερά ή ισοδύναμα μέχρι κάθε στοιχείο να μην ανατίθεται σε μια νέα ομάδα. Η παραπάνω διαδικασία απεικονίζεται στο Σχήμα 2.3.1. Στην εικόνα (a) είναι το αρχικό σύνολο δεδομένων και στη (b) γίνεται ο καθορισμός των δύο αρχικών κεντροειδών. Στα επόμενα βήματα ακολουθούνται επαναληπτικά τα βήματα του K-Μέσων μέχρι να συγκλίνει στη λύση που φαίνεται στη εικόνα (f) [5].



Σχήμα 2.3.1 Επαναλήψεις του K-Μέσων έως τη σύγκλιση

Ο ψευδοκώδικας του αλγορίθμου K-Μέσων παρατίθεται στον Πίνακα 2.3.1.

Βασικός Αλγόριθμος K-Μέσων (K-Means)
1. Αρχικοποίηση των K κέντρων c_j, $\forall j = 1...\mathrm{K}$ 2. *Επανάλαβε* 3. $C_j = \{\}$, $\forall j = 1...\mathrm{K}$ 4. Εύρεση της ομάδας που ανήκει το κάθε στοιχείο x_i, $\forall i = 1...\mathrm{N}$ ως εξής:

$j^* : \min_{j=1...K}\{\mathrm{D}(\mathrm{x}_i, \mathrm{c}_j)\}$ και έτσι $C_{j^*} = C_{j^*} \cup \{\mathrm{x}_i\}$ αναθέτοντάς το με αυτό τον τρόπο στο κοντινότερο κεντροειδές.

5. Ανανέωσε τα κεντροειδή κάθε ομάδας $c_j^{(new)} = mean\{\mathrm{C}_j\}$.
6. *Έως ότου* να παραμένουν σταθερά τα κεντροειδή κάθε ομάδας, δηλαδή μέχρι να ικανοποιηθεί η συνθήκη $\left|c_j^{(new)} - c_j^{(old)}\right| < \varepsilon$.

Πίνακας 2.3.2 Ψευδοκώδικας αλγορίθμου K-Μέσων

2.3.2. Μέτρα Ομοιότητας - Απόστασης

Τι εννοούμε όταν λέμε πως κάθε ομάδα περιλαμβάνει αντικείμενα που μοιάζουν μεταξύ τους; Πως ορίζουμε το κοντινότερο όταν αναθέτουμε τα σημεία στα κεντροειδή τους; Γενικά τα μέτρα ομοιότητας (πόσο όμοια είναι δύο αντικείμενα μεταξύ τους με βάση κάποιο κριτήριο) ή αντίστροφα τα μέτρα απόστασης (πόσο απέχουν δυο αντικείμενα μεταξύ τους) όπως αλλιώς καλούνται που χρησιμοποιούνται είναι πολλά. Έτσι για διανύσματα x, y του n-διάστατου διανυσματικού χώρου $\mathbb{R}^n$ θα έχουμε:

- Ευκλείδεια απόσταση:

$$D(x, y) = \sqrt{\sum_{i=1}^{n}(\mathrm{x}_i - \mathrm{y}_i)^2} \qquad \text{Εξ. 2.3.1}$$

- Απόσταση Manhattan:

$$D(x, y) = \sum_{i=1}^{n}\left|x_i - y_i\right| \qquad \text{Εξ. 2.3.2}$$

- Μέγιστη διαφορά μεταξύ όλων των διαστάσεων:

$$D(x,y) = \max_{i=1}^{n} |x_i - y_i| \qquad \text{Εξ. 2.3.3}$$

- Συνημιτονοειδής ομοιότητα:

$$D(x,y) = \cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \qquad \text{Εξ. 2.3.4}$$

Το πιο γνωστό από τα παραπάνω μέτρα απόστασης είναι η Ευκλείδεια απόσταση και είναι αυτό που χρησιμοποιείται για χάρη απλότητας και κυρίως όταν τα δεδομένα αναπαριστούν σημεία στο χώρο. Οι τιμές που μπορεί να έχει η Ευκλείδεια απόσταση είναι θετικές ή 0 όταν δύο σημεία ταυτίζονται (τετριμμένη περίπτωση). Αντίθετα, η συνημιτονοειδής ομοιότητα χρησιμοποιείται κυρίως κατά την ομαδοποίηση κειμένων και οι πιθανές τιμές που μπορεί να πάρει ανήκουν στο διάστημα [-1,1]. Συνημιτονοειδής ομοιότητα ίση με -1 σημαίνει τελείως ανόμοια ενώ ίση με 1 ακριβώς όμοια.

Η συνάρτηση απόστασης είναι απαραίτητη για τον ορισμό της αντικειμενικής συνάρτησης. Η αντικειμενική συνάρτηση εκφράζει το στόχο του αλγορίθμου ο οποίος είναι η βελτιστοποίηση της, είτε με την ελαχιστοποίηση είτε με τη μεγιστοποίησή της ανάλογα με τη χρησιμοποιούμενη συνάρτηση απόστασης. Στην συνήθη περίπτωση που χρησιμοποιείται η Ευκλείδεια απόσταση, η αντικειμενική συνάρτηση που θέλουμε να ελαχιστοποιηθεί είναι η *συνάρτηση τετραγωνικού σφάλματος (Sum Of Squared Errors – SSE)*. Αυτή εκφράζεται ως εξής:

$$\text{SSE} = \sum_{i=1}^{N} \min_{j=1\ldots k} \left\| x_i - c_j \right\|^2 = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - c_j \right\|^2 \qquad \text{Εξ. 2.3.5}$$

και θέλουμε την ελαχιστοποίησή της, δηλαδή το $\min_{c_j} \text{SSE}$. Το 4ο βήμα του αλγορίθμου κατά το οποίο τα σημεία ανατίθενται στα κοντινότερα προς αυτά κέντρα ελαχιστοποιεί το τοπικό

SSE κάθε ομάδας το οποίο αναπαρίσταται από την ποσότητα $\sum_{x_i \in C_j} \left\| x_i - c_j \right\|^2 \quad \forall j = 1...\mathrm{K}$. Στη συνέχεια το 5ο βήμα κατά το οποίο υπολογίζονται εκ νέου τα κεντροειδή ελαχιστοποιεί το ολικό SSE. Τα βήματα του αλγορίθμου K-Μέσων δεν εγγυώνται πως θα καταλήξουν στην εύρεση του ολικού ελάχιστου του συνολικού SSE, καθώς εξαρτώνται από την αρχικοποίηση του αλγορίθμου και μπορεί να σταματήσουν σε κάποιο τοπικό ελάχιστο.

Αυτό αποτελεί ένα από τα προβλήματα του βασικού αλγορίθμου K-Μέσων τα οποία θα αναλυθούν στη συνέχεια και τα οποία έδωσαν έναυσμα στη δημιουργία μιας οικογένειας αλγορίθμων βασισμένων στον αλγόριθμο K-Μέσων με σκοπό την αντιμετώπισή τους. Παρ' όλα αυτά ο κλασσικός αλγόριθμος χρησιμοποιείται ακόμη ευρέως στις μέρες μας κυρίως λόγω της μικρής χρονικής και χωρικής πολυπλοκότητάς του. Αν συμβολίσουμε με d το πλήθος των χαρακτηριστικών του συνόλου δεδομένων και με I τον αριθμό των επαναλήψεων που απαιτούνται για τη σύγκλιση του αλγορίθμου, τότε και από τον πίνακα 2.3 θα έχουμε:

Χρονική Πολυπλοκότητα : $O(I * N * K * d)$ Εξ. 2.3.6

Χωρική Πολυπλοκότητα: $O((N + K) * d)$ Εξ. 2.3.7

2.4. Αλγόριθμος K-Μέσων Αυξητικής Ομαδοποίησης

Μια παραλλαγή του κλασσικού αλγορίθμου κ-μέσων αποτελεί ο *K-Μέσων αυξητικής ομαδοποίησης (incremental – bisecting k-means)*. Η παραλλαγή αυτή διαφέρει από την κλασσική προσέγγιση στο ότι οι ομάδες κατασκευάζονται με αυξητικό τρόπο και στη βιβλιογραφία αναφέρεται πως παρουσιάζει καλύτερη συμπεριφορά [6] [7]. Παρακάτω παρατίθεται ο ψευδοκώδικας του αλγορίθμου:

Αλγόριθμος K-Μέσων Αυξητικής Ομαδοποίησης (Incremental Bisecting K-Means)
1. Αρχικά έχουμε $\kappa = 1$ ομάδα με κέντρο το c_1, το οποίο αποτελεί το κεντροειδές όλων των στοιχείων x_i του συνόλου δεδομένων X, $\forall i = 1...\mathrm{N}$.

2. *Επανάλαβε όσο* κ < K
3. Εύρεση της ομάδας j που θα διαχωριστεί με τη χρήση κριτηρίου.
4. Διαχωρίζουμε την ομάδα j σε 2 ομάδες με τη χρήση του κλασσικού αλγορίθμου κ-μέσων (μόνο στα σημεία $x_i \in C_j$ και δημιουργούνται 2 νέα κέντρα τα c_j και $c_{\kappa+1}$).
5. Προαιρετικά εφαρμογή του K-Μέσων σε όλα τα δεδομένα για βελτίωση των υπαρχουσών αναθέσεων.
6. κ = κ +1.
7. *Τέλος επανάληψης.*

Πίνακας 2.4 Ψευδοκώδικας αλγορίθμου K-Μέσων Αυξητικής Ομαδοποίησης

Κατά το 3ο βήμα του αλγορίθμου για τον καθορισμό της ομάδας προς διαχωρισμό απαιτείται η χρήση ενός κριτηρίου. Το κριτήριο αυτό μπορεί να είναι απλό όπως η επιλογή της ομάδας C_j με το μεγαλύτερο πλήθος στοιχείων N_j $\forall j = 1...\kappa$ με $\kappa < \mathrm{K}$. Μπορεί επίσης να επιλεχθεί η ομάδα με το μεγαλύτερο τετραγωνικό σφάλμα, δηλαδή $C_j : \max_j SSE\{\mathrm{C}_j\}$ $\forall j = 1...\kappa$ με $\kappa < \mathrm{K}$. Επιπρόσθετα μπορεί να χρησιμοποιηθεί οποιοσδήποτε συνδυασμός των δυο αυτών κριτηρίων.

Η κλασική εκδοχή του αλγορίθμου K-Μέσων προσπαθεί να βελτιώσει τη λύση στην οποία έχει καταλήξει το προηγούμενο βήμα και αυτό τη λύση του προηγούμενου βήματος κ.ο.κ. οπότε προκύπτει μεγάλη εξάρτηση από την αρχική διαμέριση των κεντροειδών στα δεδομένα η οποία πολλές φορές είναι τυχαία. Στην εκδοχή αυξητικής ομαδοποίησης όμως, η εξάρτηση προκύπτει από τη σειρά επεξεργασίας των δεδομένων από τον αλγόριθμο και τους διαχωρισμούς κάθε βήματος οπότε και αντιμετωπίζεται σε μεγάλο βαθμό αυτό το πρόβλημα του.

Ανάλογη ιδέα ακολουθεί και ο *αλγόριθμος K-Μέσων συσσωρευτικής ομαδοποίησης (agglomerative k-means)* ο οποίος ακολουθεί την αντίθετη πορεία με τον αλγόριθμο αυξητικής ομαδοποίησης. Αρχικά ο αλγόριθμος ξεκινά με πλήθος ομάδων κ ίσο με το πλήθος των στοιχείων του συνόλου των δεδομένων ή με μικρές ομάδες στοιχείων με κ > K και σε κάθε βήμα μειώνονται κατά 1 οι ομάδες έως ότου να γίνουν K στον αριθμό. Σε κάθε βήμα

συντελείται μια ένωση μεταξύ δυο ομάδων C_i, C_j με $i \neq j < \kappa$ στην ομάδα C_i με τη χρήση των κριτηρίων που παρουσιάστηκαν παραπάνω με μια μικρή παραλλαγή ώστε να ταιριάζουν με τη συσσωρευτική φύση του αλγορίθμου. Πιο συγκεκριμένα, θα επιλεχθεί η ένωση που θα παράξει την ομάδα με το μικρότερο πλήθος στοιχείων ή με το χαμηλότερο SSE ή συνδυασμούς αυτών.

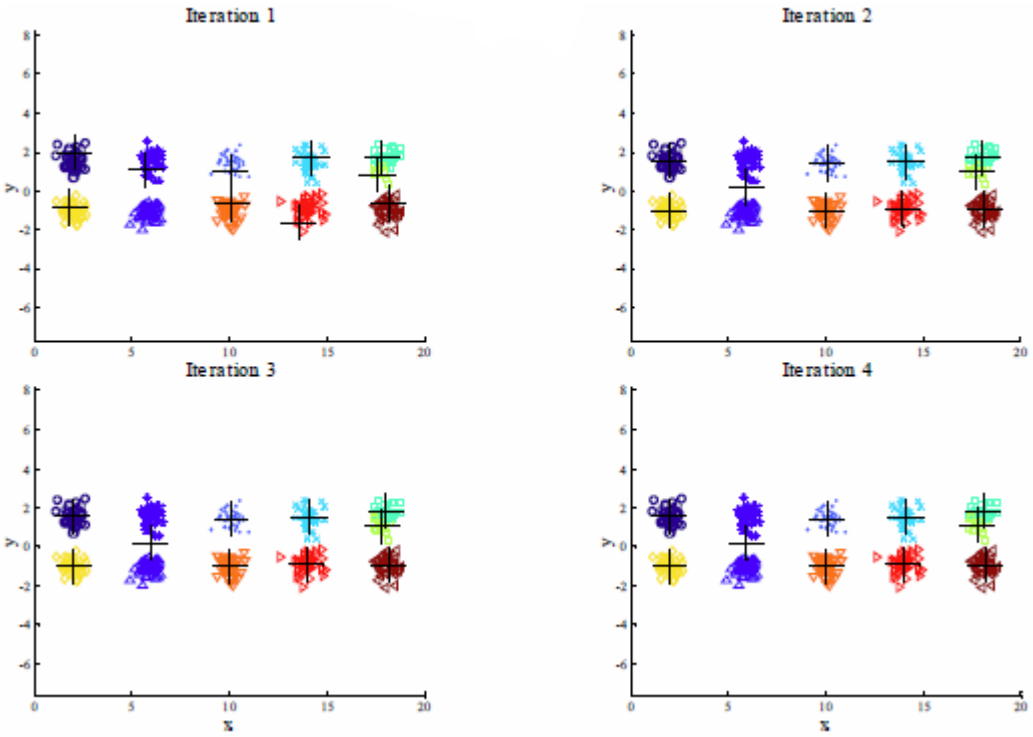
2.5. Αδυναμίες – Προβλήματα του Αλγορίθμου K-Μέσων

Παρόλη τη δημοτικότητά του και την ευρεία χρήση του σε ποικίλα πεδία εφαρμογών, ο κλασσικός αλγόριθμος K-Μέσων παρουσιάζει κάποιες αδυναμίες και περιορισμούς οι οποίες έχουν μεγάλη επίδραση στην ποιότητα της λύσης ομαδοποίησης που παράγεται. Σε κάποιες έχουν βρεθεί τρόποι αντιμετώπισής τους, όμως άλλες αδυναμίες παραμένουν και σήμερα αντικείμενο έρευνας και μελέτης ώστε να αντιμετωπιστούν.

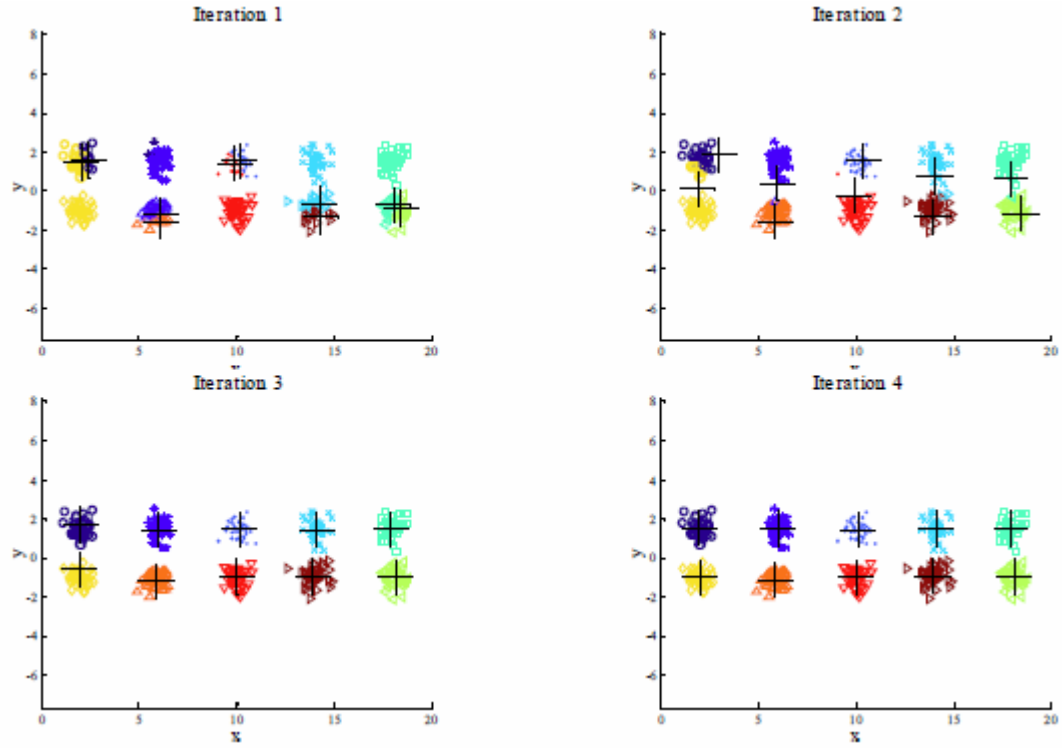
Το πρώτο και ένα από τα βασικότερα προβλήματα του αλγορίθμου K-Μέσων όπως αναφέρθηκε και παραπάνω είναι η εξάρτησή του από την αρχικοποίηση. Σε περιπτώσεις που έχουμε γνώση και εποπτεία της δομής του συνόλου των δεδομένων αυτό μπορεί να γίνει χειροκίνητα και να έχει ως αποτέλεσμα την αναμενόμενη λειτουργία του αλγορίθμου. Στις περισσότερες περιπτώσεις όμως αυτό δεν είναι εφικτό και χρησιμοποιούνται οι παρακάτω μέθοδοι:

- **Τυχαία Αρχικοποίηση**. Επιλέγονται K αρχικά κέντρα από τα αντικείμενα του συνόλου δεδομένων με τυχαίο τρόπο. Απαιτείται η εκτέλεση πολλών επαναλήψεων για την επιλογή της βέλτιστης αρχικής διαμέρισης.
- **Μέσες Τιμές**. Το σύνολο των δεδομένων χωρίζεται σε K ισομεγέθη διαστήματα ή περιοχές ενός πλέγματος ανάλογα με τη διάσταση των δεδομένων και επιλέγονται οι μέσες τιμές αυτών των διαστημάτων/περιοχών για κέντρα.
- **Απομακρυσμένα Σημεία**. Πολλές φορές επιλέγονται τα πιο απομακρυσμένα σημεία από τα υπάρχοντα κέντρα. Δηλαδή αρχικά επιλέγεται ένα στοιχείο με ακραίες τιμές για το 1^{ο} κέντρο και στη συνέχεια επιλέγονται τα υπόλοιπα έτσι ώστε να μεγιστοποιείται η απόσταση από τα ήδη υπάρχοντα κέντρα του συνόλου δεδομένων.

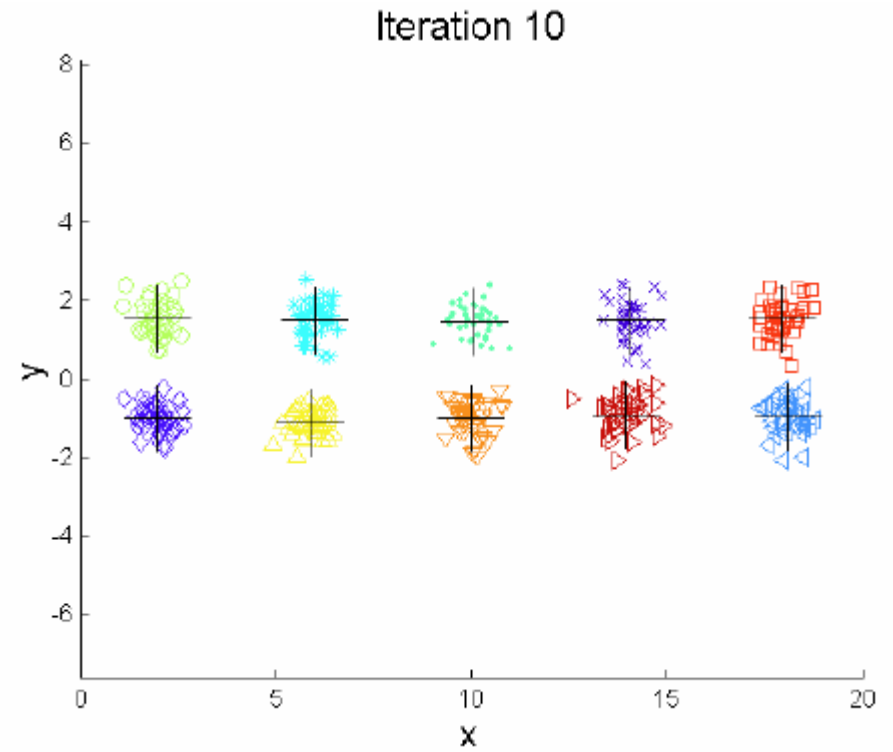
Όπως αναφέρθηκε νωρίτερα, μια λύση στο πρόβλημα της αρχικοποίησης δίνει η χρησιμοποίηση της αυξητικής ή της συσσωρευτικής εκδοχής του αλγορίθμου. Στα παρακάτω σχήματα (Σχήμα 2.5.1 & Σχήμα 2.5.2) παρουσιάζεται το πρόβλημα της αρχικοποίησης. Πιο συγκεκριμένα στο Σχήμα 2.5.1 αρχικοποιείται ένα σύνολο δεδομένων με δέκα κέντρα και σε 4 επαναλήψεις ο αλγόριθμος K-Μέσων έχει συγκλίνει σε λύση που δεν περιγράφει καλά στα δεδομένα, ενώ στο Σχήμα 2.5.2 παρουσιάζεται η ίδια περίπτωση με τη διαφορά πως έχει γίνει διαφορετική αρχικοποίηση ώστε ο αλγόριθμος να βρει το ολικό ελάχιστο. Τέλος στο Σχήμα 2.5.3 παρουσιάζεται το αποτέλεσμα του αυξητικού αλγόριθμου K-Μέσων και γίνεται εμφανές πως δεν επηρεάζεται από την αρχικοποίηση.



Σχήμα 2.5.1 K-Μέσων με μη βέλτιστη αρχικοποίηση των κέντρων



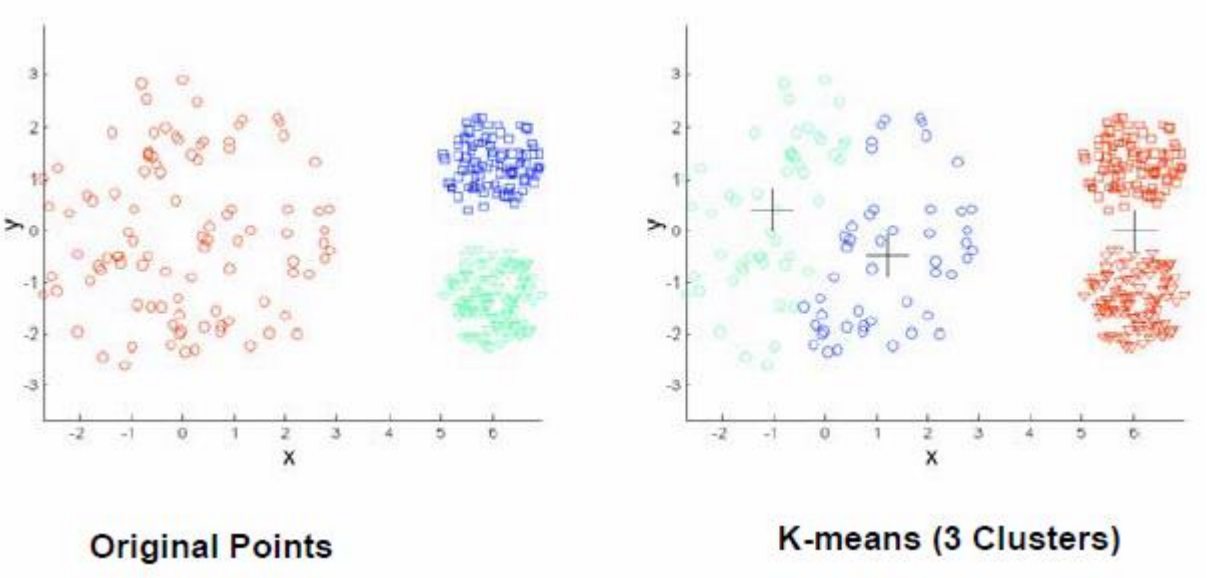
Σχήμα 2.5.2 K-Μέσων με βέλτιστη αρχικοποίηση των κέντρων



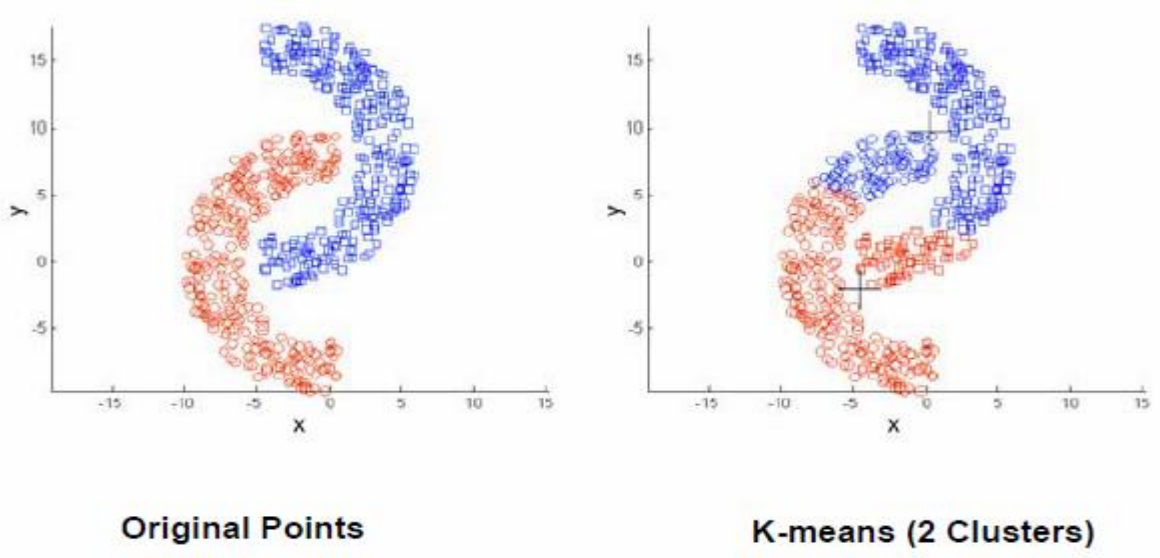
Σχήμα 2.5.3 K-Μέσων Αυξητικής Ομαδοποίησης ανεξάρτητος αρχικοποίησης

Ένα δεύτερο πρόβλημα και ίσως το πιο βασικό του κλασσικού αλγορίθμου K-Μέσων σχετίζεται με τον αριθμό των ομάδων. Ο χρήστης θα πρέπει να έχει πρότερη γνώση ή μια εποπτεία του συνόλου δεδομένων ώστε εκτός από την αρχικοποίηση των κέντρων να επιλέξει και τον κατάλληλο αριθμό των ομάδων. Το φαινόμενο αυτό μπορεί να μετριαστεί εν μέρει με *μετεπεξεργασία (post-processing)* της λύσης σε μερικές περιπτώσεις και όταν ο αριθμός των ομάδων που θεωρήσαμε είναι μεγαλύτερος από τον βέλτιστο και ενώσουμε τις ομάδες οι οποίες βρίσκονται στην ίδια περιοχή, όπου είναι εφικτό αυτό. Όμως αυτό είναι σπάνιο και ειδικά όταν έχουμε δεδομένα σε περισσότερες από 3 διαστάσεις πρακτικά αδύνατο.

Επόμενη σοβαρή αδυναμία του αλγορίθμου K-Μέσων είναι η υπόθεση σφαιρικών ομάδων ίσου μεγέθους. Αυτή η υπόθεση καθιστά δύσκολο να ανιχνευτούν "φυσικές" ομάδες, όταν οι ομάδες αυτές δεν έχουν σφαιρικό σχήμα ή ποικίλουν σε μέγεθος και πυκνότητα. Αυτό οφείλεται στη μορφή που έχει η αντικειμενική συνάρτηση που χρησιμοποιεί ο αλγόριθμος K-Μέσων αφού αυτή ελαχιστοποιείται από σφαιρικές, ίσου μεγέθους και καλά διαχωρισμένες ομάδες. Αυτό μπορεί όπως και στην προηγούμενη περίπτωση να μετριαστεί σε μερικές περιπτώσεις με ένωση των ομάδων στη λύση κατά τη μετεπεξεργασία. Στα σχήματα 2.5.4 και 2.5.5 παρουσιάζονται οι περιπτώσεις των ομάδων με διαφορετικό μέγεθος – πυκνότητα και των ομάδων που δεν έχουν σφαιρικό σχήμα αντίστοιχα.



Σχήμα 2.5.4 K-Μέσων με ομάδες διαφορετικού μεγέθους



Σχήμα 2.5.5 K-Μέσων σε μη σφαιρικές ομάδες

Αδυναμία επίσης του αλγορίθμου μπορεί να χαρακτηριστεί και η περίπτωση εμφάνισης άδειων ομάδων κατά τη διαδικασία της ομαδοποίησης όταν στο βήμα της ανάθεσης σημείων στα κέντρα δεν ανατεθεί κανένα σημείο σε κάποιο κέντρο και θα πρέπει να βρεθεί αντικαταστάτης του κέντρου αυτού. Μια προσέγγιση θα μπορούσε να είναι η επιλογή του σημείου που βρίσκεται πιο μακριά από τα υπάρχοντα κέντρα ως τον αντικαταστάτη. Επίσης μπορεί να βρεθεί αντικαταστάτης του σημείου αυτού στην ομάδα με το μεγαλύτερο SSE ώστε αυτή να χωριστεί σε 2 μικρότερες καταλήγοντας σε μικρότερο συνολικό SSE.

Τέλος η ύπαρξη ακραίων τιμών (outliers) σε ένα σύνολο δεδομένων μπορεί να έχει αρκετά αρνητικά αποτελέσματα. Αυτές οι ακραίες τιμές δημιουργούνται πολλές φορές από λάθη είτε στις μετρήσεις, είτε στην είσοδο των δεδομένων ή είναι απλά αδιάφορες τιμές που όμως μπορούν να επηρεάσουν την τοποθέτηση των κεντροειδών και να καταλήξουν σε μεγαλύτερο SSE. Οι ακραίες αυτές τιμές μπορούν να απαλειφθούν κατά τη διαδικασία της προεπεξεργασίας, πολλές φορές όμως ανάλογα με το πρόβλημα είναι απαραίτητη η ύπαρξή τους όπως για παράδειγμα η ύπαρξη πελατών που μπορεί να αποφέρουν ιδιαίτερα ασυνήθιστο αλλά υψηλό κέρδος σε μια επιχείρηση. Ένας τρόπος αντιμετώπισής τους σε αυτή την περίπτωση θα μπορούσε να είναι η χρήση ενδιάμεσων (medoids) αντί κεντροειδών ως αντιπροσώπων όπου αυτό όμως είναι εφικτό.

Οι αλγόριθμοι που αναπτύχθηκαν κατά την παρούσα εργασία έχουν ως σκοπό την εξάλειψη των κυριότερων προβλημάτων και περιορισμών του αλγορίθμου Κ-Μέσων. Πιο συγκεκριμένα, στο κεφάλαιο 4 προτείνονται και μελετώνται τρεις αλγόριθμοι της οικογένειας Κ-Μέσων οι οποίοι αντιμετωπίζουν το πρόβλημα της αρχικοποίησης, του πλήθους Κ των ομάδων και τους περιορισμούς ως προς το σχήμα και την πυκνότητα της κάθε ομάδας. Προηγουμένως, στο κεφάλαιο 3 θα παρουσιαστούν υπάρχουσες συναφείς εργασίες που προσπαθούν να επιλύσουν κάποια από τα παραπάνω ζητήματα.

ΚΕΦΑΛΑΙΟ 3. ΟΜΑΔΟΠΟΙΗΣΗ ΚΑΙ ΕΚΤΙΜΗΣΗ ΤΟΥ ΑΡΙΘΜΟΥ ΤΩΝ ΟΜΑΔΩΝ

3.1. Εισαγωγή

Τα προβλήματα του αλγορίθμου Κ-Μέσων που παρουσιάστηκαν στο προηγούμενο κεφάλαιο έχουν απασχολήσει αρκετά την επιστημονική κοινότητα. Τα τελευταία 15 χρόνια έχει ξεκινήσει να αναπτύσσεται ο σχεδιασμός αλγορίθμων της οικογένειας Κ-Μέσων που έχουν ως σκοπό τη βελτίωση του κλασικού αλγορίθμου για την εξάλειψη των αδυναμιών που παρουσιάζει και κυρίως της αδυναμίας που συνίσταται στην αρχική γνώση του αριθμού των ομάδων.

Η αδυναμία αυτή του κλασσικού αλγορίθμου είναι για πολλούς η σημαντικότερη καθώς υποθέτει από το χρήστη πρότερη γνώση της δομής των δεδομένων, κάτι που πολλές φορές δεν υπάρχει. Έτσι, λόγω της παραπάνω ανάγκης, έχουν αναπτυχθεί αλγόριθμοι που εκτιμούν αυτόματα τον αριθμό των ομάδων χωρίς κάποια είσοδο ή βοήθεια από το χρήστη, με βάση κάποιο στατιστικό κριτήριο.

Οι κυριότεροι αλγόριθμοι αυτής της κατηγορίας οι οποίοι προϋπήρχαν στη βιβλιογραφία των αλγορίθμων που αναπτύχθηκαν στο πλαίσιο αυτής της μεταπτυχιακής εργασίας, είναι οι *X-Means*, *G-Means* και *Dip-Means*. Αυτοί θα αναπτυχθούν αναλυτικά

παρακάτω, παρουσιάζοντας τη φιλοσοφία της προσέγγισης του καθενός στην αντιμετώπιση της έλλειψης γνώσης του "σωστού" αριθμού των ομάδων του υποκείμενου συνόλου δεδομένων.

3.2. Αλγόριθμος X-Means

Ο πρώτος αλγόριθμος της κατηγορίας αυτής παρουσιάστηκε το 2000 από τους D. Pelleg και Andrew Moore και ονομάζεται *x-means* [8]. Το κύριο έναυσμα για την επινόησή του αποτελεί όπως αναφέρθηκε και παραπάνω η εξάλειψη της ανάγκης του κλασικού αλγορίθμου K-Μέσων για την αρχική εισαγωγή του αριθμού των ομάδων από το χρήστη.

Ο x-means χρησιμοποιεί το στατιστικό κριτήριο *BIC (Bayesian Information Criterion)* [9], το οποίο προσπαθεί να βελτιστοποιήσει σε κάθε επανάληψη του αλγορίθμου καθώς αναζητά στο χώρο των ομάδων για την εύρεση του κατάλληλου αριθμού των ομάδων που περιγράφει καλύτερα τη δομή τους. Επιπλέον έχουν χρησιμοποιηθεί και άλλα κριτήρια [10] όπως η μετρική *AIC (Akaike Information Criterion)* σε συνδυασμό με τον αλγόριθμο x-means τα οποία όμως δίνουν ελαφρώς κατώτερα αποτελέσματα από το κριτήριο BIC. Στην ενότητα αυτή θα παρουσιαστεί ο x-means με τη χρήση του BIC κριτηρίου το οποίο χρησιμοποιήθηκε κατά τη διεξαγωγή των πειραμάτων. Ο παρακάτω πίνακας (Πίνακας 3.2.1) περιέχει τους συμβολισμούς που θα χρησιμοποιηθούν στην ενότητα αυτή για την περιγραφή του αλγορίθμου.

Συμβολισμός	Περιγραφή
D	Το σύνολο δεδομένων
x	Ένα σημείο/αντικείμενο του συνόλου δεδομένων
D_i	Η i-οστή ομάδα η οποία έχει ως κέντρο το $\mu_{(i)}$
$\mu_{(i)}$	Το κέντρο που σχετίζεται με το i-οστό στοιχείο
μ_j	Οι συντεταγμένες του j-οστού κέντρου
R_i	Το πλήθος των στοιχείων της i-οστής ομάδας
R	Το πλήθος των στοιχείων όλου του συνόλου δεδομένων

K	Ο αριθμός των ομάδων
M	Το πλήθος των διαστάσεων
M_j	Το j-οστό μοντέλο
$\Sigma = diag(\sigma^2)$	Ο Γκαουσιανός πίνακας συνδιακύμανσης

Πίνακας 3.2 Συμβολισμοί αλγορίθμου X-Means

3.2.1. Εύρεση του Κατάλληλου Μοντέλου

Ο x-means είναι ένας επαναληπτικός αλγόριθμος αυξητικής φύσης που αποτελεί ‘‘κάλυμμα’’ γύρω από τον κλασσικό αλγόριθμο κ-μέσων κάτι που σημαίνει πως περιέχει αυτόν ή τμήμα του κατά τη διαδικασία των επαναλήψεων. Αρχικά δίνουμε σαν είσοδο στον αλγόριθμο εκτός του συνόλου δεδομένων και ένα εύρος τιμών μέσα στο οποίο θα ψάξει για τον αριθμό των ομάδων *K* που περιγράφει καλύτερα τα δεδομένα, ξεκινώντας από το κάτω όριο του εύρους αυτού και προσθέτοντας αυξητικά με διαχωρισμούς καινούρια κεντροειδή έως ότου να φτάσει το άνω όριο. Από την παραπάνω διαδικασία ο αλγόριθμος παράγει τον αριθμό *K* των ομάδων καθώς και το σύνολο των κεντροειδών που αντιστοιχούν σε αυτό, τα οποία έχουν επιτύχει το καλύτερο σκορ κατά τις επαναλήψεις του αλγορίθμου. Η παραπάνω έξοδος του αλγορίθμου αποτελεί στην ουσία το κατάλληλο μοντέλο το οποίο έχει δημιουγήσει ως λύση και το οποίο επιλέγεται μέσα από ένα πλήθος $\left|K_{\max} - K_{\min}\right|$ μοντέλων. Παρακάτω παρατίθεται ο ψευδοκώδικας που περιγράφει τον x-means:

Αλγόριθμος X-Μέσων (X-Means)
1. Είσοδος του εύρους $[K_{\min}, K_{\max}]$ στο οποίο θα γίνει αναζήτηση του K. 2. Αρχικοποίηση των $K_{\min}$ κέντρων μ_i, $\forall i = 1...\mathrm{K}_{\min}$ με $K_{\min} >= 1$. 3. *Επανάλαβε* 4. Βελτίωση Παραμέτρων Μοντέλου. 5. Βελτίωση Δομής Μοντέλου. 6. *Έως ότου* $K > K_{\max}$.

7. Επέστρεψε το μοντέλο με το καλύτερο σκορ.

Πίνακας 3.2.1 Ψευδοκώδικας αλγορίθμου X-Means

Όπως φαίνεται στον παραπάνω πίνακα, κάθε επανάληψη του αλγορίθμου αποτελείται από δύο βήματα, τη βελτίωση παραμέτρων του μοντέλου και τη βελτίωση της δομής του έως ότου να εξεταστούν τα παραγόμενα μοντέλα για όλες τις πιθανές τιμές του K στο εύρος $[K_{\min}, K_{\max}]$. Η βελτίωση παραμέτρων του μοντέλου είναι μια απλή διαδικασία καθώς αποτελείται από μια εκτέλεση του κλασσικού αλγορίθμου K-Μέσων μέχρι αυτός να συγκλίνει.

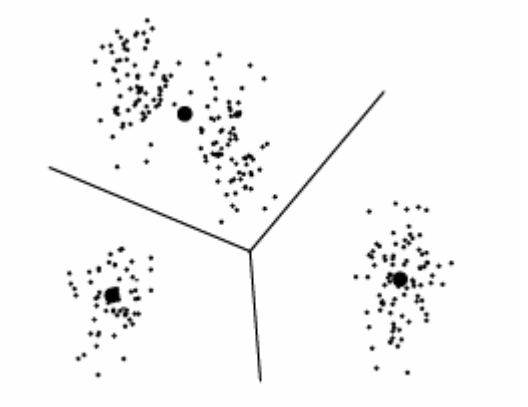
Έπειτα, στο βήμα 5 συντελείται βελτίωση της δομής του μοντέλου, αποφασίζεται δηλαδή εάν θα προστεθούν νέα κεντροειδή και σε ποιες θέσεις. Αυτό επιτυγχάνεται χρησιμοποιώντας κάποια στρατηγική ώστε να αποφασιστεί ποια από τα υπάρχοντα κεντροειδή θα χωριστούν και ποια όχι. Η στρατηγική που χρησιμοποιήθηκε στον x-means δανείστηκε ιδέες από δύο βασικές στρατηγικές διαχωρισμού κεντροειδών ώστε να αποκομίσουν τα οφέλη από καθεμιά και να αποφύγουν τις αδυναμίες τους.

Η πρώτη βασική στρατηγική για το διαχωρισμό κεντροειδών προτείνει έναν διαχωρισμό τη φορά. Αυτό σημαίνει πως επιλέγεται ένα κεντροειδές κάθε φορά και τοποθετείται ένα νέο σε κοντινή θέση. Στη συνέχεια εκτελείται ο αλγόριθμος K-Μέσων και βλέπουμε εάν το παραγόμενο αποτέλεσμα πετυχαίνει καλύτερο σκορ σε σχέση με το προηγούμενο μοντέλο. Εάν ισχύει κάτι τέτοιο, τότε κρατάμε το νέο κεντροειδές, αλλιώς επιστρέφουμε στην προηγούμενη δομή. Αυτή η προσέγγιση θα χρειαζόταν $O(K_{\max})$ χρόνο μέχρι την ολοκλήρωση του x-means. Επιπλέον εάν ένας διαχωρισμός δεν παράγει καλύτερο σκορ από το ήδη υπάρχον, αυτό σημαίνει πως θα πρέπει να επιλεχθεί ένα άλλο κεντροειδές για διαχωρισμό, κάτι που στη χειρότερη περίπτωση θα απαιτούσε την εξέταση όλων δηλαδή τη δημιουργία τετραγωνικής πολυπλοκότητας σε αυτή την περίπτωση.

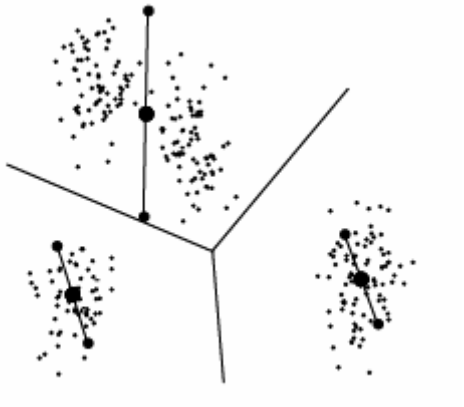
Η δεύτερη βασική στρατηγική ακολουθεί μια πιο επιθετική προσέγγιση για το διαχωρισμό των κεντροειδών, η οποία χρησιμοποιήθηκε από τους Wasserman και Moore. Εδώ επιλέγονται αρχικά τα μισά κεντροειδή από το υπάρχον μοντέλο χρησιμοποιώντας κάποιο ευρετικό κριτήριο, στη συνέχεια διαχωρίζονται όλα, τρέχει μια φορά ο αλγόριθμος K-

Μέσων σε αυτά και εξετάζεται αν το παραγόμενο μοντέλο παράγει καλύτερο σκορ από το προηγούμενο και αν ναι τότε αποδεχόμαστε το νέο μοντέλο. Η προσέγγιση αυτή απαιτεί μόνο $O(\log K_{\max})$ χρόνο μέχρι την ολοκλήρωση του x-means. Οι αδυναμίες της όμως προκύπτουν αφενός από την προϋπόθεση της ύπαρξης του ευρετικού κριτηρίου και αφετέρου από την αδυναμία εξέτασης της περίπτωσης που σε μια επανάληψη είναι επιθυμητός ο διαχωρισμός ενός ή δύο κεντροειδών αντί για τα μισά.

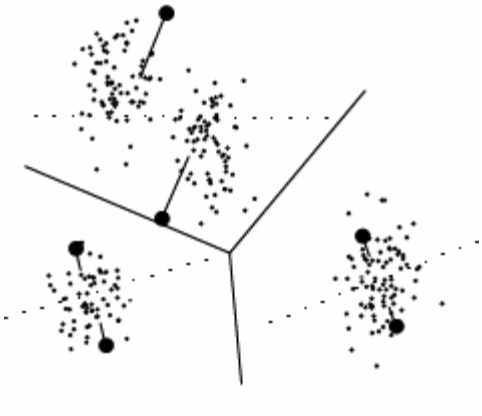
Η στρατηγική που δοκιμάστηκε από τις δύο παραπάνω ακολουθεί τα εξής βήματα: αρχικά, έστω ότι έχουμε μια διαμέριση με 3 κεντροειδή (Σχήμα 3.2.1). Επιλέγουμε κάθε κεντροειδές αυτής και το χωρίζουμε σε άλλα δύο τα οποία ισαπέχουν από αυτό και έχουν απόσταση ίση με το μέγεθος της αρχικής ομάδας (οποιασδήποτε από τις δυο κατευθύνσεις) προς μια τυχαία κατεύθυνση (Σχήμα 3.2.2). Έπειτα, σε κάθε μια από τις περιοχές που υπάρχουν, τρέχει τοπικά ο αλγόριθμος K-Μέσων με (K=2), δηλαδή χρησιμοποιούνται μόνο τα δυο νέα παραγόμενα κεντροειδή. Αυτό το βήμα μπορεί να παραλληλοποιηθεί μιας και ο κάθε K-Μέσων τρέχει τοπικά. Στο σχήμα 3.2.3 απεικονίζεται το αποτέλεσμα που παράγει ο αλγόριθμος K-Μέσων με τις γραμμές να συμβολίζουν τη μετακίνηση των νέων κεντροειδών. Στο σχήμα 3.2.4 παρουσιάζεται το μοντέλο που παράγεται από τον αλγόριθμο κ-μέσων με τα νέα κεντροειδή μόνο. Στη συνέχεια υπολογίζεται το σκορ που προκύπτει από το τεστ σε κάθε τέτοια περιοχή για K=1 (το αρχικό κεντροειδές) και για K=2 (τα δύο τελικά), επιλέγεται το τοπικό μοντέλο με το υψηλότερο σκορ, το οποίο αποτελεί και μέρος πλέον του συνολικού (Σχήμα 3.2.5). Έτσι, εξετάζονται όλες οι 2^K πιθανές περιπτώσεις και διαχωρίζονται μόνο τα κεντροειδή που δεν αντιπροσωπεύουν “καλά” το υποκείμενο υποσύνολο δεδομένων.



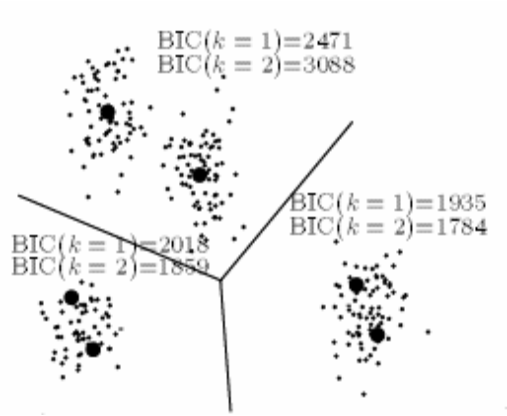
Σχήμα 3.2.1 Διαχωρισμός κέντρων X-Means 1



Σχήμα 3.2.2 Διαχωρισμός κέντρων X-Means 2



Σχήμα 3.2.3 Διαχωρισμός κέντρων X-Means 3



Σχήμα 3.2.4 Διαχωρισμός κέντρων X-Means 4

Σχήμα 3.2.5 Διαχωρισμός κέντρων X-Means 5

3.2.2. Στατιστικό Κριτήριο BIC

Για τον υπολογισμό του σκορ του κάθε μοντέλου M_j που παράγεται σε κάθε επανάληψη σε κάθε τοπική περιοχή του συνόλου δεδομένων είναι απαραίτητη η γνώση της εκ των υστέρων πιθανότητας $\Pr[M_j \mid D]$. Ο x-means υποθέτει πως όλα τα μοντέλα περιέχουν ομάδες που ακολουθούν σφαιρικές Γκαουσιανές κατανομές. Έτσι, για τη προσέγγιση της παραπάνω εκ των υστέρων πιθανότητας $\Pr[M_j \mid D]$ χρησιμοποιείται το στατιστικό κριτήριο BIC (Bayesian Information Criterion), μιας και αποφέρει καλύτερα αποτελέσματα από άλλα κριτήρια αυτής της κατηγορίας.

Το BIC σκορ κάθε μοντέλου M_j δίνεται από τον παρακάτω τύπο (Εξ. 3.2.1) όπου $\hat{l}_j(D)$ είναι η λογαριθμική πιθανοφάνεια των δεδομένων όσον αφορά το j-οστό μοντέλο και η οποία υπολογίστηκε στο σημείο μέγιστης πιθανοφάνειας. p_j είναι ο αριθμός των παραμέτρων που εμφανίζονται στο M_j, δηλ. οι συντεταγμένες των *M* * K κεντροειδών καθώς και η διακύμανσή.

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2} \cdot \log R \qquad \text{Εξ. 3.2.1}$$

Η μέγιστη πιθανοφάνεια *MLE (Maximum Likelihood Estimate)* για τη διακύμανση, υποθέτοντας σφαιρικές Γκαουσιανές κατανομές είναι:

$$\hat{\sigma}^2 = \frac{1}{R-K} \sum_i (\mathrm{x}_i - \mu_{(\mathrm{i})})^2 \qquad \text{Εξ. 3.2.2}$$

Οι πιθανότητες στα σημεία ορίζονται ως εξής:

$$\hat{P}(x_i) = \frac{R_{(i)}}{R} \cdot \frac{1}{\sqrt{2\pi}\hat{\sigma}^{\mathrm{M}}} \exp\left(-\frac{1}{2\hat{\sigma}^2} \left\| \mathrm{x}_i - \mu_{(\mathrm{i})} \right\|^2 \right) \qquad \text{Εξ. 3.2.3}$$

Έτσι η λογαριθμική πιθανοφάνεια όλου του συνόλου δεδομένων δίνεται από την:

$$l(D) = \log \prod_i P(x_i) =$$

$$\sum_i \left(\log \frac{1}{\sqrt{2\pi}\hat{\sigma}^{\mathrm{M}}} + \log \frac{R_{(i)}}{R} - \frac{1}{2\hat{\sigma}^2} \left\| \mathrm{x}_i - \mu_{(\mathrm{i})} \right\|^2 \right) \qquad \text{Εξ. 3.2.4}$$

Οπότε για τον υπολογισμό της λογαριθμικής πιθανοφάνειας μιας ομάδας D_n του συνόλου δεδομένων, με $n \in [1, \mathrm{K}]$, το οποίο έχει για κεντροειδές το $\mu_{(n)}$, όπως προκύπτει από τις παραπάνω σχέσεις θα είναι:

$$\hat{l}(D_n) = -\frac{R_n}{2}\log(2\pi) - \frac{R_n \cdot M}{2}\log(\hat{\sigma}^2)$$
$$-\frac{R_n - K}{2} + R_n \log R_n - R_n \log R \qquad \text{Εξ. 3.2.5}$$

Τέλος γίνεται φανερό ότι το κριτήριο BIC μπορεί να εφαρμοστεί και σε όλο το σύνολο των δεδομένων, πέρα από κάθε μεμονωμένο υποσύνολο D_n, με τον ίδιο τρόπο. Αυτό σημαίνει πως υπολογίζεται το σκορ κάθε μοντέλου συνολικά ώστε να βρεθεί το καλύτερο με τον ίδιο τρόπο με τον οποίο υπολογίζεται το σκορ σε για κάθε τοπικό διαχωρισμό.

3.2.3. Πλεονεκτήματα και αδυναμίες του αλγορίθμου x-means

Συνολικά, ο αλγόριθμος x-means έχει σημαντικά πλεονεκτήματα αλλά και σοβαρές αδυναμίες λόγω του τρόπου λειτουργίας του και των υποθέσεων που κάνει. Η ανάγκη για πολύ γρήγορη ομαδοποίηση μεγάλου πλήθους δεδομένων χαμηλής διάστασης (συνήθως M<=5), καθόρισε σε μεγάλο βαθμό τη δομή και τη φιλοσοφία του αλγορίθμου.

Οι συγκρίσεις που γινόταν σε πειράματα μεταξύ του x-means και του κλασσικού αλγορίθμου κ-μέσων ήταν ενθαρρυντικές ως προς το χρόνο εκτέλεσης του. Θεωρήθηκε ένα εύρος στο οποίο γινόταν αναζήτηση από τον x-means για την πραγματική τιμή του K που περιγράφει καλύτερα τα δεδομένα και έγιναν συγκρίσεις με διαδοχικές επαναλήψεις του κλασσικού K-Μέσων για κάθε μία από αυτές τις τιμές του αριθμού ομάδων. Σε σύνολα δεδομένων με μεγάλο πλήθος στοιχείων ο x-means είχε σημαντικά μικρότερους χρόνους εκτέλεσης (έως και 2 φορές γρηγορότερος) κυρίως λόγω της τοπικής ιδιότητας που έχει για τον διαχωρισμό των κεντροειδών και της χρήσης caching τεχνικών ώστε να αποθηκεύονται και να επαναχρησιμοποιούνται παλαιότερα σκορ για τα μοντέλα όπου αυτό ήταν εφικτό.

Οι κύριες αδυναμίες που εμφανίζει ο αλγόριθμος x-means είναι δύο. Η πρώτη προκύπτει από τις υποθέσεις που κάνει για τη δομή του συνόλου των δεδομένων. Πιο συγκεκριμένα, υποθέτει πως κάθε ομάδα θα ακολουθεί αυστηρά μια σφαιρική Γκαουσιανή κατανομή. Αυτό έχει ως αποτέλεσμα όπως θα φανεί και σε επόμενες ενότητες, πως δεδομένα που ακολουθούν άλλες κατανομές όπως για παράδειγμα ομοιόμορφη, δεν μπορούν να μοντελοποιηθούν καλά από τον αλγόριθμο αυτό και σε πολλές περιπτώσεις δημιουργούνται σοβαρά λάθη στην ανάθεση των ομάδων και παρατηρούνται φαινόμενα overfitting. Η δεύτερη αδυναμία του έχει να κάνει και με τον τρόπο λειτουργίας του και τον τρόπο με τον οποίο εκμεταλλεύεται κοντινές σε απόσταση ομάδες για να κάνει το διαχωρισμό των κεντροειδών. Χρειάζεται λοιπόν, οι ομάδες εκτός από το να είναι σφαιρικές, να είναι καλά διαχωρισμένες, κάτι το οποίο δεν ισχύει σε πολλές περιπτώσεις. Έτσι όταν οι ομάδες δεν είναι καλά διαχωρισμένες, παρατηρούνται φαινόμενα overfitting τα οποία κατά περίπτωση μπορεί να οδηγήσουν στην εύρεση αριθμού ομάδων Κ πολλαπλάσιου του πραγματικού.

3.3. Αλγόριθμος G-Means

Ο αλγόριθμος *g-means* [12]. παρουσιάστηκε από τους G. Hamerly και C. Elkan το 2003 και, όπως ο x-means, προσπαθεί να επιλύσει το σοβαρότερο πρόβλημα του αλγορίθμου K-Μέσων, το οποίο είναι η εύρεση με αυτόματο τρόπο του κατάλληλου αριθμού των ομάδων Κ ο οποίος περιγράφει καλύτερα ένα σύνολο δεδομένων.

Ο g-means βασίζεται σε μια σχετικά απλή φιλοσοφία: αν περισσότερα του ενός κέντρα χρησιμοποιούνται για να περιγράψουν μια unimodal κατανομή (κατανομή της οποίας η πυκνότητα πιθανότητας εμφανίζει μια κορυφή, πχ. Γκαουσιανή), τότε το μοντέλο αυτό θεωρείται περίπλοκο και τη θέση αυτών των κέντρων την παίρνει ένα και μοναδικό κέντρο, το οποίο περιγράφει καλύτερα την δομή αυτής της ομάδας. Το στατιστικό κριτήριο το οποίο χρησιμοποιείται σε κάθε επανάληψή του g-means είναι το *Anderson-Darling* σε αντιδιαστολή με το κριτήριο BIC (Bayesian Information Criterion) το οποίο χρησιμοποιήθηκε στον αλγόριθμο x-means [13]. Ακολουθώντας λοιπόν σε ένα βαθμό τα αρχικά βήματα του αλγορίθμου x-means, ο g-means προσπαθεί να βελτιώσει τις όποιες αδυναμίες εμφανίστηκαν στον x-means και οι οποίες περιορίζουν την αποτελεσματικότητά του x-means μόνο σε

συγκεκριμένης μορφής και ιδιοτήτων σύνολα δεδομένων: σύνολα δεδομένα που αποτελούνται από "φυσικές" ομάδες που ακολουθούν αυστηρά σφαιρικές Γκαουσιανές κατανομές, κάτι το οποίο εμφανίζεται σπάνια σε πραγματικά δεδομένα. Στον πίνακα που παρατίθεται παρακάτω (Πίνακας 3.3.1) συγκεντρώνονται οι κυριότεροι συμβολισμοί που θα χρησιμοποιηθούν για την παρουσίαση του αλγορίθμου g-means στη συνέχεια.

Συμβολισμός	Περιγραφή
X	Το σύνολο δεδομένων
x_i	Ένα σημείο/αντικείμενο του συνόλου δεδομένων
C	Το σύνολο των κέντρων c_j με $j=1...K$
K	Το πλήθος των ομάδων
$\bar{X}$	Το κεντροειδές του συνόλου δεδομένων
$\{\bar{X}_j\}$	Το σύνολο των κεντροειδών
$x_{(i)}$	Το i-οστό διατεταγμένο στοιχείο
D	Το πλήθος των διαστάσεων
n	Το μέγεθος ενός υποσυνόλου του συνόλου δεδομένων

Πίνακας 3.3.1 Συμβολισμοί αλγορίθμου G-Means

3.3.1. Δομή του αλγορίθμου

Ο g-means όπως και ο x-means αποτελεί έναν επαναληπτικό αλγόριθμο αυξητικής φύσης ο οποίος βασίζεται στον κλασσικό αλγόριθμο K-Μέσων. Αυτό σημαίνει πως ξεκινά από ένα μικρό αριθμό κέντρων, και διαχωρίζει τα κέντρα όποτε κριθεί απαραίτητο βάσει του στατιστικού κριτηρίου που χρησιμοποιεί, αυξάνοντας έτσι τον αριθμό των ομάδων. Εάν δεν υπάρχει κάποια αρχική γνώση για τη δομή των ομάδων του συνόλου δεδομένων, τότε αρχικά θα έχουμε $K = 1$ κέντρο, δηλαδή ξεκινάμε το διαχωρισμό από το κεντροειδές όλων των σημείων. Σε κάθε επανάληψης μετά το διαχωρισμό μιας ομάδας, μεσολαβεί μια επανάληψη του αλγορίθμου K-Μέσων καθολικά σε όλο το σύνολο δεδομένων, ώστε να βελτιωθεί η τρέχουσα λύση. Παρακάτω παρατίθενται τα βήματα του αλγορίθμου με τη μορφή ψευδοκώδικα:

Αλγόριθμος G-Μέσων (G-Means)
1. Είσοδος του αρχικού συνόλου C των κέντρων και ενός επιπέδου εμπιστοσύνης α. Αν για κέντρα χρησιμοποιούνται κεντροειδή τότε $C \leftarrow \{\bar{X}_j\}$. Αν δεν υπάρχει πρότερη γνώση του συνόλου δεδομένων, τότε $C \leftarrow \bar{X}$. 2. *Επανάλαβε* 3. $C \leftarrow$ K -Μέσων (C, X). 4. Έλεγχος στατιστικού τεστ $\forall x_i$ με $x_i \in$ ομάδα j που έχει για κέντρο το c_j, αν ακολουθεί Γκαουσιανή κατανομή με επίπεδο εμπιστοσύνης a. 5. Εάν τα x_i αυτά ακολουθούν Γκαουσιανή κατανομή, τότε παραμένει το c_j κέντρο τους. Αλλιώς διαχώρισε το σε δύο νέα $c_j \rightarrow \{c_{j'}, c_{j''}\}$ και αντικατέστησε το παλιό. 6. *Έως ότου δεν προστίθενται νέα κέντρα.*

Πίνακας 3.3.2 Ψευδοκώδικας αλγορίθμου G-Means

Όπως ειπώθηκε νωρίτερα κατά την εκτέλεση του g-means εκτελείται και ο αλγόριθμος K-Μέσων σε κάθε επανάληψη καθολικά (σε αντίθεση με τον τοπικό χαρακτήρα του στον x-means). Αυτό έχει ως αποτέλεσμα η χρονική πολυπλοκότητά του να εξαρτάται άμεσα από τον αλγόριθμο K-Μέσων και πιο συγκεκριμένα να είναι K φορές πολλαπλάσιά του.

Η υπόθεση που κάνει ο g-means για την κατανομή των στοιχείων των ομάδων του συνόλου δεδομένων είναι λίγο διαφορετική από αυτή του κλασσικού αλγορίθμου K-Μέσων και του x-means. Χρησιμοποιείται η υπόθεση ότι τα δεδομένα κάθε ομάδας ακολουθούν πολυδιάστατη Γκαουσιανή κατανομή με πίνακα συνδιακύμανσης ο οποίος μπορεί να μην είναι σταθερός. Αυτό έχει ως συνέπεια να μοντελοποιούνται και ελλειπτικές δομές ομάδων κάτι που δεν είναι τόσο περιοριστικό όσο είναι οι αυστηρά σφαιρικές δομές. Επιπλέον λαμβάνεται υπόψη και το μέγεθος n του συνόλου δεδομένων κάτι που αποτρέπει ενδεχόμενη μη αποδεκτή συμπεριφορά του αλγορίθμου σε μικρά σύνολα.

3.3.2. Στατιστικό Κριτήριο Anderson-Darling

Το στατιστικό κριτήριο που εφαρμόζεται από τον g-means ονομάζεται Anderson-Darling (Stephens, 1974)[14]. Αποτελεί μια τροποποιημένη εκδοχή του γνωστού *Kolmogorov-Smirnov (K-S)* στατιστικού τεστ (Chakravart, Laha and Roy, 1967) με τη διαφορά ότι δίνει μεγαλύτερο βάρος στις ουρές της κατανομής από ότι το K-S [15].

Οι δύο εναλλακτικές υποθέσεις H_0 και H_1 του τεστ χρησιμοποιούνται για να καθοριστεί αν το δείγμα των δεδομένων που ελέγχεται ακολουθεί ή όχι Γκαουσιανή κατανομή. Οι υποθέσεις αυτές λοιπόν είναι:

- H_0: Τα σημεία x_i που αντιστοιχούν στο κέντρο c_j περιγράφονται από μια Γκαουσιανή κατανομή
- H_1: Τα σημεία x_i που αντιστοιχούν στο κέντρο c_j δεν μπορούν να περιγραφούν από μια Γκαουσιανή κατανομή

Αν το τεστ αποδεχθεί τη μηδενική υπόθεση H_0, τότε υποθέτουμε πως τα στοιχεία x_i γύρω από το κέντρο c_j περιγράφονται ικανοποιητικά από αυτό και παραμένει ως έχει. Αν όμως απορριφθεί η μηδενική υπόθεση και γίνει αποδεκτή η H_1, τότε το κέντρο αυτό δεν περιγράφει ικανοποιητικά τα δεδομένα και το διαχωρίζουμε σε δύο νέα $c_j \to \{c_{j'}, c_{j''}\}$.

Πριν την εφαρμογή του Anderson-Darling τεστ τα δεδομένα προβάλλονται σε μια διάσταση με διάφορους εναλλακτικούς τρόπους που αναφέρονται παρακάτω. Έπειτα, τα δεδομένα κανονικοποιούνται για να έχουν μέση τιμή ίση με 0 και διακύμανση 1 ώστε να μπορεί να εφαρμοστεί το τεστ σε αυτά. Τότε, υποθέτοντας ότι $z_i = F(\mathrm{x}_{(i)})$, με F να είναι η αθροιστική συνάρτηση κατανομής $N(0,1)$ ελέγχεται αν το αποτέλεσμα $A_*^2(Z)$ του τεστ ανήκει ή όχι στη μη κρίσιμη περιοχή τιμών που καθορίστηκε στην αρχή του αλγορίθμου με την εισαγωγή του επιπέδου εμπιστοσύνης α και ανάλογα κατατάσσεται στην υπόθεση H_0 ή H_1 αντίστοιχα. Οι παρακάτω σχέσεις (Εξ. 3.3.1 και Εξ. 3.3.2) μας δίνουν τη συνάρτηση που χρησιμοποιείται από το Anderson-Darling τεστ:

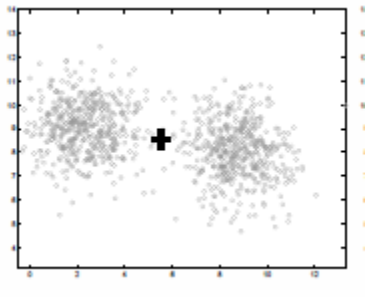
$$A^2(Z) = -\frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log(z_i) + \log(1 - z_{n+1-i})] - n \qquad \text{Εξ. 3.3.1}$$

$$A_*^2(Z) = A^2(Z)\left(1 + \frac{4}{n} - \frac{25}{n^2}\right) \qquad \text{Εξ. 3.3.2}$$

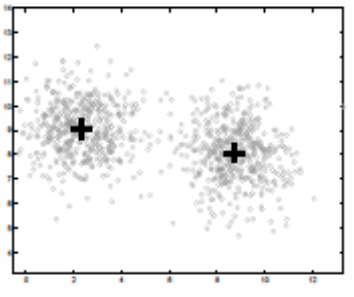
Όσον αφορά το διαχωρισμό των κέντρων που αναφέρθηκε νωρίτερα θεωρούνται δύο μέθοδοι. Η βασική τους ομοιότητα είναι πως και οι δυο προσεγγίσεις αρχικοποιούν τα κέντρα με τον εξής τρόπο: $\{c_{j'}, c_{j''}\} \leftarrow c_j \pm m$, όπου m υπολογίζεται σε κάθε περίπτωση με διαφορετικό τρόπο. Η πρώτη μέθοδος είναι η απλούστερη από τις δύο. Επιλέγεται το m ως ένα τυχαίο D-διάστατο (όσο και τα δεδομένα) διάνυσμα τέτοιο ώστε το μέτρο του να μη ξεπερνά το πιο απομακρυσμένο σημείο της ομάδας που έχει ως κέντρο το c_j προς την ίδια ή την αντίθετη κατεύθυνση. Η δεύτερη μέθοδος είναι αυτή που χρησιμοποιείται συχνότερα κατά την εκτέλεση του g-means και στηρίζεται στην ανάλυση κυρίων συνιστωσών (Principal Component Analysis - PCA). Τότε, αν s είναι η κύρια συνιστώσα που περιγράφει τα δεδομένα και λ η ιδιοτιμή της, θα είναι $m = s\sqrt{2\lambda/\pi}$. Με τη χρήση της PCA καταφέρνει να αιχμαλωτίσει την κατεύθυνση προς την οποία είναι "απλωμένα" τα δεδομένα και να τοποθετήσει τα νέα κέντρα σε αντιπροσωπευτικότερα σημεία. Να σημειωθεί πως μετά το πέρας κάθε διαχωρισμού δύναται να εκτελεσθεί ο αλγόριθμος κ-μέσων τοπικά σε αυτά τα δυο κέντρα ώστε να βελτιωθεί το τοπικό μοντέλο που αποτελείται από αυτά και στη συνέχεια να εφαρμοστεί το Anderson-Darling τεστ.

Στα παρακάτω σχήματα (Σχήμα 3.3.1 – 3.3.4) παρουσιάζεται πως ο αλγόριθμος g-means εκτελεί τον διαχωρισμό των κέντρων. Στο 1ο σχήμα παρουσιάζεται ένα κέντρο που αντιπροσωπεύει τα στοιχεία ενός συνόλου δεδομένων πριν εφαρμοστεί το Anderson-Darling τεστ. Στη συνέχεια, εφαρμόζεται το τεστ και ο αλγόριθμος απορρίπτει τη μηδενική υπόθεση H_0 και όπως φαίνεται στο σχήμα 3.3.2 ο g-means χωρίζει το κέντρο αυτό σε δύο νέα που αντιπροσωπεύουν ορθότερα βάσει των υποθέσεών του τα δεδομένα. Στα επόμενα σχήματα, 3.3.3 και 3.3.4 παρουσιάζονται περιπτώσεις στις οποίες παρατηρείται το φαινόμενο overfitting, δηλαδή περιπτώσεις που χρησιμοποιούνται περισσότερα κέντρα για την

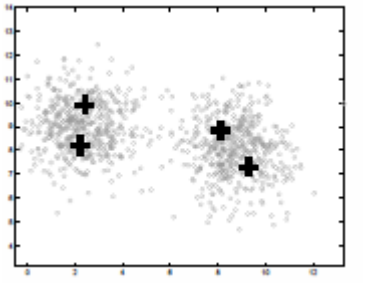
περιγραφή των δεδομένων και κατά συνέπεια περισσότερες ομάδες από όσες πραγματικά υπάρχουν.



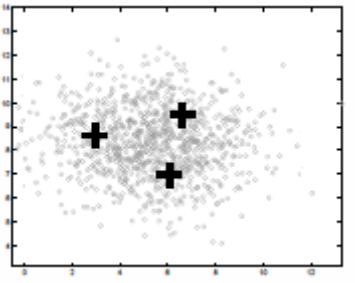
Σχήμα 3.3.1 Διαχωρισμός G-Means 1



Σχήμα 3.3.2 Διαχωρισμός G-Means 2



Σχήμα 3.3.3 Λάθος διαχωρισμός G-Means 1



Σχήμα 3.3.4 Λάθος διαχωρισμός G-Means 2

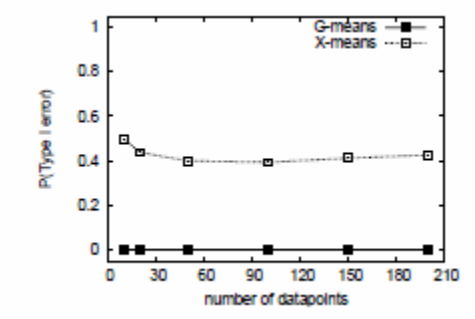
3.3.3. Πλεονεκτήματα και αδυναμίες του αλγορίθμου g-means

Ο αλγόριθμος g-means όπως και ο προκάτοχός του, ο x-means, παρουσιάζει αρκετά πλεονεκτήματα σε σχέση με τον κλασικό αλγόριθμο K-Μέσων, όμως έχει και κάποιες αδυναμίες που τον αποτρέπουν από το να είναι επιτυχής σε διάφορα ρεαλιστικά σύνολα δεδομένων. Στην ανάπτυξή του οδήγησε σε μεγάλο βαθμό η ανάγκη για βελτίωση των αδυναμιών που παρουσιάζει ο x-means. Παρακάτω λοιπόν συγκρίνονται οι δύο αλγόριθμοι ώστε να γίνει φανερή η βελτίωση που έχει επέλθει στις αδυναμίες του πρώτου.

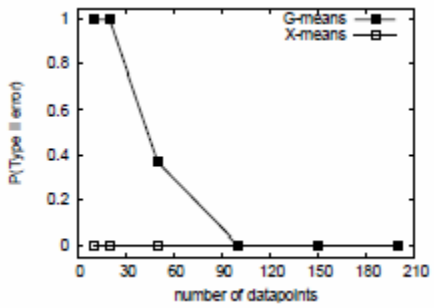
Το κριτήριο BIC που χρησιμοποιήθηκε στον x-means μπορεί να θεωρηθεί ως ένα τεστ λόγου πιθανοφάνειας με επίπεδο εμπιστοσύνης α το οποίο παραμένει σταθερό. Αυτό έχει ως

συνέπεια να μη λαμβάνεται υπόψη το πλήθος *n* των στοιχείων μιας ομάδας. Έτσι όσο το *n* μειώνεται, τόσο περισσότερο αυξάνεται το επίπεδο εμπιστοσύνης και το κριτήριο BIC γίνεται πιο αδύναμο κάτι που δεν εμφανίζεται στο Anderson-Darling τεστ. Αυτό σημαίνει πως σε μικρά σύνολα δεδομένων ο g-means λειτουργεί αποδοτικότερα από τον προκάτοχό του. Επιπλέον αναγνωρίζει ελλειπτικές δομές σε ομάδες κάτι που δίνει μεγαλύτερη ελευθερία από τον αυστηρό περιορισμό των καθαρά σφαιρικών Γκαουσιανών ομάδων που υποθέτει ο κλασσικός K-Μέσων και ο x-means.

Στα σχήματα 3.3.5 και 3.3.6 γίνεται σύγκριση της πιθανότητας εμφάνισης σφάλματος *Τύπου I* και *Τύπου II* ανάλογα με το πλήθος των αντικειμένων κάθε ομάδας. Σφάλμα Τύπου I σημαίνει στην περίπτωση αυτή την πραγματοποίηση λανθασμένου (αχρείαστου) διαχωρισμού ενώ σφάλμα Τύπου II τη μη πραγματοποίηση διαχωρισμού μιας ομάδας σε δύο που διαχωρίζονται με απόσταση 5σ, όπου σ είναι η τυπική απόκλιση των σημείων της ομάδας.

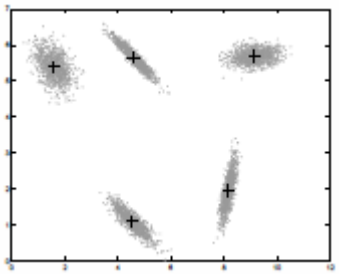


Σχήμα 3.3.5 Σύγκριση σφάλματος τύπου I

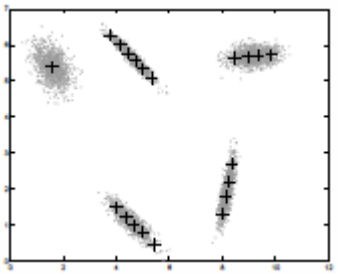


Σχήμα 3.3.6 Σύγκριση σφάλματος τύπου II

Στα παρακάτω σχήματα (Σχήμα 3.3.7 & 3.3.8) παρουσιάζεται οπτικά η υπεροχή του g-means να περιγράψει τις υποκείμενες ομάδες σε σύνολα δεδομένων με ελλειπτικές δομές και η αδυναμία του αλγορίθμου x-means αντίστοιχα στο 2ο σχήμα. Ο x-means ταυτοποίησε ως μία ομάδα μόνο την πρώτη από επάνω αριστερά καθώς ακολουθεί σφαιρική Γκαουσιανή κατανομή, ενώ στις υπόλοιπες ελλειπτικές παρατηρείται το φαινόμενο του overfitting και χρησιμοποιεί μικρές σφαιρικές ομάδες για να περιγράψει τις ελλειπτικές δομές που υπάρχουν.



Σχήμα 3.3.7 Γκαουσιανές G-Means



Σχήμα 3.3.8 Γκαουσιανές X-Means

Εκτός όμως από τις βελτιώσεις που εισήγαγε ο αλγόριθμος σε σχέση με τον x-means, παρουσιάζει και δύο σημαντικές αδυναμίες που θα επιλυθούν από τους αλγορίθμους που αναπτύσσονται στο πλαίσιο της μεταπτυχιακής αυτής εργασίας. Η πρώτη, έχει να κάνει με την αντιμετώπιση δεδομένων υψηλής διάστασης. Ο g-means εισάγει δυο τρόπους μείωσης της διάστασης των δεδομένων, με τον πρώτο να αποτελεί την προβολή των σημείων πάνω σε ένα τυχαίο διάνυσμα κάτι που δεν είναι αποτελεσματικό. Ο δεύτερος τρόπος ο οποίος είναι αυτός που χρησιμοποιείται συνήθως, κάνει χρήση της (πρώτης) κύριας συνιστώσας των δεδομένων κάτι που αποτελεί αποδεκτό τρόπο, όμως δεν είναι αποδοτικός σε όλες τις περιπτώσεις. Όπως θα αναλυθεί και σε επόμενη ενότητα, ο παραπάνω τρόπος δεν είναι ικανός να αιχμαλωτίσει όλες τις περιπτώσεις κατανομής των δεδομένων σε μεγαλύτερες διαστάσεις και πολλές φορές χάνεται χρήσιμη πληροφορία κατά την προβολή των στοιχείων σε μικρότερη διάσταση. Η δεύτερη και σημαντικότερη αδυναμία έχει να κάνει με την αρχική θεώρηση του αλγορίθμου ως προς το είδος της κατανομής που μπορεί να ακολουθεί μια ομάδα του συνόλου δεδομένων. Ο g-means υποθέτει την ύπαρξη σφαιρικών και ελλειπτικών Γκαουσιανών ομάδων κάτι που απέχει αρκετά από τη δομή που μπορεί να έχουν οι ομάδες ενός συνόλου δεδομένων. Γενικά, κατά τη διαδικασία της ομαδοποίησης μια ομάδα μπορεί να μοντελοποιηθεί ικανοποιητικά από κατανομές με μια κορυφή (unimodal) και οι σφαιρικές και οι ελλειπτικές είναι μόνο μερικές από αυτού του είδους τις κατανομές. Οι αλγόριθμοι που θα αναπτυχθούν στη συνέχεια δίνουν μεγαλύτερη ελευθερία ως προς τη θεώρηση αυτή για την περιγραφή των ομάδων. Αυτό έχει σαν αποτέλεσμα να αποφεύγονται φαινόμενα overfitting σε περιπτώσεις εμφάνισης κάποιας κατανομής διαφορετικής των δύο παραπάνω.

3.4. Αλγόριθμος Dip-Means

Ο αλγόριθμος *dip-means* [16], παρουσιάστηκε από τους Α. Καλογεράτο και Α. Λύκα το 2012 και αποτελεί μια ακόμη προσέγγιση στο δύσκολο πρόβλημα της αυτόματης εύρεσης του αριθμού *k* των ομάδων κατά τη διαδικασία της ομαδοποίησης ενός συνόλου δεδομένων.

Ο dip-means όπως και οι δύο προηγούμενοι αλγόριθμοι που παρουσιάστηκαν, είναι ένας επαναληπτικός αλγόριθμος που έχει ως βάση του τον κλασσικό αλγόριθμο K-Μέσων, και προσπαθεί με αυξητικό τρόπο να ανακαλύψει τον αριθμό των ομάδων που περιγράφει καλύτερα τα δεδομένα προς ομαδοποίηση. Ακολουθεί το κλασσικό μοντέλο των αλγορίθμων αυτής της κατηγορίας, δηλαδή αρχικά στοχεύει στη βελτίωση των παραμέτρων του τρέχοντος μοντέλου με την εκτέλεση ενός συμβατικού αλγορίθμου ομαδοποίησης για το τρέχον *k* (όπου αυτός είναι συνήθως ο κλασσικός αλγόριθμος K-Μέσων). Στο επόμενο βήμα βελτιώνεται η δομή του μοντέλου μέσω κάποιας μορφής διαχωρισμού στα κέντρα και τα παραπάνω δύο βήματα επαναλαμβάνονται επαναληπτικά έως ότου το μοντέλο να παραμένει σταθερό και να μην αλλάζει περαιτέρω. Όπως και στους προηγούμενους αλγορίθμους που περιγράψαμε έτσι και εδώ για τη πραγματοποίηση του διαχωρισμού των ομάδων χρησιμοποιείται το σκορ που προκύπτει από ένα στατιστικό κριτήριο. Αυτό στην περίπτωση του dip-means είναι το dip τεστ που θα αναλυθεί εκτενέστερα στη συνέχεια. Παρακάτω ακολουθεί ένας πίνακας με τους κυριότερους συμβολισμούς οι οποίοι θα χρησιμοποιηθούν για την περιγραφή του dip-means.

Συμβολισμός	Περιγραφή
X	Το σύνολο δεδομένων
x_i	Ένα στοιχείο του συνόλου δεδομένων
C	Το σύνολο των στοιχείων κάθε ομάδας δηλαδή $C = \{c_j\}_{j=1}^{k}$
c_j	Το σύνολο των αντικειμένων της j-οστής ομάδας
M	Τα μοντέλα, δηλαδή $M = \{m_j\}_{j=1}^{k}$
m_j	Το j-οστό κεντροειδές που αντιστοιχεί στην ομάδα c_j
k	Το πλήθος των ομάδων
n_j	Το πλήθος των στοιχείων της j-οστής ομάδας

Πίνακας 3.4.1 Συμβολισμοί αλγορίθμου Dip-Means

3.4.1. Στατιστικό Κριτήριο Dip

Μια από τις πιο βασικές και ίσως η θεμελιώδης θεώρηση που πρέπει να κάνει ένας αλγόριθμος αυτής της κατηγορίας, έχει να κάνει με τη δομή των ομάδων. Αυτό σημαίνει πως απαιτείται να οριστούν εκ των προτέρων οι κατανομές οι οποίες είναι δυνατόν να παράγουν τις ομάδες του συνόλου δεδομένων ή με πιο απλά λόγια το πώς θα μοιάζει μια ομάδα που θα αναγνωρίζεται από τον αλγόριθμο. Το ρόλο της αναγνώρισης αυτής τον αναλαμβάνει το στατιστικό κριτήριο που θα χρησιμοποιηθεί, το οποίο ελέγχει εάν ένα σύνολο δεδομένων αποτελεί μια μοναδική ομάδα ή πρέπει να γίνει διαχωρισμός του σε δύο ή περισσότερα μέρη.

Γενικά όσο πιο χαλαρό είναι ένα τέτοιο κριτήριο σημαίνει πως μπορούν να αναγνωριστούν περισσότερες από μια δομές ομάδων κάτι που αποτελεί το επιθυμητό αποτέλεσμα καθώς κάθε σύνολο δεδομένων μπορεί να περιέχει ομάδες που προέρχονται από διαφορετική κατανομή. Γενικά μπορούμε να πούμε ότι οι κατανομές οι οποίες περιγράφουν καλύτερα τα δεδομένα μιας ομάδας είναι οι *unimodal ή αλλιώς μονοτροπικές (με μία κορυφή)* κατανομές. Τέτοιες κατανομές είναι οι student-t, η Cauchy, η σφαιρική Γκαουσιανή, η ελλειπτική που είναι μια περίπτωση Γκαουσιανής, η ομοιόμορφη, η οποία αποτελεί ακραίας μορφής unimodal κατανομή κ.α. Πιο αναλυτικά, ως unimodal ορίζεται να είναι μια συνάρτηση κατανομής $F(t)$ και η οποία παρουσιάζει κορυφή στην περιοχή της $s_m = \{(\mathrm{t}_L, \mathrm{t}_U) : \mathrm{t}_L \le \mathrm{t}_U\}$, εάν είναι κυρτή στην περιοχή $s_L(-\infty, t_L]$, σταθερή στην περιοχή $[\mathrm{t}_L, \mathrm{t}_U]$ και κοίλη στην $s_U = [\mathrm{t}_U, \infty)$. Πρακτικά αυτό σημαίνει πως απομακρυνόμενοι από την κορυφή βρίσκουμε μη αυξανόμενη πυκνότητα πιθανότητας.

Ο έλεγχος για unimodality (μονοτροπικότητα) κάθε ομάδας του συνόλου των δεδομένων γίνεται με τη χρήση του Hartigan's dip statistic [17], το οποίο προτιμήθηκε από άλλα στατιστικά κριτήρια, όπως τη μέθοδο του Silverman [18], λόγω της απλότητας και της αποτελεσματικότητάς του. Το Hartigan's dip statistic αποδεικνύεται ιδιαίτερα ισχυρό κυρίως σε μονοδιάστατα δεδομένα κάτι το οποίο εφαρμόζεται και εδώ για την εξαγωγή καλύτερων αποτελεσμάτων. Πιο συγκεκριμένα το dip τεστ εφαρμόζεται στον πίνακα αποστάσεων των σημείων μιας ομάδας και όχι στα ίδια τα σημεία. Τη θέση του πίνακα αποστάσεων μπορεί να πάρει κάλλιστα και ο πίνακας ομοιότητας των σημείων. Η παραπάνω προσέγγιση έχει το

πλεονέκτημα πως δεν είναι αναγκαία η γνώση των δεδομένων μιας ομάδας, αλλά μόνο του πίνακα αποστάσεών/ομοιότητάς τους.

Διαισθητικά, το κριτήριο dip που χρησιμοποιείται θεωρεί κάθε σημείο μιας ομάδας ως ένα *θεατή (viewer)*. Δημιουργείται έτσι ένα διάνυσμα αποστάσεων κάθε θεατή με τα υπόλοιπα σημεία της ομάδας, η κατανομή των τιμών του οποίου φανερώνει τη δομή της ομάδας που ελέγχεται. Έτσι, από την κατανομή των αποστάσεων αυτών, γίνεται φανερή η ύπαρξη μιας μόνο ομάδας αν υπάρχει μια κορυφή (unimodal ομάδα), ή εάν βρεθούν δύο κορυφές (bimodal) τότε υποδηλώνεται η ύπαρξη δύο ξεχωριστών ομάδων. Είναι σημαντικό όμως να αναφερθεί πως υπάρχει εξάρτηση ως προς τη θέση του κάθε θεατή. Αυτό σημαίνει πως θεατές που βρίσκονται στα όρια των ομάδων ανιχνεύουν ευκολότερα την ύπαρξη δύο κορυφών. Απαιτείται η ύπαρξη ενός ποσοστού (π.χ. 1% τουλάχιστον) θεατών σε σχέση με το πλήθος των σημείων που εξετάζονται κάθε φορά, οι οποίοι να προτείνουν διαχωρισμό για να θεωρηθεί η ομάδα πολυτροπική (multimodal). Αυτοί οι θεατές καλούνται *θεατές διαχωρισμού (split viewers)*.

Η λειτουργία του αλγορίθμου ως προς την εφαρμογή του κριτηρίου dip έχει ως εξής: θεωρούμε δύο φραγμένες συναρτήσεις F, G και έστω $\rho(F,G) = \max_t |F(t) - G(t)|$. Επίσης συμβολίζουμε με U^* την κλάση όλων των unimodal κατανομών. Τότε η τιμή dip μιας συνάρτησης κατανομής F θα δίνεται από τον παρακάτω τύπο.

$$dip(F) = \min_{G \in U^*} \rho(F,G) \qquad \text{Εξ. 3.4.1}$$

Μια ενδιαφέρουσα ιδιότητα του κριτηρίου dip είναι πως, εάν F_n είναι ένα δείγμα από n παρατηρήσεις της κατανομής F, τότε θα ισχύει η ισότητα $\lim_{n\to\infty} dip(F_n) = dip(F)$. Επιπλέον έχει αποδειχθεί από τους J.A. Hartigan και P.M. Hartigan [17] πως η κλάση των ομοιόμορφων κατανομών U είναι η καταλληλότερη για μηδενική υπόθεση H_0 στο dip τεστ, καθώς οι dip τιμές που εμφανίζει είναι μεγαλύτερες αυτών από άλλες unimodal κατανομές. Δοθέντος ενός διανύσματος $f = \{f_i : f_i \in \mathbb{R}\}_{i=1}^n$, ο αλγόριθμος τότε εκτελεί το στατιστικό τεστ στην $F_n(t) = \frac{1}{n}\sum_n I(f_i \le t)$, με I να είναι συνάρτηση απόστασης ή ομοιότητας.

Ο υπολογισμός της p-value για ένα unimodality τεστ χρησιμοποιεί bootstrap δείγματα από ομοιόμορφες κατανομές *U* στο διάστημα [0,1] και εκφράζει την πιθανότητα η τιμή dip της F_n να είναι μικρότερη από την τιμή dip της U_n^r που αποτελείται από *n* στοιχεία. Θα είναι:

$$P = \#[dip(F_n) \le dip(U_n^r)] / b \ , \ r = 1,...,b \qquad \text{Εξ. 3.4.2}$$

Οι δύο εναλλακτικές υποθέσεις H_0 και H_1 του dip τεστ που χρησιμοποιούνται για να καθοριστεί αν το δείγμα των δεδομένων που ελέγχεται ακολουθεί ή όχι unimodal κατανομή θα είναι:

- H_0: Η F_n περιγράφεται από μια unimodal κατανομή. Αυτό ισχύει όταν η p-value είναι μεγαλύτερη από ένα επίπεδο εμπιστοσύνης *α*, p>*a* .
- H_1: Τα στοιχεία της F_n δεν υποδεικνύουν την ύπαρξη μίας και μόνο κορυφής. Αυτό σημαίνει πως υπάρχουν θεατές διαχωρισμού και κατά συνέπεια η F_n είναι multimodal.

Συνολικά, χρησιμοποιώντας όλα τα παραπάνω η διαδικασία της εκτέλεσης του κριτηρίου dip απαρτίζεται από τα 3 παρακάτω βήματα:

1) Υπολογίζονται οι U_n^r, $r = 1,...,b$ των δειγμάτων καθώς και οι dip τιμές τους.
2) Υπολογίζεται η $F_n^{(x_i)}(t) = \frac{1}{n}\sum_{x_j \in s}\{Dist(x_i, x_j) \le t\}$, *i=1,…n* και οι dip τιμές του πίνακα αποστάσεών της αφού πρώτα ταξινομηθεί.
3) Υπολογίζονται οι p-values $P^{(x_i)}$, *i=1,…,n* με επίπεδο εμπιστοσύνης *α* και τέλος υπολογίζεται το ποσοστό των θεατών διαχωρισμού.

3.4.2. Δομή του αλγορίθμου

Έχοντας αναλύσει τον τρόπο λειτουργίας του κριτηρίου dip, θα παρουσιαστεί στη συνέχεια ο αλγόριθμος dip-means, ο οποίος αποτελείται από 3 βασικά τμήματα. Το πρώτο είναι η

εκτέλεση ενός αλγορίθμου ομαδοποίησης τοπικά και πιο συγκεκριμένα του K-Μέσων, σε ένα μοντέλο M αποτελούμενο από k ομάδες για τη βελτίωση των παραμέτρων του. Το δεύτερο και σημαντικότερο κομμάτι είναι η εφαρμογή του στατιστικού κριτηρίου dip σε μια ομάδα του συνόλου δεδομένων και το τρίτο περιλαμβάνει μια διαδικασία διαχωρισμού μιας ομάδας σε δύο ανάλογα με τα αποτελέσματα που προέρχονται από το προηγούμενο κομμάτι. Στον πίνακα 3.4.2 ακολουθεί μια αναλυτικότερη περιγραφή του dip-means σε μορφή ψευδοκώδικα:

Αλγόριθμος Dip-Μέσων (Dip-Means)
1. Είσοδος του αρχικού αριθμού των ομάδων k_{init}, του επιπέδου εμπιστοσύνης α για το unimodality (dip) τεστ και του ποσοστού v_{thd} των θεατών διαχωρισμού μιας ομάδας ώστε αυτή να θεωρηθεί υποψήφια για διαχωρισμό.
2. $k \leftarrow k_{init}$.
3. $\{C, M\} \leftarrow$ K-Μέσων (X, k).
4. *Επανάλαβε*
5. *Για* j = 1 έως k
6. $dip\_score_j =$ unimodalityTest (c_j, a, v_{thd}).
7. *Τέλος για*
8. *Εάν* $\max_j(dip\_score_j) > 0$
9. $\{m_L, m_R\} \leftarrow$ Διαχωρισμός του κεντροειδούς m_w της ομάδας c_w η οποία έχει το μεγαλύτερο dip_score $\forall j = 1,...,k$.
10. Αντικατάσταση του m_w από τα δύο νέα στο μοντέλο M δηλαδή $M \leftarrow \{M - m_w, m_L, m_R\}$.
11. Βελτίωση της τρέχουσας λύσης $\{C, M\} \leftarrow$ K-Μέσων (X, M).
12. *Τέλος Εάν*
13. *Έως ότου δεν προστίθενται νέα κέντρα.*
14. Επέστρεψε την τελική λύση, δηλαδή τα C και M.

Πίνακας 3.4.2 Ψευδοκώδικας αλγορίθμου Dip-Means

Ο dip-means εκτός από το σύνολο δεδομένων παίρνει ως είσοδο δύο επιπλέον παραμέτρους που χρησιμοποιούνται από το στατιστικό κριτήριο dip. Η πρώτη είναι το

επίπεδο εμπιστοσύνης α το οποίο χρησιμοποιείται για την αποδοχή της μηδενικής υπόθεσης του τεστ H_0 ή την απόρριψη της και την αποδοχή της εναλλακτικής H_1. Η δεύτερη είναι το ποσοστό των θεατών διαχωρισμού v_{thd} των στοιχείων μιας ομάδας πέρα από το οποίο θα ταξινομείται ως υποψήφια για διαχωρισμό. Ο dip-means έχει τη δυνατότητα να ξεκινήσει την ομαδοποίηση έχοντας 1 ομάδα ή με την εισαγωγή ενός υπάρχοντος μοντέλου με $k_{init} = k$ ομάδες. Έπειτα, σε κάθε επανάληψη εξετάζονται όλες οι k ομάδες από το τεστ για unimodality και μια ομάδα c_j χαρακτηρίζεται υποψήφια για διαχωρισμό εάν $|v_j|/n_j \geq v_{thd}$ με v_j να συμβολίζει το σύνολο των θεατών διαχωρισμού της ομάδας j. Τότε σύμφωνα με τη σχέση Εξ. 3.4.3 ανατίθεται σε αυτήν ένα μη μηδενικό σκορ. Αντίθετα, στην περίπτωση που μια ομάδα δεν είναι υποψήφια για διαχωρισμό της ανατίθεται μηδενικό σκορ. Η σχέση που μας δίνει το σκορ του dip τεστ είναι:

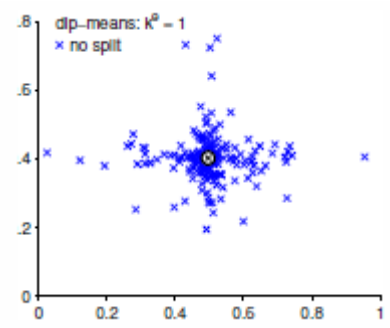
$$score_j = \frac{1}{|v_j|}\sum_{x \in v_j} dip(F^{(x)}), \quad \text{αν } \frac{|v_j|}{n_j} \geq v_{thd}$$

$$score_j = 0, \quad \text{αλλιώς} \qquad \text{Εξ.3.4.3}$$

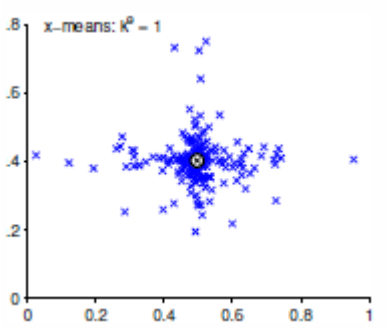
Στη συνέχεια, αφού έχει υπολογιστεί το σκορ κάθε ομάδας επιλέγεται η ομάδα με το μεγαλύτερο σκορ κάθε φορά για διαχωρισμό. Τότε τη θέση του παλιού κέντρου m_w παίρνουν τα δύο νέα τα οποία ορίζονται ως εξής : $\{m_L, m_R\} \leftarrow \{x, m_w - (x - m_w)\}$, όπου x είναι ένα τυχαίο σημείο της ομάδας. Αυτό τοποθετεί τα νέα κέντρα σε αντίθετες κατευθύνσεις ίσης απόστασης από το παλιό κέντρο. Αντί της μεθόδου αυτής μπορεί να χρησιμοποιηθεί η *principal direction divisive partitioning (PDDP)* [19] που όμως προϋποθέτει τον υπολογισμό της κύριας συνιστώσας, κάτι που προσθέτει επιπλέον κόστος στο χρόνο εκτέλεσης του αλγορίθμου. Το μοντέλο που προκύπτει από τον διαχωρισμό μπορεί να βελτιώνεται στο τέλος κάθε επανάληψης με την εκτέλεση του αλγορίθμου K-Μέσων. Τέλος, ο dip means σταματάει την εκτέλεσή του όταν βάσει του κριτηρίου δεν εντοπίζονται πλέον υποψήφιες ομάδες για διαχωρισμό, οπότε δεν προστίθενται πλέον νέα κέντρα.

Ο dip-means καταφέρνει λοιπόν να καταπολεμήσει επιτυχώς τα προβλήματα του κλασσικού αλγορίθμου K-Μέσων καθώς και των προγενέστερων αλγορίθμων του

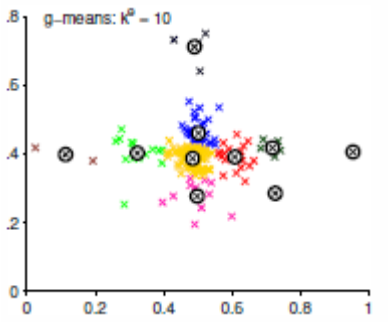
χρησιμοποιώντας μια διαφορετική φιλοσοφία, αποφεύγοντας την αυστηρή κατηγοριοποίηση των ομάδων των δεδομένων ως προς την κατανομή από την οποία έχουν παραχθεί κάτι το οποίο παράγει ποιοτικά καλύτερες λύσεις σε ρεαλιστικά σύνολα δεδομένων. Στα παρακάτω σχήματα γίνεται εμφανής η ανωτερότητά του κατά την ταυτοποίηση student-t (Σχήμα 3.4.1 – 3.4.3), ομοιόμορφων (Σχήμα 3.4.4 – 3.4.6) και διαφόρων κατανομών ομάδων στο ίδιο σύνολο δεδομένων (Σχήμα 3.4.7 – 3.4.9).



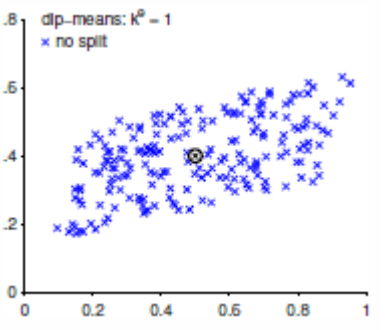
Σχήμα 3.4.1 Student-t Dip-Means



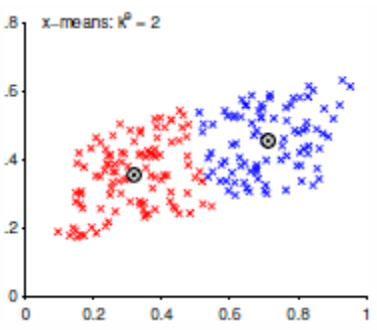
Σχήμα 3.4.2 Student-t X-Means



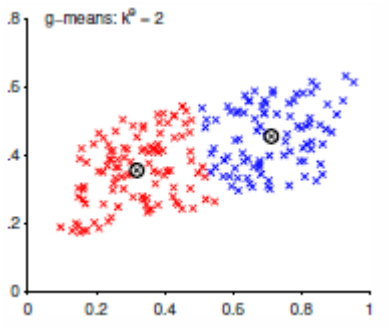
Σχήμα 3.4.3 Student-t G-Means



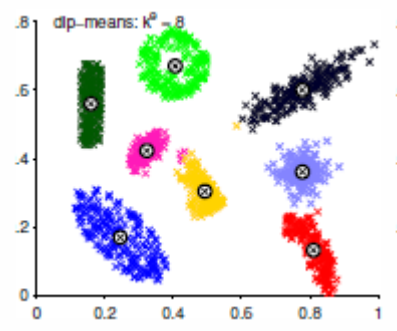
Σχήμα 3.4.4 Ομοιόμορφη Dip-Means



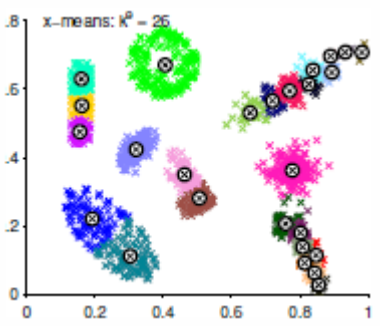
Σχήμα 3.4.5 Ομοιόμορφη X-Means



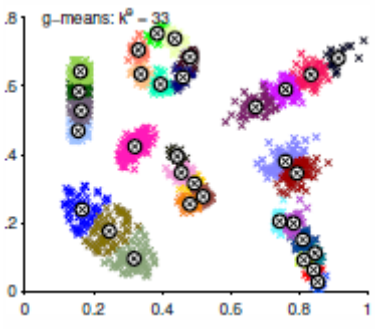
Σχήμα 3.4.6 Ομοιόμορφη G-Means



Σχήμα 3.4.7 Μονοτροπικές Dip-Means

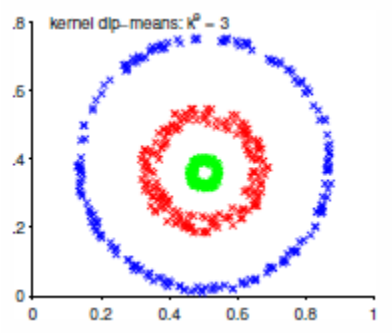


Σχήμα 3.4.8 Μονοτροπικές X-Means

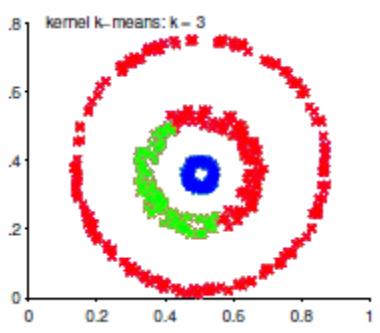


Σχήμα 3.4.9 Μονοτροπικές G-Means

Τέλος να σημειωθεί πως ο παρών αλγόριθμος μπορεί να χρησιμοποιηθεί με συναρτήσεις πυρήνα μιας και κάνει χρήση του πίνακα αποστάσεων και όχι των στοιχείων του συνόλου δεδομένων καθαυτών και τη θέση του κλασσικού αλγορίθμου κ-μέσων κατά την εκτέλεσή του μπορεί να πάρει ο kernel k-means. Η εκδοχή του αλγορίθμου αυτή ονομάζεται *kernel dip-means* και στα σχήματα 3.4.10 και 3.4.11 συγκρίνεται η αποτελεσματικότητά του σε ένα σύνολο δεδομένων με 3 ομοιόμορφους δακτυλίους σε σχέση με τον kernel k-means.



Σχήμα 3.4.10 Kernel Dip-Means



Σχήμα 3.4.11 Kernel K-Means

ΚΕΦΑΛΑΙΟ 4. ΠΡΟΤΕΙΝΟΜΕΝΕΣ ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ

4.1. Εισαγωγή

Στο προηγούμενο κεφάλαιο αναφέρθηκαν τρεις αλγόριθμοι, οι x-means, g-means και dip-means οι οποίοι αποτέλεσαν τις κυριότερες προσπάθειες για την αντιμετώπιση των μειονεκτημάτων του κλασσικού αλγορίθμου K-Μέσων όπως αυτά παρουσιάστηκαν στο κεφάλαιο 2. Όπως έγινε φανερό, η σημαντικότερη και συνάμα δυσκολότερη στην αντιμετώπισή της αδυναμία του K-Μέσων είναι η αρχική γνώση του αριθμού των ομάδων που απαιτείται ως είσοδος για την εκτέλεσή του.

Οι παραπάνω αλγόριθμοι ακολουθώντας διαφορετικές φιλοσοφίες και τεχνικές δημιούργησαν ξεχωριστές προσεγγίσεις κάθε φορά για την βελτίωση του αλγορίθμου K-Μέσων. Αρχικά ο x-means υποθέτει, όπως και ο κλασσικός αλγόριθμος, την ύπαρξη σφαιρικών Γκαουσιανών ομάδων στο εκάστοτε σύνολο δεδομένων και κάνει χρήση του στατιστικού κριτηρίου BIC για να αποφασίζει εάν μια ομάδα είναι σφαιρική ή όχι και ανάλογα να τη διαχωρίσει. Έπειτα ο g-means επέκτεινε μερικώς τη θεώρηση αυτή για τη δομή των ομάδων θεωρώντας αποδεκτές και τις ελλειπτικές ομάδες και χρησιμοποίησε το στατιστικό κριτήριο Anderson-Darling για την αναγνώριση της δομής μιας ομάδας. Οι επιδόσεις του βελτιώνονται σε σχέση με τον x-means, όμως η αποτελεσματικότητά τους

περιορίζεται σε ένα μικρό υποσύνολο των πιθανών δομών που μπορούν να πάρουν οι ομάδες "πραγματικών" συνόλων δεδομένων. Τέλος, ο dip-means βασίζεται σε πιο γενικές υποθέσεις, θεωρώντας ομάδες που ανήκουν στην κλάση των unimodal κατανομών. Για να το πετύχει αυτό κάνει χρήση του στατιστικού κριτηρίου Hartigan's Dip Statistic βάσει του οποίου μια ομάδα διαχωρίζεται εάν δεν ακολουθεί unimodal κατανομή. Όπως έγινε φανερό στο προηγούμενο κεφάλαιο, λόγω της φιλοσοφίας του αυτής, ο dip-means επιδεικνύε σημαντική βελτίωση στην ποιότητα των παραγόμενων λύσεων σε σχέση με τους άλλους δύο αλγόριθμους.

Στο κεφάλαιο αυτό, βασιζόμενοι στην υπόθεση ότι κατά τη διαδικασία της ομαδοποίησης κάθε ομάδα αντιστοιχεί στην κλάση των unimodal (μονοτροπικών) κατανομών, δηλαδή ότι περιγράφεται από μια κατανομή η οποία έχει μία κορυφή (mode), θα παρουσιάσουμε τρεις νέους αλγορίθμους που αναπτύχθηκαν στο πλαίσιο της παρούσας μεταπτυχιακής εργασίας. Έχουμε λοιπόν ως στόχο να εξετάσουμε εναλλακτικές προσεγγίσεις στη χρήση του κριτηρίου μονοτροπικότητας, με την παρουσίαση του *pdip-means* (*Projected Dip-Means*), καθώς και δύο αλγορίθμων συσσωρευτικής φύσης, του *agglodip* (*Agglomerative Dip*) και του *agglopdip* (*Agglomerative Projective Dip*).

4.2. Αλγόριθμος Pdip-Means

Ο πρώτος από τους τρεις προτεινόμενους αλγορίθμους ονομάζεται Pdip-Means (Projected Dip-Means). Ο pdip-means συνδυάζει κάποιες από τις ιδέες που παρουσιάστηκαν παραπάνω και προσθέτει επιπλέον βελτιώσεις με στόχο την ακριβέστερη εύρεση του αριθμού k των ομάδων και την εύρεση "ορθότερων" λύσεων κατά την ομαδοποίηση ενός συνόλου δεδομένων.

Όπως και οι προηγούμενοι αλγόριθμοι που παρουσιάστηκαν στο κεφάλαιο 3, έτσι και ο pdip-means ακολουθεί αυξητική πορεία στην εύρεση του κατάλληλου αριθμού k των ομάδων. Αυτό σημαίνει πως ξεκινώντας αρχικά από έναν μικρό αριθμό ομάδων ($k = k_{init}$ όταν υπάρχει πρότερη γνώση του συνόλου δεδομένων και με $k_{init} = 1$ όταν δεν υπάρχει), να αυξάνονται σε κάθε επανάληψη κατά μία οι ομάδες του τρέχοντος μοντέλου M που έχει

παράγει μέχρι στιγμής ο αλγόριθμος. Αυτό επαναλαμβάνεται έως ότου να ικανοποιείται μια συνθήκη που θα σημάνει το τέλος των επαναλήψεων και θα δώσει ως έξοδο το πλήθος των ομάδων, τα κέντρα τους καθώς και τις αναθέσεις των στοιχείων σε αυτά. Τα παραπάνω συνθέτουν τη μορφή της λύσης του αλγορίθμου ο οποίος δεν στέκεται μόνο στην παραγωγή του "σωστού" *k* αλλά στη συνολική παρουσίαση της λύσης.

Όπως και στους προηγούμενους αλγορίθμους σημαντικότατο ρόλο παίζει η απόφαση για τη διάσπαση ή όχι μιας ομάδας, η οποία γίνεται βάσει κάποιου σκορ το οποίο έχει υπολογιστεί από την εφαρμογή ενός στατιστικού κριτηρίου. Στην περίπτωσή μας, όπως και στην περίπτωση του dip-means, όπου υποθέτουμε πως οι ομάδες ενός συνόλου δεδομένων μπορούν να περιγραφούν καλύτερα από μια unimodal κατανομή, χρησιμοποιείται το στατιστικό κριτήριο Hartigan's Dip Statistic για τον έλεγχο κάθε ομάδας για unimodality.

Στον πίνακα που παρατίθεται παρακάτω (Πίνακας 4.2.1) είναι συγκεντρωμένοι οι κυριότεροι συμβολισμοί οι οποίοι θα χρησιμοποιηθούν για την περιγραφή του αλγορίθμου pdip-means.

Συμβολισμός	Περιγραφή
X	Το σύνολο δεδομένων
x_i	Ένα στοιχείο του συνόλου δεδομένων
C	Το σύνολο των στοιχείων κάθε ομάδας δηλαδή $C = \{c_j\}_{j=1}^{k}$
c_j	Το σύνολο των αντικειμένων της j-οστής ομάδας
M	Τα μοντέλα, δηλαδή $M = \{m_j\}_{j=1}^{k}$
m_j	Το j-οστό κεντροειδές που αντιστοιχεί στην ομάδα c_j
k	Το πλήθος των ομάδων
n_j	Το πλήθος των στοιχείων της j-οστής ομάδας
$projections_{j,prj}$	Το σύνολο του πλήθος prj των προβολών που αντιστοιχεί στη j-οστή ομάδα

Πίνακας 4.2.1 Συμβολισμοί αλγορίθμου Pdip-Means

4.2.1. Δομή του αλγορίθμου

Ο αλγόριθμος pdip-means παρουσιάζει ομοιότητες στη δομή του με τον αλγόριθμο dip-means όμως ακολουθεί μια διαφορετική προσέγγιση βασισμένη στη χρήση προβολών, όπως υποδηλώνει και η ονομασία του. Αποτελείται από 4 βασικά βήματα, με το πρώτο (αρχικοποίηση) να είναι ίδιο με του αλγορίθμου dip-means, δηλαδή εκτελείται ο αλγόριθμος K-Μέσων στο μοντέλο *M* που αποτελείται από *k* ομάδες και δίνεται ως είσοδος αρχικά από τον χρήστη, ώστε να βελτιωθεί ή να βρεθεί το αρχικό κεντροειδές, εάν το σύνολο δεδομένων αποτελείται από μία ομάδα. Στο δεύτερο βήμα για κάθε ομάδα που υπάρχει στο μοντέλο *M* κατά την τρέχουσα επανάληψη του αλγορίθμου προβάλλονται τα αντικείμενα όλων των c_j ομάδων $\forall j = 1,...,k$ σε ένα σύνολο ευθειών. Για κάθε ευθεία από το σύνολο αυτό παράγεται έτσι ένα διάνυσμα που εμπεριέχει τις προβολές των αντικειμένων στις ευθείες που έχουν επιλεγεί κάθε φορά. Έπειτα στο τρίτο βήμα ο pdip-means χρησιμοποιεί τα διανύσματα αυτά για την εφαρμογή του κριτηρίου dip, δηλαδή εξετάζεται για μονοτροπικότητας η κάθε ομάδα με βάση τις προβολές των σημείων της στις ευθείες. Στο τελευταίο βήμα γίνεται ο διαχωρισμός των ομάδων βάσει των αποτελεσμάτων μονοτροπικότητας που προέκυψαν στο προηγούμενο βήμα. Στον πίνακα 4.2.2 που ακολουθεί παρουσιάζεται μια περιγραφή του αλγορίθμου pdip-means με τη χρήση ψευδοκώδικα:

Αλγόριθμος PDip-Μέσων (PDip-Means)
1. Είσοδος του αρχικού αριθμού των ομάδων k_{init} και του επιπέδου εμπιστοσύνης *α* για το unimodality (dip) τεστ.
2. $k \leftarrow k_{init}$.
3. $\{C, M\} \leftarrow$ K-Μέσων (X, k).
4. *Επανάλαβε*
5. *Για* j = 1 έως *k*
6. $projections_{j,prj} \leftarrow data\_proj(c_j)$
7. $dip\_score_{j,prj} =$ unimodalityTest $(projections_{j,prj}, a)$.
8. $pdip\_score_j = \max_{prj}(dip\_score_{j,prj})$
9. *Τέλος για*
10. *Εάν* $\max_j(pdip\_score_j) > 0$

11.	$\{m_L, m_R\} \leftarrow$ Διαχωρισμός του κεντροειδούς m_w της ομάδας c_w η οποία έχει το μεγαλύτερο *pdip_score* $\forall j = 1,...,k$.
12.	Αντικατάσταση του m_w από τα δύο νέα στο μοντέλο M δηλαδή $M \leftarrow \{M - m_w, m_L, m_R\}$.
13.	Βελτίωση της τρέχουσας λύσης $\{C, M\} \leftarrow$ K-Μέσων (X, M).
14.	*Τέλος Εάν*
15.	*Έως ότου δεν προστίθενται νέα κέντρα.*
16.	Επέστρεψε την τελική λύση, δηλαδή τα *C* και *M.*

Πίνακας 4.2.2 Ψευδοκώδικας αλγορίθμου Pdip-Means

Ο pdip-means λόγω της θεώρησής του ότι κάθε ομάδα του υποκείμενου συνόλου δεδομένων θα ακολουθεί μια κατανομή από την κλάση των unimodal, δηλαδή κατανομών που εμφανίζουν μία κορυφή και σε περιοχές μακριά από αυτή υπάρχει μη αυξανόμενη πυκνότητα πιθανότητας, χρησιμοποιεί σαν στατιστικό τεστ το Hartigan's dip statistic λόγω της υπεροχής του έναντι εναλλακτικών μεθόδων [17]. Τέτοιου είδους κατανομές είναι οι Γκαουσιανές, οι student-t, οι Cauchy, οι ομοιόμορφες και άλλες.

Η διαφορά κατά την εκτέλεση του Hartigan's dip statistic με τον dip-means έγκειται στο γεγονός πως δεν είναι απαραίτητος πλέον ο υπολογισμός και η χρήση του πίνακα αποστάσεων. Το dip τεστ στον pdip-means εφαρμόζεται στα διανύσματα των προβολών $projections_{j,prj}$ των αρχικών σημείων x_i κάθε ομάδας c_j με *j=1,...,k* , *i=1,...,* n_j και με *prj* να είναι το σύνολο των προβολών που χρησιμοποιούνται κάθε φορά, που προκύπτουν από την εκτέλεση της διαδικασίας *data_proj*. Βάσει της κατανομής λοιπόν των σημείων αυτών αποφασίζεται από το τεστ εάν η ομάδα της οποίας αποτελούν μέλη έχει παραχθεί από μια unimodal κατανομή ή όχι.

Η τιμή dip στην περίπτωση αυτή θα προέρχεται από τον υπολογισμό του ελάχιστου από το σύνολο των μέγιστων αποκλίσεων της εμπειρικής αθροιστικής συνάρτησης κατανομής (ecdf) F_n των προβαλλόμενων σημείων της τρέχουσας ομάδας, με τις συναρτήσεις κατανομών της κλάσης U^* των unimodal κατανομών. Δηλαδή υπολογίζεται ως εξής:

$$dip(F_{n_j}) = \min_{G \in U^*} \rho(F_{n_j}, G_{n_j}) \quad j = 1, \ldots, k \qquad \text{Εξ 4.2.1}$$

με τη ρ να ορίζεται για δύο συναρτήσεις *F* και *G* ως εξής:

$$\rho(F, G) = \max_t \left| F(t) - G(t) \right| \qquad \text{Εξ 4.2.2}$$

Έπειτα, γίνεται ο υπολογισμός της τιμής p-value με την εφαρμογή του τύπου (Εξ. 3.4.2), αφού έχουν υπολογιστεί οι τιμές dip από τα bootstrap δείγματα από ομοιόμορφες κατανομές *U* στο διάστημα [0,1] όπως και στην περίπτωση του αλγορίθμου dip-means. Αξίζει να σημειωθεί πως ο υπολογισμός των bootstrap αυτών δειγμάτων παίρνει ως είσοδο δύο μεταβλητές, το μέγιστο πλήθος στοιχείων που μπορεί να έχει μία ομάδα καθώς και το πλήθος των κατανομών που θα παραχθούν. Είναι προφανές πως όσο μεγαλύτερες ομάδες εμφανίζονται κατά την ομαδοποίηση τόσο μεγαλύτερη θα είναι η πρώτη τιμή και πως όσο μεγαλύτερο είναι το πλήθος των bootstrap κατανομών που χρησιμοποιούνται τόσο πιο έγκυρη και ακριβής θα είναι η p-value που θα παραχθεί στη συνέχεια. Στον dip-means αυτή η διαδικασία εκτελούνταν κατά τη ροή εκτέλεσης του προγράμματος και αποφασιζόταν δυναμικά οι τιμές των bootstrap κατανομών που θα χρησιμοποιηθούν. Στον αλγόριθμο pdip αυτή η διαδικασία γίνεται πλέον μία φορά ανεξάρτητα από την εκτέλεση του αλγορίθμου και τα παραγόμενα bootstrap δείγματα μπορούν να επαναχρησιμοποιηθούν όσες φορές κριθεί απαραίτητο.

Οι εναλλακτικές υποθέσεις H_0 και H_1 του dip τεστ παραμένουν ίδιες με αυτές που χρησιμοποιούνται στον dip-means με την H_0 να χρησιμοποιείται όταν η F_n περιγράφεται καλά από μια unimodal κατανομή, δηλαδή αν η p-value της είναι μεγαλύτερη από ένα επίπεδο εμπιστοσύνης *α*, p>*a* που δόθηκε ως είσοδος αρχικά στον αλγόριθμο ενώ στην αντίθετη περίπτωση, η αποδοχή της H_1 υποδηλώνει την ύπαρξη διαχωρισμού. Αυτό σημαίνει πως η p-value θα έχει μηδενική τιμή άρα η ομάδα προς εξέταση έχει περισσότερες της μίας κορυφές και τίθεται ως υποψήφια για διαχωρισμό.

4.2.2. Προβολή των δεδομένων μιας ομάδας

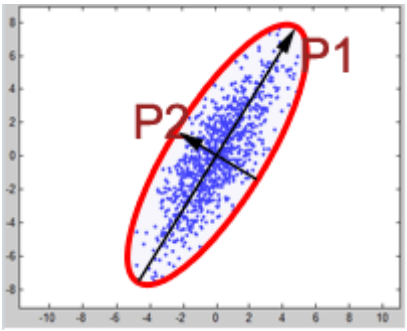
Σε προβλήματα στα οποία η μεγάλη διάσταση των δεδομένων αποτελεί πρόβλημα, είναι μεγάλης σημασίας η χρήση κάποιας μεθόδου για τη μείωση της διάστασης. Αυτό, σε περιπτώσεις που υπάρχει γνώση της δομής και της φύσης των δεδομένων, μπορεί να γίνει κατά την προεπεξεργασία του συνόλου των δεδομένων ώστε διαστάσεις με "άχρηστη" ή αλλιώς μη σημαντική πληροφορία να εξαλειφθούν. Γενικά όμως στις περισσότερες περιπτώσεις αυτό δεν είναι εφικτό κυρίως όταν τα δεδομένα είναι πολλά, μεγάλης διάστασης ή συλλέγονται με αυτόματο τρόπο. Γίνεται έτσι επιτακτική η ανάγκη για μείωση της διάστασης των δεδομένων με αυτόματο τρόπο χωρίς να χαθεί χρήσιμη πληροφορία που ενδεχομένως να περιέχεται σε κάποια διάσταση.

Όπως αναφέρθηκε νωρίτερα και παρουσιάστηκε στον ψευδοκώδικα του πίνακα 4.2.2, υπάρχει μια σημαντική διαφορά με τον αλγόριθμο dip-means, εκτός αυτών που εκφράστηκαν νωρίτερα, η οποία αποτυπώνεται στον αλγόριθμο με τη χρήση της ρουτίνας *data_proj*. Η ρουτίνα αυτή είναι υπεύθυνη για την παραγωγή ενός συνόλου ευθειών πάνω στις οποίες θα προβληθούν τα δεδομένα κάθε ομάδας c_j του συνόλου δεδομένων. Σε αντίθεση με τον αλγόριθμο g-means ο οποίος χρησιμοποιεί μόνο την κύρια συνιστώσα για την προβολή των δεδομένων, ο pdip-means χρησιμοποιεί ένα σύνολο προβολών το οποίο αυξομειώνεται δυναμικά ανάλογα με το σύνολο δεδομένων της εισόδου αλλά και την είσοδο του χρήστη αν αυτό κριθεί αναγκαίο. Σκοπός αυτού είναι η απώλεια όσο το δυνατόν λιγότερης πληροφορίας από τις ομάδες αφού στη συνέχεια επιλέγεται και χρησιμοποιείται η καλύτερη προβολή κάθε φορά.

Η *data_proj* έχει ως είσοδο τα δεδομένα μιας ομάδας κάθε φορά και προαιρετικά το είδος και το πλήθος των προβολών που θα χρησιμοποιηθούν για το εκάστοτε σύνολο δεδομένων. Οι προβολές που χρησιμοποιούνται είναι τριών ειδών έχοντας ως στόχο να περιγράψουν καλύτερα τη δομή της υποκείμενης ομάδας. Αυτές είναι :

- **Προβολές σε PCAs ευθείες**. Μια από τις πιο διαδεδομένες και αποτελεσματικές μεθόδους για την προβολή δεδομένων και τη μείωση των διαστάσεών τους είναι η ανάλυση σε κύριες συνιστώσες (Principal Component Analysis) [20] [21]. Η PCA είναι μια στατιστική μέθοδος η οποία χρησιμοποιεί ορθογώνιους μετασχηματισμούς

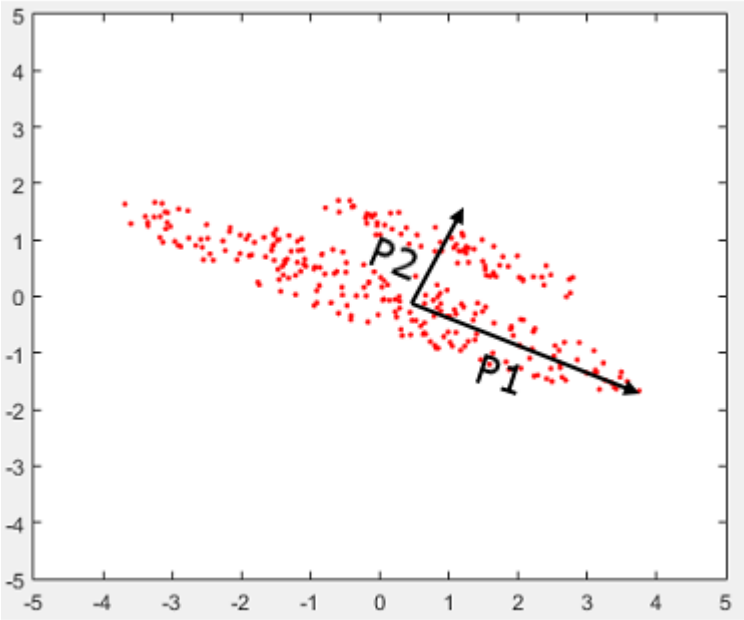
για να μετατρέψει ένα σύνολο συσχετιζόμενων παρατηρήσεων σε ένα σύνολο γραμμικά ασυσχέτιστων μεταβλητών οι οποίες ονομάζονται κύριες συνιστώσες με τον αριθμό αυτών να είναι μικρότερος ή ίσος του αριθμού των στοιχείων του συνόλου δεδομένων. Διαισθητικά, με τη χρήση των PCA συνιστωσών αιχμαλωτίζουμε τις κατευθύνσεις προς τις οποίες είναι απλωμένα τα δεδομένα, δηλαδή τις κατευθύνσεις με τις μεγαλύτερες διακυμάνσεις. Στο σχήμα 4.2.1 φαίνεται η ιδέα πίσω από τη χρησιμότητα της μεθόδου αυτής:



Σχήμα 4.2.1 Κύριες συνιστώσες PCA

Με P1 συμβολίζεται η 1η κύρια συνιστώσα ενώ με P2 η δεύτερη. Ο υπολογισμός τους βασίζεται στη χρήση γραμμικής άλγεβρας και πιο συγκεκριμένα στον υπολογισμό των ιδιοτιμών του πίνακα συνδιακύμανσης των δεδομένων. Τότε τα αντίστοιχα ιδιοδιανύσματα θα αποτελούν τις κύριες συνιστώσες της PCA.

Στις περισσότερες περιπτώσεις όπως και στην περίπτωση του αλγορίθμου g-means χρησιμοποιείται μόνο η 1η κύρια συνιστώσα κατά την οποία είναι “απλωμένα” τα δεδομένα. Κατά την εκτέλεση όμως της διαδικασίας της ομαδοποίησης δεδομένων, πολλές φορές η 1η κύρια συνιστώσα περιγράφει όχι τόσο χρήσιμη πληροφορία για τα δεδομένα καθώς η ομάδα προς εξέταση μπορεί για παράδειγμα να έχει την παρακάτω δομή (Σχήμα 4.2.2) και στην πραγματικότητα να περιγράφεται καλύτερα από δύο ομάδες.



Σχήμα 4.2.2 Χρησιμότητα 2ης κύριας συνιστώσας PCA

Εάν δεν γινόταν χρήση και της δεύτερης συνιστώσας στην παραπάνω περίπτωση θα είχε χαθεί χρήσιμη πληροφορία με την προβολή και το στατιστικό τεστ στη συνέχεια θα την ταυτοποιούσε ως μία ομάδα χωρίς να κάνει τον απαραίτητο διαχωρισμό. Στη προσέγγισή μας χρησιμοποιούμε τις κύριες συνιστώσες τις οποίες το άθροισμα των ιδιοτιμών τους δεν ξεπερνάει το 90% του συνολικού αθροίσματος αυτών. Αυτό σημαίνει πως η χρησιμοποίηση των κυρίων συνιστωσών γίνεται πλέον δυναμικά και εξαρτάται από τη δομή της κάθε ομάδας. Το 90% του συνολικού αθροίσματος είναι ικανό να κρατήσει μόνο την πληροφορία η οποία θεωρείται απαραίτητη σε κάθε ομάδα και έτσι εάν η ομάδα έχει τη δομή που έχουμε στο σχήμα 4.2.1 τότε είναι αρκετή η χρησιμοποίηση μόνο της 1ης κύριας συνιστώσας, όμως εάν εμφανίζονται πιο σύνθετες δομές όπως στο σχήμα 4.2.2 τότε χρησιμοποιείται η 2η, η 3η κ.ο.κ όπου αυτό κριθεί απαραίτητο.

- **Προβολές στους άξονες**. Μαζί με τις ευθείες που προέρχονται από την διαδικασία PCA χρησιμοποιούνται και προβολές στους άξονες x, y, z κ.ο.κ ανάλογα με τη διάσταση των αντικειμένων του συνόλου δεδομένων κάθε φορά. Οι άξονες είναι ικανοί να βρουν πληροφορία σε ομάδες οι οποίες να έχουν μορφή που να εξαρτάται από αυτούς σε μία ή περισσότερες από τις διαστάσεις του συνόλου δεδομένων.
- **Τυχαίες προβολές**. Το τελευταίο είδος προβολών που μπορεί να χρησιμοποιηθεί είναι προβολές σε τυχαίες ευθείες. Στα [22] και [23] έχει αποδειχθεί πως η χρήση επαρκούς

αριθμού τυχαίων ευθειών για προβολή των ομάδων μπορεί να ανακαλύψει επιτυχώς τη δομή Γκαουσιανών ομάδων με πιθανότητα λάθους ε = 1% αν χρησιμοποιηθούν 12 ευθείες και ε = 0.1 % αν χρησιμοποιηθούν 18. Η χρήση τους είναι προαιρετική στον αλγόριθμο pdip-means και μπορεί να καθοριστεί το πλήθος τους ανάλογα με το χρήστη.

Μετά την προβολή των στοιχείων κάθε ομάδας στις παραπάνω ευθείες με την εκτέλεση της διαδικασίας *data_proj,* το διάνυσμα $projections_{j,prj}$ που παράγεται για κάθε ομάδα c_j, είναι μεγέθους $prj \times n_j$ με *prj* να είναι το πλήθος των προβολών που έχουν επιλεχθεί να χρησιμοποιηθούν για την ομάδα c_j. Σε κάθε *prj* προβολή τότε εφαρμόζεται το Hartigan's dip τεστ και για κάθε ομάδα επιλέγεται η προβολή με το μεγαλύτερο dip σκορ, δηλαδή η ομάδα στην οποία είναι πιο πιθανό να παρατηρείται multimodality. Έπειτα, σε κάθε επανάληψη εξετάζονται όλες οι *k* ομάδες από το dip τεστ για unimodality και βρίσκονται έτσι οι υποψήφιες ομάδες για διαχωρισμό όπως και στον αλγόριθμο dip-means. Για τον υπολογισμό του dip σκορ χρησιμοποιείται ο τύπος της εξίσωσης 3.4.1 που παρουσιάστηκε στο προηγούμενο κεφάλαιο.

Από το σύνολο των υποψηφίων αυτών ομάδων για διαχωρισμό επιλέγεται αυτή με το υψηλότερο σκορ και διαχωρίζεται με τον ίδιο τρόπο όπως και στον αλγόριθμο dip-means, χρησιμοποιώντας ένα τυχαίο σημείο της ομάδας και τοποθετώντας το άλλο στην αντίθετη κατεύθυνση από αυτό. Στο επόμενο κεφάλαιο θα παρουσιαστούν αποτελέσματα πειραμάτων από την εκτέλεση του αλγορίθμου και θα γίνει σύγκριση του με τους υπόλοιπους αλγορίθμους που αναπτύχθηκαν στην παρούσα εργασία.

Συνολικά, ο αλγόριθμος pdip-means έχει αρκετά κοινά σημεία με τον dip-means και αρκετές ουσιαστικές διαφορές που σκοπεύουν είτε να βελτιώσουν τις παραγόμενες λύσεις από το στατιστικό τεστ dip, είτε να μειώσουν τον χρόνο εκτέλεσης του. Έτσι, με την εισαγωγή προβολών για τη μείωση της διάστασης των δεδομένων, αντιμετωπίζονται σύνολα δεδομένων με μεγαλύτερη διάσταση πιο αποτελεσματικά χωρίς να χάνεται σημαντική πληροφορία από τις ομάδες των δεδομένων. Επιπλέον με τη χρησιμοποίηση μόνο της προβολής που πετυχαίνει το μεγαλύτερο σκορ για τον μετέπειτα διαχωρισμό μίας εκ των υποψήφιων ομάδων, εξασφαλίζεται πως το τρέχον μοντέλο που έχει παράγει ο pdip-means

βελτιώνεται σταδιακά και χωρίς την πραγματοποίηση τυχών "λάθος" διαχωρισμών. Με τη χρήση offline υπολογισμού των δειγμάτων bootstrap, πετυχαίνουμε όχι μόνο σημαντική βελτίωση κατά την εκτέλεση του αλγορίθμου, αλλά μειώνεται αισθητά και η πιθανότητα εμφάνισης λαθών στον υπολογισμό της p-value κυρίως όταν έχουμε ομάδες με μεγάλο πλήθος στοιχείων. Τέλος επιπρόσθετη βελτίωση στο χρόνο εκτέλεσης αλλά και σε μερικές περιπτώσεις στις λύσεις μπορεί να επιτευχθεί με τη προαιρετική εφαρμογή του αλγόριθμου K-Μέσων στο τέλος κάθε επανάληψης και αντί για αυτή, να γίνεται μόνο μία στο τέλος όλων των επαναλήψεων του αλγορίθμου ή και καθόλου ώστε να παράγεται η λύση της ομαδοποίησης ακόμη γρηγορότερα. Τέλος, με τη μη χρησιμοποίηση του αλγορίθμου K-Μέσων σε μερικές περιπτώσεις αποφεύγονται και λανθασμένες αναθέσεις καθώς ο κλασσικός αλγόριθμος K-Μέσων ευνοεί τις ομάδες σφαιρικής Γκαουσιανής δομής.

4.3. Αλγόριθμος Agglodip

Ο επόμενος από τους αλγορίθμους που αναπτύχθηκαν στο πλαίσιο αυτής της εργασίας ονομάζεται Agglodip (Agglomerative Dip-means). Ο αλγόριθμος agglodip σε αντίθεση με τους αλγορίθμους που παρουσιάστηκαν έως τώρα ακολουθεί διαφορετική προσέγγιση δεδομένου ότι είναι συσσωρευτικής φύσης. Ακολουθεί δηλαδή μειωτική πορεία για την εύρεση του αριθμού των ομάδων, ξεκινώντας από έναν αριθμό $k = k_{init}$ μεγαλύτερο από τον αριθμό των ομάδων που αναμένουμε να βρούμε στο εκάστοτε σύνολο δεδομένων. Η εξασφάλιση αυτή πως $k_{init} > k_{final}$ αποτελεί μια θεωρητική αδυναμία των αλγορίθμων συσσωρευτικής φύσεως η οποία όμως μπορεί να ξεπεραστεί όπως θα παρουσιαστεί παρακάτω. Έχοντας ως αρχικό αριθμό ομάδων $k = k_{init}$, μειώνεται σταδιακά το πλήθος τους κατά 1 σε κάθε επανάληψη, βελτιώνοντας το τρέχον μοντέλο M έως ότου φτάσουμε στο επιθυμητό μοντέλο που περιγράφει καλύτερα το σύνολο δεδομένων που έχει δοθεί για ομαδοποίηση. Τη λύση εκτός από τον αριθμό των ομάδων k που περιγράφει καλύτερα το σύνολο δεδομένων συνθέτει και το μοντέλο που αντιστοιχεί στο k αυτό.

Στους αλγορίθμους που προηγήθηκαν, μεγάλο ρόλο έπαιζε ο διαχωρισμός των ομάδων σε κάθε επανάληψη. Αντίθετα, στην περίπτωση του agglodip λόγω της συσσωρευτικής του φύσης συντελείται ένωση (merging) δύο ομάδων σε μία ανάλογα με το

αποτέλεσμα που δίνει η εφαρμογή του στατιστικού κριτηρίου Hartigan's Dip Statistic σε αυτές. Στον παρακάτω πίνακα (Πίνακας 4.3.1) παρατίθεται το σύνολο των συμβολισμών που θα χρησιμοποιηθούν κατά την ενότητα αυτή:

Συμβολισμός	Περιγραφή
X	Το σύνολο δεδομένων
x_i	Ένα στοιχείο του συνόλου δεδομένων
C	Το σύνολο των στοιχείων κάθε ομάδας δηλαδή $C = \{c_j\}_{j=1}^{k}$
c_j	Το σύνολο των αντικειμένων της j-οστής ομάδας
M	Τα μοντέλα, δηλαδή $M = \{m_j\}_{j=1}^{k}$
m_j	Το j-οστό κεντροειδές που αντιστοιχεί στην ομάδα c_j
k	Το πλήθος των ομάδων
N	Το πλήθος των στοιχείων του συνόλου δεδομένων X
n_j	Το πλήθος των στοιχείων της j-οστής ομάδας

Πίνακας 4.3.1 Συμβολισμοί αλγορίθμου Agglodip

4.3.1. Δομή του αλγορίθμου

Ο αλγόριθμος agglodip πέρα από τη χρήση του στατιστικού κριτηρίου dip παρουσιάζει λίγες ομοιότητες με τον dip-means το οποίο είναι επακόλουθο της διαφορετικής φιλοσοφίας στην οποία στηρίζεται. Αποτελείται από τέσσερα βασικά βήματα τα οποία καθορίζουν τη λειτουργία του. Στο πρώτο βήμα, σε αντίθεση με τις προηγούμενες προσεγγίσεις, εκτελείται ο αλγόριθμος ομαδοποίησης *global_kmeans* [24] και παράγεται μια αρχική λύση ομαδοποίησης με μεγάλο αριθμό ομάδων. Ο *global_kmeans* παίρνει τη θέση του κλασσικού αλγορίθμου K-Μέσων, ο οποίος όμως θα μπορούσε εναλλακτικά να χρησιμοποιηθεί. Στο δεύτερο βήμα εκτελούνται δύο λειτουργίες: merge_small_clusters και unimodality_checking. Η merge_small_clusters όπως υποδεικνύει και η ονομασία της ενώνει ομάδες μικρού μεγέθους. Το ποια ομάδα θεωρείται μικρή καθορίζεται από μια παράμετρο s_{thd} η οποία υποδηλώνει πως κάθε ομάδα c_j, με $j=1,\ldots,k$, η οποία έχει σε πλήθος λιγότερα στοιχεία από την παραπάνω παράμετρο θα ενωθεί με κάποια άλλη ομάδα έως ότου να μην υπάρχουν "μικρές" ομάδες

αρχικά στο σύνολο δεδομένων. Η λειτουργία unimodality_checking εκτελείται στη συνέχεια και ελέγχει μία προς μία τις ομάδες για unimodality. Οταν κάποια ομάδα δεν είναι μονοτροπική, την διασπά έως ότου η αρχική λύση να αποτελείται μόνο από unimodal ομάδες. Έπειτα, στο επόμενο βήμα (που αποτελεί τον κυρίως κορμό του αλγορίθμου) κάθε ζεύγος ομάδων του τρέχοντος μοντέλου ελέγχεται για unimodality, υπολογίζεται το σκορ του βάσει του στατιστικού κριτηρίου dip και στο τελικό κομμάτι ενώνεται το ζεύγος ομάδων που πετυχαίνει το καλύτερο σκορ από το σύνολο όλων των υποψηφίων ζευγών. Στον πίνακα 4.3.2 που ακολουθεί στη συνέχεια υπάρχει ο ψευδοκώδικας του αλγορίθμου agglodip:

Αλγόριθμος Agglodip (Agglomerative Dip-Means)

1. Είσοδος του αρχικού αριθμού των ομάδων k_{init}, του επιπέδου εμπιστοσύνης α για το unimodality (dip) τεστ, του ποσοστού v_{thd} των θεατών διαχωρισμού μιας ομάδας ώστε αυτή να θεωρηθεί υποψήφια για διαχωρισμό καθώς και μιας παραμέτρου s_{thd} που εκφράζει το ελάχιστο πλήθος στοιχείων που θα έχει μια ομάδα.
2. $\{C, M\} \leftarrow \text{global\_kmeans}(X, k_{init})$.
3. $\{C, M, k\} \leftarrow \text{merge\_small\_clusters}(C, M, s_{thd})$.
4. $\{C, M, k\} \leftarrow \text{unimodality\_checking}(X, C, k)$.
5. *Επανάλαβε*
6. *Για* $i = 1$ έως $k - 1$
7. *Για* $j = i + 1$ έως k
8. //Η *tempc* είναι μια προσωρινή ομάδα που προκύπτει από την
9. //ένωση των ομάδων c_i και c_j
10. $tempc \leftarrow \{c_i, c_j\}$.
11. $score_j \leftarrow \text{unimodalityTest}(tempc, a, v_{thd})$.
12. *Τέλος για*
13. $score_i \leftarrow \max(score_j)$.
14. *Τέλος για*
15. *Εάν* $\max_i(score_i) > 0$
16. $\{cmerge_{i1}, cmerge_{i2}\} \leftarrow \arg\max_{i1,i2}(score_i)$.
17. $c_{cmerge_{i1}} \leftarrow \text{mergeCluster}(c_{cmerge_{i1}}, c_{cmerge_{i2}})$.

18. $m_{cmerge_{i1}} \leftarrow \text{computeCentroid}(c_{cmerge_{i1}})$.
19. $M \leftarrow \{M - m_{cmerge_{i2}}, m_{cmerge_{i1}}\}$.
20. *Τέλος Εάν*
21. *Έως ότου δεν γίνονται ενώσεις ομάδων*.
22. Βελτίωση της παραγόμενης λύσης $\{C, M\} \leftarrow \text{K-Μέσων}(X, M)$.
23. Επέστρεψε την τελική λύση, δηλαδή τα *C* και *M.*

Πίνακας 4.3.2 Ψευδοκώδικας αλγορίθμου Agglodip

Ο αλγόριθμος agglodip σε αντίθεση με τους dip-means και pdip-means ξεκινά με την εκτέλεση του αλγορίθμου global k-means αντί του κλασσικού αλγορίθμου K-Μέσων. Ο global k-means [24] αποτελεί ένα ντετερμινιστικό αλγόριθμο ομαδοποίησης αυξητικής φύσης που δοθέντος του αριθμού των ομάδων *k* όπως και στον κλασσικό αλγόριθμο K-Μέσων ομαδοποιεί το σύνολο δεδομένων σε *k* ομάδες. Η βασική ιδέα πίσω από αυτόν στηρίζεται στο γεγονός πως η βέλτιστη λύση για ένα πρόβλημα ομαδοποίησης με *k* ομάδες μπορεί να προέλθει από τη βέλτιστη λύση του προβλήματος για *k*-1 ομάδες και αυτό με τη σειρά του από τη λύση του για *k*-2 ομάδες κ.ο.κ. Η διαδικασία αυτή ξεκινά αυξητικά και κατά τη διάρκειά της εκτελείται *N* φορές ο αλγόριθμος K-Μέσων, μία για κάθε σημείο του συνόλου δεδομένων, ώστε να βρεθεί η βέλτιστη θέση των *k* κέντρων και κατά συνέπεια και το βέλτιστο μοντέλο που αντιστοιχεί σε αυτά. Ο global k-means έχει επιλεγεί έναντι του αντίστοιχου κλασσικού αλγορίθμου K-Μέσων λόγω της μη εξάρτησης του από την αρχικοποίηση των κέντρων η οποία αποτελεί ένα από τα προβλήματα του αλγορίθμου K-Μέσων και μειώνει σημαντικά την απόδοση του. Στη θέση του μπορεί βεβαίως να χρησιμοποιηθεί ο κλασσικός αλγόριθμος K-Μέσων για μείωση του χρόνου εκτέλεσης.

Στη συνέχεια εκτελείται η λειτουργία merge_small_clusters παίρνοντας ως είσοδο το αποτέλεσμα του αλγορίθμου global k-means καθώς και την παράμετρο s_{thd}. Η λειτουργία της έγκειται όπως υποδηλώνει το όνομά της στην ένωση "μικρών" ομάδων. Το ποια ομάδα λογίζεται ως μικρή καθορίζεται από την παράμετρο s_{thd}. Ομάδες στις οποίες ισχύει $n_j < s_{thd}$ ενώνονται με τις κοντινότερες σε αυτές ομάδες και το νέο κέντρο της ομάδας θα είναι το κεντροειδές που προκύπτει μετά την ένωσή τους. Η διαδικασία αυτή εκτελείται επαναληπτικά έως ότου όλες οι ομάδες που έχουν διατηρηθεί να μην θεωρούνται πλέον "μικρές".

Ακολουθεί η εκτέλεση της λειτουργίας unimodality_checking κατά την οποία κάθε ομάδα που προήλθε από την merge_small_clusters εξετάζεται για unimodality. Ο έλεγχος αυτός γίνεται με τη χρήση του Hartigan's Dip Test με την εφαρμογή του τοπικά σε κάθε ομάδα. Εάν κάποια ομάδα δεν είναι unimodal, διαχωρίζεται έως ότου οι επιμέρους συνιστώσες που την απαρτίζουν να προέρχονται από unimodal κατανομές. Η χρησιμότητα της συγκεκριμένης διαδικασίας είναι διττή. Από τη μία αποφεύγονται σπάνιες περιπτώσεις στις οποίες ενδέχεται να παραχθεί μια μη unimodal ομάδα λόγω της merge_small_clusters, Από την άλλη, σε περιπτώσεις που δοθεί αρχικό πλήθος ομάδων k_{init} μικρότερο από αυτό που περιγράφει πραγματικά τα δεδομένα, παρέχεται μια δικλείδα ασφαλείας η οποία διορθώνει λάθη αυτής της μορφής. Εκτελώντας διαχωρισμούς η unimodality_checking αυξάνει τον αριθμό των ομάδων με τη βοήθεια του dip-means ο οποίος εφαρμόζεται τοπικά, και δημιουργεί έτσι ένα μοντέλο όπου κάθε ομάδα είναι unimodal και το οποίο συνολικά θα έχει $k_{init} \geq k$ ώστε στη συνέχεια να εφαρμοστεί επαναληπτικά το κυρίως μέρος του αλγορίθμου agglodip.

4.3.2. Ένωση των ομάδων

Το σύνολο των αλγορίθμων που έχει παρουσιαστεί έως τώρα αποτελείται από αλγορίθμους αυξητικούς, που ξεκινούν με αρχικό αριθμό ομάδων k_{init} ίσο με τη μονάδα και στη συνέχεια με διαδοχικούς διαχωρισμούς καταλήγουν στο ιδανικό *k* ανάλογα με τα αποτελέσματα των στατιστικών τους τεστ κάθε φορά. Ο αλγόριθμος agglodip ανήκει στην κατηγορία των συσσωρευτικών αλγορίθμων και λειτουργεί με την αντίστροφη φιλοσοφία. Ξεκινάει με μια διαμέριση από *k* ομάδες του συνόλου δεδομένων, με το *k* αυτό να είναι μεγαλύτερο του πραγματικού και επαναληπτικά καταλήγει στη λύση που περιγράφει καλύτερα τη δομή των δεδομένων βάσει του κριτηρίου dip.

Αρχικά, δημιουργείται μια προσωρινή ομάδα *tempc*, η οποία αποτελείται από τα μέλη δύο άλλων ομάδων c_i, c_j και αυτή η νέα ομάδα εξετάζεται για unimodality με τον ίδιο ακριβώς τρόπο με τον οποίο εφαρμόζεται το dip τεστ στον αλγόριθμο dip-means που παρουσιάστηκε σε προηγούμενο κεφάλαιο, δηλαδή υπολογίζεται ο πίνακας αποστάσεων των

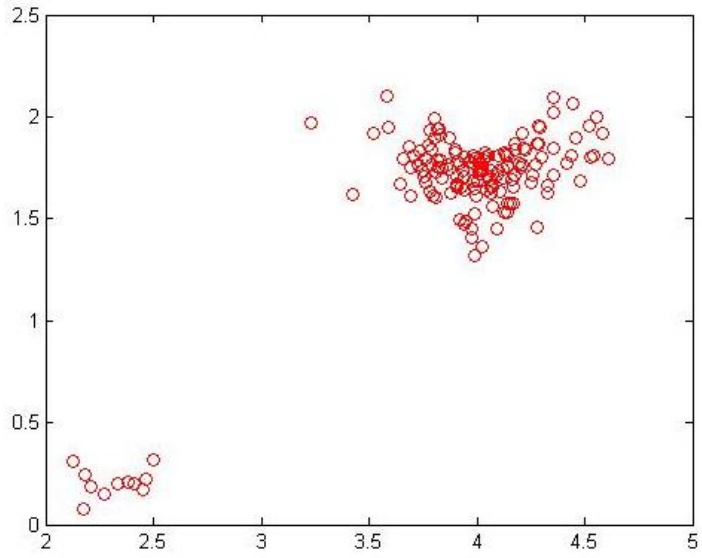
στοιχείων των ομάδων και με τη χρήση του dip τεστ υπολογίζεται το σκορ της προσωρινής ομάδας *tempc*. Αυτό εφαρμόζεται για όλους τους πιθανούς συνδυασμούς ζευγών ομάδων και τα σκορ τους ενημερώνουν ένα δυναμικό πίνακα μεγέθους k^2 ο οποίος αλλάζει μέγεθος κάθε φορά που γίνεται μια νέα ένωση. Περιπτώσεις όπως η δημιουργία ομάδας με τον εαυτό της και εξέταση της ένωσης των αντιστρόφων ζευγαριών δεν υπολογίζονται και συνεπώς ο παραπάνω πίνακας περιέχει στοιχεία πάνω από την κύρια διαγώνιό του, δηλαδή είναι άνω τριγωνικός.

Έπειτα, από τον παραπάνω πίνακα επιλέγεται το ζευγάρι των ομάδων το οποίο έχει πετύχει το καλύτερο σκορ και τα στοιχεία τους ενώνονται πλέον σε μία ενιαία ομάδα. Υπολογίζεται στη συνέχεια το νέο κεντροειδές που θα αντικαταστήσει τα δύο παλαιότερα και θα αντιστοιχεί στη νέα αυτή ομάδα και ανανεώνεται έτσι το τρέχον μοντέλο το οποίο θα περιέχει πλέον μία ομάδα λιγότερη. Στην επόμενη επανάληψη ο δυναμικός πίνακας που κρατάει τα σκορ των ενώσεων των ζευγών των ομάδων λόγω της προηγούμενης ένωσης θα μειωθεί κατά μία γραμμή και μία στήλη σε μέγεθος και θα υπολογιστούν μόνο τα σκορ της νέας ομάδας με τις υπόλοιπες. Συνολικά λοιπόν θα έχουμε πως κατά την πρώτη επανάληψη του αλγορίθμου θα γίνουν $\frac{k_{init}^2 - k_{init}}{2}$ έλεγχοι για unimodality ενώ σε κάθε επόμενη επανάληψη θα γίνονται $k_{current} - 1$ τέτοιοι έλεγχοι, με $k_{current} = k_{init} - 1, ..., k_{final}$ και $k_{final} < k_{init}$. Χρησιμοποιώντας τις παραπάνω σχέσεις προκύπτει πως συνολικά κατά την εκτέλεση του agglodip θα γίνουν $\frac{k_{init}^2 - k_{init}}{2} + \sum_{k=k_{final}}^{k=k_{init}} (k-1)$ εφαρμογές του κριτηρίου συνένωσης ομάδων, το οποίο αποτελεί το μεγαλύτερο σε κόστος τμήμα του αλγορίθμου.

Αντίθετα με τους αλγορίθμους dip-means και pdip-means που υπολογίζουν το σκορ για unimodality μιας ομάδας με τη χρήση της τιμής dip και του πλήθους των θεατών διαχωρισμού (Εξ. 3.4.3), ο agglodip λαμβάνει υπόψη τόσο την τιμή p-value που αντιστοιχεί στο dip σκορ της προσωρινά ενωμένης ομάδας όσο και το μέγεθος καθεμιάς από τις δύο προς ένωση ομάδες για τον υπολογισμό του σκορ που χρησιμοποιεί. Μεγαλύτερη τιμή p-value σημαίνει πως η ομάδα που εξετάζεται περιγράφεται καλύτερα από μία unimodal κατανομή και κατά συνέπεια το ζευγάρι αυτό των ομάδων αποτελεί υποψήφιο για ένωση, ενώ μηδενική p-value υποδηλώνει την ύπαρξη περισσότερων της μίας κορυφών στην ομάδα προς εξέταση.

Οι εναλλακτικές υποθέσεις H_0 και H_1 αντιστρέφονται σε σχέση με τους δύο προηγούμενους αλγορίθμους κάτι αναμενόμενο αν αναλογιστούμε πως εφαρμόζεται η αντίστροφη φιλοσοφία λόγω της συσσωρευτικής φύσης του αλγορίθμου agglodip.

Κατά την εφαρμογή του Hartigan's dip statistic κριτηρίου στον agglodip παρατηρείται σε μερικές περιπτώσεις το εξής φαινόμενο. Έστω ότι το ζεύγος των ομάδων που εξετάζεται προς ένωση είναι το $\{c_i, c_j\}$ με $i \neq j$ και η προσωρινή ομάδα που συντίθεται από τα στοιχεία τους είναι η c_b. Τότε εάν οι c_i και c_j είναι καλά διαχωρισμένες μεταξύ τους και το πλήθος των στοιχείων n_j της c_j είναι κατά πολύ μεγαλύτερο του πλήθους των στοιχείων της c_i ή και το αντίστροφο, ενδέχεται η προσωρινή ομάδα που ελέγχεται με το dip τεστ να χαρακτηριστεί unimodal χωρίς στην πραγματικότητα να είναι. Αυτές οι περιπτώσεις είναι σπάνιες και συμβαίνουν συνήθως όταν μια ομάδα έχει επιβληθεί σε συνεχόμενες ενώσεις, έχει αυξηθεί πολύ το μέγεθός της και συγκρίνεται με μια ομάδα η οποία δεν έχει υποστεί καμία ένωση και το πλήθος των στοιχείων τους διαφέρει σημαντικά. Στο σχήμα 4.3.1 φαίνεται ένα παράδειγμα αυτού του φαινομένου όπου υπάρχει μεγάλη διαφορά μεγέθους μεταξύ των δύο ομάδων:



Σχήμα 4.3.1 Φαινόμενο μεγάλης διαφοράς μεγέθους ομάδων

Για την αντιμετώπιση λοιπόν αυτού του φαινομένου, είναι απαραίτητη η χρήση του πλήθους n_j των στοιχείων κάθε ομάδας που απαρτίζει τα ζεύγη ώστε να δίνεται μεγαλύτερη

προτεραιότητα σε ομάδες που έχουν μικρή διαφορά στο μέγεθος τους και ταυτόχρονα το ζεύγος τους χαρακτηρίζεται unimodal ώστε να μπορούν να ενωθούν. Η σχέση που μας δίνει το σκορ στον αλγόριθμο θα είναι:

$$score_b = pval_b * \frac{n_i}{n_j} \text{ , αν } pval > 0$$

$$score_b = 0 \text{ , αλλιώς} \qquad \text{Εξ. 4.3.1}$$

Στην παραπάνω σχέση η n_i συμβολίζει το πλήθος ομάδας με το μικρότερο αριθμό στοιχείων και αντίστοιχα η n_j αυτή με το μεγαλύτερο από το ζεύγος $\{c_i, c_j\}$ των ομάδων. $pval_b$ είναι η p-value που προκύπτει από το dip τεστ για την προσωρινή ομάδα που αποτελείται από τα στοιχεία των c_i και c_j ομάδων. Γνωρίζουμε επιπλέον πως η p-value $\in [0,1]$ με τις θετικές τιμές να υποδηλώνουν unimodality της ομάδας και όσο μεγαλύτερη τιμή παίρνει, τόσο ισχυρότερος είναι ο ισχυρισμός αυτός από το τεστ. Επίσης, λόγω της παραπάνω υπόθεσης το πηλίκο $\frac{n_i}{n_j}$ ανήκει και αυτό στο ίδιο διάστημα, με τις τιμές κοντά στη μονάδα να είναι ιδανικές, και συνολικά θα έχουμε πως το $score \in [0,1]$. Το σύνολο αυτό των σκορ αποθηκεύεται στο δυναμικό πίνακα που αναφέρθηκε νωρίτερα και επιλέγεται από αυτόν το ζεύγος των ομάδων στο οποίο αντιστοιχεί το μέγιστο. Το ζεύγος $\{i, j\}$ τότε ενώνεται σε μια ομάδα με δείκτη b η οποία θα πάρει τη θέση της ομάδας με δείκτη i στον πίνακα των σκορ και οι εγγραφές που αντιστοιχούν στην j διαγράφονται. Τέλος, υπολογίζεται το κεντροειδές της νέας ομάδας και ανανεώνεται το τρέχον μοντέλο ομαδοποίησης.

4.3.3. Τεχνικές επιτάχυνσης

Οι συσσωρευτικοί αλγόριθμοι είναι γενικά από τη φύση τους πιο ακριβοί σε χρόνο και μνήμη σε σχέση με τους αυξητικούς κάτι που είναι λογικό αν αναλογιστεί κανείς ότι στη γενική περίπτωση ξεκινούν με κάθε αντικείμενο να αποτελεί μια ομάδα έως ότου να καταλήξουν στη λύση. Επιπλέον στις περισσότερες περιπτώσεις το πλήθος των ομάδων της τελικής λύσης

είναι κατά πολύ μικρότερο του αριθμού των στοιχείων του συνόλου δεδομένων. Αυτό έχει ως συνέπεια υπολογιστικά δαπανηρές διαδικασίες όπως είναι η εφαρμογή στατιστικών κριτηρίων να εφαρμόζονται περισσότερες φορές σε σχέση με τις αντίστοιχες αυξητικές εκδοχές των αλγορίθμων.

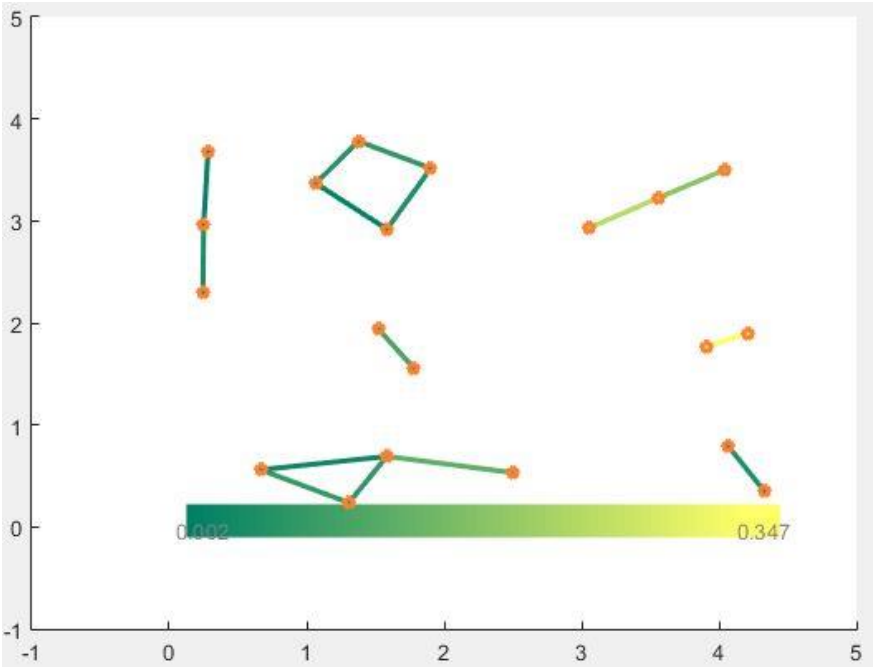
Οι μέθοδοι που εφαρμόστηκαν στον αλγόριθμο pdip-means για τη μείωση του χρόνου εκτέλεσής του αξιοποιήθηκαν και στον agglodip. Ετσι το δείγμα bootstrap των ομοιόμορφων κατανομών και οι αντίστοιχες dip τιμές υπολογίζονται μία φορά offline, πριν την εκτέλεση του αλγορίθμου, αποθηκεύονται και χρησιμοποιούνται κατά την εκτέλεση. Επιπλέον, η εφαρμογή του dip τεστ στον πίνακα αποστάσεων κάθε φορά γίνεται με τη χρήση παραλληλίας κάτι που σε πολυπύρηνες επεξεργαστικές μονάδες μειώνει δραματικά τον χρόνο εκτέλεσης. Μια ακόμη μέθοδος που χρησιμοποιήθηκε στον pdip-means και χρησιμοποιείται και εδώ είναι η προαιρετική εφαρμογή του κλασσικού αλγορίθμου K-Μέσων πριν την εξαγωγή της τελικής λύσης η οποία σε αρκετές περιπτώσεις βελτιώνει το τελικό μοντέλο αλλά μπορεί να παραληφθεί αν κριθεί απαραίτητο για αύξηση της ταχύτητας.

Επιπρόσθετα με τις παραπάνω μεθόδους, εξετάσαμε δύο εναλλακτικές προσεγγίσεις οι οποίες μεταξύ άλλων οδηγούν σημαντική επιτάχυνση του αλγορίθμου. Η πρώτη τεχνική αρχικά αποφεύγει την εφαρμογή του κριτηρίου dip σε όλες τις γραμμές του πίνακα αποστάσεων των δεδομένων των ομάδων που εξετάζονται για συνένωση. Αντί αυτού υπολογίζονται *δύο μόνο* διανύσματα αποστάσεων, θεωρώντας ως θεατές μόνο τα κεντροειδή της καθεμίας από τις δύο ομάδες που εξετάζονται για συνένωση. Έπειτα, το dip τεστ εφαρμόζεται μόνο σε αυτά τα δύο διανύσματα, μειώνοντας δραματικά το πλήθος των ελέγχων που γίνονται σε κάθε επανάληψη. Αυτό έχει σαν συνέπεια, από τον αρχικό υπολογισμό $n_i * n_j$ αποστάσεων για δυο ομάδες c_i και c_j να υπολογίζονται πλέον μόνο $2(n_i + n_j - 1)$ αποστάσεις και έτσι το dip τεστ να έχει πολύ λιγότερες αποστάσεις προς επεξεργασία.

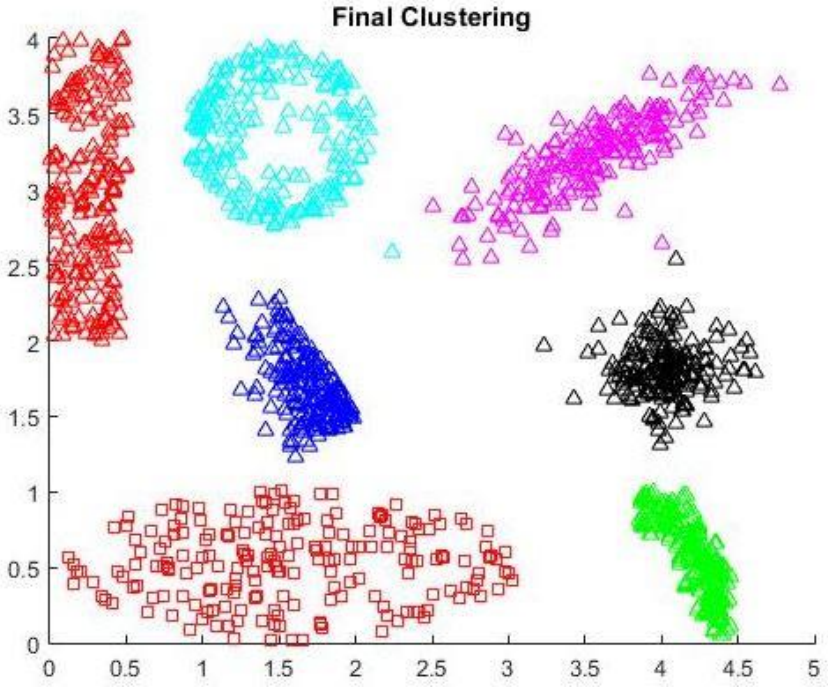
Η δεύτερη τεχνική στηρίζεται στη χρήση γραφήματος και των ιδιοτήτων του για την πραγματοποίηση ενώσεων μεταξύ των ομάδων. Η τεχνική αυτή διαφέρει σημαντικά από την προηγούμενη καθώς και από την κλασσική προσέγγιση καθώς με τη χρήση της συντελείται *ομαδοποίηση σε ένα βήμα* και ο αλγόριθμος χάνει την επαναληπτική φύση του. Η λειτουργία

του αλγορίθμου κατά το πρώτο βήμα εκτελείται κανονικά, δηλαδή υπολογίζονται τα σκορ για όλα τα ζεύγη του *αρχικού συνόλου* ομάδων. Στη συνέχεια από τα σκορ αυτά ορίζεται ένας *πίνακας γειτνίασης μεταξύ των ομάδων του αρχικού μοντέλου* ο οποίος έχει τιμή ίση με 1 για τα ζεύγη ομάδων που πέτυχαν θετική τιμή p-value (δηλαδή επιτρέπεται η συνένωσή τους) και μηδέν για αυτά με μηδενική τιμή p-value (δηλ. δεν επιτρέπεται η συνένωσή τους). Ο πίνακας γειτνίασης μπορεί να αναπαρασταθεί με ένα γράφημα στο οποίο οι κόμβοι αντιστοιχούν στις αρχικές ομάδες και οι ακμές δηλώνουν εάν επιτρέπεται η συνένωσή τους. Η τελική λύση ομαδοποίησης προκύπτει εύκολα βρίσκοντας τις *συνεκτικές συνιστώσες* του γραφήματος αυτού και συνενώνοντας τις ομάδες των οποίων οι αντίστοιχοι κόμβοι ανήκουν στην ίδια συνεκτική συνιστώσα.

Στο σχήμα 4.3.2 παρουσιάζεται μια τέτοια λύση με τη χρήση γραφήματος. Οι κόμβοι που απεικονίζονται είναι 20 και αποτελούν το πλήθος των ομάδων του αρχικού μοντέλου. Οι ακμές μεταξύ τους συμβολίζουν τις ενώσεις τους και κάθε συνεκτική συνιστώσα που αποτελείται από ένα σύνολο τέτοιων κόμβων και ακμών αποτελεί μια ομάδα στην τελική λύση ομαδοποίησης. Επιπλέον, συμβολίζουμε με χρώμα στις ακμές τις τιμές p-value που οδήγησαν στην ένωση, με το κίτρινο να συμβολίζει τις μεγαλύτερες τιμές και με σκούρο πράσινο τις μικρότερες. Στο σχήμα 4.3.3 που παρατίθεται στη συνέχεια παρουσιάζεται η οπτικοποίηση της λύσης για την ίδια περίπτωση.



Σχήμα 4.3.2 Συνεκτικές συνιστώσες γραφήματος γειτνίασης



Σχήμα 4.3.3 Τελική λύση ομαδοποίησης με γράφημα

Τέλος, αξίζει να σημειωθεί πως η τεχνική αυτή μπορεί να συνδυαστεί με την προηγούμενη που κάνει χρήση μόνο των κεντροειδών για τον υπολογισμό του σκορ κάθε ζεύγους υποψηφίων για συνένωση ομάδων επιτυγχάνοντας έτσι ακόμη μεγαλύτερη επιτάχυνση του αλγορίθμου. Είναι προφανές ότι οι παραπάνω δύο τροποποιήσεις μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα σε σχέση με τον βασικό αλγόριθμο agglodip. Το ζήτημα αυτό μελετάται κατά την πειραματική αξιολόγηση των μεθόδων στο κεφάλαιο 5.

4.4. Αλγόριθμος Agglopdip

Ο τελευταίος από τη σειρά των αλγορίθμων ομαδοποίησης που αναπτύχθηκαν κατά την παρούσα εργασία ονομάζεται Agglopdip (Agglomerative Projective Dip-means). Ο agglopdip συνδυάζει στοιχεία από τη φιλοσοφία των αλγορίθμων που έχουν παρουσιαστεί έως τώρα καθώς και μεγάλο μέρος των βελτιώσεων και τεχνικών που χρησιμοποιήθηκαν έχοντας ως στόχο την αποδοτικότερη λειτουργία και την παραγωγή καλύτερων λύσεων από τους τρέχοντες αλγόριθμους που έχουν παρουσιαστεί.

Όπως και οι προηγούμενοι, έτσι και αυτός αποτελεί μία προσπάθεια για τη διόρθωση των ελαττωμάτων του κλασσικού αλγορίθμου κ-μέσων και κυρίως της ανάγκης της χειροκίνητης εισαγωγής του πλήθους k των ομάδων από το χρήστη. Ανήκει στην ίδια κατηγορία αλγορίθμων με τον agglodip καθώς αποτελεί έναν επαναληπτικό αλγόριθμο συσσωρευτικής φύσης που είναι κάλυμμα γύρω από τον αλγόριθμο κ-μέσων. Ξεκινά δηλαδή από έναν αριθμό ομάδων k_{init} μεγαλύτερο από τον τελικό k_{final} που θα παραχθεί στο μοντέλο της λύσης της ομαδοποίησης και σταδιακά σε κάθε επανάληψη μειώνεται κατά ένα το πλήθος τους. Αυτό βελτιώνει το τρέχον μοντέλο κάθε επανάληψης έως ότου να υπάρξει ένα μοντέλο το οποίο να μην επιδέχεται περαιτέρω βελτίωση σύμφωνα με τη λειτουργία του αλγορίθμου.

Πολύ σημαντικό ρόλο και εδώ παίζει το στατιστικό κριτήριο που χρησιμοποιείται. Λόγω της ευελιξίας και της αυξημένης αποδοτικότητας που έχει επιδείξει κατά την αναγνώριση χρήσιμων δομών σε ομάδες δεδομένων, επιλέχθηκε το Hartigan's Dip Statistic και σε αυτή την προσέγγιση. Στα πρότυπα του agglodip λοιπόν, έπειτα από κάθε εφαρμογή του dip τεστ, ακολουθεί η ένωση (merging) ή όχι δυο ομάδων του συνόλου δεδομένων σε μία, ανάλογα με τα αποτελέσματα που δίνει ως έξοδο το κριτήριο dip, ώστε να καταλήξει στο τελικό μοντέλο που περιγράφει όσο το δυνατόν καλύτερα τα δεδομένα του προβλήματος.

Ο πίνακας που παρατίθεται στη συνέχεια (Πίνακας 4.4.1) περιέχει τους βασικότερους συμβολισμούς που θα χρησιμοποιηθούν σε αυτή την ενότητα για την συνοπτική παρουσίαση του αλγορίθμου agglopdip:

Συμβολισμός	Περιγραφή
X	Το σύνολο δεδομένων
x_i	Ένα στοιχείο του συνόλου δεδομένων
C	Το σύνολο των στοιχείων κάθε ομάδας δηλαδή $C = \{c_j\}_{j=1}^{k}$
c_j	Το σύνολο των αντικειμένων της j-οστής ομάδας
M	Τα μοντέλα, δηλαδή $M = \{m_j\}_{j=1}^{k}$
m_j	Το j-οστό κεντροειδές που αντιστοιχεί στην ομάδα c_j
k	Το πλήθος των ομάδων
n_j	Το πλήθος των στοιχείων της j-οστής ομάδας

$projections_{j,prj}$	Το σύνολο του πλήθος *prj* των προβολών που αντιστοιχεί στη *j*-οστή ομάδα

Πίνακας 4.4.1 Συμβολισμοί αλγορίθμου Agglopdip

4.4.1. Δομή και λειτουργία του αλγορίθμου

Η δομή και η λειτουργία του αλγορίθμου agglopdip ακολουθούν τη φιλοσοφία των δύο προηγούμενων αλγορίθμων που παρουσιάστηκαν. Ο agglopdip μπορούμε να πούμε πως αποτελεί την συσσωρευτική εκδοχή του αλγορίθμου pdip-means αφού έχει δανειστεί τα κυριότερα χαρακτηριστικά του από τους pdip-means και agglodip με σκοπό την εφαρμογή της μεθόδου χρήσης προβολών ευθειών του πρώτου σε μια συσσωρευτική προσέγγιση ακολουθώντας τις τεχνικές που έχουν εφαρμοστεί στον δεύτερο.

Αποτελείται από 5 βασικά κομμάτια, με τα δεδομένα που απαιτούνται για είσοδο από το χρήστη να είναι τα ίδια με αυτά που απαιτούνται και στον agglodip, δηλαδή εκτός του συνόλου δεδομένων, απαιτείται και η εισαγωγή του αρχικού αριθμού των ομάδων k_{init} από όπου θα ξεκινήσει τη διαδικασία ομαδοποίησης ο αλγόριθμος, το επίπεδο εμπιστοσύνης α που θα χρησιμοποιηθεί από το κριτήριο dip, το ποσοστό των θεατών διαχωρισμού v_{thd} που απαιτείται για να χαρακτηριστεί μια ομάδα multimodal και το ελάχιστο μέγεθος s_{thd} που επιτρέπεται να έχει μια ομάδα. Το 1ο και το 2ο κομμάτι του αλγορίθμου περιέχουν τις διαδικασίες που χρησιμοποιούνται και στον αλγόριθμο agglodip. Έτσι, εφαρμόζεται αρχικά ο ντετερμινιστικός αλγόριθμος ομαδοποίησης global k-means ο οποίος είναι υπεύθυνος για τον καθορισμό του αρχικού μοντέλου M_{init} ή τη βελτίωσή του, αν δοθεί ένα αρχικό μοντέλο του συνόλου δεδομένων ως είσοδος. Η ανθεκτικότητα που επιδεικνύει ο αλγόριθμος αυτός στο πρόβλημα της αρχικοποίησης είναι η κύρια αιτία που οδήγησε στο να προτιμηθεί αντί του κλασσικού αλγορίθμου κ-μέσων.

Έπειτα, η ρουτίνα merge_small_clusters εφαρμόζεται στο αρχικό μοντέλο M_{init} και εξαλείφει τις ομάδες μικρού μεγέθους από αυτό. Μια ομάδα λογίζεται μικρή με βάση την τιμή της s_{thd} που έχει δοθεί ως είσοδος και αν θεωρηθεί μικρή τότε πραγματοποιείται ένωση

μεταξύ αυτής και της κοντινότερης ομάδας προς αυτήν. Ακολουθεί η ρουτίνα unimodality_checking της οποίας σκοπός είναι η απαλλαγή του μοντέλου από τυχόν multimodal ομάδες. Η διαδικασία αυτή εκτελεί τοπικά σε κάθε υπάρχουσα ομάδα του μοντέλου τον αλγόριθμο pdip-means και με αυτόν τον τρόπο όλες οι ομάδες που θα χρησιμοποιηθούν στο κύριο κομμάτι του agglodip για την διεξαγωγή της ομαδοποίησης προέρχονται από unimodal κατανομές. Στον πίνακα 4.4.2 που ακολουθεί, παρατίθεται υπό τη μορφή ψευδοκώδικα ο αλγόριθμος agglodip και στη συνέχεια περιγράφεται ο βασικός κορμός του.

Αλγόριθμος Agglopdip (Agglomerative Projective Dip-Means)

1. Είσοδος του αρχικού αριθμού των ομάδων k_{init}, του επιπέδου εμπιστοσύνης α για το unimodality (dip) τεστ καθώς και μιας παραμέτρου s_{thd} που εκφράζει το ελάχιστο πλήθος στοιχείων που θα έχει μια ομάδα.
2. $\{C, M\} \leftarrow \text{global\_kmeans}(X, k_{init})$.
3. $\{C, M, k\} \leftarrow \text{merge\_small\_clusters}(C, M, s_{thd})$.
4. $\{C, M, k\} \leftarrow \text{unimodality\_checking}(X, C, k)$.
5. *Επανάλαβε*
6. *Για* i = 1 έως k – 1
7. *Για* $j = i + 1$ έως k
8. //Η *tempc* είναι μια προσωρινή ομάδα που προκύπτει από την
9. //ένωση των ομάδων c_i και c_j
10. $tempc \leftarrow \{c_i, c_j\}$.
11. $projections_L \leftarrow data\_proj(\text{tempc})$.
12. $dip\_score_{\text{tempc},L} \leftarrow \text{unimodalityTest}(projections_L, a)$.
13. $pdip\_score_{tempc} = \max_L(dip\_score_{\text{tempc},L})$
14. *Τέλος για*
15. *Τέλος για*
16. *Εάν* $\max_{tempc}(pdip\_score_{tempc}) > 0$
17. $\{cmerge_{i1}, cmerge_{i2}\} \leftarrow \arg\max_{i1,i2}(pdip\_score_{tempc})$.
18. $c_{cmerge_{i1}} \leftarrow \text{mergeCluster}(c_{cmerge_{i1}}, c_{cmerge_{i2}})$.

19. $m_{cmerge_{i1}} \leftarrow \text{computeCentroid}\,(c_{cmerge_{i1}})$.
20. $M \leftarrow \{M - m_{cmerge_{i2}}, m_{cmerge_{i1}}\}$.
21. *Τέλος Εάν*
22. *Έως ότου δεν γίνονται ενώσεις ομάδων.*
23. Βελτίωση της παραγόμενης λύσης $\{C, M\} \leftarrow \text{K-Μέσων}\,(X, M)$.
24. Επέστρεψε την τελική λύση, δηλαδή τα *C* και *M.*

Πίνακας 4.4.2 Ψευδοκώδικας αλγορίθμου Agglopdip

Στο επόμενο κομμάτι του αλγορίθμου, πραγματοποιείται μείωση της διάστασης των στοιχείων του τρέχοντος μοντέλου μέσω της ρουτίνας data_proj. Η ρουτίνα αυτή, όπως παρουσιάστηκε και σε προηγούμενη ενότητα χρησιμοποιεί ένα δυναμικό πλήθος ευθειών, ώστε να προβληθούν τα αντικείμενα κάθε ομάδας στις ευθείες αυτές και στη συνέχεια να ομαδοποιηθούν βάσει των προβολών τους. Το σύνολο των ευθειών που χρησιμοποιούνται αποτελείται όπως και στον αλγόριθμο pdip-means από τις σημαντικότερες συνιστώσες που προκύπτουν από PCA, από τους άξονες συντεταγμένων κάθε διάστασης καθώς και από ένα πλήθος τυχαίων κάθε φορά ευθειών. Έτσι εξασφαλίζεται πως η χρήσιμη πληροφορία των αρχικών δεδομένων έχει αποτυπωθεί στις προβολές τους, με το διάνυσμα $projections_L$ να τις περιέχει, για κάθε προσωρινή ομάδα *tempc* που δημιουργείται.

Το 4ο κομμάτι αφορά το στατιστικό κριτήριο Hartigan's Dip statistic. Πιο συγκεκριμένα, το dip τεστ εφαρμόζεται σε καθένα από τα παραπάνω διανύσματα προβολών που έχουν παραχθεί από τη ρουτίνα data_proj και υπολογίζεται το σκορ κάθε προβολής, για κάθε προσωρινή ομάδα *tempc*. Ο υπολογισμός του σκορ αυτού όπως και στον αλγόριθμο agglodip, εξαρτάται από τις τιμές p-values που αντιστοιχούν στα dip σκορ κάθε προσωρινής ομάδας και επίσης λαμβάνεται υπόψη και το μέγεθος n_j των ομάδων c_j με $j=1,\ldots,k$. Έτσι, το σκορ για κάθε προβολή κάθε προσωρινής ομάδας υπολογίζεται από τη σχέση 4.3.1 ακολουθώντας την ίδια λογική. Έτσι, από το σύνολο των σκορ αυτών για κάθε ομάδα επιλέγεται το μέγιστο το οποίο θα την αντιπροσωπεύει. Θετικές τιμές του σκορ κατατάσσουν την ομάδα στο σύνολο των υποψηφίων ομάδων για ένωση.

Από το παραπάνω σύνολο των υποψηφίων προσωρινών ομάδων προς ένωση επιλέγεται αυτή με το μεγαλύτερο σκορ, και οι δύο επιμέρους ομάδες του μοντέλου M_k,

$k = k_{init}, ..., k_{final}$ με $k_{final} \leq k_{init}$, που την αποτελούν ενώνονται μόνιμα πλέον στο τελευταίο κομμάτι του αλγορίθμου. Τα δύο κεντροειδή m_{old1} και m_{old2} που υπήρχαν στις επιμέρους ομάδες διαγράφονται και το κέντρο m_{new} γίνεται πλέον το κεντροειδές της νέας ομάδας, το οποίο υπολογίζεται από τη ρουτίνα computeCentroid. Έπειτα, το μοντέλο ενημερώνεται ώστε να περιέχει τη νέα ομάδα στη θέση των δύο παλαιότερων επιμέρους που την αποτελούσαν. Η παραπάνω διαδικασία γίνεται επαναληπτικά μέχρι να μην υπάρχει πλέον υποψήφια προς ένωση ομάδα, δηλαδή το μοντέλο να μην επιδέχεται περεταίρω βελτίωση σύμφωνα με τον αλγόριθμο.

Να σημειωθεί πως οι βελτιώσεις που εφαρμόζονται στον αλγόριθμο agglodip εφαρμόζονται και εδώ, ώστε να πετυχαίνουμε όσο το δυνατό μικρότερο χρόνο εκτέλεσης και ποιοτικότερες λύσεις σε σχέση με τις κλασσικές προσεγγίσεις των προγενέστερων αλγορίθμων. Η διαφορά όμως με αυτόν έγκειται στο γεγονός πως η τεχνική υπολογισμού των αποστάσεων από δύο κέντρα προς τα στοιχεία των ομάδων των ζευγών τους δεν μπορεί να εφαρμοστεί καθώς απαιτείται ο υπολογισμός του πίνακα αποστάσεων κάτι που αποφεύγεται στον agglopdip αφού χρησιμοποιούνται οι προβολές των στοιχείων του συνόλου δεδομένων από το Hartigan’s Dip statistic.

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ

5.1. Εισαγωγή

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν τρεις αλγόριθμοι ομαδοποίησης που βασίζονται στο στατιστικό κριτήριο dip και ακολουθούν διαφορετικές μεταξύ τους προσεγγίσεις. Στόχος των αλγορίθμων αυτών είναι η αντιμετώπιση των αδυναμιών του κλασσικού αλγορίθμου K-μέσων και κυρίως της ανάγκης για εισαγωγή του αριθμού *k* των ομάδων από τον χρήστη κάτι που απαιτεί πρότερη γνώση της δομής που παρουσιάζουν τα δεδομένα στο εκάστοτε πρόβλημα. Επιπλέον στόχο αποτελεί η βελτίωση της ποιότητας των παραγόμενων λύσεων σε σχέση με προγενέστερες προσπάθειες που έχουν γίνει από αλγόριθμους ομαδοποίησης που εκτιμούν αυτόματα των αριθμό των ομάδων.

Στη συνέχεια αυτού του κεφαλαίου θα παρουσιαστούν τα αποτελέσματα από πειράματα σε διάφορα σύνολα δεδομένων που έχουν γίνει με τη χρήση των τριών αλγορίθμων που αναπτύχθηκαν στην παρούσα εργασία. Επιπλέον θα γίνει σύγκριση με τους ήδη υπάρχοντες αλγορίθμους της κατηγορίας αυτής τονίζοντας έτσι την ανωτερότητα που παρέχει η χρήση του στατιστικού τεστ dip χρησιμοποιώντας δείκτες εκτίμησης της ποιότητας της ομαδοποίησης που παράγεται.

5.2. Δείκτες Εκτίμησης Ποιότητας Ομαδοποίησης

Το τι ορίζουμε ως καλή λύση ομαδοποίησης ενός προβλήματος είναι πολλές φορές υποκειμενικό καθώς υπάρχει η θεωρητική αδυναμία του αυστηρού ορισμού της έννοιας της ομάδας. Η εκτίμηση της ποιότητας των αποτελεσμάτων αποτελεί από μόνη της έναν ερευνητικό ζήτημα με ιδιαίτερο ενδιαφέρον. Εχουν προταθεί αρκετοί δείκτες για την μέτρηση της ποιότητας μιας λύσης ομαδοποίησης βάσει διάφορων χαρακτηριστικών.

Οι δείκτες αυτοί χωρίζονται σε δυο κατηγορίες, ανάλογα με τα αρχικά δεδομένα που έχουμε στην κατοχή μας. Έτσι θα έχουμε:

- **Δείκτες εκτίμησης ποιότητας ομαδοποίησης με επίβλεψη (Supervised quality measures)**. Οι δείκτες εκτίμησης ποιότητας ομαδοποίησης με επίβλεψη είναι οι δείκτες εκείνοι οι οποίοι εκμεταλλεύονται τη γνώση της σωστής λύσης ομαδοποίησης, η οποία παρέχεται μαζί με το σύνολο δεδομένων του προβλήματος. Σε αυτή τη λύση δηλαδή συμπεριλαμβάνονται οι πραγματικές κατηγορίες του κάθε αντικειμένου και κατά συνέπεια οι δείκτες αυτοί είναι πιο ακριβείς σε σχέση με αυτούς χωρίς επίβλεψη.
- **Δείκτες εκτίμησης ποιότητας ομαδοποίησης χωρίς επίβλεψη (Unsupervised quality measure)**. Οι δείκτες αυτοί χρησιμοποιούνται σε πραγματικά προβλήματα όπου δεν έχουμε πληροφορία για τη λύση που περιγράφει καλύτερα το πρόβλημα. Αυτοί είναι συνήθως λιγότερο ακριβείς σε σχέση με τους προηγούμενους, όμως αποτελούν ένα μέτρο πέρα από την οπτικοποίηση των δεδομένων για την εκτίμηση της ποιότητας των λύσεων.

5.2.1. Άθροισμα Τετραγωνικών Σφαλμάτων (Sum of Squared Errors – SSE)

Ο πρώτος δείκτης εκτίμησης της ποιότητας της ομαδοποίησης είναι το άθροισμα τετραγωνικών σφαλμάτων ή SSE όπως αλλιώς αποκαλείται. Αποτελεί ένα τρόπο να εκτιμηθεί η ποιότητα της λύσης ενός προβλήματος ομαδοποίησης χωρίς επίβλεψη και βασίζεται στην ιδέα πως καλύτερες λύσεις ομαδοποίησης είναι αυτές που αποτελούνται από πιο συμπαγείς ομάδες, δηλαδή τα στοιχεία κάθε ομάδας να βρίσκονται κοντά στο κέντρο της. Μικρότερες

τιμές του SSE σημαίνουν πιο συμπαγείς λύσεις. Η παρακάτω σχέση (Εξ 5.2.1) δίνει το συνολικό SSE όλων των ομάδων του συνόλου δεδομένων:

$$\text{SSE} = \sum_{j=1}^{k} \sum_{x_i \in c_j} \left\| x_i - m_j \right\|^2 \qquad \text{Εξ. 5.2.1}$$

Το εσωτερικό άθροισμα υποδεικνύει το τοπικό σφάλμα σε κάθε ομάδα, το οποίο θέλουμε να είναι όσο το δυνατό μικρότερο για να χαρακτηριστεί πιο ποιοτική η λύση του εκάστοτε προβλήματος. Σε περιπτώσεις όπου χρησιμοποιείται κάποιο μέτρο ομοιότητας, ιδανικές θεωρούνται πλέον οι μεγαλύτερες τιμές του SSE και τη θέση της ευκλείδειας απόστασης μπορεί να πάρει η συνάρτηση ομοιότητας. Να σημειωθεί πως το SSE δεν αντικατοπτρίζει πάντα την αλήθεια στη μέτρηση της ποιότητας ομαδοποίησης καθώς ευνοεί λύσεις με μεγαλύτερο αριθμό ομάδων.

5.2.2. Δείκτης Rand (Rand Index - RI)

Ο δείκτης Rand [33] όπως και οι δύο επόμενοι δείκτες που θα παρουσιαστούν στη συνέχεια αποτελεί ένα δείκτη εκτίμησης της ποιότητας της ομαδοποίησης με επίβλεψη. Ο δείκτης Rand μετρά την ομοιότητα μεταξύ δύο λύσεων ομαδοποίησης και παίρνει τιμές στο κλειστό διάστημα [0,1] με το 0 να υποδηλώνει τελείως διαφορετικά σύνολα και το 1 ακριβώς ίδια. Έτσι, όταν από ένα σύνολο δεδομένων *X* έχουμε δύο διαμερίσεις του, την *A* και τη *B*, τότε αν ως *S* ορίσουμε το πλήθος των συμφωνιών μεταξύ των *A* και *B*, δηλαδή το πλήθος των στοιχείων τα οποία έχουν κατηγοριοποιηθεί στην ίδια ομάδα στη διαμέριση *A* και στη διαμέριση *B* και με *D* ορίσουμε το σύνολο των διαφωνιών, δηλαδή τα ίδια στοιχεία της διαμέρισης *A* να ανήκουν σε άλλη ομάδα στη διαμέριση *B*, τότε ο δείκτης Rand θα είναι:

$$\text{RI} = \frac{S}{S + D} \qquad \text{Εξ 5.2.2}$$

5.2.3. Προσαρμοσμένος Δείκτης Rand (Adjusted Rand Index – ARI)

Ο προσαρμοσμένος δείκτης Rand [33] αποτελεί μια παραλλαγή του κλασσικού δείκτη Rand και ο ορισμός του βασίζεται σε μια ελαφρώς πολυπλοκότερη προσέγγιση. Δοθέντων των παραπάνω συνόλων *X, A* και *B,* αν είναι $A=\{A_1, A_2, ..., A_{k1}\}$ και $B=\{B_1, B_2, ..., B_{k2}\}$ οι ομάδες κάθε διαμέρισης αντίστοιχα και με S_{ij} το πλήθος των στοιχείων που συμφωνούν στα δύο σύνολα για το ζεύγος των *ij* ομάδων, τότε ο πίνακας γειτνίασης των *A* και *B* θα είναι:

A/B	B_1	B_2	**...**	B_{k2}	**Άθροισμα**
A_1	S_{11}	S_{12}	…	S_{1k2}	a_1
A_2	S_{21}	S_{22}	...	S_{2k2}	a_2
⋮	⋮	⋮	⋱	⋮	⋮
A_{k1}	S_{k11}	S_{k12}	…	S_{k1k2}	a_{k1}
Άθροισμα	b_1	b_2	…	b_{k2}	

Πίνακας 5.2.1 Πίνακας γειτνίασης δείκτη ARI

Με βάση τον Πίνακα 5.2.1 ο δείκτης ARI ορίζεται ως εξής:

$$\text{ARI} = \frac{\sum_{ij}\binom{S_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{S}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{S}{2}} \qquad \text{Εξ 5.2.3}$$

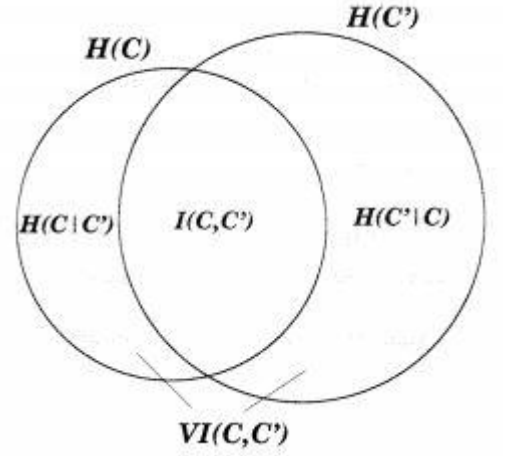
5.2.4. Δείκτης Διακύμανσης της Πληροφορίας (Variance of Information – VI)

Ο τελευταίος δείκτης που θα χρησιμοποιηθεί για την εκτίμηση της ποιότητας της ομαδοποίησης είναι η διακύμανση της πληροφορίας (variance of information) [34]. Όπως και οι δύο προηγούμενοι δείκτες που παρουσιάστηκαν, έτσι και ο VI ανήκει στην κατηγορία δεικτών εκτίμησης της ποιότητας της ομαδοποίησης με επίβλεψη. Ο ορισμός του βασίζεται

στην εντροπία H(C) ενός συνόλου C και στην αμοιβαία πληροφορία I(C,C') δύο συνόλων *C* και *C'*. Στο σχήμα 5.2.1 παρουσιάζεται σχηματικά η ιδέα πίσω από τον VI και η σχέση 5.2.4 μας δίνει τον ορισμό της βάσει της εντροπίας και της αμοιβαίας πληροφορίας.

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \qquad \text{Εξ 5.2.4}$$

Να σημειωθεί επίσης ότι μικρότερες τιμές του VI σημαίνουν γενικά καλύτερα ποιότητα αποτελεσμάτων.



Σχήμα 5.2.1 Ορισμός αμοιβαίας πληροφορίας με τη χρήση της εντροπίας

5.3. Πειραματικά Αποτελέσματα

Στην ενότητα αυτή θα παρουσιαστούν πειραματικά αποτελέσματα από την εκτέλεση των αλγορίθμων pdip-means, agglodip και agglopdip, οι οποίοι αναπτύχθηκαν στην παρούσα εργασία πάνω σε σύνολα δεδομένων με διαφορετικές δομές ομάδων. Οι αλγόριθμοι θα συγκριθούν με τους προγενέστερους αλγορίθμους της κατηγορίας τους με βάση τους δείκτες εκτίμησης της ποιότητας των λύσεών τους που αναπτύχθηκαν στην προηγούμενη ενότητα και με την οπτικοποίηση των αποτελεσμάτων, όπου αυτό είναι εφικτό. Να σημειωθεί πως σε όλους τους αλγορίθμους που κάνουν χρήση του στατιστικού κριτηρίου dip, θεωρούμε ότι ισχύει πάντα $a = 0,\ v_{thd} = 1\%$.

Στους πίνακες των αποτελεσμάτων παρουσιάζονται οι τιμές των δεικτών RI, ARI και VI αν υπάρχουν οι πραγματικές κατηγορίες του κάθε αντικειμένου για το συγκεκριμένο σύνολο δεδομένων, το SSE, ο τελικό αριθμός των ομάδων *K* που περιέχει το μοντέλο της λύσης του αλγορίθμου, ο αρχικός αριθμός των ομάδων (ο οποίος για τους αλγόριθμους συσσωρευτικής φύσης δοκιμάζεται ως 2 και 3 φορές πολλαπλάσιος του πραγματικού, ενώ για τους υπόλοιπους είναι ίσος με 1), καθώς και η τεχνική που χρησιμοποιείται από τους συσσωρευτικούς αλγορίθμους. Η All to all είναι η τεχνική στην οποία εξετάζονται όλα τα σημεία των δύο υποψηφίων για συνένωση ομάδων, ενώ η Cent to all είναι αυτή στην οποία εξετάζονται μόνο οι αποστάσεις των δύο κεντροειδών από τα σημεία των δύο ομάδων. Η Graph είναι η τεχνική όπου χρησιμοποιείται το γράφημα με τις συνεκτικές του συνιστώσες για την ομαδοποίηση και η Cent & Graph αναφέρεται στο συνδυασμό των δύο παραπάνω. Στα κελιά των πινάκων τοποθετείται "-" για να υποδηλώσει την έλλειψη της αντίστοιχης πληροφορίας. Επίσης παρουσιάζονται και οπτικά τα αποτελέσματα των πειραμάτων όπου είναι δυνατό, με τις λύσεις που είναι ίδιες για περισσότερες από ένα αλγορίθμους να παρουσιάζονται μια μόνο φορά.

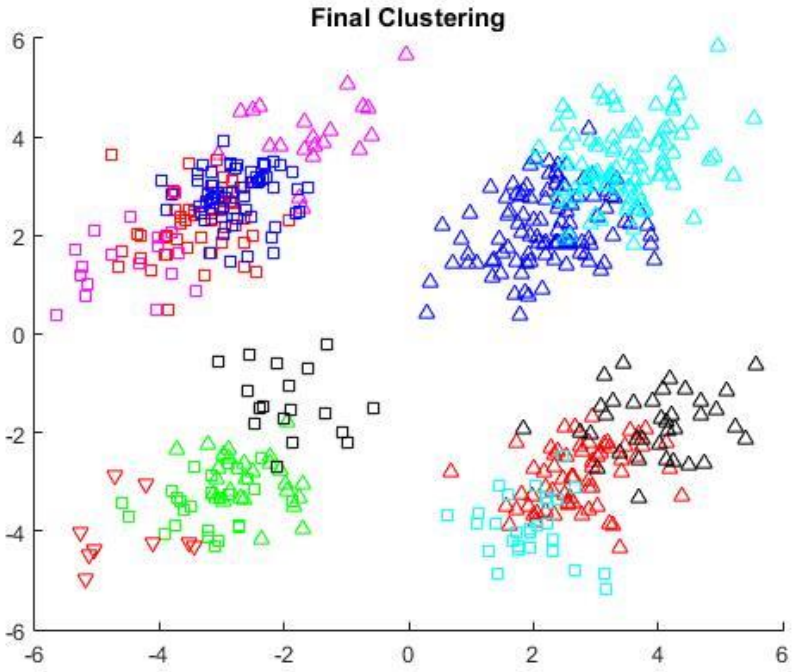
1) Kmeansdata Dataset

Αυτό το σύνολο δεδομένων έχει 4 καλά διαχωρισμένες ομάδες που ακολουθούν Γκαουσιανές κατανομές και θεωρείται συγκριτικά το απλούστερο σύνολο δεδομένων που εξετάστηκε.

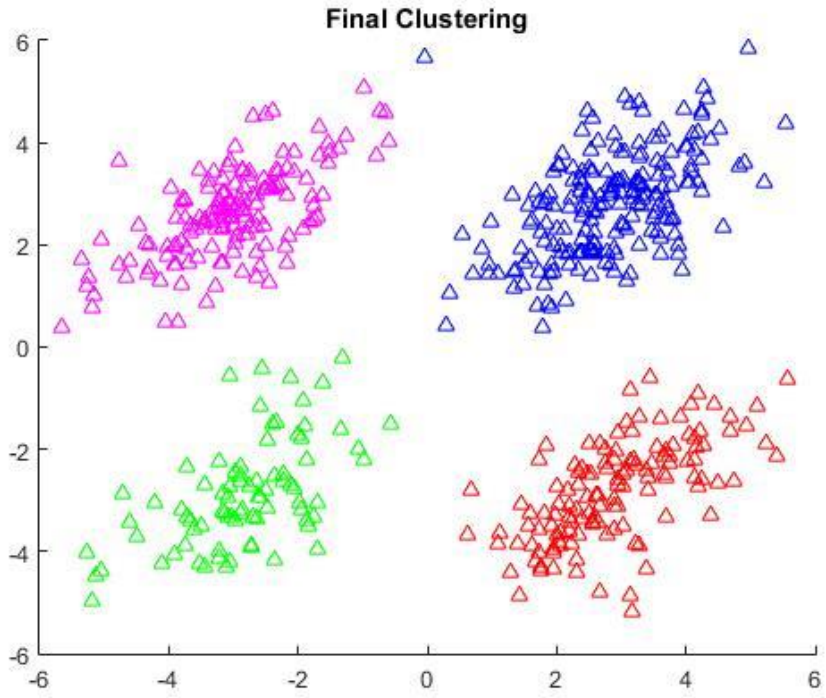
Αλγόριθμος	**Τελικό *K***	**RI**	**ARI**	**VI**	**SSE**	**Αρχικό *K***	**Τεχνική**
X-means	13	-	-	-	698.91	1	-
G-means	4	-	-	-	1000.55	1	-
Dip-means	4	-	-	-	1000.55	1	-
Pdip-means	4	-	-	-	1000.55	1	-
Agglodip	4	-	-	-	1000.66	8	All to all
Agglodip	4	-	-	-	1000.66	8	Cent to all
Agglodip	4	-	-	-	1000.66	8	Graph
Agglodip	4	-	-	-	1000.66	8	Cent & Graph
Agglodip	4	-	-	-	1000.66	12	All to all
Agglodip	4	-	-	-	1000.66	12	Cent to all

Agglodip	4	-	-	-	1000.66	12	Graph
Agglodip	4	-	-	-	1000.66	12	Cent & Graph
Agglopdip	4	-	-	-	1000.55	8	All to all
Agglopdip	4	-	-	-	1000.55	8	Graph
Agglopdip	4	-	-	-	1000.55	12	All to all
Agglopdip	4	-	-	-	1000.55	12	Graph

Πίνακας 5.3.1 Αποτελέσματα kmeansdata, 4 πραγματικές ομάδες

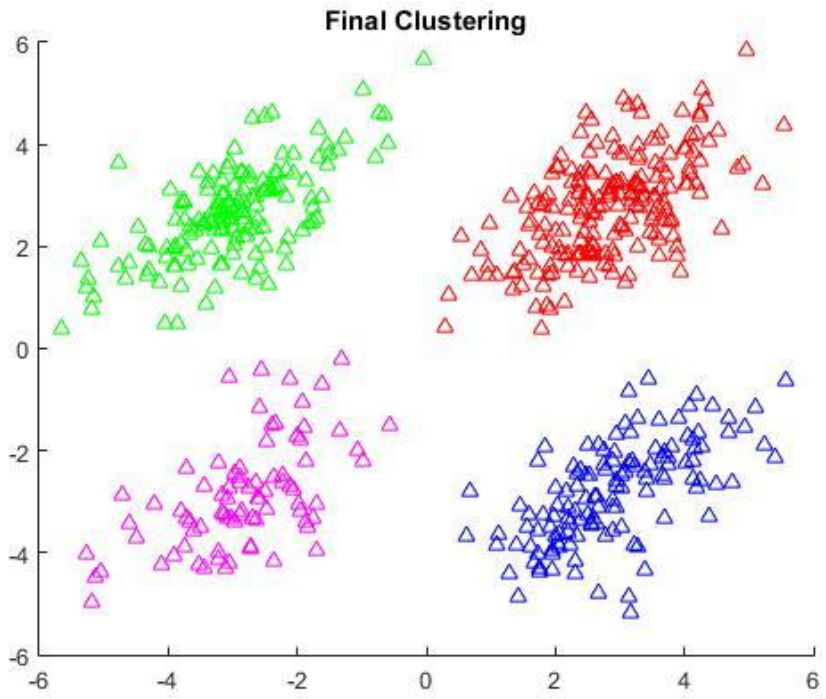


Σχήμα 5.3.1 Kmeansdata: X-Means



Σχήμα 5.3.2 Kmeansdata: G-Means

Agglodip(All to all, Init K 8)



Σχήμα 5.3.3 Kmeansdata: Agglodip (All to all, Αρχικό K = 8)

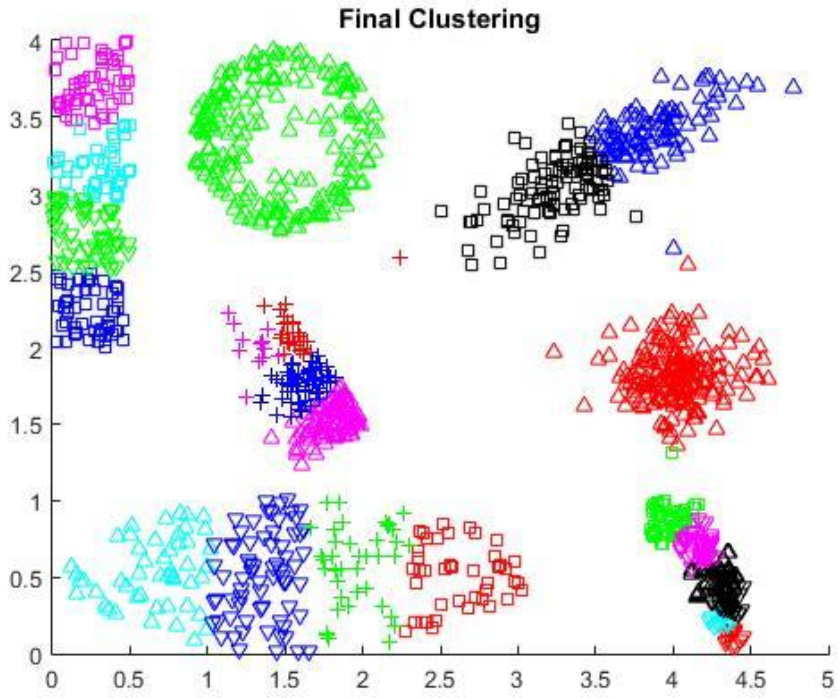
Οι περιπτώσεις των dip-means, pdip-means και agglopdip παραλείφθηκαν καθώς παράγουν την ίδια λύση με τον αλγόριθμο g-means. Επιπλέον παραλείφθηκαν και οι παραλλαγές των agglodip και agglopdip μιας και είναι και αυτές ίδιες με τις παραπάνω όπως φαίνεται και από τον πίνακα 5.3.1.

2) Combosetting Dataset

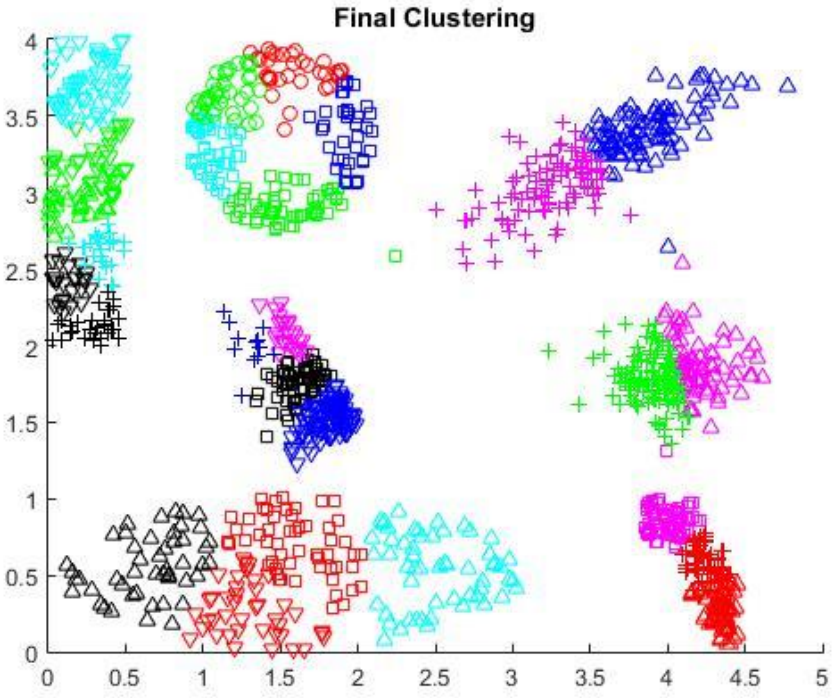
Αυτό το σύνολο δεδομένων έχει 7 ομάδες διαφορετικού σχήματος και μεγέθους και οι οποίες ακολουθούν διαφορετικές unimodal κατανομές.

Αλγόριθμος	**Τελικό *K***	**RI**	**ARI**	**VI**	**SSE**	**Αρχικό *K***	**Τεχνική**
X-means	22	0.9291	0.6341	0.9285	335.492	1	-
G-means	26	0.9046	0.4587	1.2210	267.632	1	-
Dip-means	8	0.9889	0.9535	0.1211	518.765	1	-
Pdip-means	7	0.9948	0.9786	0.0820	567.926	1	-
Agglodip	7	0.9987	0.9950	0.0250	568.840	14	All to all
Agglodip	7	0.9987	0.9950	0.0250	568.840	14	Cent to all
Agglodip	9	0.9815	0.9203	0.2003	496.543	14	Graph
Agglodip	9	0.9815	0.9203	0.2003	496.543	14	Cent & Graph
Agglodip	7	0.9987	0.9950	0.0250	568.840	21	All to all
Agglodip	7	0.9987	0.9950	0.0250	568.840	21	Cent to all
Agglodip	6	0.9579	0.8460	0.2230	685.937	21	Graph
Agglodip	6	0.9579	0.8460	0.2230	685.937	21	Cent & Graph
Agglopdip	7	0.9948	0.9786	0.0820	567.926	14	All to all
Agglopdip	7	0.9948	0.9786	0.0820	567.926	14	Graph
Agglopdip	7	0.9948	0.9786	0.0820	567.926	21	All to all
Agglopdip	6	0.9544	0.8316	0.3184	686.619	21	Graph

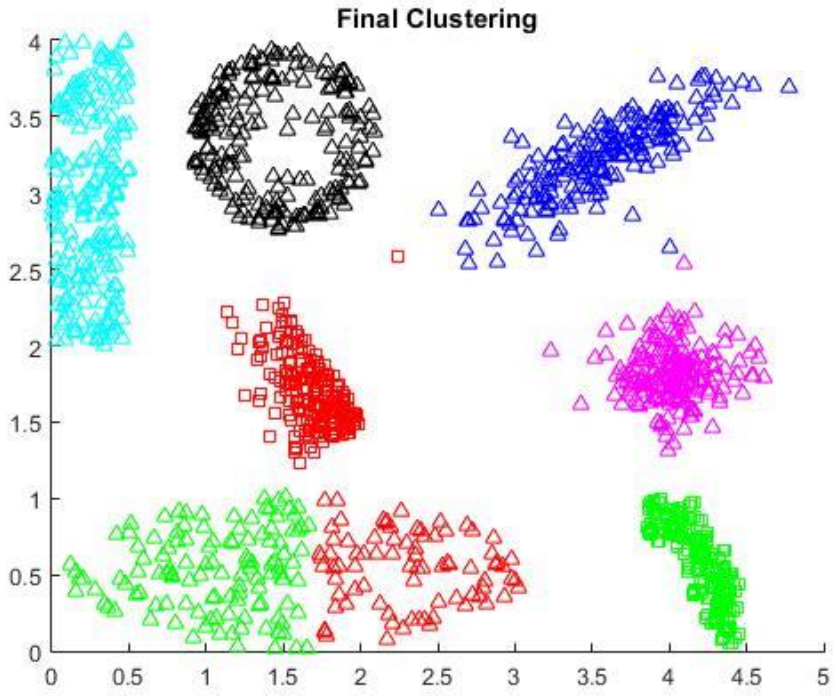
Πίνακας 5.3.2 Αποτελέσματα Combosetting, 7 πραγματικές ομάδες



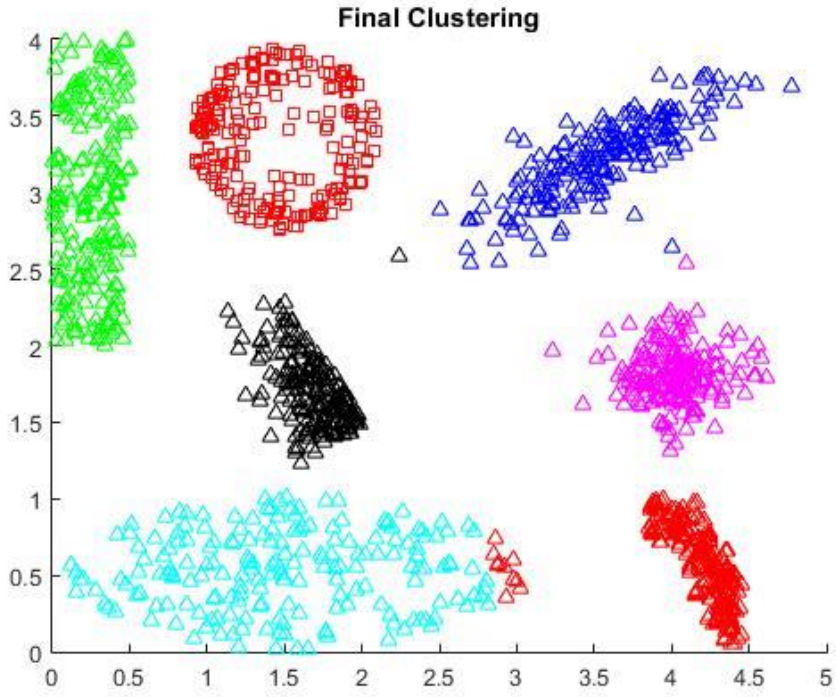
Σχήμα 5.3.4 Combosetting: X-Means



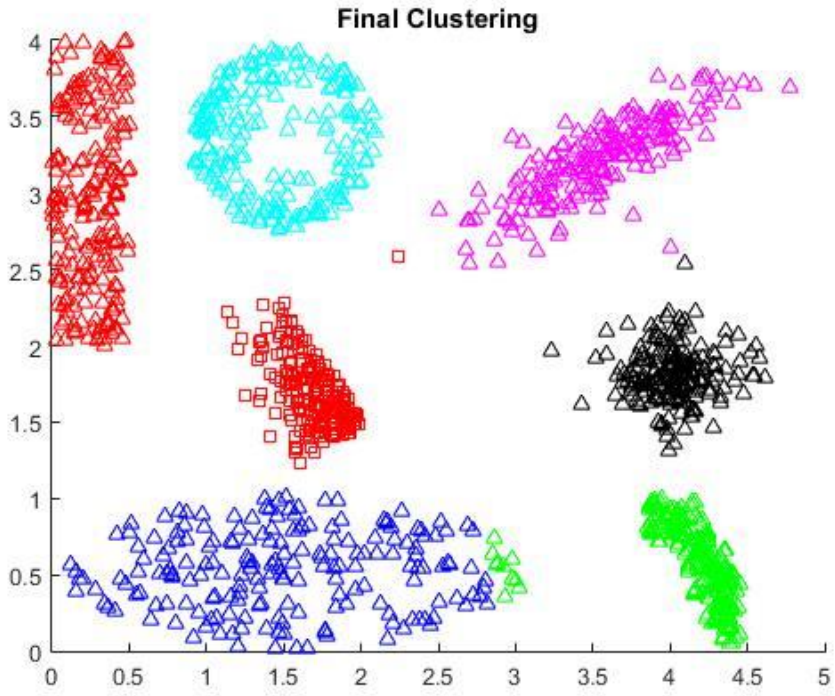
Σχήμα 5.3.5 Combosetting: G-Means



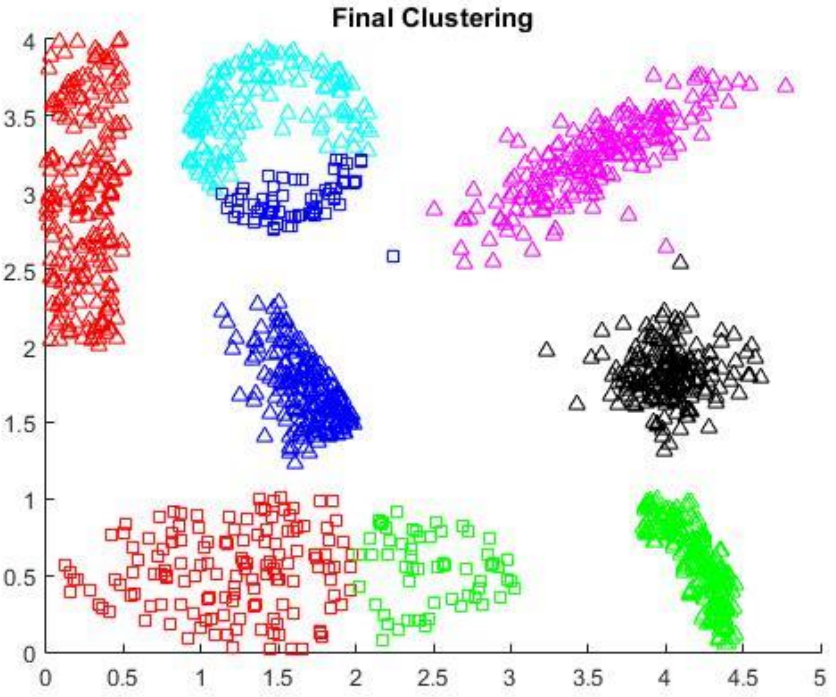
Σχήμα 5.3.6 Combosetting: Dip-Means



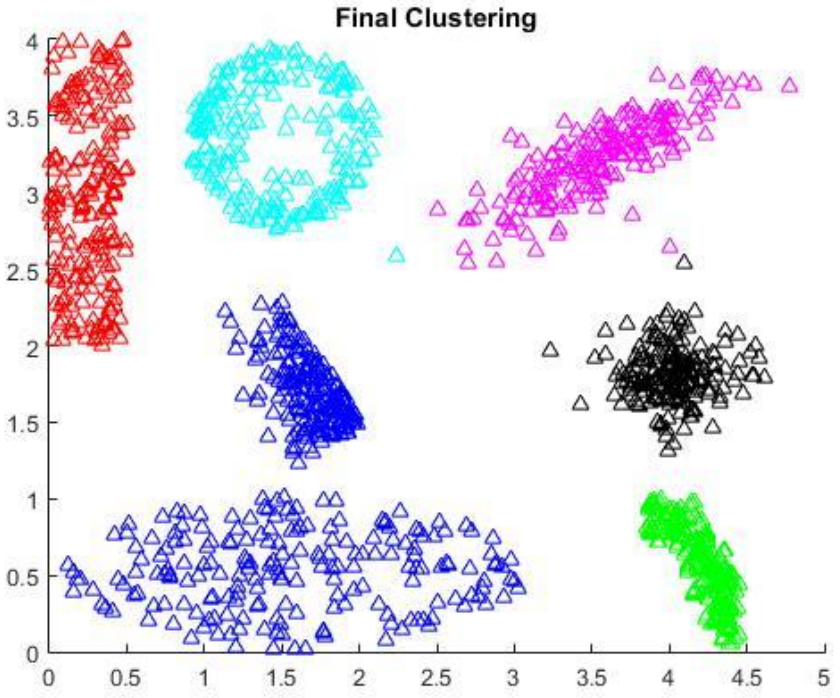
Σχήμα 5.3.7 Combosetting: Pdip-Means



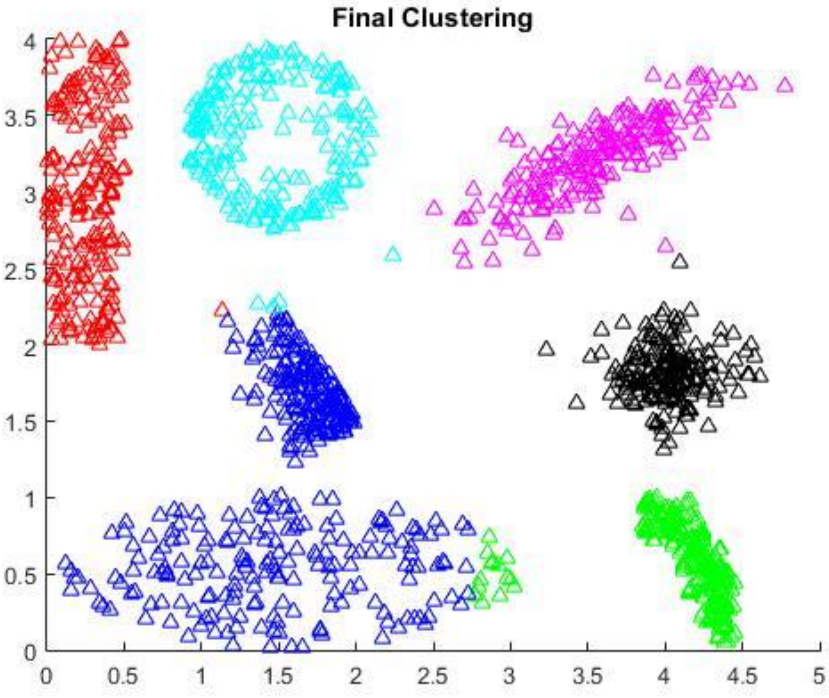
Σχήμα 5.3.8 Combosetting: Agglodip (All to all, Αρχικό K = 14)



Σχήμα 5.3.9 Combosetting: Agglodip (Graph, Αρχικό K = 14)



Σχήμα 5.3.10 Combosetting: Agglodip (Graph, Αρχικό K = 21)



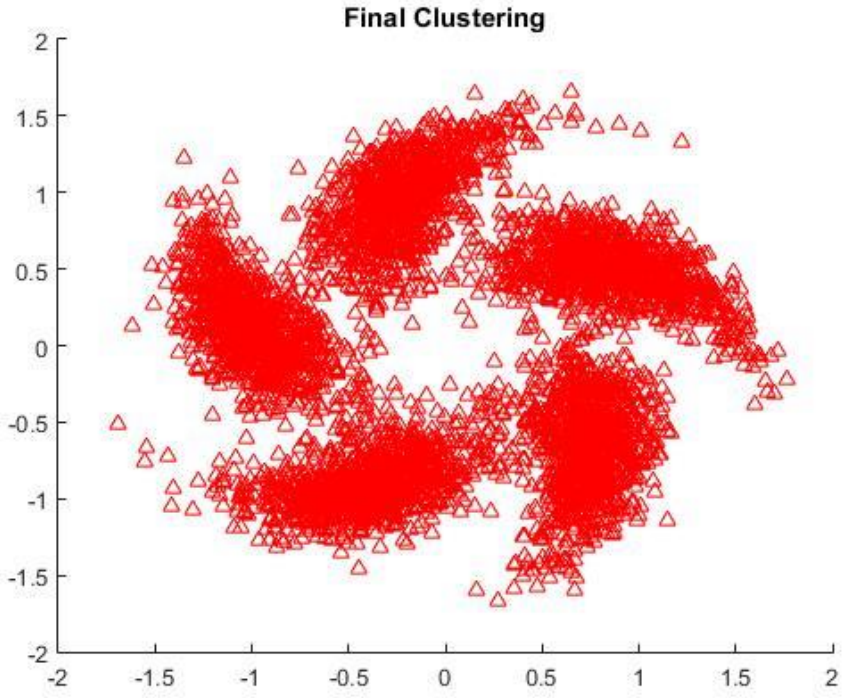
Σχήμα 5.3.11 Combosetting: Agglopdip (Graph, Αρχικό K = 21)

3) Pinwheel Dataset

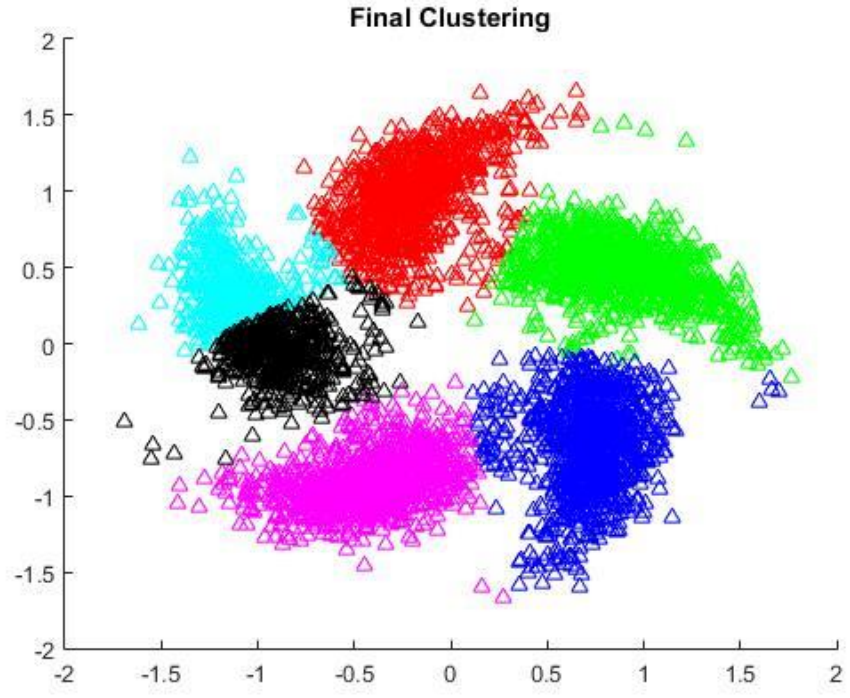
Αυτό το σύνολο δεδομένων έχει 5 ομάδες ελικοειδούς μορφής οι οποίες ξεκινούν από το ίδιο κέντρο.

Αλγόριθμος	**Τελικό *K***	**RI**	**ARI**	**VI**	**SSE**	**Αρχικό *K***	**Τεχνική**
X-means	1	-	-	-	5111.59	1	-
G-means	65	-	-	-	495.193	1	-
Dip-means	6	-	-	-	1448.58	1	-
Pdip-means	6	-	-	-	1448.58	1	-
Agglodip	5	-	-	-	1537.63	10	All to all
Agglodip	5	-	-	-	1537.63	10	Cent to all
Agglodip	5	-	-	-	1537.63	10	Graph
Agglodip	5	-	-	-	1537.63	10	Cent & Graph
Agglodip	5	-	-	-	1546.57	15	All to all
Agglodip	5	-	-	-	1546.57	15	Cent to all
Agglodip	4	-	-	-	2199.12	15	Graph
Agglodip	4	-	-	-	2199.12	15	Cent & Graph
Agglopdip	5	-	-	-	1529.98	10	All to all
Agglopdip	5	-	-	-	1529.98	10	Graph
Agglopdip	5	-	-	-	1529.98	15	All to all
Agglopdip	4	-	-	-	2168.58	15	Graph

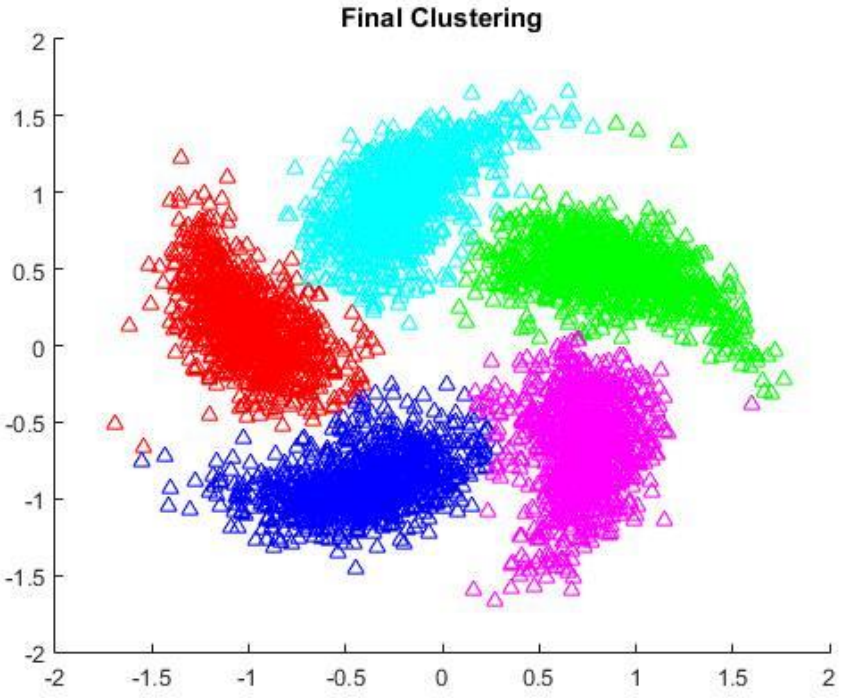
Πίνακας 5.3.3 Αποτελέσματα Pinwheel, 5 πραγματικές ομάδες



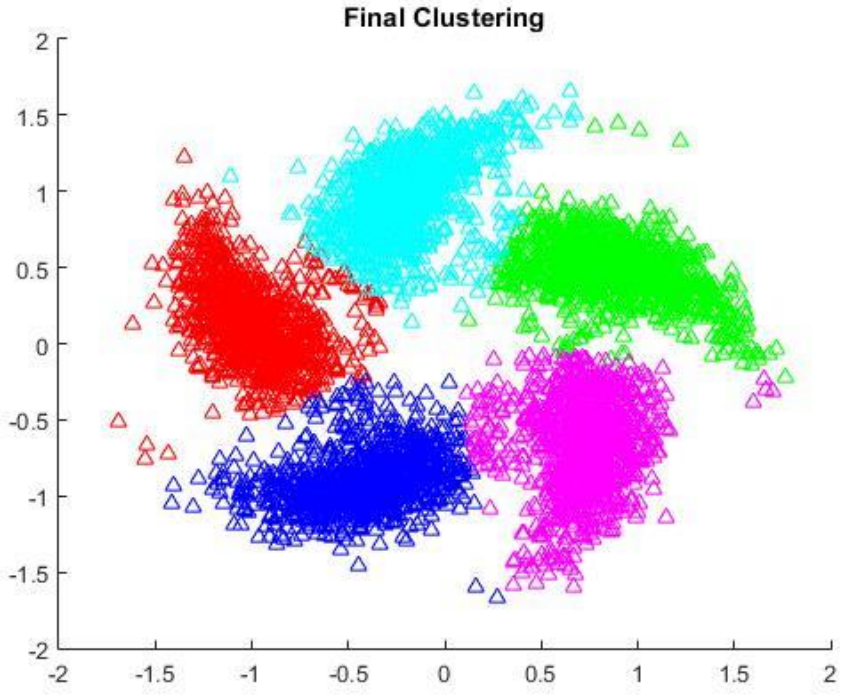
Σχήμα 5.3.12 Pinwheel: X-means



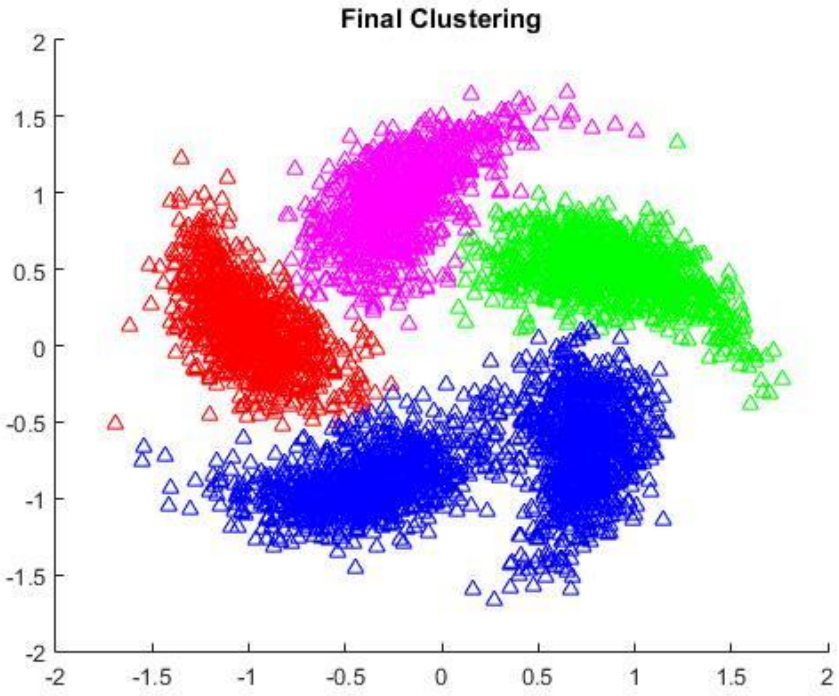
Σχήμα 5.3.13 Pinwheel: Dip-means



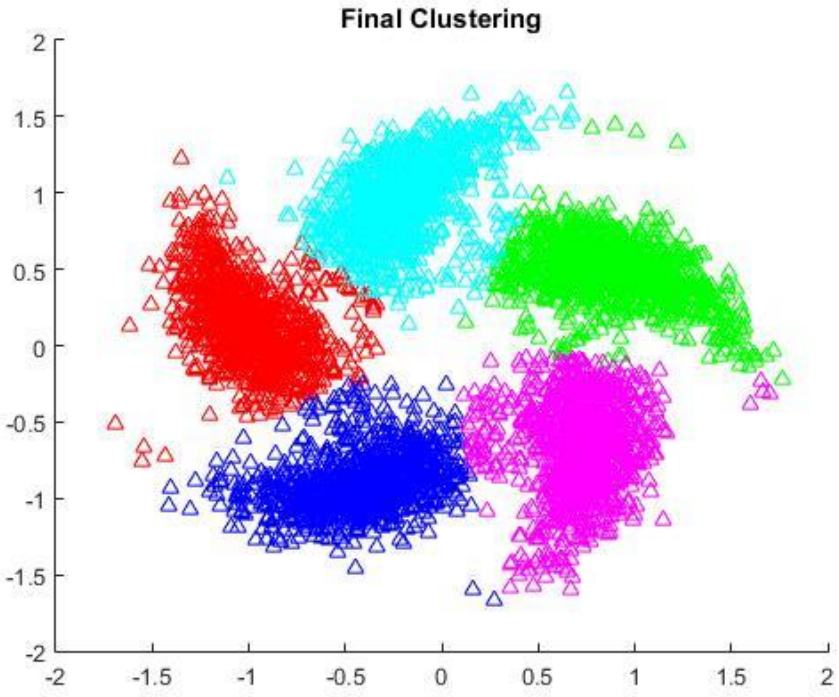
Σχήμα 5.3.14 Pinwheel: Agglodip (All to all, Αρχικό K = 10)



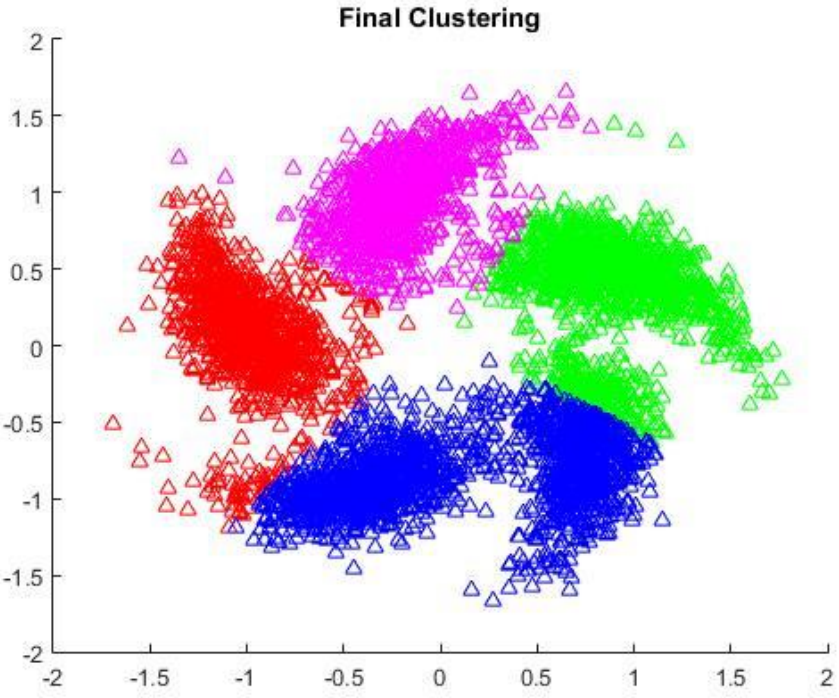
Σχήμα 5.3.15 Pinwheel: Agglodip (All to all, Αρχικό K = 15)



Σχήμα 5.3.16 Pinwheel: Agglodip (Graph, Αρχικό K = 15)



Σχήμα 5.3.17 Pinwheel: Agglopdip (All to all, Αρχικό K = 10)



Σχήμα 5.3.18 Pinwheel: Agglopdip (Graph, Αρχικό K = 15)

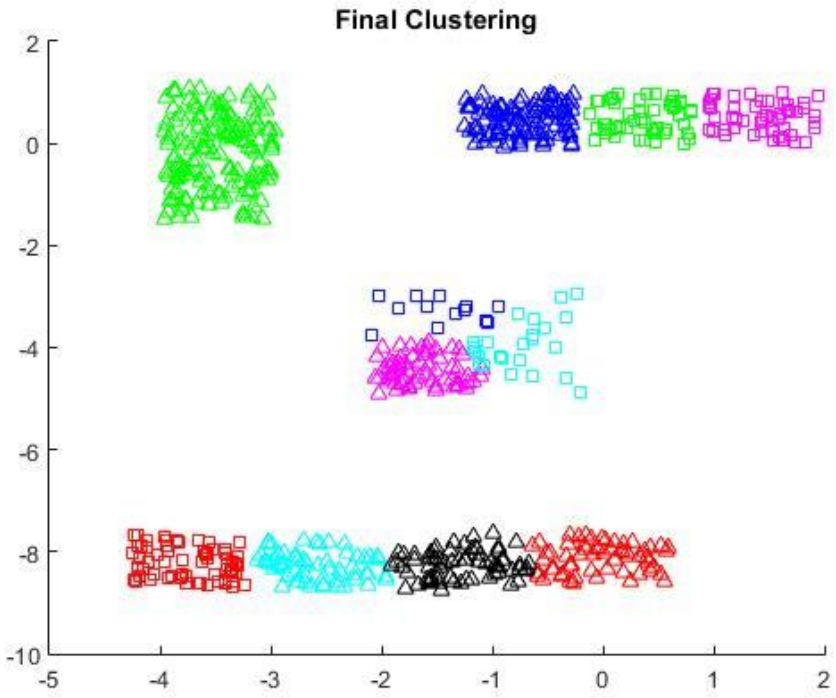
4) Dtest Rectangles Dataset

Το σύνολο δεδομένων Dtest Rectangles Dataset αποτελείται από 4 ομάδες ορθογώνιου σχήματος που ακολουθούν ομοιόμορφη κατανομή.

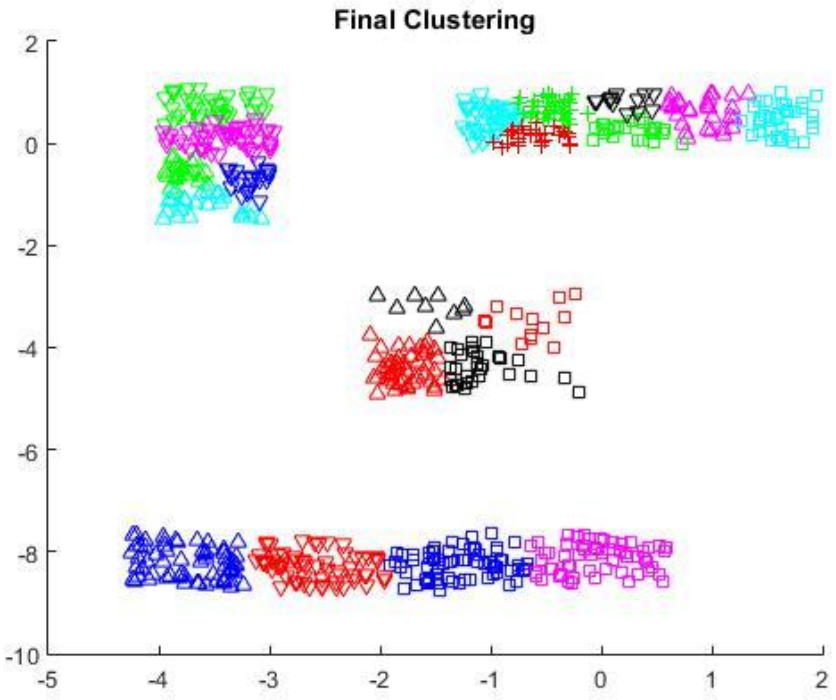
Αλγόριθμος	**Τελικό *K***	**RI**	**ARI**	**VI**	**SSE**	**Αρχικό *K***	**Τεχνική**
X-means	11	-	-	-	326.364	1	-
G-means	20	-	-	-	228.939	1	-
Dip-means	5	-	-	-	588.938	1	-
Pdip-means	4	-	-	-	673.694	1	-
Agglodip	5	-	-	-	588.957	8	All to all
Agglodip	5	-	-	-	588.957	8	Cent to all
Agglodip	5	-	-	-	588.957	8	Graph
Agglodip	5	-	-	-	588.957	8	Cent & Graph
Agglodip	5	-	-	-	588.957	12	All to all
Agglodip	5	-	-	-	588.957	12	Cent to all

Agglodip	2	-	-	-	1429.12	12	Graph
Agglodip	2	-	-	-	1429.12	12	Cent & Graph
Agglopdip	4	-	-	-	673.694	8	All to all
Agglopdip	4	-	-	-	673.694	8	Graph
Agglopdip	4	-	-	-	673.694	12	All to all
Agglopdip	2	-	-	-	1380.24	12	Graph

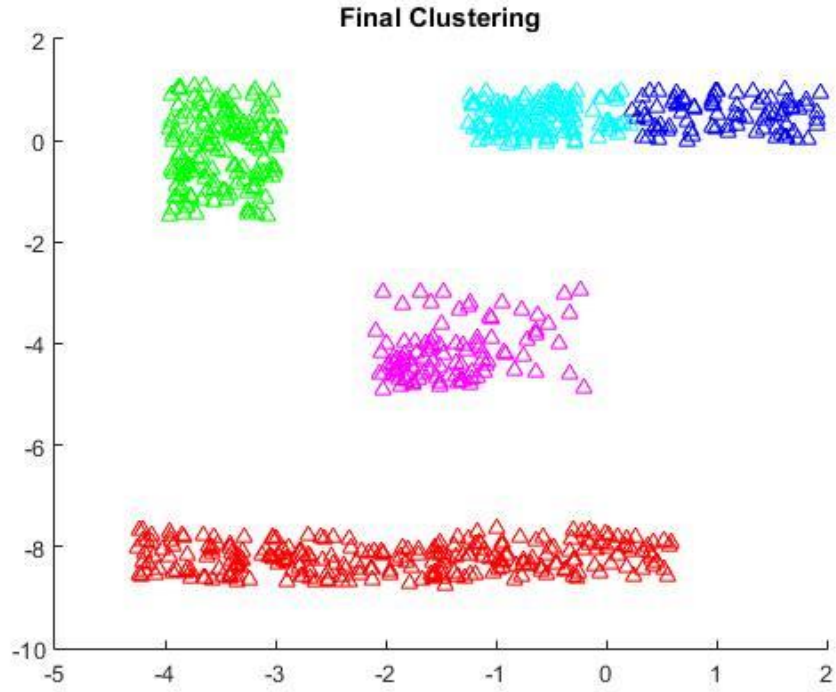
Πίνακας 5.3.4 Αποτελέσματα Dtest Rectangles, 4 πραγματικές ομάδες



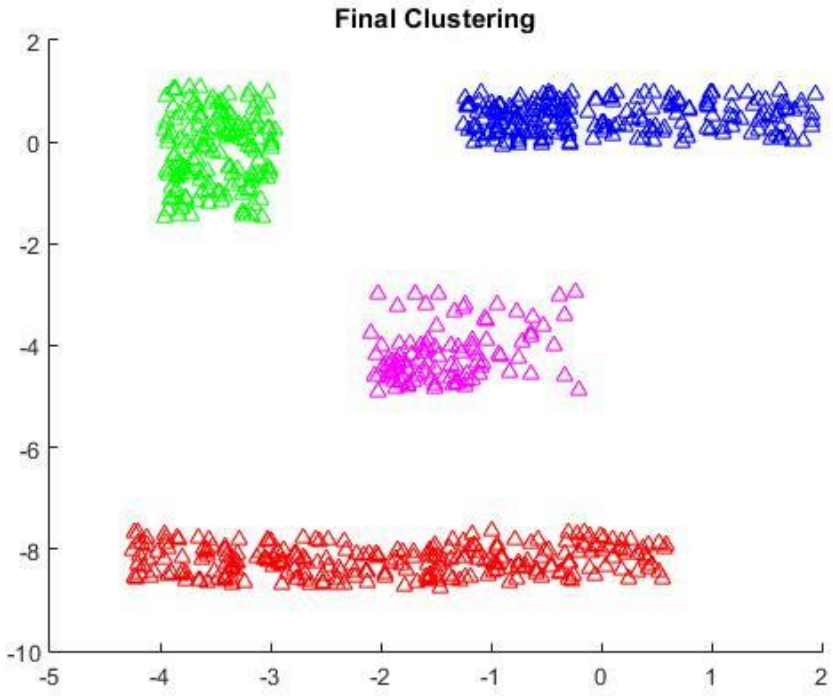
Σχήμα 5.3.19 Dtest Rectangles: X-Means



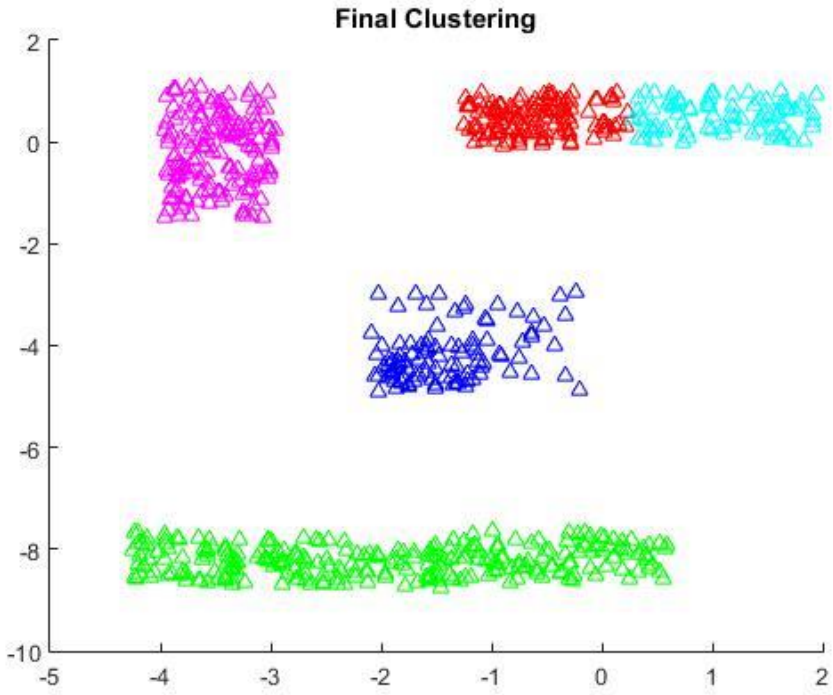
Σχήμα 5.3.20 Dtest Rectangles: G-Means



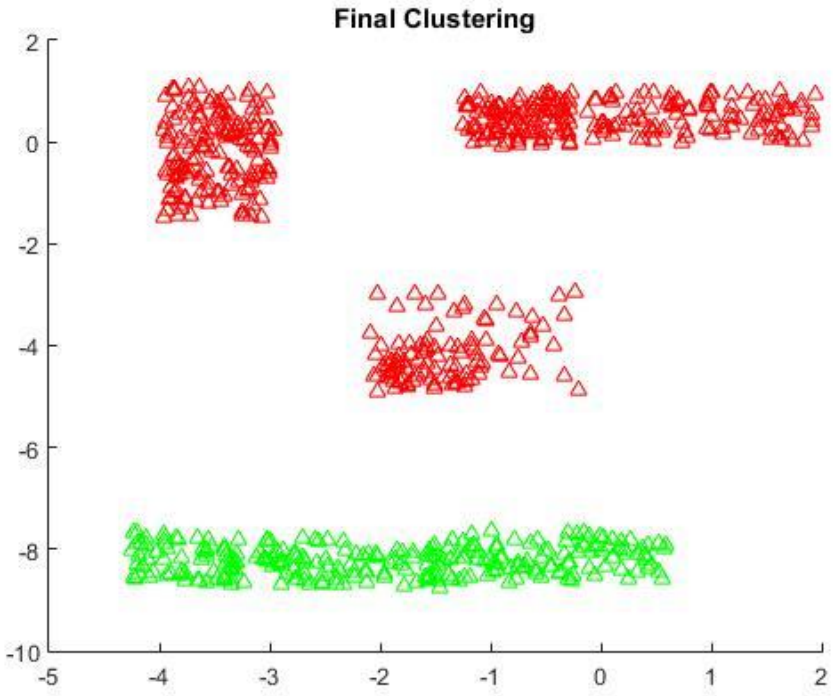
Σχήμα 5.3.21 Dtest Rectangles: Dip-Means



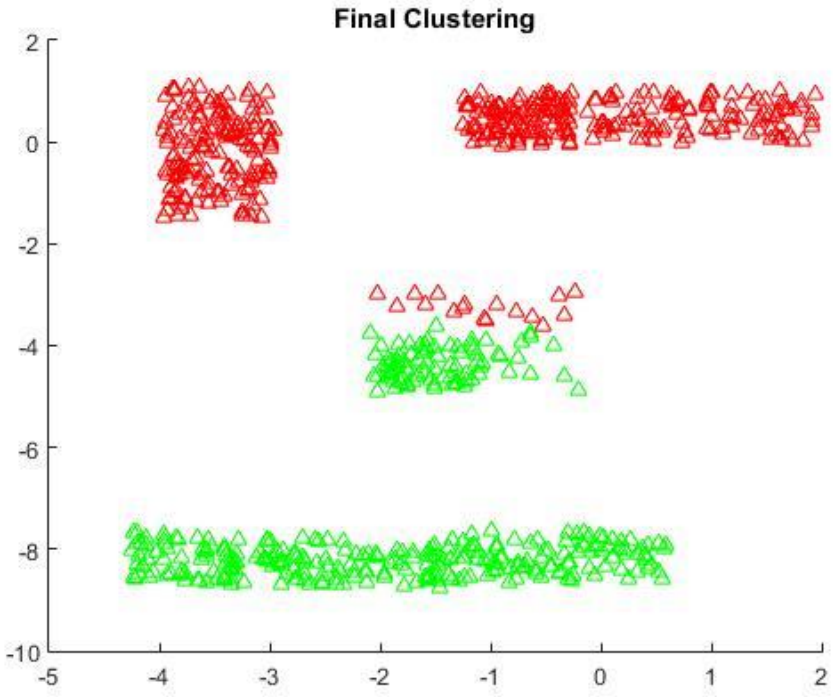
Σχήμα 5.3.22 Dtest Rectangles: Pdip-Means



Σχήμα 5.3.23 Dtest Rectangles: Agglodip (All to all, Αρχικό K = 8)



Σχήμα 5.3.24 Dtest Rectangles: Agglodip (Graph, Αρχικό K = 12)



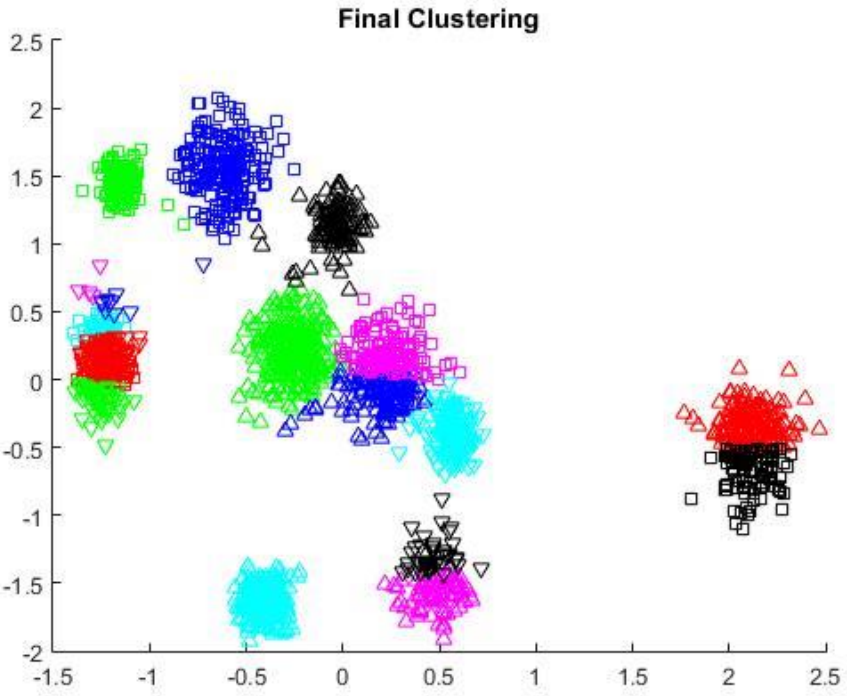
Σχήμα 5.3.25 Dtest Rectangles: Agglodip (Graph, Αρχικό K = 12)

5) X10 Dataset

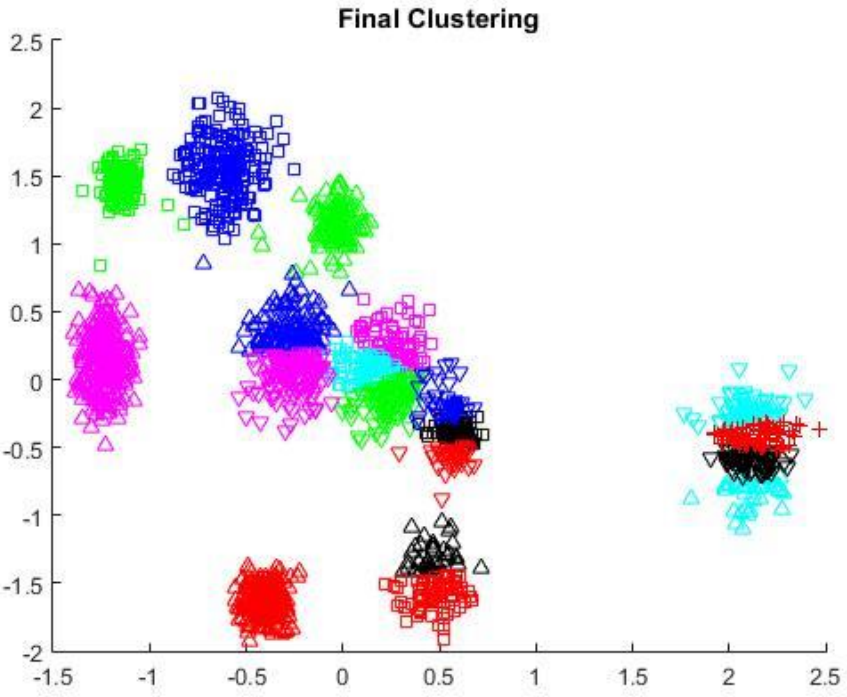
Το σύνολο δεδομένων αυτό αποτελείται από 10 ομάδες μερικές από τις οποίες βρίσκονται σε μικρή απόσταση μεταξύ τους κάτι που δυσκολεύει την ομαδοποίησή του.

Αλγόριθμος	**Τελικό *K***	**RI**	**ARI**	**VI**	**SSE**	**Αρχικό *K***	**Τεχνική**
X-means	18	-	-	-	280.007	1	-
G-means	19	-	-	-	269.188	1	-
Dip-means	10	-	-	-	343.311	1	-
Pdip-means	9	-	-	-	394.155	1	-
Agglodip	10	-	-	-	345.766	20	All to all
Agglodip	9	-	-	-	403.858	20	Cent to all
Agglodip	8	-	-	-	500.343	20	Graph
Agglodip	8	-	-	-	500.343	20	Cent & Graph
Agglodip	10	-	-	-	344.699	30	All to all
Agglodip	9	-	-	-	403.709	30	Cent to all
Agglodip	9	-	-	-	403.709	30	Graph
Agglodip	9	-	-	-	403.709	30	Cent & Graph
Agglopdip	10	-	-	-	343.292	20	All to all
Agglopdip	7	-	-	-	514.638	20	Graph
Agglopdip	10	-	-	-	343.292	30	All to all
Agglopdip	8	-	-	-	445.686	30	Graph

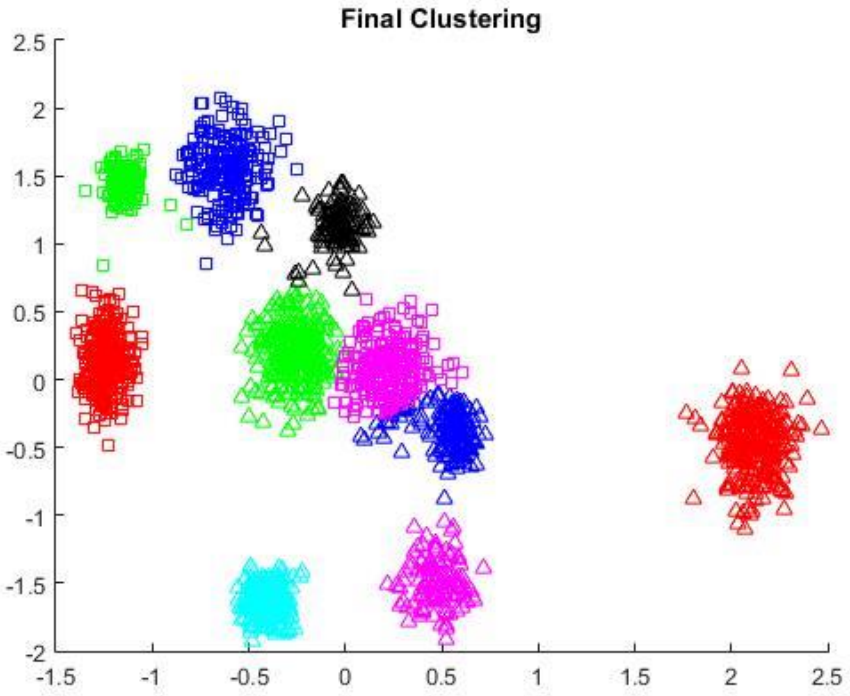
Πίνακας 5.3.5 Αποτελέσματα X10, 10 πραγματικές ομάδες



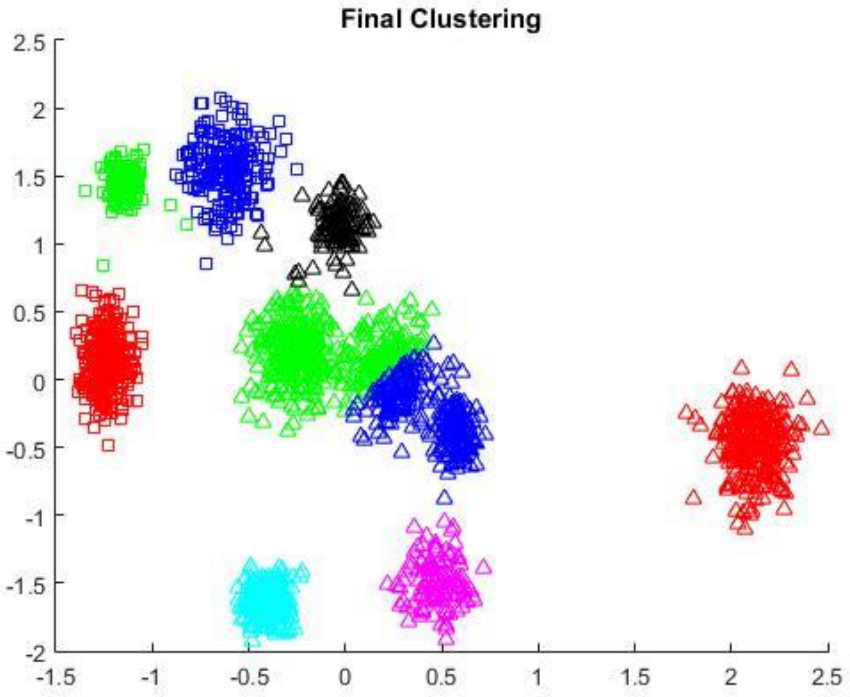
Σχήμα 5.3.26 X10: X-Means



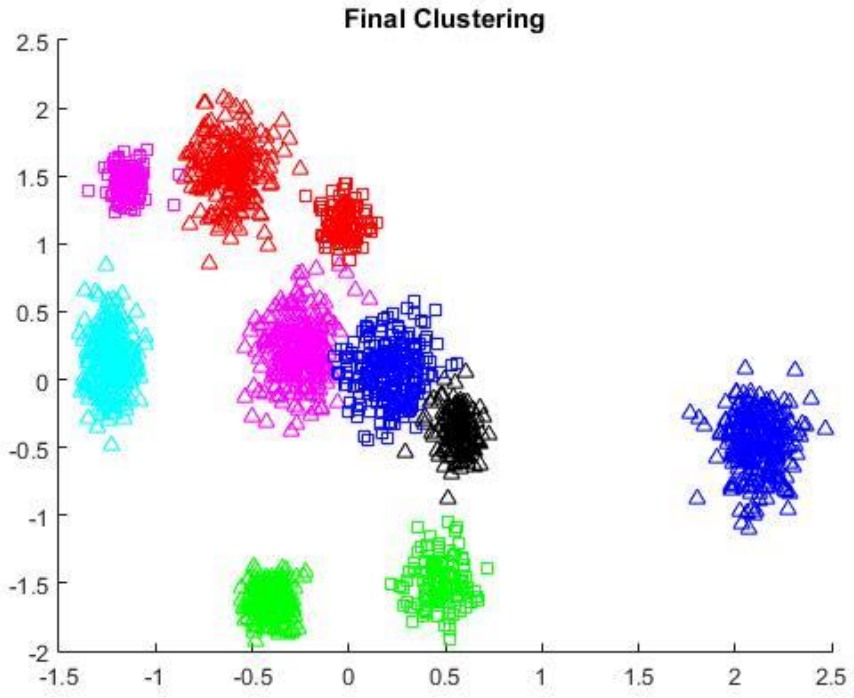
Σχήμα 5.3.27 X10: G-Means



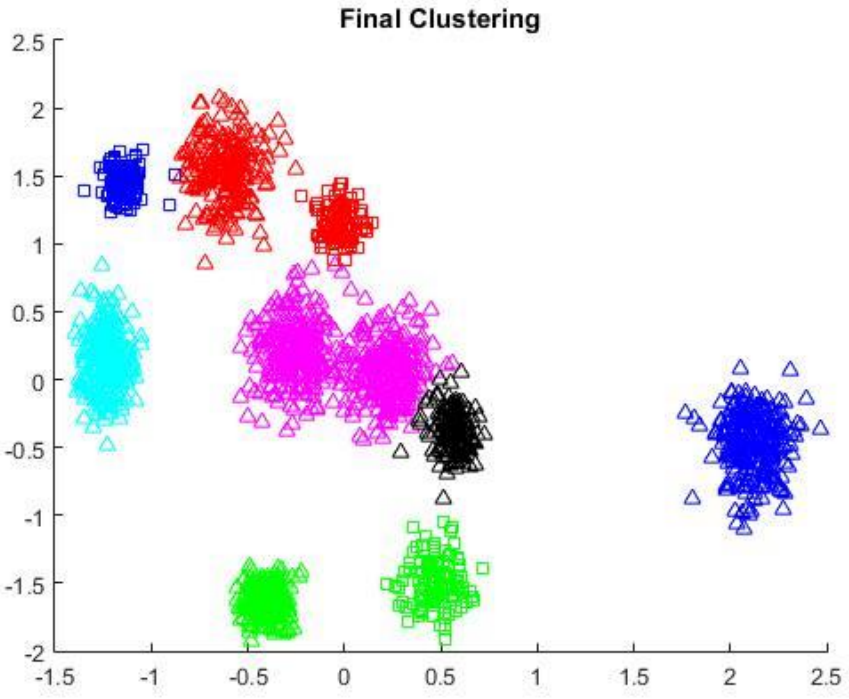
Σχήμα 5.3.28 X10: Dip-Means



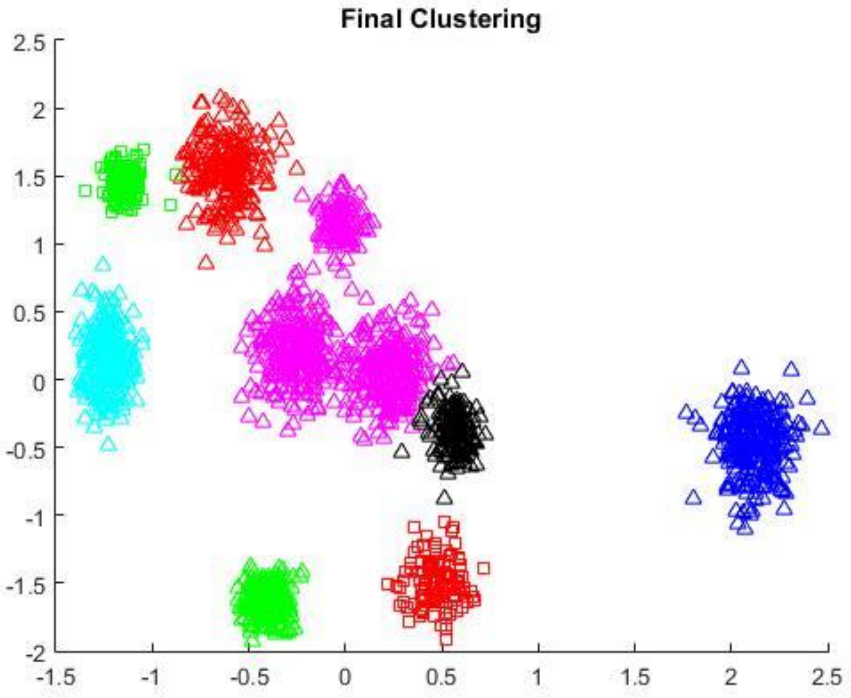
Σχήμα 5.3.29 X10: Pdip-Means



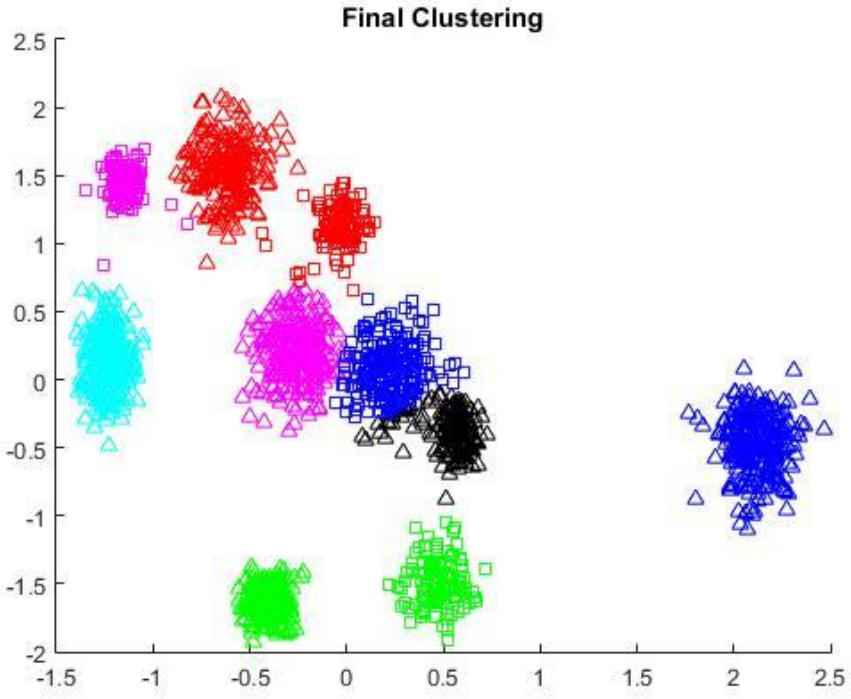
Σχήμα 5.3.30 X10: Agglodip (All to all, Αρχικό K = 20)



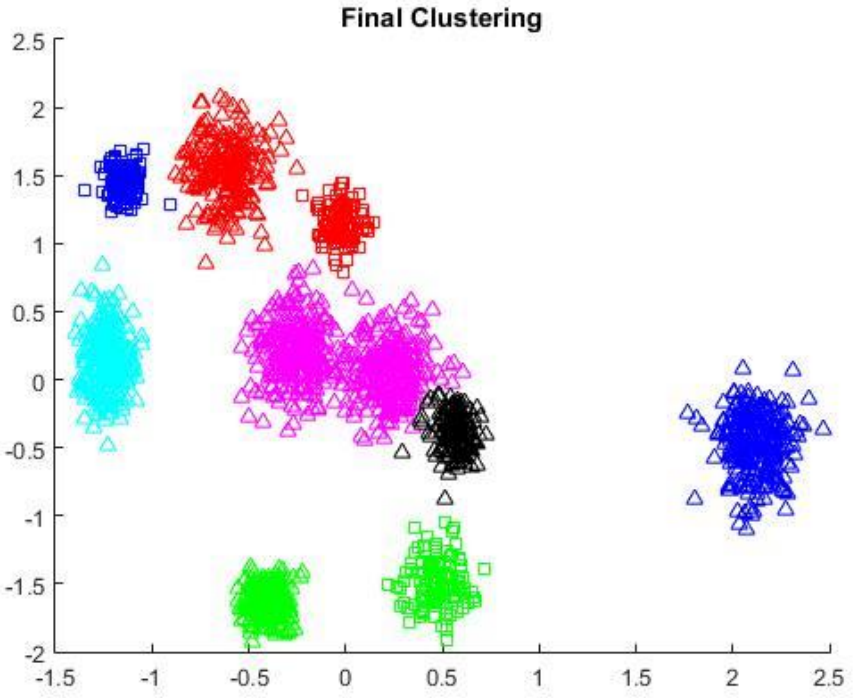
Σχήμα 5.3.31 X10: Agglodip (Cent to all, Αρχικό K = 20)



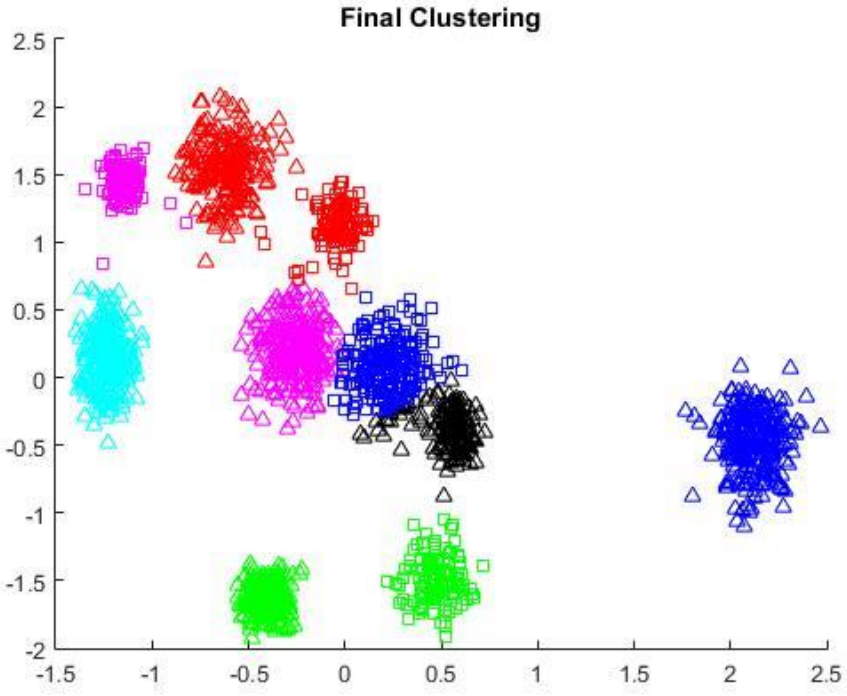
Σχήμα 5.3.32 X10: Agglodip (Graph, Αρχικό K = 20)



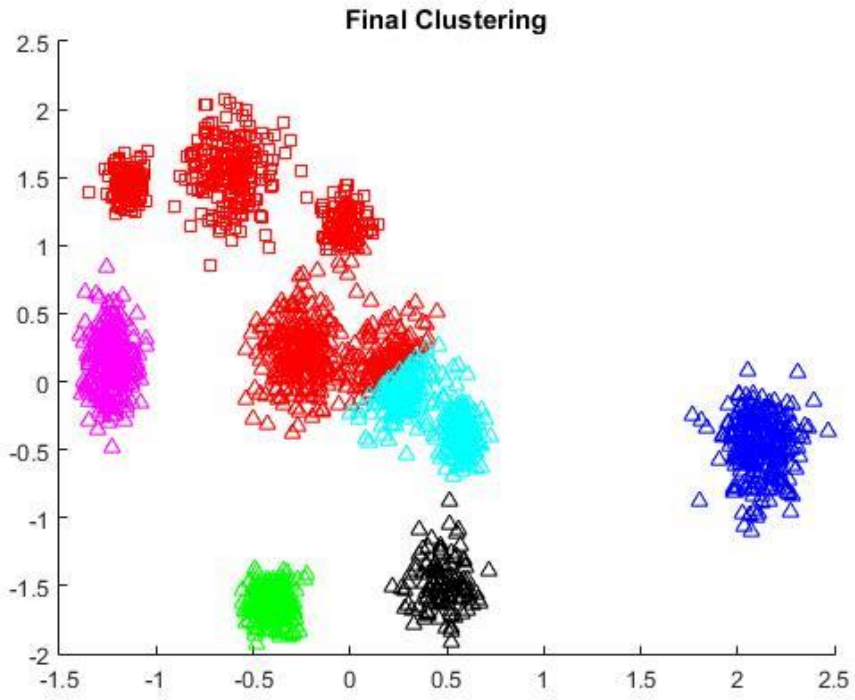
Σχήμα 5.3.33 X10: Agglodip (All to all, Αρχικό K = 30)



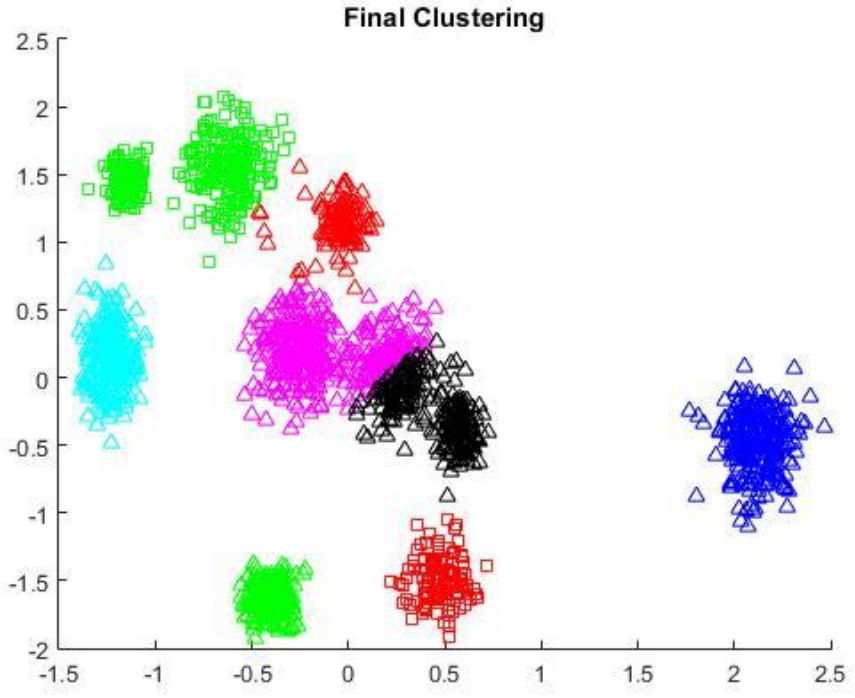
Σχήμα 5.3.34 X10: Agglodip (Cent to all, Αρχικό K = 30)



Σχήμα 5.3.35 X10: Agglopdip (All to all, Αρχικό K = 20)



Σχήμα 5.3.36 X10: Agglopdip (Graph, Αρχικό K = 20)



Σχήμα 5.3.37 X10: Agglopdip (Graph, Αρχικό K = 30)

6) Multidimensional Gaussians

Αυτό το σύνολο δεδομένων αποτελείται από 20 ομάδες που ακολουθούν Γκαουσιανή κατανομή και είναι c-διαχωρισμένες σε d διαστάσεις. Το c θέτουμε να είναι ίσο με την τιμή 1.5 στους πρώτους 4 πίνακες (5.3.6-5.3.9) και στους υπόλοιπους 4 ίσο με την τετραγωνική ρίζα της διάστασης d ενώ οι διαστάσεις που δοκιμάζονται είναι d = 2,4,16,32. Το σύνολο δεδομένων αποτελείται από n = 6000 σημεία με κάθε ομάδα να έχει διαφορετική εκκεντρότητα κάθε φορά και παράγονται έτσι με τις παραπάνω ιδιότητες 30 διαφορετικά σύνολα δεδομένων και υπολογίζεται η μέση τιμή των δεικτών. Στους παρακάτω πίνακες είναι συγκεντρωμένα τα αποτελέσματα από το σύνολο των 30 δοκιμών.

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	8.8	0.1692	2.4883	1	-
G-means	27.4	0.9309	0.2122	1	-
Dip-means	19.7	0.9755	0.0469	1	-
Pdip-means	19.7	0.9755	0.0469	1	-
Agglodip	19.9	0.9878	0.0351	40	All to all
Agglodip	19.7	0.9836	0.0296	40	Cent to all
Agglodip	19.6	0.9770	0.0414	40	Graph
Agglodip	19.6	0.9753	0.0497	40	Cent & Graph
Agglodip	19.8	0.9841	0.0388	60	All to all
Agglodip	19.6	0.9749	0.0469	60	Cent to all
Agglodip	17.1	0.7840	0.2819	60	Graph
Agglodip	12.3	0.5449	1.1109	60	Cent & Graph
Agglopdip	19.7	0.9795	0.0450	40	All to all
Agglopdip	17.9	0.8840	0.1849	40	Graph
Agglopdip	19.4	0.9593	0.0613	60	All to all
Agglopdip	15.6	0.7766	0.3971	60	Graph

Πίνακας 5.3.6 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 2 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	40.2	0.4451	1.6425	1	-
G-means	20.4	0.9969	0.0096	1	-
Dip-means	20	1	0	1	-
Pdip-means	20	1	0	1	-
Agglodip	20	1	0	40	All to all
Agglodip	20	1	0	40	Cent to all
Agglodip	20	1	0	40	Graph
Agglodip	20	0.9999	0.0004	40	Cent & Graph
Agglodip	20	1	0	60	All to all
Agglodip	20	1	0	60	Cent to all
Agglodip	19.6	0.9723	0.0330	60	Graph
Agglodip	16.3	0.7553	0.5166	60	Cent & Graph
Agglopdip	20	1	0	40	All to all
Agglopdip	20	1	0	40	Graph
Agglopdip	20	1	0	60	All to all
Agglopdip	19.2	0.9592	0.0624	60	Graph

Πίνακας 5.3.7 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 4 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	62.2	0.6005	0.9629	1	-
G-means	20.4	0.9949	0.0134	1	-
Dip-means	20	1	0	1	-
Pdip-means	20	1	0	1	-
Agglodip	20	1	0	40	All to all
Agglodip	20	1	0	40	Cent to all
Agglodip	19.8	0.9861	0.0165	40	Graph
Agglodip	18.3	0.9024	0.2601	40	Cent & Graph
Agglodip	20	1	0	60	All to all
Agglodip	20	1	0	60	Cent to all
Agglodip	19	0.9197	0.1319	60	Graph

Agglodip	17.6	0.9723	0.0330	60	Cent & Graph
Agglopdip	20	1	0	40	All to all
Agglopdip	20	1	0	40	Graph
Agglopdip	20	1	0	60	All to all
Agglopdip	20	1	0	60	Graph

Πίνακας 5.3.8 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 16 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	37.5	0.7547	0.5454	1	-
G-means	20.1	0.9987	0.0035	1	-
Dip-means	20	1	0	1	-
Pdip-means	20	1	0	1	-
Agglodip	20	1	0	40	All to all
Agglodip	20	1	0	40	Cent to all
Agglodip	20	1	0	40	Graph
Agglodip	20	1	0	40	Cent & Graph
Agglodip	20	1	0	60	All to all
Agglodip	20	1	0	60	Cent to all
Agglodip	20	1	0	60	Graph
Agglodip	18.9	0.9158	0.1491	60	Cent & Graph
Agglopdip	20	1	0	40	All to all
Agglopdip	19.6	0.9642	0.0402	40	Graph
Agglopdip	20	1	0	60	All to all
Agglopdip	19.6	0.9642	0.0402	60	Graph

Πίνακας 5.3.9 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 32 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	1	0.0000	2.9957	1	-
G-means	31.1	0.9204	0.2661	1	-
Dip-means	19.9	0.9729	0.0639	1	-

Pdip-means	19.6	0.9637	0.0694	1	-
Agglodip	19.7	0.9772	0.0600	40	All to all
Agglodip	19.5	0.9695	0.0609	40	Cent to all
Agglodip	19.3	0.9533	0.1032	40	Graph
Agglodip	19.1	0.9415	0.1410	40	Cent & Graph
Agglodip	19.6	0.9713	0.0520	60	All to all
Agglodip	19.5	0.9650	0.0628	60	Cent to all
Agglodip	18	0.8622	0.1910	60	Graph
Agglodip	15	0.7294	0.7152	60	Cent & Graph
Agglopdip	19.5	0.9661	0.0635	40	All to all
Agglopdip	19.2	0.9548	0.0839	40	Graph
Agglopdip	19.4	0.9601	0.0695	60	All to all
Agglopdip	17.3	0.8547	0.2245	60	Graph

Πίνακας 5.3.10 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 2 διαστάσεις, c = $\sqrt{2}$

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	58.83	0.7890	0.6283	1	-
G-means	20.2	0.9974	0.0068	1	-
Dip-means	20	1	0	1	-
Pdip-means	20	1	0	1	-
Agglodip	20	1	0	40	All to all
Agglodip	20	1	0	40	Cent to all
Agglodip	20	1	0	40	Graph
Agglodip	20	1	0	40	Cent & Graph
Agglodip	20	1	0	60	All to all
Agglodip	20	1	0	60	Cent to all
Agglodip	19.4	0.9622	0.0472	60	Graph
Agglodip	18.3	0.8717	0.2139	60	Cent & Graph
Agglopdip	20	1	0	40	All to all
Agglopdip	20	1	0	40	Graph
Agglopdip	20	1	0	60	All to all

Agglopdip	19.8	0.9904	0.0139	60	Graph

Πίνακας 5.3.11 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 4 διαστάσεις, c = 2

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	62.6	0.5887	0.9892	1	-
G-means	20.3	0.9961	0.0101	1	-
Dip-means	20	1	0	1	-
Pdip-means	20	1	0	1	-
Agglodip	20	1	0	40	All to all
Agglodip	20	1	0	40	Cent to all
Agglodip	20	1	0	40	Graph
Agglodip	20	1	0	40	Cent & Graph
Agglodip	20	1	0	60	All to all
Agglodip	20	1	0	60	Cent to all
Agglodip	19.7	0.9754	0.0277	60	Graph
Agglodip	17.7	0.8273	0.2989	60	Cent & Graph
Agglopdip	20	1	0	40	All to all
Agglopdip	20	1	0	40	Graph
Agglopdip	20	1	0	60	All to all
Agglopdip	19.7	0.9970	0.0277	60	Graph

Πίνακας 5.3.12 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 16 διαστάσεις, c = 4

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	36.8	0.7674	0.5226	1	-
G-means	20.2	0.9974	0.0068	1	-
Dip-means	20	1	0	1	-
Pdip-means	20	1	0	1	-
Agglodip	20	1	0	40	All to all
Agglodip	20	1	0	40	Cent to all
Agglodip	20	1	0	40	Graph

Agglodip	20	1	0	40	Cent & Graph
Agglodip	20	1	0	60	All to all
Agglodip	20	1	0	60	Cent to all
Agglodip	20	1	0	60	Graph
Agglodip	20	1	0	60	Cent & Graph
Agglopdip	20	1	0	40	All to all
Agglopdip	20	1	0	40	Graph
Agglopdip	20	1	0	60	All to all
Agglopdip	20	1	0	60	Graph

Πίνακας 5.3.13 Αποτελέσματα Gaussian-Dimensionality, 20 πραγματικές ομάδες, 2 διαστάσεις, c = $\sqrt{32}$

7) Multidimensional Mixed Distributions

Αυτό το σύνολο δεδομένων (όπως και το προηγούμενο) αποτελείται από 20 ομάδες οι οποίες όμως ακολουθούν Γκαουσιανες (40%), Student-t (20%), καθώς και ομοιόμορφες κατανομές με σχήμα παραλληλόγραμμου (20%) και οβάλ (20%) και είναι c-διαχωρισμένες στις d διαστάσεις. Το c θέτουμε και πάλι να είναι ίσο με 1.5 αρχικά και ίσο με $\sqrt{d}$ στους πίνακες 5.3.18-5.3.21, με τις διαστάσεις που δοκιμάζονται να είναι d = 2,4,16,32. Το σύνολο δεδομένων αποτελείται από n = 6000 σημεία με κάθε ομάδα να έχει διαφορετική εκκεντρότητα κάθε φορά και παράγονται έτσι με τις παραπάνω ιδιότητες 30 διαφορετικά σύνολα δεδομένων και υπολογίζεται η μέση τιμή των δεικτών. Στους παρακάτω πίνακες είναι συγκεντρωμένα τα αποτελέσματα από το σύνολο των 30 δοκιμών.

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	6.3	0.1376	2.5271	1	-
G-means	101.1	0.6118	1.1507	1	-
Dip-means	19.9	0.9629	0.1272	1	-
Pdip-means	19.9	0.9703	0.1182	1	-
Agglodip	20.2	0.9697	0.1286	40	All to all
Agglodip	19.7	0.9517	0.1717	40	Cent to all
Agglodip	5.3	0.1617	2.2998	40	Graph

Agglodip	2	0.0130	2.8552	40	Cent & Graph
Agglodip	21	0.9630	0.1467	60	All to all
Agglodip	19.6	0.9361	0.1926	60	Cent to all
Agglodip	2	0.0022	2.9581	60	Graph
Agglodip	1.6	0.0011	2.9769	60	Cent & Graph
Agglopdip	19.9	0.9703	0.1187	40	All to all
Agglopdip	3.4	0.1325	2.4676	40	Graph
Agglopdip	19.9	0.9628	0.1292	60	All to all
Agglopdip	1.3	0.0268	2.8003	60	Graph

Πίνακας 5.3.14 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 2 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	10.2	0.1425	2.5152	1	-
G-means	120.9	0.6187	1.1145	1	-
Dip-means	19.9	0.9661	0.1049	1	-
Pdip-means	18.4	0.8853	0.2874	1	-
Agglodip	20	0.9864	0.0661	40	All to all
Agglodip	20.1	0.9824	0.0768	40	Cent to all
Agglodip	2.1	0.0098	2.9810	40	Graph
Agglodip	1.6	0.0016	2.9679	40	Cent & Graph
Agglodip	20	0.9838	0.0786	60	All to all
Agglodip	20	0.9764	0.1090	60	Cent to all
Agglodip	2	0.0038	2.9325	60	Graph
Agglodip	1.5	0.0017	2.9674	60	Cent & Graph
Agglopdip	20.2	0.9845	0.0690	40	All to all
Agglopdip	1.9	0.0693	2.5173	40	Graph
Agglopdip	20.3	0.9804	0.0939	60	All to all
Agglopdip	1.8	0.0653	2.5354	60	Graph

Πίνακας 5.3.15 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 4 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	128	0.5348	1.3006	1	-
G-means	114	0.8298	0.6532	1	-
Dip-means	20	0.9974	0.0159	1	-
Pdip-means	20.1	0.9974	0.0160	1	-
Agglodip	20.3	0.9941	0.0308	40	All to all
Agglodip	20.7	0.9895	0.0430	40	Cent to all
Agglodip	7.3	0.3010	2.0607	40	Graph
Agglodip	2.2	0.0060	2.9005	40	Cent & Graph
Agglodip	20.5	0.9892	0.0450	60	All to all
Agglodip	20.7	0.9875	0.0504	60	Cent to all
Agglodip	4.1	0.1011	2.6800	60	Graph
Agglodip	2.2	0.0035	2.9355	60	Cent & Graph
Agglopdip	20.6	0.9944	0.0242	40	All to all
Agglopdip	5.6	0.2075	2.2736	40	Graph
Agglopdip	20.3	0.9941	0.0269	60	All to all
Agglopdip	2	0.0262	2.7258	60	Graph

Πίνακας 5.3.16 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 16 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	102.8	0.6247	0.9719	1	-
G-means	152.5	0.8538	0.6675	1	-
Dip-means	20	0.9973	0.0166	1	-
Pdip-means	20	0.9973	0.0166	1	-
Agglodip	20	0.9970	0.0176	40	All to all
Agglodip	20	0.9962	0.0216	40	Cent to all
Agglodip	12.8	0.5989	1.1997	40	Graph
Agglodip	10.3	0.4522	1.5586	40	Cent & Graph
Agglodip	20.3	0.9934	0.0352	60	All to all
Agglodip	20.4	0.9900	0.0451	60	Cent to all
Agglodip	7.9	0.3046	2.0189	60	Graph

Agglodip	7.5	0.3046	2.0049	60	Cent & Graph
Agglopdip	20.1	0.9974	0.0164	40	All to all
Agglopdip	10.6	0.5009	1.4745	40	Graph
Agglopdip	20.3	0.9949	0.0166	60	All to all
Agglopdip	4.1	0.1135	2.4657	60	Graph

Πίνακας 5.3.17 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 32 διαστάσεις, c = 1.5

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	1	0	2.9957	1	-
G-means	116.3	0.5308	1.3826	1	-
Dip-means	19.7	0.9272	0.2053	1	-
Pdip-means	19.8	0.9304	0.2022	1	-
Agglodip	19.7	0.9421	0.1915	40	All to all
Agglodip	19.6	0.9361	0.2095	40	Cent to all
Agglodip	7.3	0.2931	2.0177	40	Graph
Agglodip	1.7	0.0014	2.9732	40	Cent & Graph
Agglodip	19.9	0.9418	0.1722	60	All to all
Agglodip	19.6	0.9416	0.1944	60	Cent to all
Agglodip	2.11	0.0012	2.9767	60	Graph
Agglodip	1.2	0.0002	2.9959	60	Cent & Graph
Agglopdip	19.7	0.9447	0.1840	40	All to all
Agglopdip	1.6	0.0522	2.5718	40	Graph
Agglopdip	19.7	0.9432	0.1879	60	All to all
Agglopdip	1.4	0.0315	2.6645	60	Graph

Πίνακας 5.3.18 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 2 διαστάσεις, c = $\sqrt{2}$

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	22.8	0.1995	2.2737	1	-
G-means	112.1	0.6432	1.0518	1	-
Dip-means	20.11	0.9905	0.0575	1	-

Pdip-means	20.10	0.9893	0.0575	1	-
Agglodip	20.2	0.9874	0.0610	40	All to all
Agglodip	20.4	0.9811	0.0825	40	Cent to all
Agglodip	3.2	0.0749	2.7043	40	Graph
Agglodip	1.4	0.0010	2.9966	40	Cent & Graph
Agglodip	20.1	0.9876	0.0575	60	All to all
Agglodip	20.2	0.9804	0.0906	60	Cent to all
Agglodip	2.2	0.0024	2.9553	60	Graph
Agglodip	1.6	0.0009	2.9796	60	Cent & Graph
Agglopdip	20.2	0.9881	0.1840	40	All to all
Agglopdip	1.6	0.0466	2.6630	40	Graph
Agglopdip	20.2	0.9853	0.0711	60	All to all
Agglopdip	1.8	0.0627	2.5800	60	Graph

Πίνακας 5.3.19 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 4 διαστάσεις, c = 2

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	149.7	0.5543	1.3229	1	-
G-means	139	0.8055	0.7708	1	-
Dip-means	20	0.9965	0.0217	1	-
Pdip-means	20	0.9965	0.0217	1	-
Agglodip	20.2	0.9942	0.0317	40	All to all
Agglodip	20.2	0.9935	0.0322	40	Cent to all
Agglodip	1.9	0.0023	2.6492	40	Graph
Agglodip	2.4	0.0546	2.9879	40	Cent & Graph
Agglodip	20.3	0.9910	0.0403	60	All to all
Agglodip	20.33	0.9862	0.0559	60	Cent to all
Agglodip	2.2	0.0077	2.8647	60	Graph
Agglodip	1.6	0.0009	2.9560	60	Cent & Graph
Agglopdip	20.1	0.9964	0.0220	40	All to all
Agglopdip	5.3	0.1869	2.2391	40	Graph
Agglopdip	20.22	0.9946	0.0231	60	All to all

Agglopdip	3	0.0329	2.5772	60	Graph

Πίνακας 5.3.20 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 16 διαστάσεις, c = 4

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	114.3	0.6819	0.8858	1	-
G-means	206.4	0.7940	0.8960	1	-
Dip-means	20	0.9965	0.0210	1	-
Pdip-means	20.1	0.9966	0.0211	1	-
Agglodip	20.2	0.9931	0.0342	40	All to all
Agglodip	20.2	0.9951	0.0249	40	Cent to all
Agglodip	13.3	0.6016	1.1584	40	Graph
Agglodip	10.9	0.4983	1.5086	40	Cent & Graph
Agglodip	20.2	0.9923	0.0381	60	All to all
Agglodip	20.6	0.9890	0.0460	60	Cent to all
Agglodip	2.3	0.0050	2.9091	60	Graph
Agglodip	2	0.0049	2.9114	60	Cent & Graph
Agglopdip	20.2	0.9963	0.0225	40	All to all
Agglopdip	10.7	0.5004	1.4699	40	Graph
Agglopdip	18.7	0.8971	0.2903	60	All to all
Agglopdip	2.5	0.0176	2.5061	60	Graph

Πίνακας 5.3.21 Αποτελέσματα Mixed Distributions-Dimensionality, 20 πραγματικές ομάδες, 32 διαστάσεις, c = $\sqrt{32}$

8) Handwritten Pendigits (UCI)

Το παρακάτω σύνολο δεδομένων περιέχει δεδομένα χειρόγραφων χαρακτήρων ενός πραγματικού προβλήματος [35]. Πιο συγκεκριμένα, καταγράφτηκε ο γραφικός χαρακτήρας 14 συγγραφέων για τη γραφή των ψηφίων 0-9 με τη χρήση ψηφιακής γραφίδας σε οθόνη ηλεκτρονικού υπολογιστή. Τα δεδομένα αυτά αποτελούνται από μετρήσεις που γινόταν κάθε 100 millisecond για τη θέση (x,y) της γραφίδας σε ένα τετράγωνο πλευράς 500 pixels στην οθόνη και τα οποία είναι 16 διαστάσεων. Το πλήθος των δεδομένων αποτελείται συνολικά

από 3498 εγγραφές με την κάθε κατηγορία να περιέχει περίπου 350 εγγραφές κατά μέσο όρο. Από το σύνολο δεδομένων αυτό θα μελετηθούν 3 υποσύνολά του, τα PD_1, PD_2 και PD_3 τα οποία αφορούν τα ψηφία {0,4,7}, {1,3,5,7,9} και {0,2,4,6,8} αντίστοιχα. Παρακάτω παρατίθενται οι πίνακες (Πίνακας 5.3.22-5.3.24) που περιέχουν όλες τις μεθόδους που μελετήθηκαν και στα προηγούμενα πειράματα για τα υποσύνολα αυτά.

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	200	0.0230	4.0448	1	-
G-means	29	0.2032	2.0089	1	-
Dip-means	5	0.5468	0.9735	1	-
Pdip-means	5	0.5468	0.9735	1	-
Agglodip	3	0.7201	0.6590	6	All to all
Agglodip	3	0.7201	0.6590	6	Cent to all
Agglodip	1	0	1.0986	6	Graph
Agglodip	1	0	1.0986	6	Cent & Graph
Agglodip	3	0.8259	0.4041	9	All to all
Agglodip	3	0.8013	0.4527	9	Cent to all
Agglodip	1	0	1.0986	9	Graph
Agglodip	1	0	1.0986	9	Cent & Graph
Agglopdip	8	0.4997	0.9672	6	All to all
Agglopdip	7	0.5029	1.0406	6	Graph
Agglopdip	8	0.4587	1.0858	9	All to all
Agglopdip	6	0.5817	0.9443	9	Graph

Πίνακας 5.3.22 Αποτελέσματα Handwrittern Pendigits PD_1**,** 3 πραγματικές ομάδες, 16 διαστάσεις

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	297	0.0323	3.9467	1	-
G-means	40	0.2487	2.0555	1	-
Dip-means	6	0.4854	1.4077	1	-
Pdip-means	7	0.2569	1.6608	1	-
Agglodip	6	0.4334	1.4136	10	All to all

Agglodip	4	0.2096	1.4010	10	Cent to all
Agglodip	1	0	1.6086	10	Graph
Agglodip	1	0	1.6086	10	Cent & Graph
Agglodip	6	0.4589	1.4581	15	All to all
Agglodip	3	0.1895	1.5479	15	Cent to all
Agglodip	3	0.0019	1.7176	15	Graph
Agglodip	3	0.0019	1.7176	15	Cent & Graph
Agglopdip	13	0.5246	1.3618	10	All to all
Agglopdip	4	0.3310	1.7796	10	Graph
Agglopdip	14	0.3644	1.8366	15	All to all
Agglopdip	4	0.3329	1.7722	15	Graph

Πίνακας 5.3.23 Αποτελέσματα Handwrittern Pendigits PD_2, 5 πραγματικές ομάδες, 16 διαστάσεις

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	329	0.0274	4.0379	1	-
G-means	38	0.2524	1.9069	1	-
Dip-means	7	0.6495	0.8729	1	-
Pdip-means	5	0.5773	0.9860	1	-
Agglodip	6	0.8907	0.3365	10	All to all
Agglodip	5	0.8022	0.4879	10	Cent to all
Agglodip	1	0	1.6087	10	Graph
Agglodip	1	0	1.6087	10	Cent & Graph
Agglodip	6	0.9192	0.2367	15	All to all
Agglodip	5	0.7753	0.5567	15	Cent to all
Agglodip	1	0	1.6087	15	Graph
Agglodip	1	0	1.6087	15	Cent & Graph
Agglopdip	10	0.6001	0.9505	10	All to all
Agglopdip	9	0.6781	0.8202	10	Graph
Agglopdip	16	0.4610	1.1844	15	All to all

Agglopdip	2	0.1940	1.7162	15	Graph

Πίνακας 5.3.24 Αποτελέσματα Handwrittern Pendigits PD_3, 5 πραγματικές ομάδες, 16 διαστάσεις

9) Handwritten Digits (USPS)

Όπως και το προηγούμενο, έτσι και αυτό το σύνολο δεδομένων περιέχει δεδομένα χειρόγραφων χαρακτήρων ενός πραγματικού προβλήματος. Πιο συγκεκριμένα, χρησιμοποιήθηκε ο γραφικός χαρακτήρας για τη γραφή των ψηφίων 0-9 πάνω στους ταχυδρομικούς φακέλους από την USPS (United States Postal Service). Τα δεδομένα αυτά προήλθαν από ασπρόμαυρες εικόνες μεγέθους 8x8 pixels των παραπάνω ψηφίων και κατά συνέπεια περιγράφονται από 64 διαστάσεις. Τα σύνολα που θα χρησιμοποιηθούν για τη διεξαγωγή των πειραμάτων, αποτελούν υποσύνολα του συνόλου ελέγχου που αποτελείται από 2000 στοιχεία. Τα υποσύνολα αυτά είναι τα H_1, H_2, H_3 και αφορούν τα ψηφία {0,4,7}, {1,3,5,7,9} και {0,2,4,6,8} αντίστοιχα.

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	29	0.0901	2.4909	1	-
G-means	9	0.2334	1.4720	1	-
Dip-means	1	0	0.6931	1	-
Pdip-means	3	0.6951	0.4969	1	-
Agglodip	1	0	0.6931	6	All to all
Agglodip	1	0	0.6931	6	Cent to all
Agglodip	1	0	0.6931	6	Graph
Agglodip	1	0	0.6931	6	Cent & Graph
Agglodip	1	0	0.6931	9	All to all
Agglodip	1	0	0.6931	9	Cent to all
Agglodip	1	0	0.6931	9	Graph
Agglodip	1	0	0.6931	9	Cent & Graph
Agglopdip	6	0.3409	1.1035	6	All to all
Agglopdip	6	0.3883	1.0612	6	Graph
Agglopdip	6	0.3310	1.1190	9	All to all

Agglopdip	6	0.5763	0.9039	9	Graph

Πίνακας 5.3.25 Αποτελέσματα Handwrittern Digits H_1, 3 πραγματικές ομάδες, 64 διαστάσεις

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	66	0.1165	2.8187	1	-
G-means	18	0.3370	1.7803	1	-
Dip-means	1	0	1.6094	1	-
Pdip-means	7	0.5483	1.4107	1	-
Agglodip	1	0	1.6094	10	All to all
Agglodip	1	0	1.6094	10	Cent to all
Agglodip	1	0	1.6094	10	Graph
Agglodip	1	0	1.6094	10	Cent & Graph
Agglodip	1	0	1.6094	15	All to all
Agglodip	1	0	1.6094	15	Cent to all
Agglodip	1	0	1.6094	15	Graph
Agglodip	1	0	1.6094	15	Cent & Graph
Agglopdip	10	0.3930	1.7488	10	All to all
Agglopdip	9	0.4351	1.596	10	Graph
Agglopdip	13	0.3689	1.8023	15	All to all
Agglopdip	17	0.3372	1.8008	15	Graph

Πίνακας 5.3.26 Αποτελέσματα Handwrittern Digits H_2, 5 πραγματικές ομάδες, 64 διαστάσεις

Αλγόριθμος	**Τελικό *K***	**ARI**	**VI**	**Αρχικό *K***	**Τεχνική**
X-means	109	0.0859	3.2106	1	-
G-means	14	0.3240	1.8776	1	-
Dip-means	2	0.1430	1.8222	1	-
Pdip-means	6	0.3774	1.5116	1	-
Agglodip	2	0.1552	1.7180	10	All to all
Agglodip	1	0	1.6094	10	Cent to all
Agglodip	1	0	1.6094	10	Graph
Agglodip	1	0	1.6094	10	Cent & Graph

Agglodip	2	0.2039	1.6187	15	All to all
Agglodip	1	0	1.6094	15	Cent to all
Agglodip	1	0	1.6094	15	Graph
Agglodip	1	0	1.6094	15	Cent & Graph
Agglopdip	10	0.3547	1.8307	10	All to all
Agglopdip	10	0.3208	1.9186	10	Graph
Agglopdip	13	0.3377	1.9203	15	All to all
Agglopdip	12	0.4010	1.6309	15	Graph

Πίνακας 5.3.27 Αποτελέσματα Handwrittern Digits H_3, 5 πραγματικές ομάδες, 64 διαστάσεις

5.4. Συμπεράσματα

Ερμηνεύοντας τα παραπάνω πειραματικά αποτελέσματα μπορούμε να ισχυριστούμε ότι οι νέοι αλγορίθμοι pdip-means, agglodip και agglopdip στις περισσότερες των περιπτώσεων λειτουργούν το ίδιο ή καλύτερα από τον dip-means προσφέροντας καλύτερης ποιότητας λύσεις τόσο βάσει των μετρικών που χρησιμοποιήθηκαν όσο και μέσω της οπτικοποίησής τους.

Εν αντιθέσει, οι x-means και g-means δεν μπορούν να ανταποκριθούν στην αναγνώριση ομάδων με μη Γκαουσιανές κατανομές, ενώ ακόμη και σε ομάδες που ακολουθούν Γκαουσιανή κατανομή τείνουν να παρουσιάσουν φαινόμενα overfitting κυρίως λόγω της διαφοράς των μεγεθών μεταξύ των ομάδων. Υπήρξαν κάποιες περιπτώσεις στις οποίες ο x-means δεν διασπούσε το αρχικό σύνολο κυρίως στα σύνολα δεδομένων mixed distributions. Το φαινόμενο αυτό εκτός από την αδυναμία του αλγορίθμου στην αναγνώριση ομάδων που προέρχονται από κατανομές διαφορετικές της Γκαουσιανής, οφείλεται και στον καθορισμό μικρής τιμής για το βαθμό διαχωρισμού των ομάδων (c), ώστε να τονιστεί η αδυναμία του αυτή στα συνθετικά αυτά πειράματα.

Επιπλέον κατέστη εμφανές πως οι προσεγγίσεις που βασίζονται σε προβολές είναι αποδοτικότερες σε ορισμένα σύνολα δεδομένων όπως το Dtest Rectangles, ενώ υστερούν ελαφρώς σε άλλα. Επιπλέον, οι συσσωρευτικοί με τους αυξητικούς αλγορίθμους διαφέρουν

και αυτοί σε κάποιες περιπτώσεις, όμως οι διαφορές μεταξύ τους είναι σχετικά μικρές και ευνοούν τους πρώτους ή τους δεύτερους ανάλογα με τη δομή του συνόλου δεδομένων.

Στα σύνολα δεδομένων που προέρχονται από προβλήματα πραγματικών εφαρμογών παρατηρείται μεγάλη βελτίωση στη λύση της ομαδοποίησης σε σχέση με τους x-means, g-means και dip-means σε κάποιες περιπτώσεις. Πιο συγκεκριμένα στο Handwritten Pendigits (UCI) ο agglodip και στο Handwritten Digits (USPS) ο pdip-means καταλήγουν σε πολύ ποιοτικότερες λύσεις σε σχέση με τους υπολοίπους με σχετικά μικρό χρόνο εκτέλεσης.

Σε ότι αφορά στις τεχνικές που χρησιμοποιήθηκαν με τους συσσωρευτικούς αλγορίθμους με σκοπό τη μείωση του χρόνου εκτέλεσης, αξίζει να αναφερθεί πως σε ορισμένες περιπτώσεις χάνεται μέρος της ακρίβειας των λύσεων κυρίως όταν συντελείται ομαδοποίηση σε ένα μόνο βήμα με τη χρήση γραφήματος. Αυτό είναι ιδιαίτερα εμφανές στο τελευταίο σύνολο δεδομένων όπου η χρήση του γραφήματος μειώνει κατά πολύ την ποιότητα των λύσεων. Τέλος, αξίζει να αναφερθεί πως οι τεχνικές αυτές επηρεάζονται από την αύξηση του αρχικού αριθμού των ομάδων και μειώνεται η απόδοσή τους σε μικρή όμως έκταση.

ΚΕΦΑΛΑΙΟ 6. ΚΑΤΑΤΜΗΣΗ ΕΙΚΟΝΩΝ

6.1. Εισαγωγή

Κατάτμηση μιας εικόνας [25] είναι η διαδικασία της διαμέρισης των εικονοστοιχείων (pixels) μιας ψηφιακής εικόνας σε υποσύνολα με ομοιογενές οπτικό περιεχόμενο Τα υποσύνολα αυτά ονομάζονται segments και σκοπός της εύρεσής τους είναι η ευκολότερη και παραστατικότερη κατανόηση των δομών που συνθέτουν σε μια εικόνα. Η ανάλυση των segments μιας εικόνας χρησιμοποιείται για την εύρεση και ταυτοποίηση αντικειμένων, περιγραμμάτων και ορίων (γραμμές, καμπύλες κ.ο.κ.) κάτι που έχει ευρεία εφαρμογή στην ανάλυση ιατρικών αποτελεσμάτων, την αναγνώριση αντικειμένων και την ταυτοποίηση βιομετρικών δεδομένων σε εικόνες. Η κατάτμηση μιας εικόνας γίνεται βάσει κάποιων οπτικών χαρακτηριστικών, όπως είναι το χρώμα, η φωτεινότητα, η υφή κλπ, ώστε τα pixels που βρίσκονται στο ίδιο segment ενός να έχουν “παρόμοιες” τιμές χαρακτηριστικών και ταυτόχρονα να διαφέρουν από τα pixels των υπολοίπων τμημάτων της εικόνας.

6.2. Κατάτμηση εικόνας με χρήση του Agglodip

Η κατάτμηση εικόνας λοιπόν, μπορεί να λογιστεί και ως μια διαδικασία ομαδοποίησης στην οποία τα αντικείμενα–pixels ανατίθενται σε ομάδες βάσει κάποιου

μέτρου ομοιότητας. Η ομαδοποίηση αυτή μπορεί να υλοποιηθεί συσσωρευτικά, ξεκινώντας με ομάδες μικρού μεγέθους ή και με κάθε μεμονωμένο pixel να αναπαριστά μια ξεχωριστή ομάδα και με συνενώσεις μεταξύ των όμοιων μεταξύ τους ομάδων να δημιουργούνται τμήματα ομογενή ως προς συγκεκριμένα χαρακτηριστικά.

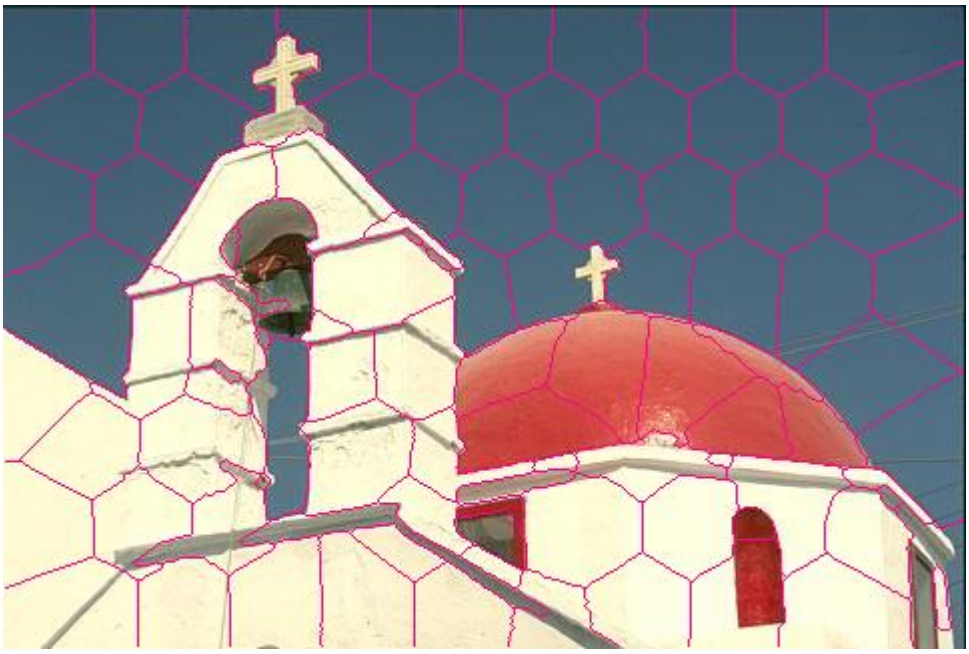
Κατάλληλοι για την επίλυση αυτού του προβλήματος ομαδοποίησης φαίνονται να είναι οι δυο αλγόριθμοι συσσωρευτικής φύσης που αναπτύχθηκαν κατά την παρούσα εργασία, ο agglodip και ο agglopdip. Από τους παραπάνω επιλέχθηκε ο agglodip για την πειραματική μελέτη που παρουσιάζεται παρακάτω. Για την έναρξη της ομαδοποίησης κρίνεται αναγκαία μια διαδικασία προεπεξεργασίας των εικόνων μέσω της οποίας καθορίζεται η μορφή του συνόλου δεδομένων που θα δοθεί ως είσοδος στον αλγόριθμο agglodip για ομαδοποίηση. Η προεπεξεργασία βασίζεται στην ιδέα των *superpixels* και περιγράφεται στη συνέχεια.

6.2.1. Προεπεξεργασία εικόνων

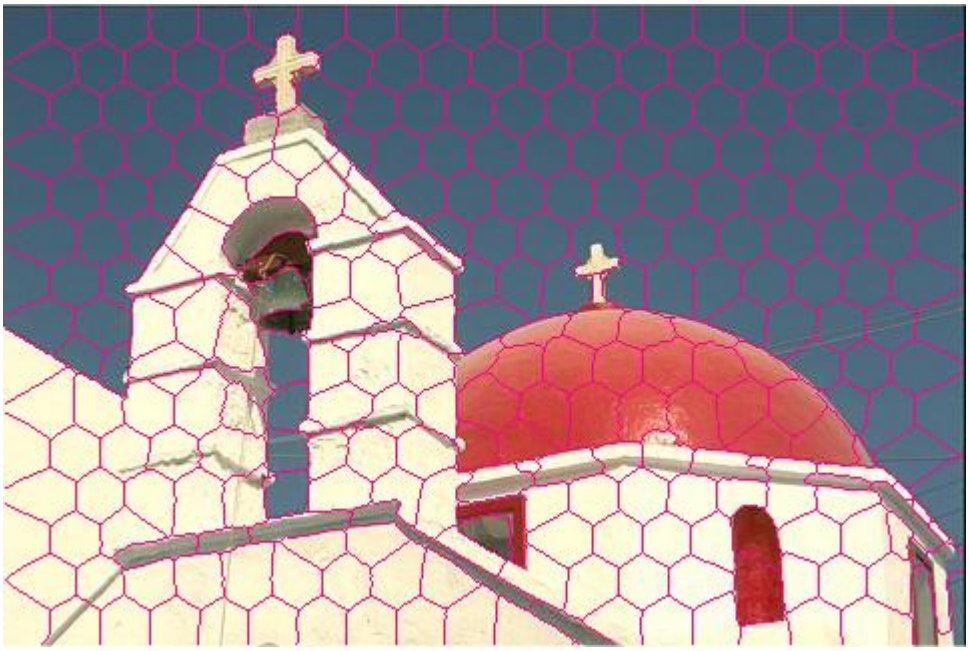
Ένα σημαντικό ζήτημα στην επίλυση προβλημάτων κατάτμησης εικόνων αποτελεί η προεπεξεργασία. Αρχικά παρέχεται ως είσοδος μια έγχρωμη εικόνα, με κάθε pixel της να περιγράφεται από τη θέση του στην εικόνα καθώς και από ένα σύνολο 3 τιμών RGB για το χρώμα που έχει. Στη συνέχεια, αυτά τα δεδομένα εισάγονται στον αλγόριθμο SLIC [26] από τον οποίο παράγεται μια αρχική διαμέριση της εικόνας σε superpixels μικρού σχετικά μεγέθους. Έπειτα εξάγουμε μόνο την πληροφορία που θεωρούμε χρήσιμη από καθένα από τα superpixels αυτά, τη μετασχηματίζουμε κατάλληλα αν κριθεί απαραίτητο και δίνεται το σύνολο δεδομένων και μια αρχική λύση ομαδοποίησης (βασισμένη στα superpixels) στον αλγόριθμο agglodip. Ο τελευταίος θα πραγματοποιήσει συνενώσεις μεταξύ superpixels-ομάδων με παρόμοια χαρακτηριστικά και θα παράγει μια τελική λύση κατάτμησης που περιγράφει τις διάφορες ομογενείς δομές που εμφανίζονται σε μια εικόνα

Ο αλγόριθμος *SLIC* (*Simple Linear Iterative Clustering*) [26] αποτελεί έναν αλγόριθμο ομαδοποίησης για την κατάτμηση εικόνων σε supepixels. Χρησιμοποιεί πενταδιάστατα δεδομένα κάθε pixel (3 για το χρώμα και 2 για τη θέση) ώστε να παράξει μια

αρχική διαμέριση της εικόνας σε ένα επιθυμητό πλήθος superpixels. Αυτού του είδους αλγόριθμοι χρησιμοποιούνται συχνά για την προεπεξεργασία δεδομένων από εικόνες τα οποία θα αποτελέσουν είσοδο σε άλλους αλγορίθμους ομαδοποίησης στη συνέχεια. Είναι σημαντική λοιπόν σε αυτό το σημείο η αποφυγή τεχνικών προβλημάτων που ενδέχεται να δημιουργηθούν από superpixels με μεγάλη διαφορά μεγέθους σε μια εικόνα καθώς και από την ύπαρξη superpixels με πολλές σχηματικές ανωμαλίες. Επιπλέον, είναι ιδανικό να καθορίζεται προσεγγιστικά το πλήθος των superpixels που θα παραχθούν και ο αλγόριθμος που θα εκτελεστεί να έχει μικρό υπολογιστικό κόστος. Ο αλγόριθμος SLIC επιλέχθηκε έναντι αντίστοιχων αλγορίθμων όπως οι *GS04* [27], *NC05* [28], *TP09* [29], *QS09* [30] λόγω της ανωτερότητας του ως προς τη μορφή των superpixels που παράγει και λόγω του μικρού συγκριτικά υπολογιστικού κόστους του σε μνήμη και σε χρόνο, καθότι είναι γραμμικής πολυπλοκότητας $O(N)$, όπου N το πλήθος των pixels της εικόνας. Στα παρακάτω σχήματα (Σχήμα 6.2.1, Σχήμα 6.2.2) παρουσιάζονται δύο παραδείγματα κατάτμησης εικόνων με τον αλγόριθμο SLIC για 100 και 300 superpixels αντίστοιχα.



Σχήμα 6.2.1 Παράδειγμα κατάτμησης SLIC με 100 superpixels



Σχήμα 6.2.2 Παράδειγμα κατάτμησης SLIC με 300 superpixels

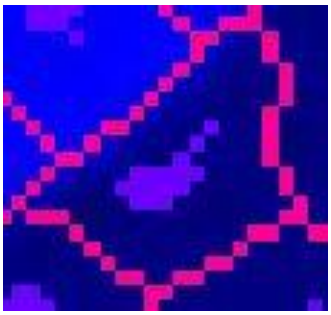
Έπειτα, έχοντας στη διάθεσή μας τη διαμέριση σε superpixels, επιλέγονται από κάθε superpixel μόνο τα κεντρικά pixels του, τα οποία θα αναπαριστούν το superpixel αυτό και θα είναι μέλη του τελικού συνόλου δεδομένων που θα δοθεί για επεξεργασία στον αλγόριθμο ομαδοποίησης agglodip. Η επιλογή των κεντρικών pixels για κάθε superpixel αποσκοπεί στη σημαντική μείωση του μεγέθους του συνόλου δεδομένων κάτι που θα συντελέσει και στη γρηγορότερη εκτέλεση του αλγορίθμου. Αυτό, σε μηχανήματα καθημερινής χρήσης με περιορισμένο μέγεθος μνήμης RAM κρίνεται απαραίτητο καθώς όσο η ανάλυση των εικόνων αυξάνεται, δημιουργούνται πίνακες μεγάλοι σε μέγεθος που είναι δύσκολο να επεξεργαστούν από προσωπικούς υπολογιστές. Επιπλέον με τη χρήση αυτής της μεθόδου, αποφεύγονται τα pixels που βρίσκονται στα όρια μεταξύ superpixels, καθώς ενδέχεται να υπάρχει αλλοίωση/ξεθώριασμα των χρωμάτων όσο περισσότερο απομακρυνόμαστε από το κέντρο ενός superpixel. Κατά συνέπεια τα κεντρικά pixels αποτελούν καλούς αντιπροσώπους των superpixels στα οποία ανήκουν.

Η επιλογή των κεντρικών pixels από κάθε superpixel είναι δυναμική και εξαρτάται όχι μόνο από το μέγεθος του κάθε superpixel αλλά και από το σχήμα του. Η μέθοδος λοιπόν για την επιλογή των κεντρικών pixels ακολουθεί τα παρακάτω 3 βήματα:

1. Από την κάθε στήλη με pixels που ανήκει στο superpixel, επιλέγεται το κεντρικό στοιχείο προσθέτοντας άλλα w σε πλήθος pixels σε απόσταση από αυτό, δηλαδή w pixels πάνω από το κεντρικό και w pixels κάτω από αυτό.
2. Η ίδια λογική εφαρμόζεται και για τις γραμμές, δηλαδή επιλέγονται το κεντρικό pixel κάθε γραμμής και w pixels αριστερά και w δεξιά από αυτά.
3. Από τα σύνολα αυτών των pixel κάθε γραμμής και κάθε στήλης που έχουν επιλεχθεί, παίρνουμε τελικά την τομή τους, δηλαδή τα pixels που ανήκουν και στα δύο σύνολα (στηλών και γραμμών) των επιλαχόντων pixels. Αυτά θα αποτελούν τα κεντρικά pixels που θα συνθέτουν το σύνολο δεδομένων προς ομαδοποίηση.

Η παραπάνω μέθοδος εξασφαλίζει την σωστή επιλογή των κεντρικών pixels από ένα superpixel ανεξάρτητα από τη σχηματική του δομή. Στα παρακάτω σχήματα (Σχήμα 6.2.3 – Σχήμα 6.2.5) παρουσιάζεται η επιλογή των κεντρικών pixels, τα οποία απεικονίζονται με μωβ χρώμα, από superpixels με διάφορες δομές.

Σχήμα 6.2.3 Κεντρικά pixels 1



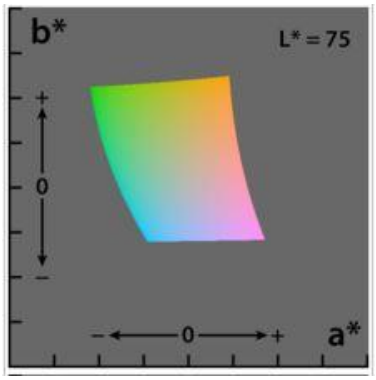
Σχήμα 6.2.4 Κεντρικά pixels 2

Σχήμα 6.2.5 Κεντρικά pixels 3

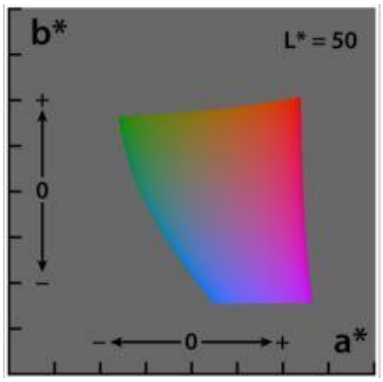
Σε αυτό το σημείο υπάρχουν 3 προσεγγίσεις για τη δημιουργία του συνόλου δεδομένων το οποίο θα χρησιμοποιηθεί από τον αλγόριθμο agglodip για συνένωση των superpixels. Η 1η συνίσταται στην χρησιμοποίηση των τιμών χρώματος του κάθε pixel χωρίς επεξεργασία. Κάθε pixel δηλαδή θα αποτελεί ένα αντικείμενο στο σύνολο δεδομένων με 3 χαρακτηριστικά, ένα για κάθε τιμή του RGB. Οι τιμές αυτές αντιστοιχούν στα βασικά χρώματα κόκκινο, πράσινο και μπλε και κυμαίνονται μεταξύ των τιμών [0-255].

Η 2η προσέγγιση χρησιμοποιεί έναν μετασχηματισμό για τη μετάβαση από τον χρωματικό χώρο RGB στον χρωματικό χώρο *CIE 1976 (L*,a*,b*)* [31] [32]. Αυτή η προσέγγιση διαφέρει από την κλασσική αναπαράσταση του χρώματος κάθε pixel βάσει των

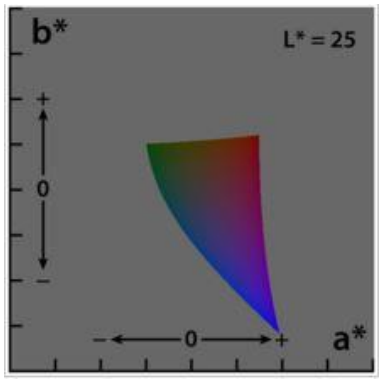
RGB τιμών του καθώς η τιμή L συμβολίζει τη φωτεινότητα κάθε pixel, με $L \in [0,100]$ και οι τιμές a και b το χρώμα από ένα φάσμα χρωμάτων που παίρνει τιμές στο διάστημα [-100,100]. Το φάσμα αυτό ξεπερνάει σε μέγεθος το κλασσικό φάσμα χρωμάτων sRGB, στην παρούσα εργασία όμως χρησιμοποιούμε εικόνες που ανήκουν στην περιοχή απεικόνισης χρωμάτων του sRGB για χάρη απλότητας. Στα σχήματα 6.2.6 έως 6.2.8 απεικονίζονται τα χρώματα τα οποία ανήκουν στο φάσμα sRGB στον χώρο χρωμάτων (L*,a*,b*) για διάφορες τιμές φωτεινότητας.



Σχήμα 6.2.6 sRGB στο χώρο (L*,a*,b*), L* = 75



Σχήμα 6.2.7 sRGB στο χώρο (L*,a*,b*), L* = 50



Σχήμα 6.2.8 sRGB στο χώρο (L*,a*,b*), L* = 25

Η 3η προσέγγιση διαφέρει σημαντικά από τις δύο που αναφέρθηκαν παραπάνω καθώς στοχεύει στην απεξάρτηση σε ένα βαθμό του συνόλου δεδομένων από το χρώμα. Πιο συγκεκριμένα, η προσέγγιση αυτή ακολουθεί τα εξής βήματα:

1. Οι χρωματικές τιμές RGB που περιγράφουν κάθε pixel μετασχηματίζονται στις αντίστοιχες ασπρόμαυρες (gray-scale) και πλέον κάθε pixel χαρακτηρίζεται από μία τιμή στην κλίμακα του γκρι (0-255).

2. Ορίζουμε ένα μέγεθος παραθύρου w.
3. Για κάθε pixel βρίσκουμε τις τιμές που έχουν τα γειτονικά από αυτό pixels σε απόσταση w προς κάθε διεύθυνση γύρω του. Οι τιμές αυτές είναι $(2\times w+1)^2$ σε πλήθος.
4. Εφαρμόζουμε στο σύνολο αυτών των τιμών που χαρακτηρίζουν τα pixels της εικόνας PCA ώστε να μειώσουμε τη διάσταση επιλέγοντας $m=2\times w+1$ πρωτεύουσες διαστάσεις.
5. Κάθε pixel της εικόνας πλέον θα αποτελείται από τα m παραπάνω χαρακτηριστικά. Το τελικό σύνολο δεδομένων θα περιέχει μόνο τα κεντρικά pixels κάθε superpixel, το οποίο θα αποτελέσει είσοδο στον αλγόριθμο agglodip για ομαδοποίηση στη συνέχεια.

Στην επόμενη ενότητα θα παρουσιαστούν πειράματα και εφαρμογές των παραπάνω μεθόδων για διαφορετικές τιμές παραμέτρων καθώς και συγκρίσεις μεταξύ τους σε έγχρωμες εικόνες, τεχνητές και μη.

6.3. Πειραματικά Αποτελέσματα

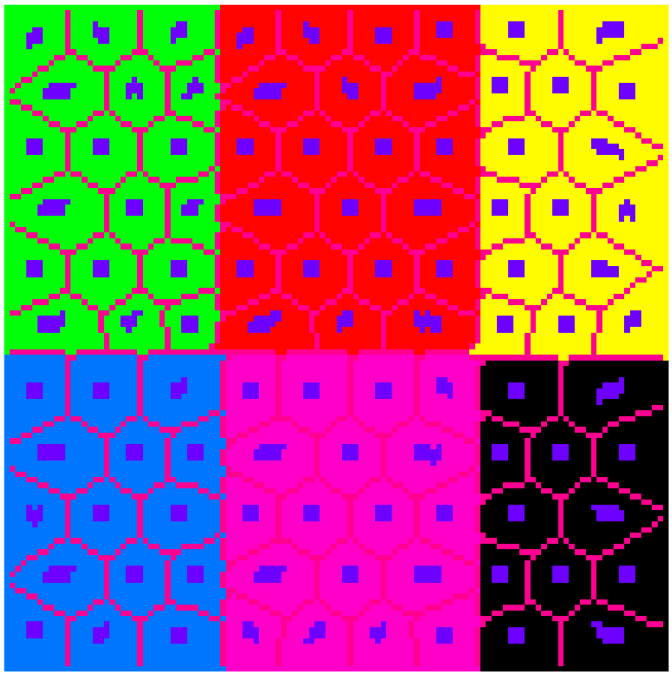
Στην ενότητα αυτή θα παρουσιαστούν τα αποτελέσματα από την εκτέλεση πειραμάτων (υπέρ)κατάτμησης εικόνων με τη χρήση του αλγορίθμου SLIC και μιας διαδικασίας προεπεξεργασίας όπως αναφέρθηκε παραπάνω και του αλγορίθμου agglodip για την εκτέλεση της συσσωρευτικής ομαδοποίησης στη συνέχεια. Οι παράμετροι που χρησιμοποιούνται από τον SLIC για την παραγωγή των αρχικών superpixels διαφέρουν από εικόνα σε εικόνα και χρησιμοποιείται αυτή που παράγει την πιο αντιπροσωπευτική απεικόνιση ώστε ο agglodip να έχει ως είσοδο όσο το δυνατόν περισσότερη χρήσιμη πληροφορία. Για μέγεθος παραθύρου ορίζουμε w = 2. Οι παράμετροι εισόδου του agglodip είναι όπως και κατά την εκτέλεση των πειραμάτων του προηγούμενου κεφαλαίου $a=0,\ v_{thd}=1\%.$

Τα αποτελέσματα θα παρουσιαστούν υπό τη μορφή σχημάτων, δηλαδή θα παρουσιαστούν εικόνες στις οποίες θα είναι εμφανής η κατάτμησή τους και κάθε ομάδα θα ξεχωρίζει από κάθε γειτονική της με τη χρήση γραμμών μωβ χρώματος που καθορίζουν τα

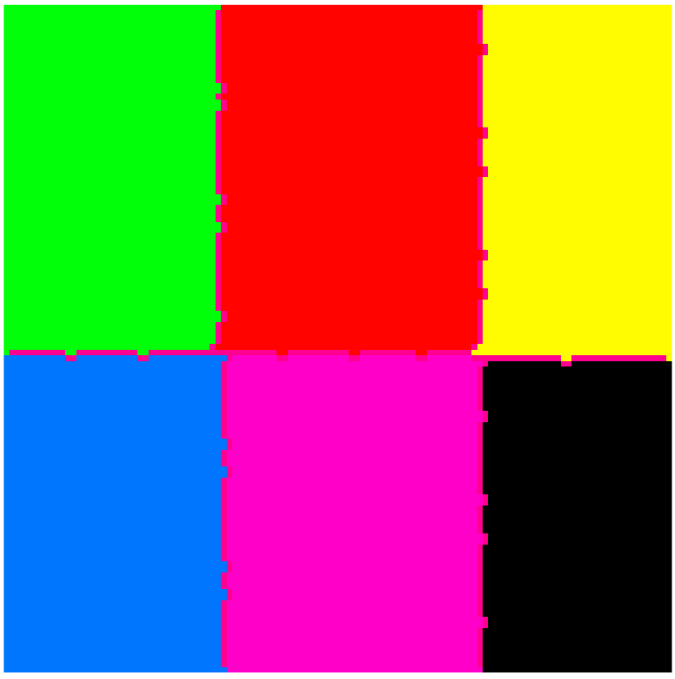
όριά της. Αρχικά θα παρουσιάζεται η αρχική διαμέριση που παράγει ο αλγόριθμος SLIC και στο εσωτερικό κάθε superpixel θα εμφανίζονται με μπλε χρώμα τα pixels που επιλέγονται ως αντιπροσωπευτικά αυτού και τα οποία χρησιμοποιούνται για να γίνει η ομαδοποίηση. Έπειτα θα παρουσιάζονται: μία εικόνα για την εκδοχή της κατάτμησης με τη χρησιμοποίηση χρωμάτων από τον χρωματικό χώρο *CIE 1976 (L*,a*,b*)* και μία με τη δεύτερη εκδοχή που χρησιμοποιεί τις ασπρόμαυρες χρωματικές τιμές των γειτονικών pixels με μετέπειτα προβολή στην PCA για την περιγραφή του καθενός μεμονωμένου pixel. Αυτές οι δυο εικόνες έχουν επιλεγεί από ένα σύνολο 4 εκδοχών διαφορετικών εκτελέσεων του αλγορίθμου agglodip οι οποίες όπως αναφέρθηκαν και σε προηγούμενα κεφάλαια είναι: η χρήση όλων των αποστάσεων των δύο ομάδων pixels για την εφαρμογή του dip τεστ, η χρήση μόνο των αποστάσεων των κεντρικών pixels προς κάθε μία ομάδα, η χρήση των συνεκτικών συνιστωσών του γραφήματος γειτνίασης και ο συνδυασμός των δύο τελευταίων, και επιλέγεται από αυτό το σύνολο η εικόνα στην οποία εμφανίζεται η καλύτερη κατάτμηση, με άξονα την εξαγωγή συμπερασμάτων με περισσότερη πληροφορία για τη δομή των σχημάτων και των χρωμάτων της υποκείμενης εικόνας. Αρχικά θα εξεταστούν συνθετικές εικόνες και στη συνέχεια θα παρουσιαστούν πειράματα κατάτμησης τα οποία έχουν εκτελεστεί σε φωτογραφίες. Τέλος να επισημανθεί πως οι συμβολισμοί (A), (C), (G), (GC) αντιστοιχούν στις τεχνικές (All to all, Centroids to all, Graph, Graph with Centroids to all) που χρησιμοποιήθηκαν για την παραγωγή της αντίστοιχης κατάτμησης.

1) 6 Colored Rectangles

Η πρώτη εικόνα αποτελείται από 6 χρώματα: πράσινο, κόκκινο, κίτρινο, μπλε, ροζ και μαύρο σε 6 παραλληλόγραμμα τα οποία εφάπτονται το ένα στο άλλο. Στα παρακάτω σχήματα (Σχήμα 6.3.1 – Σχήμα 6.3.3) παρουσιάζονται η αρχική διαμέριση με την επιλογή των σημείων κάθε superpixel που θα χρησιμοποιηθεί για την κατάτμηση της εικόνας, η καλύτερη κατάτμηση με τη χρήση χρώματος και η καλύτερη με τη χρήση γειτονικών pixels ασπρόμαυρου χρώματος για την περιγραφή καθενός εικονοστοιχείου αντίστοιχα.



Σχήμα 6.3.1 6 Colored Rectangles: Επιλογή κεντρικών pixels

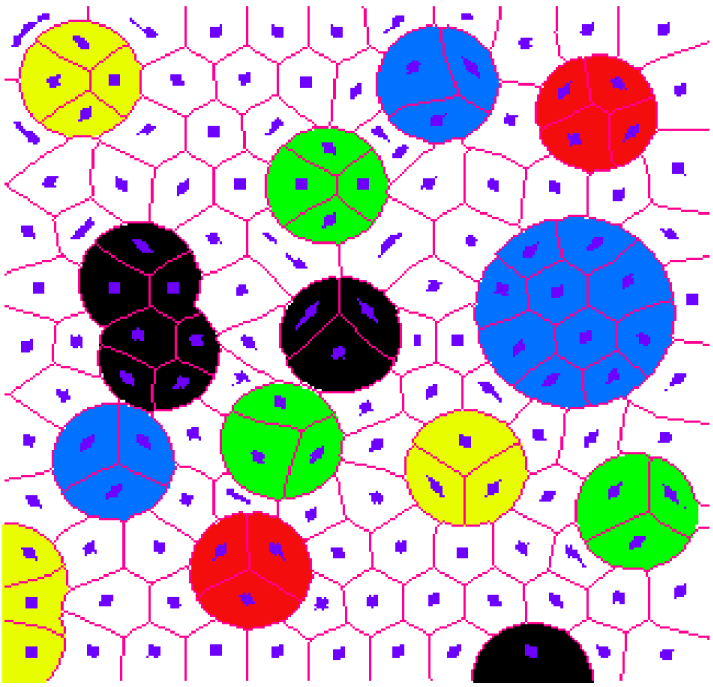


Σχήμα 6.3.2 6 Colored Rectangles: Κατάτμηση χρώματος (C)

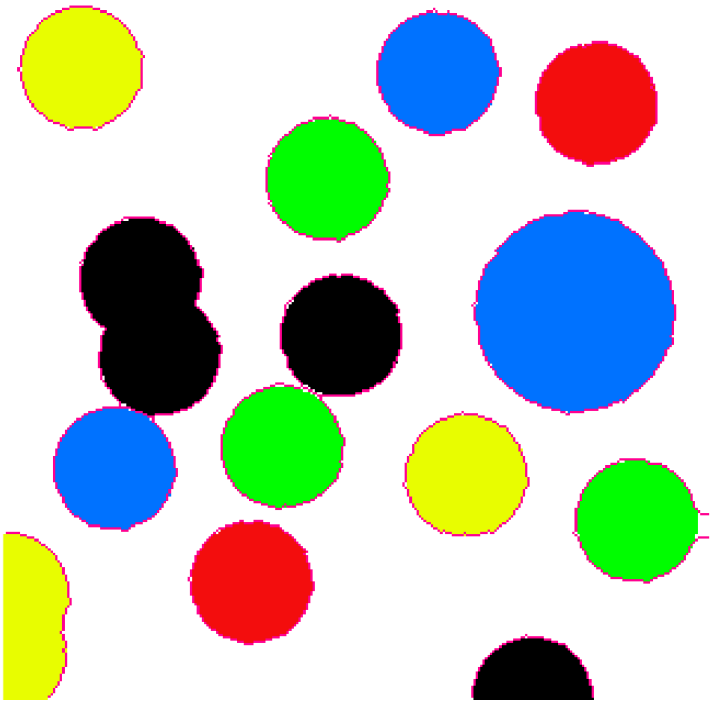
Σχήμα 6.3.3 6 Colored Rectangles: Κατάτμηση με PCA (A)

2) <u>16 Colored Circles</u>

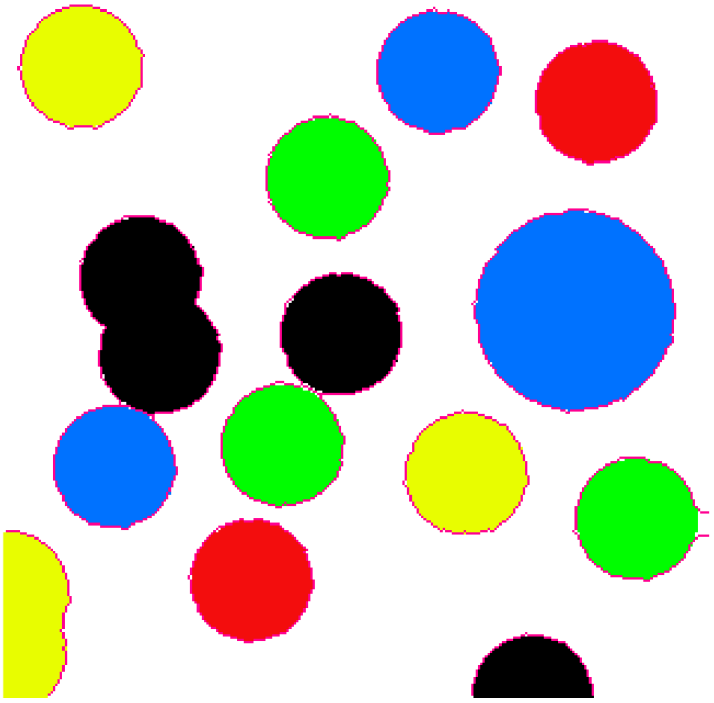
Η εικόνα αυτή αποτελείται από 16 κύκλους διαφορετικών χρωμάτων και σχήματος και από τους οποίους οι 2 είναι ενωμένοι με άλλους 2 κύκλους ίδιου χρώματος. Παρακάτω παρουσιάζονται τα αποτελέσματα.



Σχήμα 6.3.4 16 Colored Circles: Επιλογή κεντρικών pixels



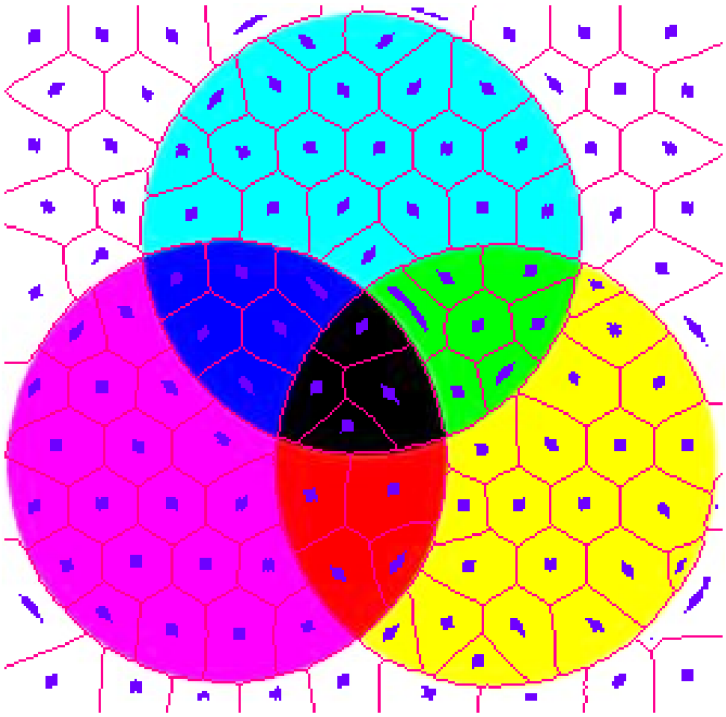
Σχήμα 6.3.5 16 Colored Circles: Κατάτμηση χρώματος (GC)



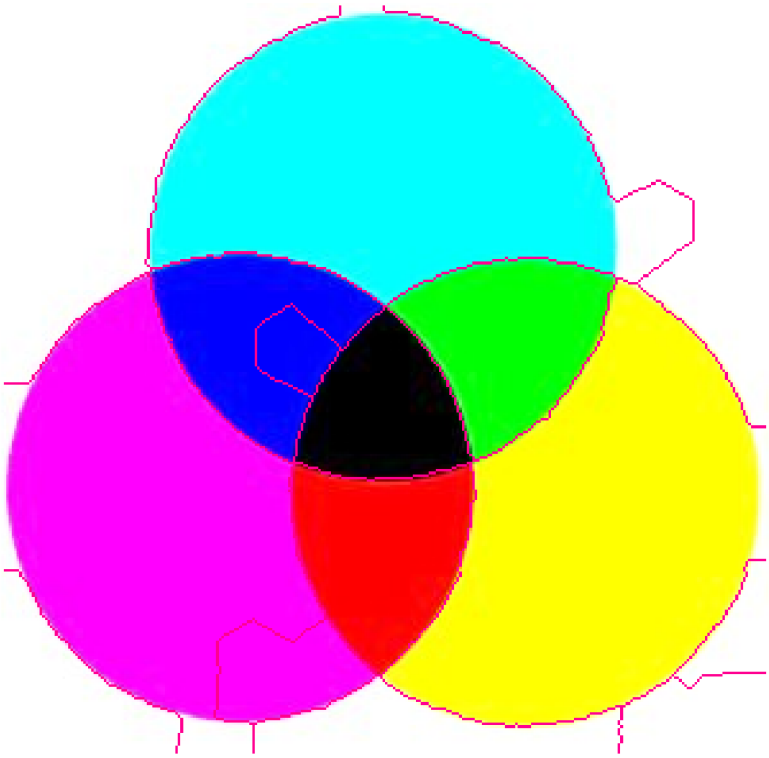
Σχήμα 6.3.6 16 Colored Circles: Κατάτμηση με PCA (GC)

3) Primary Colors

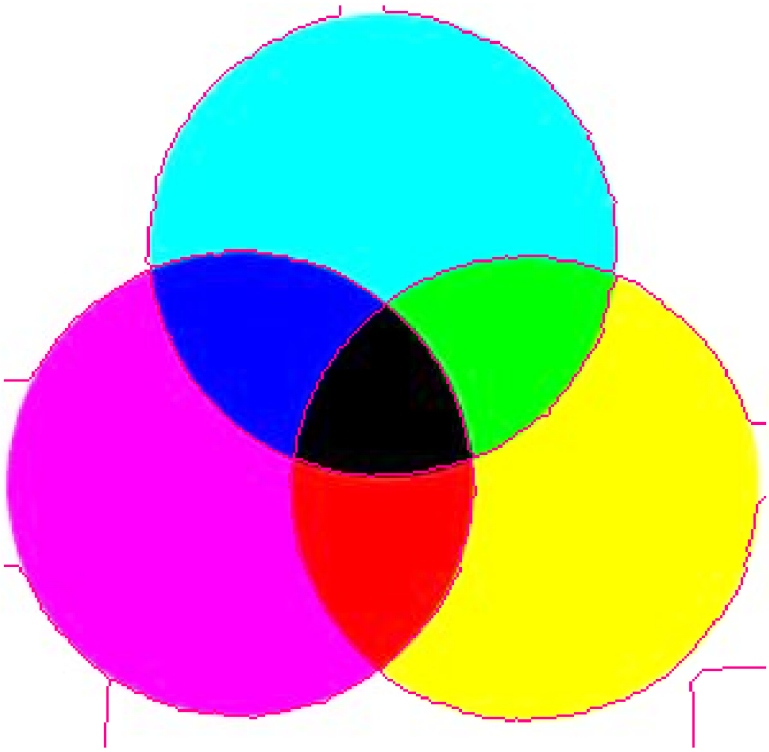
Στην παρούσα εικόνα εμφανίζονται τα 3 βασικά χρώματα και οι συνδυασμοί τους αντίστροφα. Δηλαδή οι 3 αρχικοί τεμνόμενοι κύκλοι έχουν κυανό, κίτρινο και μωβ χρώμα αντίστοιχα. Οι περιοχές στις οποίες τέμνονται αυτοί ανά δύο δημιουργούν τα 3 βασικά χρώματα και η περιοχή στην οποία τέμνονται όλοι μεταξύ τους δημιουργεί μαύρο χρώμα. Παρακάτω παρουσιάζονται τα αποτελέσματα.



Σχήμα 6.3.7 Primary Colors: Επιλογή κεντρικών pixels



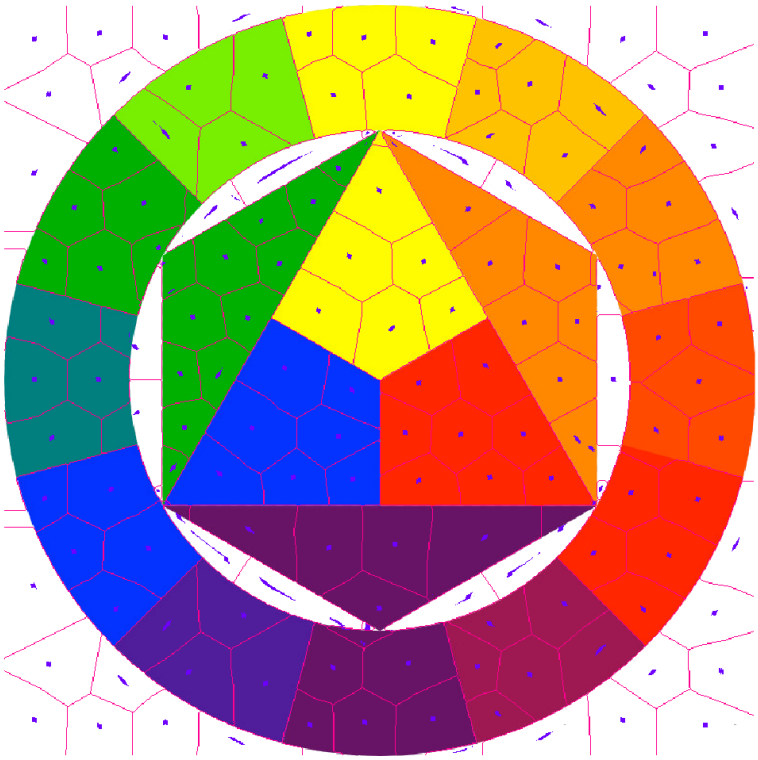
Σχήμα 6.3.8 Primary Colors: Κατάτμηση χρώματος (C)



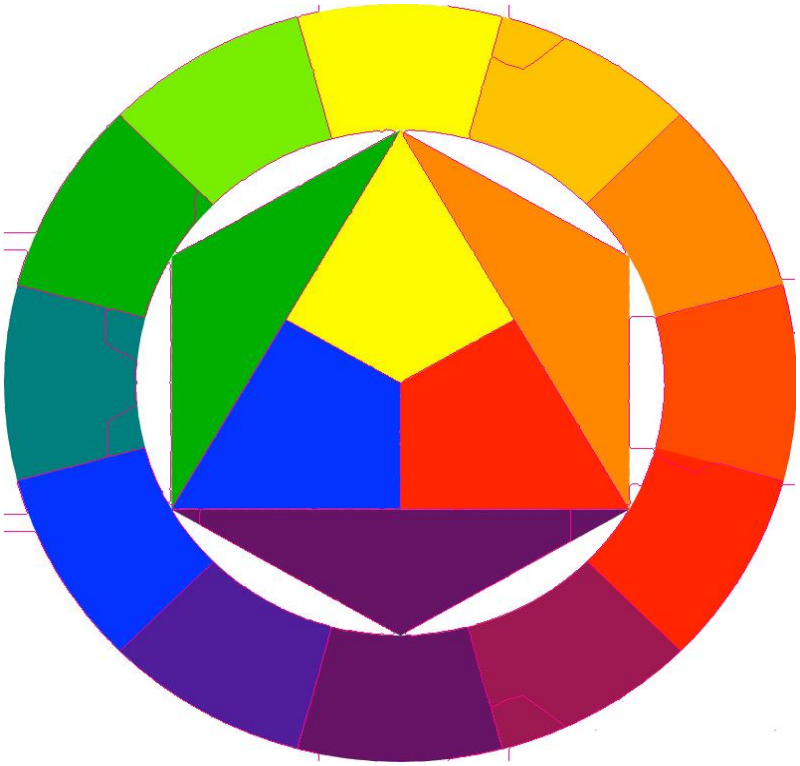
Σχήμα 6.3.9 Primary Colors: Κατάτμηση με PCA (C)

4) Color Wheel

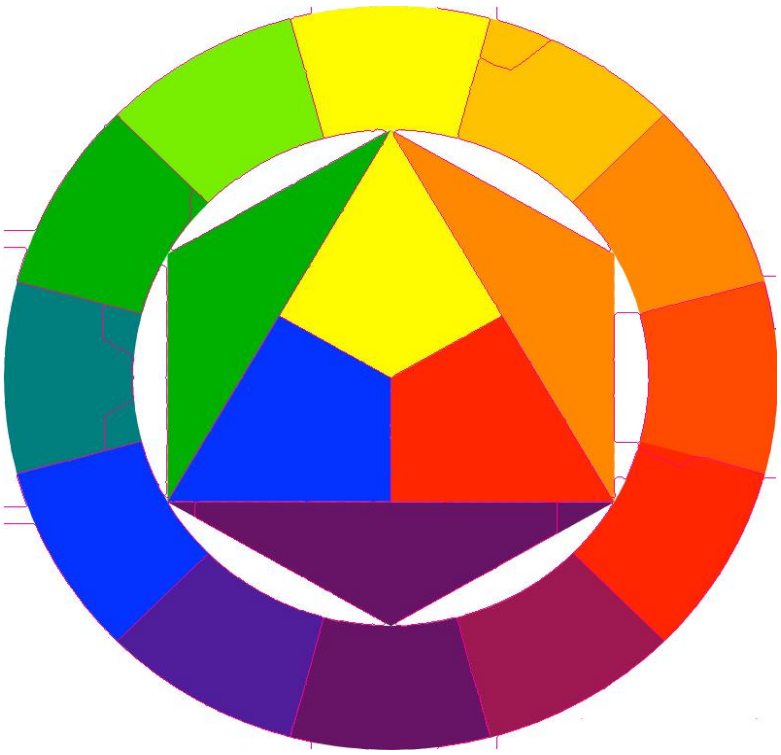
Η εικόνα που χρησιμοποιείται στην προκειμένη περίπτωση είναι η color wheel. Ξεκινά από ένα τρίγωνο με τα χρώματα κίτρινο, κόκκινο και μπλε και το οποίο είναι εσωτερικό ενός εξάγωνου το οποίο περιέχει στις 3 περιοχές που του μένουν τα χρώματα πορτοκαλί, μωβ και πράσινο που προέρχονται από τον ανά δύο συνδυασμό των 3 χρωμάτων του τριγώνου. Τέλος το εξάγωνο αυτό είναι εσωτερικό ενός δακτυλίου ο οποίος περιέχει τους τα 6 χρώματα που έχουν χρησιμοποιηθεί μέχρι στιγμής καθώς και άλλους 6 συνδυασμούς τους. Τα αποτελέσματα της κατάτμησης της εικόνας αυτής παρατίθενται παρακάτω.



Σχήμα 6.3.10 Color Wheel: Επιλογή κεντρικών pixels



Σχήμα 6.3.11 Color Wheel: Κατάτμηση χρώματος (C)



Σχήμα 6.3.12 Color Wheel: Κατάτμηση με PCA (C)

5) White Duck

Η 5η εικόνα απεικονίζει μια άσπρη πάπια να επιπλέει σε μια μικρή γαλάζια λίμνη και ένα νούφαρο, ενώ το φόντο πίσω τους αποτελείται από δύο αποχρώσεις του μπλε. Η κατάτμηση της εικόνας αυτής φαίνεται στη συνέχεια.

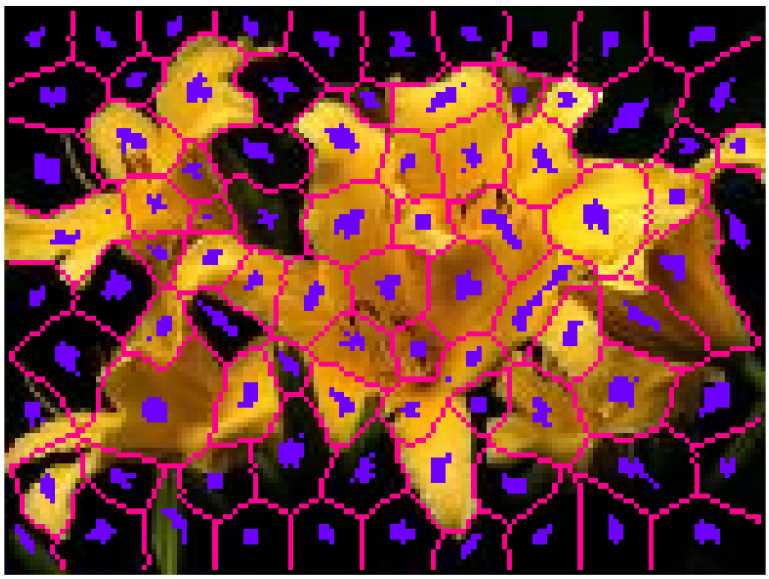
Σχήμα 6.3.12 White Duck: Επιλογή κεντρικών pixels

Σχήμα 6.3.13 White Duck: Κατάτμηση χρώματος (C)

Σχήμα 6.3.13 White Duck: Κατάτμηση με PCA (C)

6) Yellow Flowers

Το 6ο πείραμα συνίσταται από μια φωτογραφία των ανθών κίτρινων λουλουδιών. Η φωτογραφία αυτή όπως και αυτές που θα παρουσιαστούν στη συνέχεια δεν έχουν υποστεί κάποια επεξεργασία σε αντίθεση με τις προηγούμενες εικόνες που έχουν δημιουργηθεί με τη χρήση ηλεκτρονικού υπολογιστή. Η κατάτμησή της παρατίθεται στη συνέχεια.



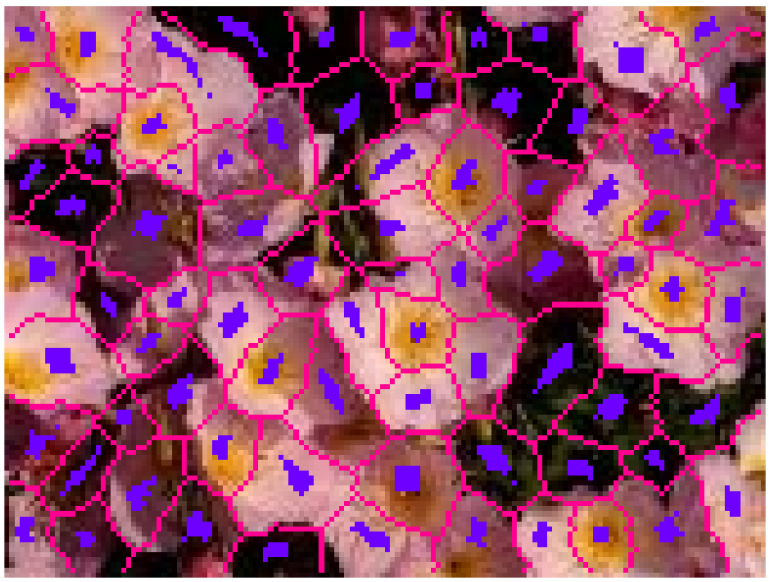
Σχήμα 6.3.14 Yellow Flowers: Επιλογή κεντρικών pixels

Σχήμα 6.3.15 Yellow Flowers: Κατάτμηση χρώματος (C)

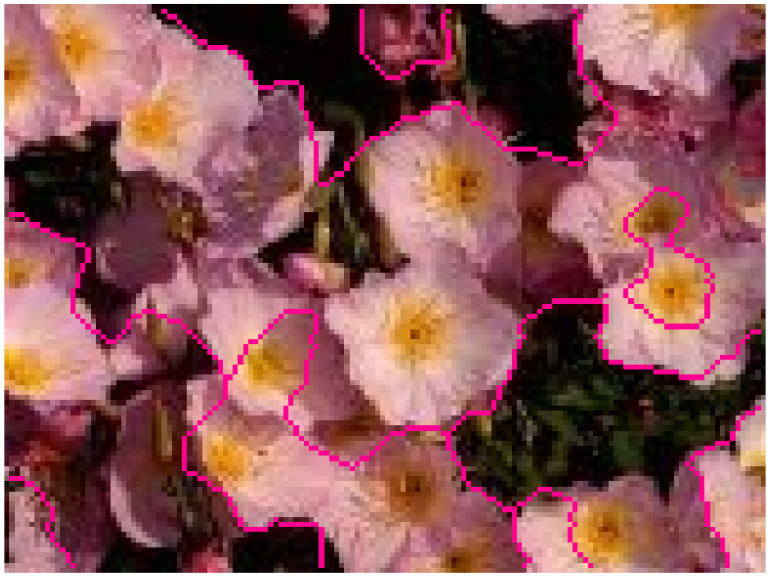
Σχήμα 6.3.16 Yellow Flowers: Κατάτμηση με PCA (C)

7) <u>Pink Flowers</u>

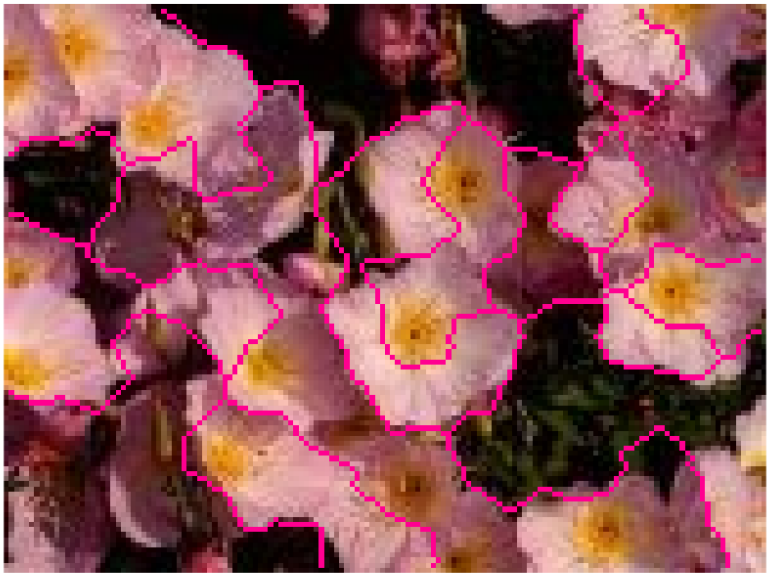
Το 7ο πείραμα συνίσταται από μία φωτογραφία των ανθών λουλουδιών όπως και στο προηγούμενο, ροζ χρώματος όμως σε αυτή την περίπτωση. Παρακάτω παρουσιάζονται τα αποτελέσματα της κατάτμησής της.



Σχήμα 6.3.17 Pink Flowers: Επιλογή κεντρικών pixels



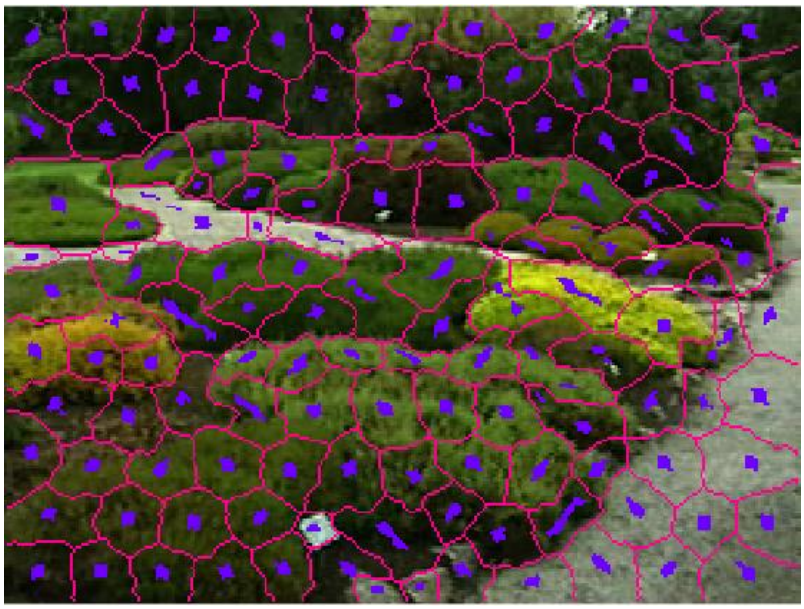
Σχήμα 6.3.18 Pink Flowers: Κατάτμηση χρώματος (GC)



Σχήμα 6.3.19 Pink Flowers: Κατάτμηση με PCA (A)

8) <u>Green Park</u>

Η φωτογραφία που χρησιμοποιείται στο πείραμα αυτό περιέχει πολλές αποχρώσεις του πράσινου καθώς περιέχει διάφορα είδη θάμνων και δέντρων καθώς και 2 μονοπάτια. Η κατάτμηση της φωτογραφίας αυτής παρουσιάζεται στη συνέχεια.



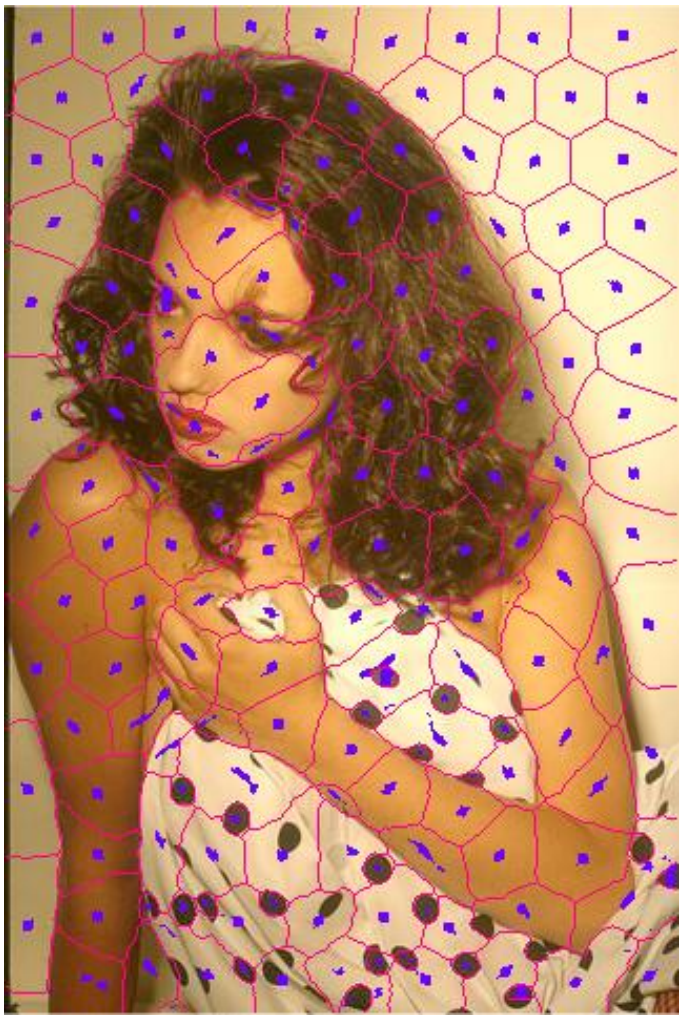
Σχήμα 6.3.20 Green Park: Επιλογή κεντρικών pixels

Σχήμα 6.3.21 Green Park: Κατάτμηση χρώματος (A)

Σχήμα 6.3.22 Green Park: Κατάτμηση με PCA (A)

9) <u>Woman Dots Dress</u>

Στο πείραμα 9 η φωτογραφία που χρησιμοποιείται απεικονίζει μια γυναίκα η οποία φορά ένα άσπρο πουά φόρεμα με μαύρες κουκκίδες και έχει ως φόντο ένα καφέ τοίχο. Η κατάτμηση της φωτογραφίας αυτής θεωρείται δύσκολη λόγω του ότι περιέχει πολλές κουκκίδες το φόρεμα καθώς και των διαβαθμίσεων του φωτισμού πράγμα που γίνεται εμφανές στον τοίχο που βρίσκεται πίσω από τη γυναίκα. Η κατάτμησή της παρατίθεται στη συνέχεια.



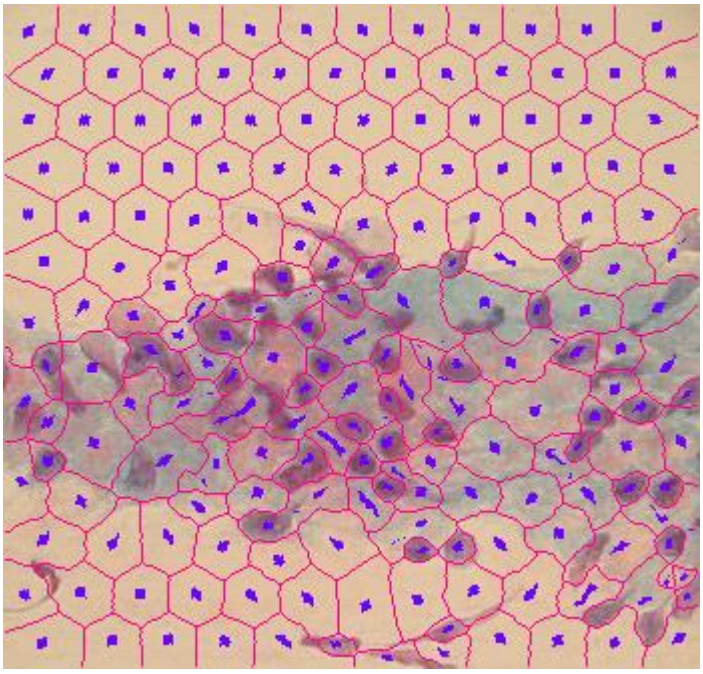
Σχήμα 6.3.23 Woman Dots Dress: Επιλογή κεντρικών pixels

Σχήμα 6.3.24 Woman Dots Dress: Κατάτμηση χρώματος (C)

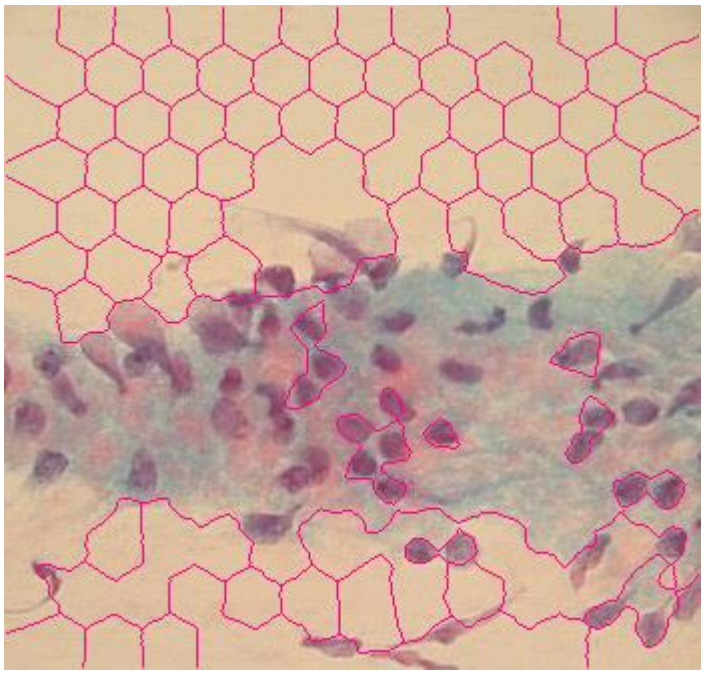
Σχήμα 6.3.25 Woman Dots Dress: Κατάτμηση με PCA (C)

10) <u>Microscope Structure</u>

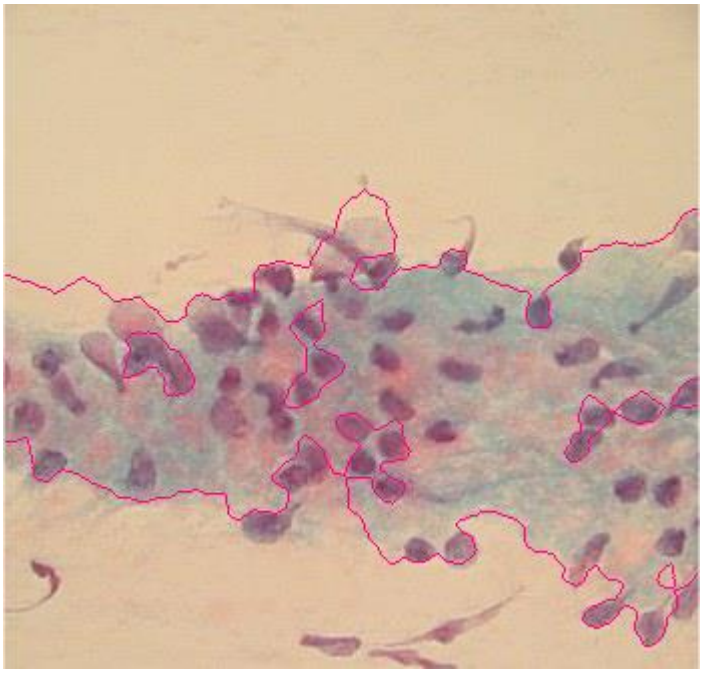
Στο πείραμα αυτό το οποίο αποτελεί και το τελευταίο της σειράς πειραμάτων κατάτμησης εικόνων, επιτελείται κατάτμηση σε μια φωτογραφία που έχει προέλθει από την εξέταση μέσω ενός μικροσκοπίου. Το μεγαλύτερο μέρος της δομής που εμφανίζεται στη φωτογραφία αυτή καταλαμβάνει το κυρίως το κεντρικό κομμάτι της και αποτελείται από μικρά σχήματα ποικίλων δομών και μεγέθους. Ακολουθεί η κατάτμησή της.



Σχήμα 6.3.26 Microscope Structure: Επιλογή κεντρικών pixels



Σχήμα 6.3.27 Microscope Structure: Κατάτμηση χρώματος (A)



Σχήμα 6.3.27 Microscope Structure: Κατάτμηση με PCA (A)

6.4. Συμπεράσματα

Ερμηνεύοντας τα παραπάνω πειραματικά αποτελέσματα της κατάτμησης των εικόνων που χρησιμοποιήθηκαν, παρατηρούμε πως σε γενικές γραμμές οι δύο μέθοδοι λειτουργούν αποδοτικά και πως δεν υπάρχουν σημαντικές διαφορές μεταξύ τους. Σε μερικές περιπτώσεις η μέθοδος που μετατρέπει την εικόνα σε ασπρόμαυρη και χρησιμοποιεί τους γείτονες κάθε εικονοστοιχείου για την περιγραφή του εμφάνισε ελαφρώς καλύτερα αποτελέσματα στις εικόνες Primary Colors, Color Wheel και White Duck και εμφανώς καλύτερη κατάτμηση στην περίπτωση της φωτογραφίας Microscope Structure.

Αυτό που παρατηρήθηκε καθ’ όλο το εύρος των πειραμάτων είναι πως κάθε τεχνική (All to all, Centroids to all, Graph, Graph with Centroids to all) έχει διαφορετική απόδοση και συμπεριφορά σε κάθε εικόνα ή φωτογραφία λόγω της φύσης της. Έτσι, ενώ σε μια εικόνα μπορεί η τεχνική All to all να κάνει oversegmentation, σε άλλη μπορεί να είναι ιδανική και για παράδειγμα η τεχνική Graph που χρησιμοποιεί το γράφημα γειτνίασης να κάνει undersegmentation και να καταλήγει σε μικρότερο αριθμό ομάδων (segments) από το επιθυμητό. Η τεχνική που παράγει στις περισσότερες περιπτώσεις το καλύτερο αποτέλεσμα

είναι η Centroids to all η οποία κυριαρχεί σε 6 από τα 10 πειράματα που εκτελέστηκαν στην προηγούμενη ενότητα.

ΚΕΦΑΛΑΙΟ 7. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

7.1. Σύνοψη Συμπερασμάτων

Στην εργασία αυτή παρουσιάστηκε και μελετήθηκε το πρόβλημα της αυτόματης εύρεσης του κατάλληλου αριθμού των ομάδων *k* σε αλγορίθμους της οικογένειας του αλγορίθμου K-Μέσων. Προτείναμε μεθόδους για τη βελτίωση και αλλά και την εξάλειψη επιπρόσθετων προβλημάτων που αυτός και οι αλγόριθμοι της οικογένειάς του παρουσιάζουν, όπως είναι η επιτυχής μοντελοποίηση περισσότερων κατανομών στα δεδομένα πέρα της Γκαουσιανής.

Οι μέθοδοι ομαδοποίησης που παρουσιάστηκαν είχαν ως βάση τους το στατιστικό κριτήριο Hartigan's Dip Statistic το οποίο χρησιμοποιήσαμε για τον έλεγχο μονοτροπικότητας (unimodality) μιας ομάδας δεδομένων το οποίο όπως αποδείχθηκε με τη χρήση του αλγορίθμου dip-means παράγει ικανοποιητικά αποτελέσματα. Οι τρεις αλγόριθμοι που αναπτύξαμε αποτελούν τροποποιήσεις αυτού, χρησιμοποιώντας διαφορετική προσέγγιση κάθε φορά. Έτσι, στους pdip-means και agglopdip προτείνεται χρήση του κριτηρίου dip σε προβολές των δεδομένων, ενώ με τους agglodip και agglopdip εισάγονται δύο αλγόριθμους συσσωρευτικής φύσης στην εργαλειοθήκη των μεθόδων που υπολογίζουν αυτόματα τον αριθμό των ομάδων.

Έπειτα, μελετήσαμε πειραματικά τη συμπεριφορά των παραπάνω αλγορίθμων σε συνθετικά αλλά και σε πραγματικά σύνολα δεδομένων, μετρήσαμε την απόδοσή τους και

τους συγκρίναμε με τους προϋπάρχοντες αλγορίθμους της κατηγορίας αυτής. Παρατηρήσαμε ότι συνήθως οι επιδόσεις είναι ισάξιες ή καλύτερες σε σχέση με τον dip-means και πολύ καλύτερες σε σχέση με τους g-means και x-means. Επιπλέον, αξίζει να σημειωθεί πως σε αρκετές περιπτώσεις κυρίως με τη χρήση τεχνικών επιτάχυνσης πετύχαμε σημαντικά μειωμένους χρόνους χωρίς απαραίτητα να επέρχεται μείωση της ποιότητας της ομαδοποίησης.

Τέλος, παρουσιάσαμε μια εφαρμογή ενός από τους συσσωρευτικούς αλγορίθμους και πιο συγκεκριμένα του agglodip, για την κατάτμηση εικόνων με την εκτέλεση κατάλληλης προεπεξεργασίας βασισμένης σε αρχική κατάτμηση σε superpixels. Τα πειράματα που εκτελέσαμε με την προτεινόμενη επέδειξαν ενθαρρυντικά αποτελέσματα κατάτμησης, ωστόσο η μεθοδολογία χρειάζεται περισσότερη αξιολόγηση.

7.2. Κατευθύνσεις Μελλοντικής Εργασίας

Μια αρχική κατεύθυνση για μελλοντική εργασία στις συσσωρευτικές μεθόδους ομαδοποίησης που αναπτύχθηκαν θα μπορούσε να είναι η εισαγωγή νέων τεχνικών επιτάχυνσης ή/και βελτίωση των υπαρχόντων. Κάτι τέτοιο θα μπορούσε να μειώσει ενδεχομένως την χρήσιμη πληροφορία που χάνεται σε ορισμένες περιπτώσεις προς όφελος της ταχύτητας και να καταστεί δυνατή η παραγωγή υψηλής ποιότητας λύσεων ομαδοποίησης με πολύ μικρό χρόνο εκτέλεσης.

Επιπρόσθετα, νέες τεχνικές μπορούν να εφαρμοστούν στο ζήτημα της προεπεξεργασίας των εικόνων με σκοπό την κατάτμησή τους, ώστε τα δεδομένα που δίνονται ως είσοδο στον αλγόριθμο ομαδοποίησης agglodip να περιέχουν περισσότερη χρήσιμη πληροφορία για την παραγωγή “ορθότερης” λύσης ομαδοποίησης. Επίσης, λόγω της σημαντικότητας του αλγορίθμου SLIC για την παραγωγή κατάλληλων αρχικών διαμερίσεων, θα ήταν πολύ χρήσιμη η εύρεση των κατάλληλων κάθε φορά παραμέτρων του SLIC με αυτόματο τρόπο, ώστε όλη η εφαρμογή αυτή της κατάτμησης να εκτελείται και να παράγει το καλύτερο δυνατό αποτέλεσμα με αυτόματο τρόπο.

Τέλος, η διεξαγωγή επιπλέον συγκριτικών πειραμάτων στο σύνολο των μεθόδων θα μπορούσε να οδηγήσει στην εξαγωγή πιο εμπεριστατωμένων συμπερασμάτων σχετικά με τη συμπεριφορά και τις επιδόσεις τους. Επίσης οι προτεινόμενες μέθοδοι μπορούν να εξεταστούν και σε άλλα πεδία εφαρμογής (όπως π.χ. η βιοπληροφορική και η ομαδοποίηση κειμένων) όπου είναι αναγκαία η επίλυση προβλημάτων ομαδοποίησης.

ΑΝΑΦΟΡΕΣ

[1] P.-N. Tan, M. Steinbach and V. Kumar: "*Introduction to Data Mining*", 2006.

[2] Vladimir Brusic and John Zeleznikow. "*Knowledge discovery and data mining in biological databases*". Cambridge University Press, 1999.

[3] https://en.wikipedia.org/wiki/Data_mining

[4] Jain A.K., Murty M.N., Flyn, P.J. "*Data Clustering: A Review*", ACM Computing Surveys, Vol.31, No3, pp 264-323, 1999.

[5] http://stanford.edu/~cpiech/cs221/handouts/kmeans.html

[6] P. Berkin. "*Survey of Clustering Data Mining Techniques*", Research paper, Accure Software, 2002.

[7] M. Teboule, P. Berkhin and I. Dhillon, Y. Guan, and J. Kogan: "*Clustering with Entropy-Like k-Means Algorithms*", Grouping Multidimentional Data, Springer, pp 127-160, February 2006.

[8] D. Pelleg and Andrew Moore. "*X-means: extending k-means with efficient estimation of the number of clusters*". International Conference on Machine Learning (ICML), pp. 727-734, 2000.

[9] R.E. Kass and L. Wasserman. "*A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion*". Journal of the American Statistical Association, 90(431), pp. 928-934, 1995.

[10] X. Hu and L. Xu. "*A comparative study of several cluster number selection criteria.*" In J. Liu et al.(eds.) Intelligent Data Engineering and Automated Learning, pp. 195–202, Springer, 2003.

[11] Provost F. and Fawcett T.: "*Data Science for Business: What you need to know about data mining and data-analytic thinking*", 2013.

[12] G. Hamerly and C. Elkan. "*Learning the k in k-means.*" Advances in Neural Information Processing Systems (NIPS), pp. 281-288, 2003.

[13] Robert E. Kass and Larry Wasserman. "*A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion.*" Journal of the American Statistical Association, 90(431):928–934, 1995.

[14] M. A. Stephens. "*EDF statistics for goodness of fit and some comparisons.*" American Statistical Association, 69(347):730–737, September 1974.

[15] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

[16] A. Kalogeratos and A. Likas, "*Dip-means: an incremental clustering method for estimating the number of clusters.*" Proc. Neural Information Processing Systems (NIPS'12), Lake Tahoe, Nevada, USA, 2012.

[17] J.A. Hartigan and P. M. Hartigan. "*The dip test of unimodality.*" The Annals of Statistics, 13(1), pp. 70-84, 1985.

[18] B.W. Silverman. "*Using Kernel density estimates to investigate multimodality.*" Journal of Royal Statistic Society B, 43(1), pp. 97-99, 1981.

[19] D.L. Boley. "*Principal direction divisive partitioning*." Data Mining and Knowledge Discovery, 2(4), pp. 344, 1998.

[20] https://en.wikipedia.org/wiki/Principal_component_analysis#Further_components

[21] http://www.peltarion.com/doc

[22] Sanjoy Dasgupta. "*Experiments with random projection.*" In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI-2000), pages 143–151. Morgan Kaufmann Publishers, 2000.

[23] Y. Feng and G. Hamerly. "*PG-means: learning the number of clusters in data.*" Advances in Neural Information Processing Systems (NIPS), pp. 393–400, 2006.

[24] A. Likas, N. Vlassis and J. Verbeek, "*The Global K-Means Clustering Algorithm, in Pattern Recognition."* vol. 36, pp. 451-461, 2003.

[25] https://en.wikipedia.org/wiki/Image_segmentation

[26] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine S¨usstrunk, "*SLIC Superpixels.*" EPFL Technical Report 149300, June 2010.

[27] Felzenszwalb, P., Huttenlocher, D.: "*Efficient graph-based image segmentation.*" IJCV (2004) 167–181.

[28] Shi, J., Malik., J.: "*Normalized cuts and image segmentation.*" PAMI (2000) 888–905.

[29] Levinshtein, A., Stere, A., Kutulakos, K., Fleet, D., Dickinson, S., Siddiqi, K.: Turbopixels: "*Fast superpixels using geometric flows.*" PAMI (2009).

[30] Vedaldi, A., Soatto, S.: "*Quick shift and kernel methods for mode seeking.*" ECCV (2008).

[31] https://en.wikipedia.org/wiki/CIELUV

[32] https://en.wikipedia.org/wiki/Lab_color_space

[33] https://en.wikipedia.org/wiki/Rand_index

[34] Marina Meilă, "*Comparing Clusterings by the Variation of Information*." Learning Theory and Kernel Machines, pp173-187, January 2003.

[35] A. Asuncion and D. Newman. UCI Machine Learning Repository. *University of California at Irvine*, Irvine, CA, 2007. Available at: http://www.mlearn.ics.uci.edu/databases/pendigits/

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Ο Θεόφιλος Χαμάλης γεννήθηκε στις Σέρρες στις 12 Μαρτίου 1990. Αποφοίτησε από το 2ο Ενιαίο Λύκειο της ίδιας πόλης και εισήχθη στο προπτυχιακό πρόγραμμα σπουδών του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων το 2008 από όπου αποφοίτησε το 2012. Το 2013 εισήχθη στο Μεταπτυχιακό πρόγραμμα σπουδών του τμήματος Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής.