Εύρεση Πολλαπλών Λύσεων σε Προβλήματα Ομαδοποίησης

Σολομωνίδου Βασιλεία

Μεταπτυχιακή Εργασία Εξειδίκευσης

—◆—

Ιωάννινα, Ιούλιος 2013

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA

ΤΙΤΛΟΣ ΔΙΑΤΡΙΒΗΣ

ΕΥΡΕΣΗ ΠΟΛΛΑΠΛΩΝ ΛΥΣΕΩΝ ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΟΜΑΔΟΠΟΙΗΣΗΣ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

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

από την

Βασιλεία Σολομωνίδου

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Ιούλιος 2013

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Λύκα Αριστείδη και τον μεταδιδάκτορα κ. Χασάνη Βασίλειο για την πολύτιμη βοήθειά τους και την υπομονή κατά τη διάρκεια της εργασίας αυτής.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Βασιλεία Σολομωνίδου του Νικολάου και της Όλγας. MSc, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιούλιος 2013. Εύρεση πολλαπλών λύσεων σε προβλήματα ομαδοποίησης. Επιβλέπων: Λύκας Αριστείδης.

Η ομαδοποίηση (clustering) ενός συνόλου δεδομένων αποτελεί ένα από τα σημαντικότερα προβλήματα στον τομέα της μηχανικής μάθησης και της εξόρυξης γνώσης από τα δεδομένα. Ωστόσο οι τυπικοί αλγόριθμοι παρέχουν μία μόνο τελική λύση ομαδοποίησης γεγονός που είναι περιοριστικό σε πολλές εφαρμογές δεδομένου ότι μπορεί να υπάρχουν πολλές λύσεις υψηλής ποιότητας, αλλά εντελώς διαφορετικές μεταξύ τους. Είναι επίσης πιθανόν ένας ερευνητής να μην ενδιαφέρεται για ήδη γνωστές ομαδοποιήσεις ενός συγκεκριμένου συνόλου δεδομένων, αλλά για άλλες δομές οι οποίες δεν έχουν ακόμα ανακαλυφθεί και το χαρακτηρίζουν εναλλακτικά. Καθώς λοιπόν τα σύνολα δεδομένων γίνονται πιο σύνθετα ενδέχεται να υπάρχουν πολλαπλές ομαδοποιήσεις που είναι εξίσου ενδιαφέρουσες και χαρακτηρίζουν από διαφορετικές οπτικές τα δεδομένα. Στην εργασία αυτή μελετάται το πρόβλημα εύρεσης πολλαπλών ομαδοποιήσεων ενός συνόλου δεδομένων. Αρχικά παρουσιάζεται με σαφήνεια το πρόβλημα της πολλαπλής ομαδοποίησης και η χρησιμότητα των αποτελεσμάτων της σε πολλές σύγχρονες εφαρμογές. Στη συνέχεια περιγράφεται μια κατηγοριοποίηση των μεθόδων ως προς τον χώρο δεδομένων στον οποίο λειτουργούν και αναλύονται μερικές από τις πιο σημαντικές μεθόδους που χρησιμοποιούν τον αρχικό χώρο δεδομένων. Έπειτα εξετάζεται αναλυτικά ο αλγόριθμος global k-means και προτείνεται μια νέα μεθοδολογία ομαδοποίησης που βασίζεται σε αυτόν. Τέλος παρουσιάζονται τα πειραματικά αποτελέσματα μετά την εφαρμογή της προτεινόμενης μεθοδολογίας πολλαπλής ομαδοποίησης σε συνθετικά δεδομένα, σε κατάτμηση εικόνας καθώς και στην εξαγωγή χαρακτηριστικών εικονοπλαισίων από ακολουθίες βίντεο.

EXTENDED ABSTRACT IN ENGLISH

Solomonidou Vasileia N. MSc, Computer Science & Engineering Department, University of Ioannina, Greece. July, 2013. Discovering Multiple Solutions of Clustering Problems. Thesis Supervisor: Likas Aristidis.

Clustering is a fundamental and widely used task in machine learning and data mining. Traditional clustering algorithms usually provide only one clustering solution. However in today's applications more than one solutions may exist when clustering complex data. In addition, in some cases the clustering solution provided by an algorithm does not lie in the area of the user's interest. Alternative clustering solutions may exist and may be interesting for different purposes and equally meaningful to various users. This project aims at creating a new approach in providing multiple clusterings. Initially, we introduce the reader to the field of multiple clustering. We present a categorisation and an overview of the most important methods that use the original data space in order to highlight their main differences. Moreover, we describe thoroughly the global k-means algorithm and we propose a new multiple clustering method based on it. Finally, we present experimental results after having applied the proposed method on synthetic data, image segmentation and key-frame extraction from video sequences.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1. Περιγραφή του Προβλήματος

Η εποχή της πληροφορίας έφερε μαζί της μια τεράστια τάση για ψηφιοποίηση και συλλογή δεδομένων. Η εξέλιξη του διαδικτύου ειδικότερα, έχει παίξει κεντρικό ρόλο στη διάδοση πληροφοριών παρέχοντας εργαλεία για τη μεταφορά και την αποθήκευση μεγάλου όγκου ψηφιακών δεδομένων. Καθώς ο όγκος των δεδομένων διευρύνεται συνεχώς αυτά αποκτούν εξαιρετικό ενδιαφέρον, εάν μέσω κατάλληλης επεξεργασίας εξάγουμε χρήσιμη γνώση. Για το λόγο αυτό έχουν αναπτυχθεί πολλές τεχνικές ανάλυσης δεδομένων. Ένας από τους πιο αποτελεσματικούς τρόπους εξαγωγής γνώσης από τα δεδομένα είναι η ομαδοποίηση. Η ομαδοποίηση αποτελεί θεμελιώδη τεχνική ανάλυσης και χρησιμοποιείται ευρέως σε πολλούς επιστημονικούς τομείς. Στόχος της ομαδοποίησης είναι ο διαμερισμός ενός συνόλου δεδομένων σε διαφορετικές ομάδες, έτσι ώστε τα μέλη μιας ομάδας να είναι όμοια μεταξύ τους, ενώ οι διαφορετικές ομάδες που δημιουργούνται να είναι ανόμοιες. Ωστόσο οι παραδοσιακοί αλγόριθμοι ομαδοποίησης έχουν αρκετούς περιορισμούς. Παράλληλα δεν είναι συνήθως δυνατή η εύρεση πολλαπλών εναλλακτικών λύσεων για το ίδιο σύνολο δεδομένων.

Τα τελευταία χρόνια έχουν δημιουργηθεί πρόσθετες ανάγκες και εναλλακτικές λύσεις ομαδοποίησης είναι επιθυμητές από πολλές σύγχρονες εφαρμογές. Για παράδειγμα, σε μια εφαρμογή ομαδοποίησης πελατών, ένας πελάτης είναι δυνατό να εμφανίζει

πολλαπλές συμπεριφορές και ιδιότητες. Το γεγονός αυτό υποδεικνύει ότι ο πελάτης μπορεί να είναι μέλος πολλών εναλλακτικών ομάδων, όταν λαμβάνεται υπόψη άλλη πτυχή της συμπεριφοράς του. Στη ανάλυση γονιδιακής έκφρασης τα αντικείμενα ανήκουν σε πολλαπλές ομάδες λόγω των διαφορετικών λειτουργιών του κάθε γονιδίου. Σε αντίθεση με τις σύγχρονες απαιτήσεις, οι παραδοσιακές τεχνικές ομαδοποίησης ανιχνεύουν μόνο μια λύση ομαδοποίησης και «χάνουν» πιθανές εναλλακτικές.

Καθώς λοιπόν τα σύνολα δεδομένων γίνονται πιο σύνθετα και αποκτούν περισσότερες διαστάσεις ενδέχεται να υπάρχουν πολλαπλές ομαδοποιήσεις που είναι εξίσου ενδιαφέρουσες και χαρακτηρίζουν από διαφορετικές οπτικές τα δεδομένα. «Διαφορετικές ομαδοποιήσεις είναι σωστές για διαφορετικούς σκοπούς, οπότε δεν μπορούμε να πούμε ποια λύση είναι καλύτερη»[1]. Δημιουργείται λοιπόν το ερώτημα:«Γιατί να διαλέγουμε μόνο μια λύση ενώ υπάρχουν πολλαπλές εναλλακτικές ομαδοποιήσεις;»[2]. Η ιδέα αυτή αποτελεί το κίνητρο του αναπτυσσόμενου ερευνητικού αντικειμένου της πολλαπλής ομαδοποίησης (multiple clustering).

1.2. Αντικείμενο της Εργασίας

Βασικός στόχος της παρούσας εργασίας είναι να περιγράψει με σαφήνεια το πρόβλημα της πολλαπλής ομαδοποίησης και να ορίσει μια νέα μεθοδολογία εύρεσης πολλαπλών λύσεων ομαδοποίησης. Αν και το ερευνητικό ενδιαφέρον για τον τομέα της πολλαπλής ομαδοποίησης είναι σχετικά πρόσφατο, αρκετές εναλλακτικές προσεγγίσεις και μέθοδοι έχουν εμφανιστεί στη βιβλιογραφία τον τελευταίο καιρό. Στην εργασία περιγράφεται μια κατηγοριοποίηση των μεθόδων αυτών ως προς τον χώρο δεδομένων στον οποίο λειτουργούν και επιχειρείται με αυτόν τον τρόπο η εύρεση των κυριότερων διαφορών τους. Παρουσιάζονται επίσης μερικές από τις πιο σημαντικές μεθόδους που χρησιμοποιούν τον αρχικό χώρο δεδομένων.

Στη συνέχεια εξετάζεται αναλυτικά ο αλγόριθμος global k-means[15] και προτείνεται μια νέα μεθοδολογία ομαδοποίησης που βασίζεται στον global k-means. Η προτεινόμενη προσέγγιση έχει σαν στόχο την επέκταση της λειτουργίας του global k-

means. Ο global k-means είναι ένας αυξητικός, ντετερμινιστικός αλγόριθμος ο οποίος προσθέτει δυναμικά μια ομάδα σε κάθε στάδιο χρησιμοποιώντας τον τυπικό k-means αλγόριθμο. Για να προστεθεί μια νέα ομάδα, ο αλγόριθμος πραγματοποιεί N εκτελέσεις του k-means (όπου N είναι το μέγεθος του συνόλου δεδομένων) από τις κατάλληλες αρχικές θέσεις. Από τις N διαφορετικές λύσεις που προκύπτουν επιλέγεται η βέλτιστη. Αυτό ακριβώς το χαρακτηριστικό εκμεταλλεύεται η μεθοδολογία που προτείνεται. Για να λύσουμε το πρόβλημα της πολλαπλής ομαδοποίησης, εκτός από την βέλτιστη εκμεταλλευόμαστε και μερικές λιγότερο «καλές» λύσεις. Οι λύσεις αυτές συντελούν τελικά στον υπολογισμό πολλαπλών εναλλακτικών ομαδοποιήσεων.

1.3. Δομή της Εργασίας

Στο 2° κεφάλαιο παρουσιάζεται μια γενική εισαγωγή στην έννοια της ομαδοποίησης. Περιγράφεται στη συνέχεια το πρόβλημα της πολλαπλής ομαδοποίησης και η χρησιμότητα των αποτελεσμάτων της σε πολλές σύγχρονες εφαρμογές. Τέλος περιγράφονται συνοπτικά οι κατηγορίες μεθόδων ως προς τον βασικό χώρο δεδομένων και μελετάται η πολλαπλή ομαδοποίηση στον αρχικό χώρο δεδομένων.

Στο 3° κεφάλαιο παρουσιάζεται η προτεινόμενη μεθοδολογία. Αρχικά παρουσιάζονται οι αλγόριθμοι ομαδοποίησης k-means και global k-means. Στη συνέχεια περιγράφεται αναλυτικά η διαδικασία που ακολουθήθηκε για την επέκταση του global k-means αλγορίθμου. Τέλος αναλύονται οι λεπτομέρειες υλοποίησης και η μέθοδος υπολογισμού της ομοιότητας των ομαδοποιήσεων.

Στο 4° κεφάλαιο παρουσιάζονται τα αποτελέσματα από τις σειρές πειραμάτων που εκτελέστηκαν. Το κεφάλαιο είναι χωρισμένο σε τρία μέρη. Στο πρώτο μέρος παρουσιάζονται τα πειραματικά αποτελέσματα μετά την εφαρμογή της μεθόδου σε συνθετικά τεχνητά δεδομένα. Στο δεύτερο μέρος περιγράφεται εφαρμογή του αλγορίθμου σε κατάτμηση εικόνων. Στο τελευταίο μέρος παρουσιάζεται η εφαρμογή του αλγορίθμου σε ακολουθίες βίντεο για την εύρεση χαρακτηριστικών εικονοπλαισιών (key-frames) ενός πλάνου.

Στο 5° κεφάλαιο αναφέρονται τα συμπεράσματα που προέκυψαν από τα πειράματα καθώς επίσης και κατευθύνσεις για μελλοντική εργασία.

ΚΕΦΑΛΑΙΟ 2. ΠΟΛΛΑΠΛΗ ΟΜΑΔΟΠΟΙΗΣΗ ΚΑΙ ΣΧΕΤΙΚΗ ΕΡΕΥΝΑ

Στο κεφάλαιο αυτό παρουσιάζεται μια γενική εισαγωγή στην ομαδοποίηση. Στη συνέχεια περιγράφεται η έννοια της πολλαπλής ομαδοποίησης καθώς και μερικές αντιπροσωπευτικές προσεγγίσεις, σύμφωνα με την ταξινόμηση που παρουσιάζεται.

2.1. Ομαδοποίηση (Clustering)

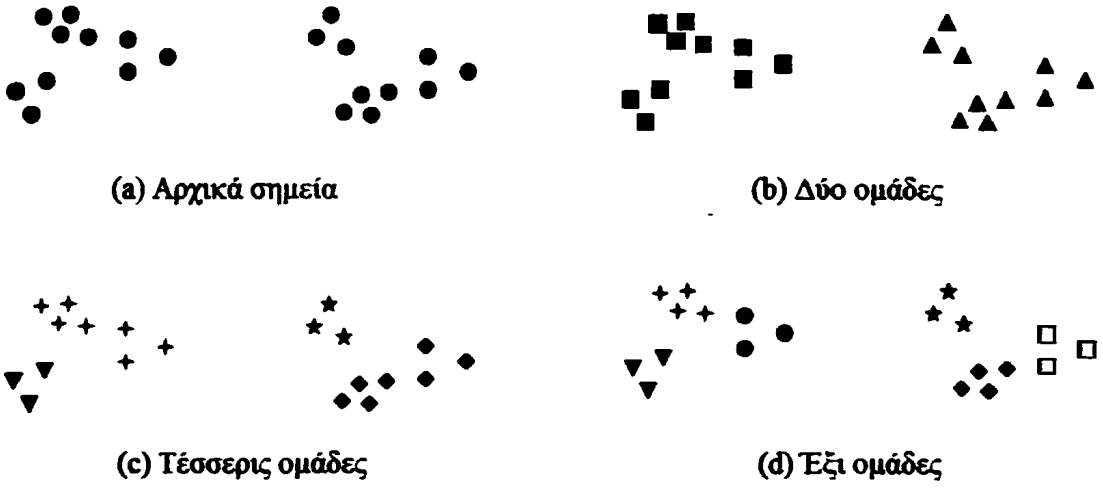
Η συσταδοποίηση ή ομαδοποίηση (clustering) είναι μία από τις βασικές τεχνικές ανάλυσης ενός συνόλου δεδομένων[3]. Η ομαδοποίηση δεδομένων (data clustering) αποτελεί μια τεχνική στατιστικής ανάλυσης δεδομένων και βρίσκει εφαρμογή σε πολλούς επιστημονικούς κλάδους. Ανάμεσα και σε αυτούς, συγκαταλέγονται η μηχανική μάθηση, η εξόρυξη δεδομένων, η αναγνώριση προτύπων, η στατιστική και η βιοπληροφορική. Εφαρμογή βρίσκει επίσης και σε άλλους τομείς, όπως η βιολογία, η ψυχολογία και η ιατρική.

Ομαδοποίηση, ορίζεται η ταξινόμηση όμοιων αντικειμένων σε διαφορετικές ομάδες ή αλλιώς καταμερισμός των δεδομένων σε υποσύνολα, clusters όπως είναι η γνωστή τους ονομασία στη βιβλιογραφία, έτσι ώστε τα δεδομένα να μοιράζονται κοινά χαρακτηριστικά. Με άλλα λόγια, βασικός στόχος της ομαδοποίησης είναι τα αντικείμενα που ανήκουν στην ίδια ομάδα να είναι όμοια (ή να σχετίζονται) μεταξύ

τους, καθώς επίσης να διαφέρουν (ή να μην σχετίζονται) με τα αντικείμενα των άλλων ομάδων [4].

Με μαθηματικούς όρους, η ομαδοποίηση μπορεί να περιγραφεί ως ένα πρόβλημα στο οποίο δίνεται ένα σύνολο n σημείων $X = \{x_1, x_2, \dots, x_N\}$ $x_i \in R^d$ και ζητείται να γίνει ομαδοποίηση σε K ομάδες $\{C_1, C_2, \dots, C_K\}$. Οι ομάδες που παράγονται θα πρέπει να είναι ξένες ανά δύο μεταξύ τους, και η ένωσή τους να αναπαριστά ολόκληρο το σύνολο.

Σε πολλές εφαρμογές η έννοια της ομάδας μπορεί να είναι αφηρημένη[4]. Για να γίνει καλύτερα κατανοητή η δυσκολία που υπάρχει στον ορισμό της σύστασης μιας ομάδας, ας δούμε το σχήμα 2.1. Στο παράδειγμα φαίνονται 20 σημεία και τρεις διαφορετικοί τρόποι διαμερισμού τους. Στο σχήμα 2.1(b) και 2.1(d) βλέπουμε τον διαχωρισμό των δεδομένων σε δύο και έξι ομάδες αντίστοιχα. Επίσης, μπορεί να ισχυριστεί κανείς ότι τα σημεία δημιουργούν τέσσερις ομάδες όπως φαίνεται στο σχήμα 2.1(c). Το παράδειγμα λοιπόν, καταδεικνύει ότι ο ορισμός μιας ομάδας είναι ασαφής και η ομαδοποίηση εξαρτάται από τη φύση των δεδομένων και τον στόχο της ανάλυσης.



(a) Αρχικά σημεία (b) Δύο ομάδες

(c) Τέσσερις ομάδες (d) Έξι ομάδες

Σχήμα 2.1 Ομαδοποίηση δεδομένων με δύο, τέσσερις και έξι ομάδες

Αν και η βάση της ομαδοποίησης είναι αυστηρά μαθηματική, τα κίνητρα ύπαρξής της πηγάζουν από την τάση του ανθρώπου να ομαδοποιεί τα στοιχεία ενός συνόλου πληροφορίας η οποία εισάγεται με οποιονδήποτε τρόπο στο νοητικό του πεδίο. Με την οργάνωση αυτήν, ο άνθρωπος καταφέρνει να διακρίνει γρήγορα τα κυριότερα χαρακτηριστικά της πληροφορίας και να αποκτήσει μια καλή αντίληψη για αυτήν.

Τυπικά η διαδικασία της ομαδοποίησης περιλαμβάνει τα παρακάτω [5]:

1. Αναπαράσταση, επιλογή και εξαγωγή χαρακτηριστικών: Η αναπαράσταση των αντικειμένων αναφέρεται στον καθορισμό του αριθμού των αντικειμένων και του αριθμού, του τύπου και της κλίμακας των χαρακτηριστικών. Στη συνέχεια με την επιλογή χαρακτηριστικών, επιλέγονται τα πιο αντιπροσωπευτικά από τα αρχικά χαρακτηριστικά. Εξαγωγή χαρακτηριστικών είναι η χρήση ενός η περισσότερων μετασχηματισμών των αρχικών χαρακτηριστικών έτσι ώστε να παραχθούν καινούργια.
2. Ομοιότητα: Η ομοιότητα των αντικειμένων υπολογίζεται με κάποια μετρική απόστασης. Η πιο συνηθισμένη μετρική απόστασης είναι η Ευκλείδεια (Euclidean Distance).
3. Ομαδοποίηση: Το βήμα αυτό μπορεί να γίνει με διάφορους τρόπους και εξαρτάται από τη μέθοδο ομαδοποίησης. Τα δεδομένα κατατάσσονται με απόλυτο τρόπο στις διάφορες ομάδες ή κατατάσσονται στις ομάδες με βαθμούς συμμετοχής.
4. Αξιολόγηση: Κρίσιμο βήμα είναι η αξιολόγηση των αποτελεσμάτων. Όλοι οι αλγόριθμοι ομαδοποίησης θα δημιουργήσουν κάποια ομαδοποίηση, χωρίς αυτό να σημαίνει ότι αυτή η λύση περιγράφει αντιπροσωπευτικά το σύνολο δεδομένων. Για αυτό το λόγο, θα πρέπει πρώτα να αποτιμηθούν τα ίδια τα δεδομένα και στη συνέχεια ο αλγόριθμος και τα αποτελέσματά του. Αυτό σημαίνει ότι θα πρέπει να διαπιστωθεί αρχικά αν ένα σύνολο αντικειμένων μπορεί να αποτελέσει την είσοδο ενός αλγορίθμου ομαδοποίησης (cluster tendency) και στη συνέχεια αξιολογούνται τα αποτελέσματά της.

2.2. Πολλαπλή Ομαδοποίηση (Multiple Clustering)

Η ομαδοποίηση αποτελεί μια από τις πιο γνωστές μεθόδους μάθησης χωρίς επίβλεψη. Στη μάθηση χωρίς επίβλεψη, το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, χρησιμοποιώντας κάποια κριτήρια. Τα κριτήρια αυτά, ενδέχεται να ποικίλουν και να παράγουν διαφορετικές ομαδοποιήσεις. Κάθε μια από αυτές τις λύσεις μπορεί να είναι πολύτιμη και ενδιαφέρουσα στο βαθμό που προσδιορίζει τη δομή των δεδομένων από διαφορετική οπτική γωνία.

Κατά την επεξεργασία δεδομένων, είναι πιθανόν η λύση ομαδοποίησης που παράγεται από έναν αλγόριθμο, να μην είναι εκείνη που επιθυμεί ο χρήστης. Με άλλα λόγια, μπορεί ένας ερευνητής να μην ενδιαφέρεται για ήδη γνωστές ομαδοποιήσεις ενός συγκεκριμένου συνόλου δεδομένων, αλλά για άλλες δομές οι οποίες δεν έχουν ακόμα ανακαλυφθεί και το χαρακτηρίζουν εναλλακτικά. Για παράδειγμα, έστω ότι δίνονται δεδομένα για μια ομάδα ανθρώπων και είναι ήδη γνωστή η ομαδοποίησή τους ως προς το φύλο. Ο χρήστης είναι δυνατό να έχει επίγνωση της ομαδοποίησης αυτής και θα προτιμούσε εναλλακτικά μια λύση που χαρακτηρίζει με διαφορετικό τρόπο το σύνολο δεδομένων και πιθανόν να είναι εξίσου σημαντική με την υπάρχουσα λύση.

Ωστόσο οι τυπικοί αλγόριθμοι παρέχουν μόνο μια λύση ομαδοποίησης. Τα τελευταία χρόνια άρχισε να παρατηρείται αξιοσημείωτο ερευνητικό ενδιαφέρον όσον αφορά στην ανάπτυξη αλγορίθμων που υπολογίζουν περισσότερες από μια ομαδοποιήσεις και ερμηνεύουν τη συμπεριφορά των δεδομένων με ποικίλους τρόπους.«Κάθε ομαδοποίηση θα πρέπει να συνοδεύεται τουλάχιστον με μια εναλλακτική λύση»[6].

Παρά το γεγονός ότι ο τομέας της πολλαπλής ομαδοποίησης είναι σχετικά πρόσφατος, υπάρχει ένας αξιόλογος αριθμός προσεγγίσεων στη βιβλιογραφία. Η διαδικασία κατά την οποία υπολογίζονται πολλαπλές ομαδοποιήσεις για ένα σύνολο δεδομένων, εμφανίζεται με διάφορες ορολογίες όπως Εναλλακτική Ομαδοποίηση (Alternative Clustering)[7][8], Ανόμοια Ομαδοποίηση (Disparate Clustering)[9], Meta-Clustering[10] και Σχεσιακή Ομαδοποίηση (Relational Clustering)[11]. Η βασική ιδέα πίσω από όλες αυτές τις μεθόδους, έγκειται στον υπολογισμό

εναλλακτικών λύσεων, όπου κάθε αντικείμενο του συνόλου δεδομένων ανήκει σε πολλαπλές ομάδες. Οι ομαδοποιήσεις αυτές αποτελούν ξεχωριστές λύσεις, και συνεπώς προσφέρουν επιπλέον πληροφορία στον τρόπο με τον οποίο συνοψίζονται τα δεδομένα.

2.2.1. Παρουσίαση του Προβλήματος

Οι παραδοσιακοί αλγόριθμοι ομαδοποίησης στοχεύουν σε μία και μοναδική ομαδοποίηση, στην οποία κάθε αντικείμενο ανατίθεται σε μία ακριβώς ομάδα. Αντίθετα, σύμφωνα με την αρχή των πολλαπλών λύσεων ομαδοποίησης[12], μπορούν να οριστούν πολλές διαφορετικές ομαδοποιήσεις (Clusterings) για ένα σύνολο δεδομένων. Οι κυριότεροι στόχοι της πολλαπλής ομαδοποίησης, μπορούν να περιγραφούν ως εξής:

- Κάθε αντικείμενο μπορεί να κατατάσσεται σε πολλαπλές ομάδες (clusters), αναπαριστώντας τα δεδομένα από διαφορετική οπτική γωνία.
- Το αποτέλεσμα μπορεί να περιλαμβάνει πολλές εναλλακτικές λύσεις. Ο χρήστης μπορεί να διαλέξει μία ή περισσότερες από τις λύσεις.
- Οι εναλλακτικές λύσεις πρέπει να διαφέρουν μεταξύ τους σε μεγάλο βαθμό, και συνεπώς κάθε μία από αυτές να παρέχει επιπρόσθετη γνώση.

Πολλές σύγχρονες εφαρμογές δημιουργούν την ανάγκη ύπαρξης εναλλακτικών λύσεων ομαδοποίησης. Για παράδειγμα στην ανάλυση γονιδιακής έκφρασης, τα γονίδια περιγράφονται από τη συμπεριφορά τους, κάτω από διαφορετικές συνθήκες. Έτσι ένα αντικείμενο μπορεί να έχει πολλαπλές λειτουργίες οπότε μια ομαδοποίηση δεν είναι αντιπροσωπευτική. Σε μια εφαρμογή ομαδοποίησης πελατών, παρατηρούνται για κάθε πελάτη εναλλακτικές συμπεριφορές οι οποίες θα πρέπει να αναγνωρίζονται ως ξεχωριστές ομάδες. Η εφαρμογή ενός παραδοσιακού αλγόριθμου, ομαδοποιεί όμοια αντικείμενα σε μια ομάδα και αντίστοιχα διαμερίζει ανόμοια αντικείμενα σε διαφορετικές ομάδες. Με τον τρόπο αυτό παράγεται μόνο μια λύση ομαδοποίησης και κάθε αντικείμενο ανατίθεται σε μια μόνο ομάδα. Στο Σχήμα 2.2 παρουσιάζεται το παράδειγμα κατηγοριοποίησης συμπεριφοράς πελατών, ύστερα από την εφαρμογή ενός απλού αλγορίθμου ομαδοποίησης. Θεωρούμε ότι κάθε γραμμή αντιπροσωπεύει έναν πελάτη ο οποίος περιγράφεται από διάφορα χαρακτηριστικά.

object ID	age	income	blood pres.	sport activ.	profession
1					
2					
3	50	59.000	130	comp. game	CS
4	51	61.000	129	comp. game	CS
5	49	58.500	...	...	...
6	47	62.000	...	...	...
7	52	60.000	...	...	...
8					
9					

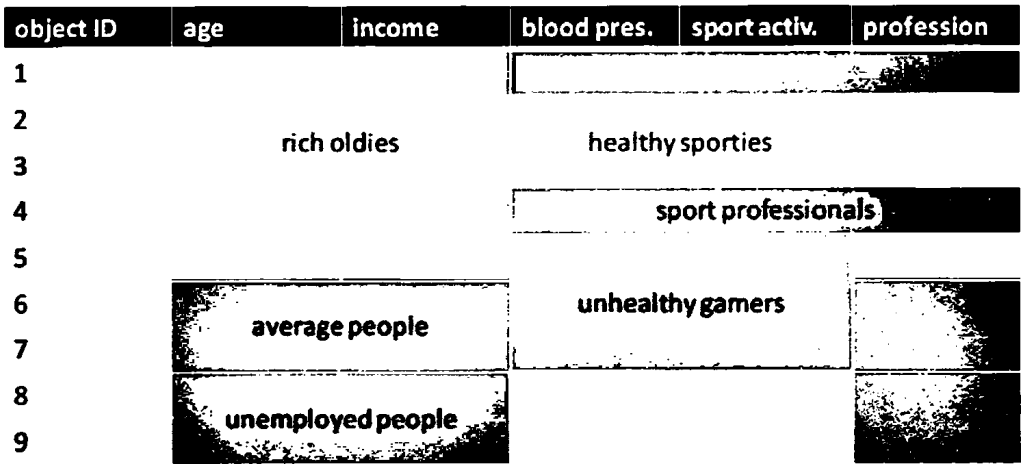
Σχήμα 2.2 Παράδειγμα, μία ομάδα πελατών με παρόμοια χαρακτηριστικά.

Αντίθετα, σύμφωνα με τη λογική των πολλαπλών λύσεων, κάθε αντικείμενο ομαδοποιείται σε σχέση με πολλαπλά κριτήρια, όπως ηλικία, υγεία και μουσικό ενδιαφέρον (Σχήμα 2.3). Τα κριτήρια αυτά εξαρτώνται από τις πιθανές δομές των ομάδων και μπορούν να οδηγήσουν σε πολλαπλές λύσεις ομαδοποίησης, εξετάζοντας τα αντικείμενα από διαφορετικές απόψεις. Για παράδειγμα:

- Ομαδοποίηση των πελατών που μοιάζουν ως προς την κατάσταση υγείας τους.
- Ομαδοποίηση των πελατών που μοιάζουν ως προς τα μουσικά ενδιαφέροντα.
- Ομαδοποίηση των πελατών που μοιάζουν ως προς την αθλητική τους δραστηριότητα.

Για τον υπολογισμό πολλαπλών λύσεων μπορεί επίσης να ληφθεί υπόψη τυχόν προηγούμενη γνώση που μπορεί να έχει ο χρήστης σχετικά με τα δεδομένα. Η γνώση αυτή είναι δυνατό να χρησιμοποιηθεί για την εξαγωγή νέων αποτελεσμάτων τα οποία προσθέτουν επιπλέον πληροφορία σε σχέση με τη δομή των δεδομένων. Το ερώτημα που προκύπτει για ένα σύνολο δεδομένων είναι εάν όντως υπάρχει εναλλακτικός τρόπος ομαδοποίησης ο οποίος να είναι τουλάχιστον τόσο σημαντικός όσο ο υπάρχων, δηλαδή αν μια νέα εναλλακτική ομαδοποίηση προσφέρει επιπλέον πληροφορία στον χρήστη. Με άλλα λόγια ο τομέας της πολλαπλής ομαδοποίησης έχει

σαν κίνητρο να εξετάσει αν υπάρχει μόνο μια βέλτιστη λύση για ένα πρόβλημα ομαδοποίησης, ή αν οι διάφορες εναλλακτικές ομαδοποιήσεις είναι εξίσου σημαντικές για το σύνολο δεδομένων.



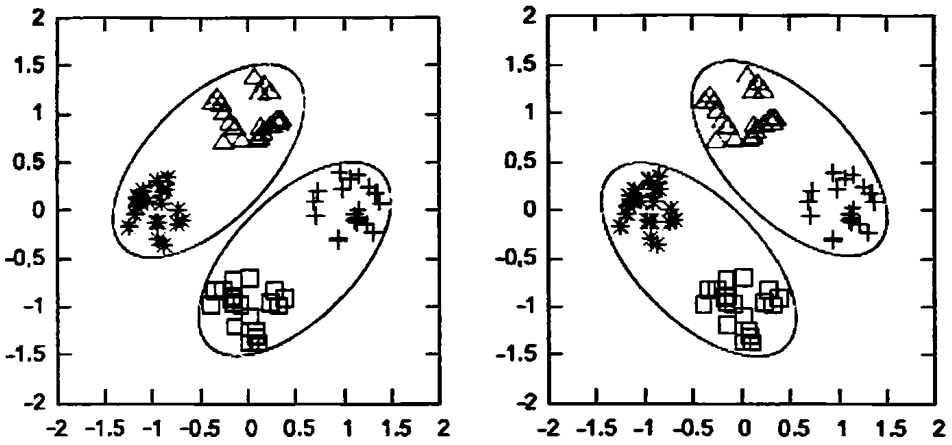
Σχήμα 2.3 Παράδειγμα, πολλαπλή ομαδοποίηση συμπεριφοράς πελατών.

2.2.2. Ορισμός του Προβλήματος

Ας υποθέσουμε ότι θέλουμε να χωρίσουμε ένα διδιάστατο σύνολο δεδομένων σε δύο ομάδες. Υπάρχουν δύο διαφορετικές λύσεις οι οποίες περιγράφουν εναλλακτικά τα δεδομένα, όπως φαίνεται στο Σχήμα 2.4. «Αν δίνεται μια ομαδοποίηση $Clust_1$ με k ομάδες, στόχος είναι να βρεθεί μια δεύτερη ομαδοποίηση $Clust_2$ με k ομάδες, τέτοια ώστε να διαφέρει επαρκώς σε σχέση με την $Clust_1$, ικανοποιώντας παράλληλα τις απαιτήσεις ποιότητας»[8].

Έστω N το σύνολο αντικειμένων, $Clust_i$ μια ομαδοποίηση του συνόλου N και Clusterings το σύνολο όλων των πιθανών ομαδοποιήσεων. Η συνάρτηση Q(Clust) χρησιμοποιείται ως δείκτης ποιότητας μιας ομαδοποίησης και η συνάρτηση $Diss(Clust_i, Clust_j)$ ως μέτρο ανομοιότητας μεταξύ δύο ομαδοποιήσεων. Η πολλαπλή ομαδοποίηση έχει σαν στόχο την εύρεση διαμερισμών $Clust_1,\ldots, Clust_m$ τέτοιων ώστε:

- Η συνάρτηση ποιότητας $Q(Clust_i)$ να έχει υψηλή τιμή για κάθε $i \in \{1,\ldots, m\}$
- Η συνάρτηση ανομοιότητας $Diss(Clust_i, Clust_j)$ να έχει υψηλή τιμή για κάθε $i,j \in \{1,\ldots, m\}$, $i \neq j$



Σχήμα 2.4 Δύο εναλλακτικές ομαδοποιήσεις σε ένα διδιάστατο σύνολο δεδομένων.

Σύμφωνα με τον ορισμό του προβλήματος όπως περιγράφηκε παραπάνω, δημιουργούνται νέα ερωτήματα:

- Πώς μπορεί η ποιότητα μιας ομαδοποίησης να οριστεί αντικειμενικά;
- Ποιος είναι ο κατάλληλος συνδυασμός ποιότητας και ανομοιότητας;
- Οι διάφορες εναλλακτικές λύσεις θα πρέπει να είναι εντελώς διαφορετικές μεταξύ τους για να χαρακτηρίζουν εναλλακτικά τα δεδομένα;

2.3. Προσεγγίσεις Πολλαπλής Ομαδοποίησης

Οι μέθοδοι πολλαπλής ομαδοποίησης μπορούν να κατηγοριοποιηθούν με βάση διάφορους παράγοντες. Μπορεί να παράγουν τα εναλλακτικά αποτελέσματα ταυτόχρονα ή επαναληπτικά. Να εξάγουν μία ή πολλές εναλλακτικές λύσεις. Να ανιχνεύουν ένα σταθερό αριθμό λύσεων ή ο χρήστης να ορίζει τον αριθμό εναλλακτικών που θέλει. Κύριο χαρακτηριστικό της ταξινόμησης που παρουσιάζεται στο[12] αποτελεί ο βασικός χώρος δεδομένων, δημιουργώντας τρεις υποκατηγορίες:

- Προσεγγίσεις που χρησιμοποιούν τον αρχικό χώρο δεδομένων (original data space)

  Προσεγγίσεις που χρησιμοποιούν ορθογώνιο μετασχηματισμό του χώρου δεδομένων (orthogonal data space)
- Προσεγγίσεις που χρησιμοποιούν προβολές σε διαφορετικούς υπο-χώρους (subspace projections)

Στη πρώτη κατηγορία χρησιμοποιείται ο αρχικός χώρος δεδομένων και οι περισσότερες μέθοδοι διαφέρουν στον τρόπο με τον οποίο ορίζουν την αντικειμενική συνάρτηση ομαδοποίησης. Στόχος είναι η βελτιστοποίηση τόσο της ποιότητας όσο και της ανομοιότητας της εναλλακτικής λύσης. Στη δεύτερη κατηγορία, οι περισσότερες μέθοδοι αναζητούν πολλαπλές λύσεις ομαδοποίησης εφαρμόζοντας ορθογώνιο μετασχηματισμό στο χώρο δεδομένων. Ακολουθούν την γενική ιδέα ότι, αν οι μετασχηματισμένοι χώροι είναι ανεξάρτητοι τότε και οι αντίστοιχες λύσεις ομαδοποίησης που προέρχονται από αυτούς τους χώρους, θα είναι επίσης ανεξάρτητες. Τέλος, οι μέθοδοι που ανήκουν στην τρίτη κατηγορία βασίζονται στη λογική ότι διαφορετικές ομάδες είναι «κρυμμένες» σε διαφορετικές προβολές του χώρου δεδομένων. Καλούνται να απαντήσουν στο βασικό ερώτημα: «Πώς είναι δυνατό να βρεθούν ομάδες σε αυθαίρετες προβολές των δεδομένων;»

2.4. Ομαδοποίηση στον Αρχικό Χώρο Δεδομένων

Παρουσιάζουμε στη συνέχεια, τις μεθόδους πολλαπλής ομαδοποίησης όπως ταξινομούνται στον αρχικό χώρο δεδομένων και μερικά αντιπροσωπευτικά παραδείγματα των προσεγγίσεων.

2.4.1. Ομαδοποίηση Χωρίς Πρότερη Γνώση

Η προσέγγιση αυτή αναφέρεται στη δημιουργία πολλαπλών λύσεων ομαδοποίησης και δεν προϋποθέτει την ύπαρξη κάποιας αρχικής ομαδοποίησης (σχήμα 2.5a). Ο αλγόριθμος Meta Clustering χρησιμοποιεί την λογική της συγκεκριμένης προσέγγισης.

Ο όρος Meta Clustering[10] αναφέρεται στη διαδικασία κατά την οποία παράγονται πολλές εναλλακτικές ομαδοποιήσεις και ο χρήστης έχει τη δυνατότητα να διαλέξει τις χρήσιμες γι αυτόν. Η μέθοδος αναλύεται σε τρία βασικά βήματα:

- Δημιουργία πολλών καλών και διαφορετικών (base level) ομαδοποιήσεων για ένα σύνολο δεδομένων.
- Υπολογισμός ομοιότητας των base level λύσεων που δημιουργήθηκαν στο πρώτο βήμα, έτσι ώστε παρόμοιες λύσεις να μπορούν να ομαδοποιηθούν.
- Οργάνωση των λύσεων σε ένα «μετά-επίπεδο» είτε χρησιμοποιώντας ομαδοποίηση είτε με προβολή σε χαμηλότερη διάσταση, και παρουσίαση των αποτελεσμάτων στον χρήστη.

Για τη δημιουργία των λύσεων εφαρμόζονται δύο προσεγγίσεις. Σύμφωνα με την πρώτη προσέγγιση χρησιμοποιείται ο αλγόριθμος k-means με τυχαία αρχικοποίηση, αφού με αυτόν τον τρόπο παράγονται κάθε φορά διαφορετικά αποτελέσματα. Σύμφωνα με τη δεύτερη προσέγγιση εφαρμόζονται διαφορετικά βάρη στα διανύσματα χαρακτηριστικών πριν την ομαδοποίηση με τον k-means αλγόριθμο, έτσι ώστε να δοθεί έμφαση σε διαφορετικές πτυχές των δεδομένων.

Η μέτρηση της ανομοιότητας των base-level ομαδοποιήσεων σχετίζεται με τον δείκτη Rand Index[16] και παίρνει την παρακάτω μορφή:

$$ClusterDifference = \frac{\Sigma_{i<j} I_{ij}}{N(N-1)/2} \qquad \text{Εξ.21}$$

Η ποσότητα I_{ij} παίρνει την τιμή 1 αν τα αντικείμενα i και j ανήκουν στην ίδια ομάδα για την μια ομαδοποίηση και σε διαφορετική ομάδα για την εναλλακτική ομαδοποίηση, ενώ σε διαφορετική περίπτωση παίρνει την τιμή 0. Η ποσότητα N αντιπροσωπεύει τον συνολικό αριθμό των αντικειμένων.

Βασικό πλεονέκτημα της μεθόδου είναι η ανίχνευση πολλαπλών λύσεων ταυτόχρονα καθώς επίσης και η δυνατότητα επιλογής που προσφέρει στο χρήστη. Το μειονέκτημα ωστόσο του αλγορίθμου, βρίσκεται στη «τυφλή» και μη ελεγχόμενη δημιουργία

λύσεων. Η συγκεκριμένη λογική είναι συχνά αναποτελεσματική και είναι δυνατό να παράγει πολλές, όμοιες ομαδοποιήσεις.

2.4.2. Ομαδοποίηση Χρησιμοποιώντας Πρότερη Γνώση

Σύμφωνα με αυτή τη προσέγγιση, ο υπολογισμός εναλλακτικής λύσης βασίζεται σε υπάρχουσα ομαδοποίηση και εφαρμόζεται επαναληπτικά. Το βασικό χαρακτηριστικό της βρίσκεται στο να ληφθεί υπόψη οποιαδήποτε γνώση έχει ο χρήστης σχετικά με τα δεδομένα, και να χρησιμοποιηθεί για να παραχθεί μια νέα εξίσου ενδιαφέρουσα εναλλακτική λύση (Σχήμα 2.5b). Την παραπάνω λογική εφαρμόζει ο αλγόριθμος COALA[8]. Αντικείμενο είναι η βελτιστοποίηση μιας αντικειμενικής συνάρτησης που συνδυάζει τόσο την απαίτηση για ανομοιότητα όσο και για ποιότητα της παραγόμενης εναλλακτικής λύσης.

Πιο αναλυτικά ο αλγόριθμος COALA χρησιμοποιεί έναν συσσωρευτικό (agglomerative), ιεραρχικό αλγόριθμο ομαδοποίησης και κριτήριο ομοιότητας τη μέση απόσταση (average-link) ανάμεσα σε ζεύγη ομάδων. Η τεχνική αποτελείται από δύο βήματα. Στο πρώτο βήμα χρησιμοποιείται η αρχική ομαδοποίηση για να δημιουργηθούν οι περιορισμοί (cannot-link constraints), οι οποίοι απαγορεύουν τη σύνδεση αντικειμένων τα οποία ανήκουν στην ίδια ομάδα. Στο δεύτερο βήμα χρησιμοποιούνται οι περιορισμοί για να παραχθεί η εναλλακτική ομαδοποίηση. Ο αλγόριθμος ξεκινά με κάθε αντικείμενο να αποτελεί μια ξεχωριστή ομάδα και στη συνέχεια επαναληπτικά, συγχωνεύει τα ζεύγη λύσεων. Για τη συγχώνευση η μέθοδος χρησιμοποιεί ένα κατώφλι w και αποφασίζει σε κάθε επανάληψη αν θα γίνει συγχώνευση βάσει της ποιότητας ή βάσει της ανομοιότητας.

Για τον υπολογισμό της ανομοιότητας μεταξύ δύο ομαδοποιήσεων Clust_i και Clust_j ο αλγόριθμος χρησιμοποιεί τον δείκτη Jaccard Index[13]. Βασίζεται στην καταμέτρηση ζευγών αντικειμένων που ανατίθενται σε κάθε ομάδα μεταξύ των δύο ομαδοποιήσεων:

$$J(\text{Clust}_i, \text{Clust}_j) = \frac{M_{11}}{M_{11} + M_{01} + M_{10}} \qquad \text{Εξ. 2.2}$$

- M_{11}, ο αριθμός των ζευγών των αντικειμένων που ανήκουν στην ίδια ομάδα και για τις δύο ομαδοποιήσεις $Clust_i$ και $Clust_j$.
- M_{10}, ο αριθμός των ζευγών των αντικειμένων που ανήκουν στην ίδια ομάδα στην $Clust_i$ και σε διαφορετικές ομάδες στην $Clust_j$.
- M_{01},ο αριθμός των ζευγών των αντικειμένων που ανήκουν διαφορετικές ομάδες στην $Clust_i$ και στην ίδια ομάδα στην $Clust_j$.

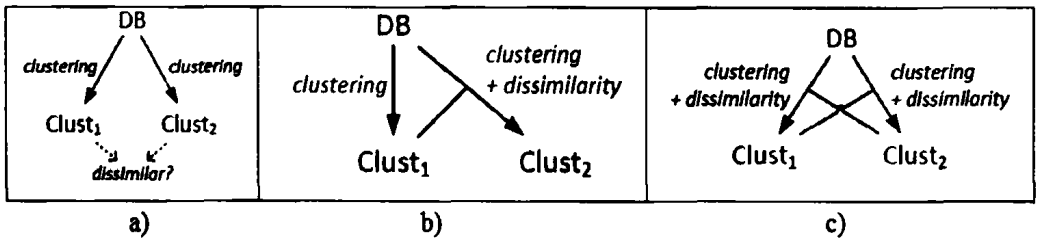
Ο δείκτης Jaccard επιστρέφει μια τιμή μεταξύ 0 και 1, όπου υψηλότερη τιμή υποδεικνύει μεγάλη ανομοιότητα.

Για την μέτρηση της ποιότητας χρησιμοποιείται ο δείκτης Dunn Index[14]. Για μια ομαδοποίηση $C = \{c_1,\ldots,c_k\}$ ισχύει ότι:

$$DI(C) = \frac{\min_{i \neq j} \left\{ \delta(c_i, c_j) \right\}}{\max_{1 \leq l \leq k} \left\{ \Delta_{(c_l)} \right\}} \qquad \text{Εξ. 2.3}$$

Όπου $\delta(c_i,c_j)$ είναι η απόσταση των κέντρων των ομάδων c_i και c_j και η ποσότητα $\Delta(c_l)$ μετρά τη διάμετρο της ομάδας l (μέγιστη απόσταση μεταξύ δύο οποιονδήποτε αντικειμένων της ομάδας). Υψηλές τιμές του δείκτη υποδηλώνουν υψηλότερη ποιότητα ομαδοποίησης.

Πρόκειται για μια ευρετική προσέγγιση εναλλακτικής ομαδοποίησης. Η λύση που παράγεται συνήθως χαρακτηρίζεται από υψηλή ποιότητα ομαδοποίησης καθώς επίσης και ανομοιότητα σε σχέση με την υπάρχουσα λύση. Μειονέκτημα, ωστόσο αποτελεί το γεγονός ότι υπολογίζεται μόνο μια εναλλακτική ομαδοποίηση χωρίς να είναι δυνατή η εξαγωγή πολλαπλών ομαδοποιήσεων.



Σχήμα 2.5 Πολλαπλή ομαδοποίηση στον αρχικό χώρο δεδομένων.

2.4.3. Ταυτόχρονη Ομαδοποίηση με Πολλαπλές Λύσεις

Η προσέγγιση αυτή έχει σαν στόχο την *ταυτόχρονη* δημιουργία πολλών διαφορετικών διαμερισμών $Clust_1,\ldots,Clust_m$ (Σχήμα 2.5c) οι οποίες χαρακτηρίζονται τόσο από ποιότητα όσο και ανομοιότητα ομαδοποίησης. Χρησιμοποιεί μια συνδυαστική αντικειμενική συνάρτηση, βελτιστοποιώντας τόσο την ποιότητα ομαδοποίησης όσο και την ανομοιότητα μεταξύ των εναλλακτικών λύσεων. Δεδομένου ότι δεν χρειάζεται προηγούμενη λύση ως είσοδο, η παραπάνω λογική είναι πιο κοντά στη φύση της ομαδοποίησης, που αποτελεί ένα πρόβλημα μάθησης χωρίς επίβλεψη.

Ο αλγόριθμος Decorrelated k-means[9], υπολογίζει δύο εναλλακτικές ομαδοποιήσεις ταυτόχρονα, χρησιμοποιώντας μια αντικειμενική συνάρτηση που συνδυάζει το τετραγωνικό σφάλμα ομαδοποίησης και τη συσχέτιση (correlation) μεταξύ των δύο ομαδοποιήσεων. Εισάγεται ένα μέτρο συσχέτισης μεταξύ των ομαδοποιήσεων και προτείνεται ένας αλγόριθμος με αντικειμενική συνάρτηση παρόμοια με τον k-means.

Ας υποθέσουμε ότι έχουμε ένα σύνολο από σημεία $X = \{x_1,\ldots,x_n\} \in R^d$ και προσπαθούμε να ανακαλύψουμε δύο διαφορετικές ομαδοποιήσεις. Ειδικότερα θέλουμε να διαμερίσουμε τα δεδομένα σε k ομάδες για την πρώτη λύση, και *m* ομάδες για τη δεύτερη λύση. Τα διανύσματα $r_1,\ldots,r_k$ αναπαριστούν τις ομάδες C_1, $\ldots,C_k$ της πρώτης λύσης, ενώ αντίστοιχα τα διανύσματα $s_1,\ldots,s_m$ τις ομάδες D_1, $\ldots,D_m$ της δεύτερης λύσης. Τα κέντρα των ομάδων συμβολίζονται με $\alpha_1,\ldots,\alpha_k$ για την πρώτη και $\beta_1,\ldots,\beta_m$ για την δεύτερη (Σχήμα 2.6). Για τον υπολογισμό των λύσεων προτείνεται η ελαχιστοποίηση της ακόλουθης αντικειμενικής συνάρτησης:

$$G(r_1,\ldots,r_k,s_1,\ldots,s_m) = \sum_i \sum_{x\in C_i} \| x - r_i \|^2 + \sum_j \sum_{x\in D_j} \| x - s_j \|^2 + \lambda \sum_{i,j} (\beta_j^T \bullet r_i)^2 + \lambda \sum_{i,j} (a_i^T \bullet s_j)^2 \quad \text{Εξ. 2.4}$$

Οι δύο πρώτοι όροι αντιστοιχούν στο τετραγωνικό σφάλμα του αλγόριθμου k-means, με την κρίσιμη διαφορά ότι το διάνυσμα αναπαράστασης μπορεί να μην είναι το μέσο της ομάδας. Οι δύο τελευταίοι όροι αποτελούν ένα νέο μέτρο συσχέτισης (deccorelation) μεταξύ των ομαδοποιήσεων. Η ποσότητα $\lambda > 0$ είναι μια παράμετρος που ορίζει το trade-off του σφάλματος ομαδοποίησης με τη συσχέτιση των λύσεων. Διαισθητικά, οι δύο τελευταίοι όροι υπαγορεύουν ότι αν τα διανύσματα

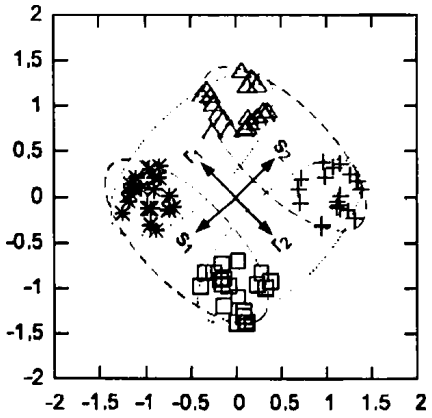
αναπαράστασης των ομαδοποιήσεων είναι κάθετα μεταξύ τους, τότε οι ταμπέλες που παράγονται με χρήση του κοντινότερου γείτονα, είναι ανεξάρτητες.

Η προσέγγιση αρχικοποιεί μία από τις δύο ομαδοποιήσεις χρησιμοποιώντας τον k-means με k ομάδες και την άλλη με τυχαίο τρόπο. Τα κέντρα των ομάδων συμβολίζονται με $\alpha_1,\ldots,\alpha_k$ για την πρώτη και $\beta_1,\ldots,\beta_m$ για την δεύτερη. Για την βελτιστοποίηση της αντικειμενικής συνάρτησης επαναληπτικά, οι ποσότητες r_i, s_j θα πρέπει να ενημερώνονται σε κάθε βήμα με τον ακόλουθο τρόπο:

$$r_i = (I - \xi_i VQ(I + \xi_i \Sigma)^{-1} Q' V')a_i \qquad \text{Εξ.2.5}$$

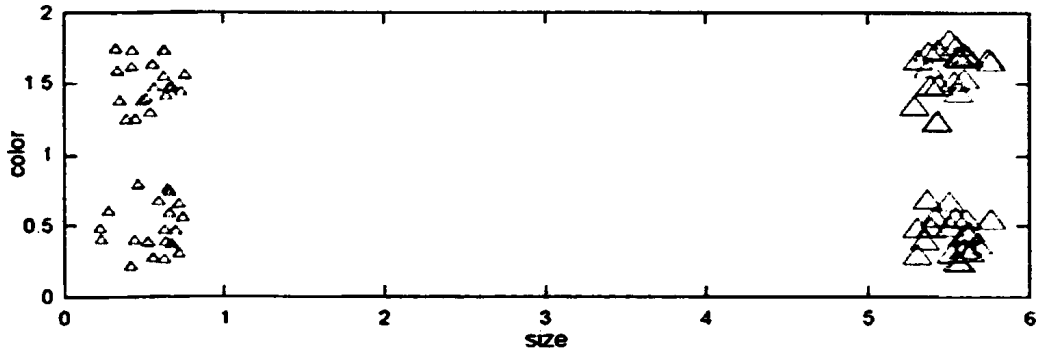
$$s_j = (I - \zeta_j MU(I + \zeta_j \Lambda)^{-1} U' M')\beta_j \qquad \text{Εξ. 2.6}$$

όπου $\xi_i = \frac{\lambda}{\Sigma_j n_{ij}}$, $V = [\beta_1,\ldots,\beta_m]$, $V'V=Q\Lambda Q'$ είναι η αποσύνθεση ιδιοτιμών για την πρώτη εξίσωση και αντίστοιχα $\zeta_j = \frac{\lambda}{\Sigma_i n_{ij}}$, $M = [\alpha_1,\ldots,\alpha_k]$, $M'M = U\Lambda U'$ για την δεύτερη εξίσωση και n_{ij} είναι ο αριθμός των αντικειμένων που ανήκουν στην ομάδα C_i και στην D_j. Αξίζει να τονίσουμε ότι η μέθοδος αυτή επιτρέπει την παραμετροποίηση του επιθυμητού αριθμού ομαδοποιήσεων ($T \geq 2$ εναλλακτικές ομαδοποιήσεις μπορούν να παραχθούν). Δεν προϋποθέτει την ύπαρξη αρχικής λύσης και υπολογίζει ταυτόχρονα τις εναλλακτικές ομάδες.



Σχήμα 2.6 Decorrelated k-means, αναπαράσταση ομάδων με διανύσματα.

Οι μέθοδοι που περιγράφηκαν παραπάνω έχουν σχεδιαστεί για συγκεκριμένους αλγορίθμους. Η ομαδοποίηση εφαρμόζεται στον αρχικό χώρο δεδομένων, ελέγχοντας την ανομοιότητα λύσεων και δημιουργούνται εναλλακτικές ομαδοποιήσεις σύμφωνα με διαφορετικά χαρακτηριστικά. Υπάρχουν ωστόσο περιπτώσεις, για τις οποίες η προσέγγιση στον αρχικό χώρο δεδομένων δεν παράγει «καλές» εναλλακτικές λύσεις. Για παράδειγμα στο Σχήμα 2.7, ας υποθέσουμε ότι θέλουμε να ομαδοποιήσουμε τα δεδομένα σύμφωνα με το χρώμα. Η ομαδοποίηση ενδέχεται να αποτύχει, αφού η ανίχνευση των ομάδων που δημιουργούνται με το συγκεκριμένο κριτήριο είναι αρκετά δύσκολη στον αρχικό χώρο δεδομένων. Για το λόγο αυτό αναπτύχθηκαν και άλλες προσεγγίσεις που αντιμετωπίζουν αυτά τα προβλήματα. Όπως αναφέρθηκε, μπορούμε να υπολογίσουμε πολλαπλές λύσεις εφαρμόζοντας ορθογώνιους μετασχηματισμούς στο χώρο δεδομένων. Με αυτή την προσέγγιση οποιοσδήποτε αλγόριθμος μπορεί να εφαρμοστεί σε ένα σύνολο δεδομένων που έχει υποστεί μετασχηματισμό. Άλλες προσεγγίσεις προτείνουν διαφορετικές προβολές στον υπόχωρο, όπου κάθε αντικείμενο ομαδοποιείται διαφορετικά σε κάθε προβολή. Ωστόσο οι μορφές αυτές ξεφεύγουν από το αντικείμενο εργασίας και δεν θα αναλυθούν περαιτέρω.



Σχήμα 2.7 Ομαδοποίηση συνόλου δεδομένων με βάση το χρώμα.

ΚΕΦΑΛΑΙΟ 3. ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΑ ΟΜΑΔΟΠΟΙΗΣΗΣ

Το κεφάλαιο αυτό περιλαμβάνει την πλήρη και αναλυτική περιγραφή του αλγορίθμου που υλοποιήθηκε. Αρχικά, περιγράφεται ο αλγόριθμος global k-means καθώς και ο απλός k-means. Στη συνέχεια παρουσιάζεται η μεθοδολογία ομαδοποίησης και οι λεπτομέρειες υλοποίησης.

3.1. Αλγόριθμος k-means

Ο αλγόριθμος k-means[4] αποτελεί έναν από τους απλούστερους αλγορίθμους ομαδοποίησης και ανήκει στην κατηγορία των διαιρετικών αλγορίθμων. Αρχικά, ο χρήστης ορίζει έναν συγκεκριμένο αριθμό ομάδων (k clusters). Βασικός στόχος του αλγορίθμου, είναι ο ορισμός k κέντρων, ένα για κάθε μία από τις k ομάδες. Κάθε σημείο στη συνέχεια, ανατίθεται στο κοντινότερό του κέντρο, και το σύνολο των σημείων που έχει ανατεθεί σε μία ομάδα συγκροτεί την ομάδα αυτή. Στη συνέχεια, το κέντρο κάθε ομάδας ενημερώνεται και προσδιορίζεται με βάση τα νέα σημεία που έχουν ανατεθεί στην ομάδα. Η διαδικασία της ανάθεσης και επαναπροσδιορισμού των κέντρων επαναλαμβάνεται ωσότου τα σημεία δεν αλλάζουν ομάδες ή με άλλα λόγια, τα κέντρα των ομάδων παραμένουν αμετάβλητα.

Συνοπτικά, τα βασικά βήματα του αλγορίθμου παρουσιάζονται παρακάτω:

Αλγόριθμος 1: k-means

1. Επιλογή του αριθμού ομάδων.
2. Τυχαία δημιουργία k ομάδων και ορισμός των κέντρων των ομάδων (αρχικοποίηση).
3. Ανάθεση του κάθε σημείου στο κέντρο της κοντινότερης ομάδας.
4. Υπολογισμός των νέων κέντρων των ομάδων.
5. Επανάληψη των βημάτων 3 και 4 ωσότου να μην αλλάζουν τα κέντρα.

Αν υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων $X = \{x_1,...,x_N\}$, $x_i \in R^d$, το πρόβλημα των K ομάδων αποσκοπεί στον διαχωρισμό ενός συνόλου δεδομένων σε K διαφορετικές ομάδες $C_1,...C_K$ έτσι ώστε να βελτιστοποιείται κάποιο κριτήριο ομαδοποίησης. Για την μέτρηση της ποιότητας ομαδοποίησης χρησιμοποιείται το τετραγωνικό σφάλμα ομαδοποίησης (Sum of Squared Errors, SSE). Υπολογίζεται το σφάλμα κάθε σημείου x_i, δηλαδή η Ευκλείδεια απόστασή του από το κέντρο m_k της ομάδας C_k και στη συνέχεια υπολογίζεται το άθροισμα των τετραγώνων όλων των σφαλμάτων για κάθε ομάδα.

$$E(m_1,...,m_K) = \sum_{i=1}^{N}\sum_{k=1}^{K} I(x_i \in C_k) \| x_i - m_k \|^2 \quad \text{Εξ 3.1}$$

όπου I(x) = 1 αν x είναι αληθές και 0 όταν είναι ψευδές. Σκοπός του αλγορίθμου είναι να ελαχιστοποιεί τη συνάρτηση τετραγωνικού σφάλματος. Με άλλα λόγια, αν έχουν παραχθεί δύο διαφορετικές ομαδοποιήσεις με τον αλγόριθμο, τότε επιλέγεται η ομαδοποίηση με το χαμηλότερο τετραγωνικό σφάλμα.

Αξίζει να σημειωθεί ότι ο αλγόριθμος k-means, χάριν στην απλότητα και την ταχύτητά του έχει χρησιμοποιηθεί σε πολλές εφαρμογές ομαδοποίησης. Βασικό μειονέκτημα της μεθόδου ωστόσο, αποτελεί η εξάρτηση από την αρχικοποίηση των κέντρων των ομάδων. Όσον αφορά στην απόδοσή του ο k-means δεν εγγυάται ότι θα συγκλίνει στη καλύτερη λύση ως προς το σφάλμα ομαδοποίησης. Για το λόγο αυτό, λύσεις που προσεγγίζουν το βέλτιστο μπορούν να επιτευχθούν μόνο μετά από πολλαπλές αρχικοποιήσεις των κέντρων.

3.2. Αλγόριθμος global k-means

Στη μέθοδο που υλοποιήθηκε στην παρούσα εργασία χρησιμοποιήθηκε σαν βάση ο αλγόριθμος global k-means. Ο global k-means[15] είναι ένας ντετερμινιστικός αλγόριθμος, ο οποίος χρησιμοποιεί τον k-means σαν εσωτερική διαδικασία και είναι βασισμένος στην ελαχιστοποίηση του τετραγωνικού σφάλματος ομαδοποίησης. Σε αντίθεση με τον k-means, δεν αρχικοποιεί τυχαία τα κέντρα των ομάδων, αλλά λειτουργεί αυξητικά, προσπαθώντας να εισάγει βέλτιστα μια ομάδα σε κάθε στάδιο. Η λογική που προτείνεται βασίζεται στην ιδέα, ότι η επίλυση ενός προβλήματος ομαδοποίησης σε K ομάδες επιτυγχάνεται δια μέσου της τοπικής αναζήτησης, χρησιμοποιώντας τον απλό k-means αλγόριθμο.

Ο αλγόριθμος global k-means ακολουθεί την ακόλουθη διαδικασία: Αρχικά, θεωρούμε μια μόνο ομάδα, όπου βέλτιστη λύση ορίζεται το κέντρο του συνόλου X (k=1). Για να προστεθεί μια νέα ομάδα k=2, εκτελείται ο αλγόριθμος k-means N φορές (όπου N είναι το πλήθος των δεδομένων). Το πρώτο κέντρο της ομάδας τοποθετείται πάντα στην βέλτιστη θέση για k=1, ενώ το δεύτερο κέντρο για την n-οστή εκτέλεση αρχικά τοποθετείται στη θέση του στοιχείου x_n (n=1,...,N). Η καλύτερη λύση (αυτή με το μικρότερο σφάλμα ομαδοποίησης), μετά από N εκτελέσεις του αλγορίθμου k-means, είναι και η αποδεκτή για το πρόβλημα των k=2 ομάδων. Χρησιμοποιώντας επαναληπτικά την παραπάνω μέθοδο, μπορούμε να υπολογίσουμε τις βέλτιστες λύσεις για κάθε k=2, ...,K.

Γενικά, αν θέλουμε να λύσουμε ένα πρόβλημα με K ομάδες εφαρμόζουμε τοπική αναζήτηση με τον απλό k-means αλγόριθμο, ξεκινώντας από τις ακόλουθες αρχικές θέσεις:

- τα k-1 κέντρα είναι τοποθετημένα στις βέλτιστες θέσεις τους, που προέκυψαν λύνοντας το k-1 πρόβλημα ομαδοποίησης.
- το k-οστό κέντρο τοποθετείται στην κατάλληλη θέση μετά από N εκτελέσεις.

Η μέθοδος βασίζεται στην ιδέα ότι μια καλή λύση για το k πρόβλημα ομαδοποίησης είναι εφικτή διαμέσου της τοπικής αναζήτησης από το (k-1) πρόβλημα ομαδοποίησης, εάν το επιπρόσθετο k-οστό κέντρο τοποθετηθεί στην κατάλληλη θέση

του συνόλου δεδομένων. Είναι επίσης λογικός ο περιορισμός των πιθανών θέσεων του k οστού κέντρου στα διαθέσιμα σημεία του συνόλου X.

Ένα μεγάλο πλεονέκτημα του αλγορίθμου είναι ότι παρέχει τις λύσεις για όλα τα k-προβλήματα ομαδοποίησης με $k \leq K$. Το χαρακτηριστικό αυτό μπορεί να φανεί χρήσιμο σε πολλές εφαρμογές όπου ο στόχος είναι επίσης να ανακαλυφθεί ο «σωστός» αριθμός ομάδων. Για να επιτευχθεί κάτι τέτοιο πρέπει να λυθεί το πρόβλημα των k ομάδων για διαφορετικό αριθμό ομάδων και στη συνέχεια να εφαρμοστούν διάφορα κριτήρια για την επιλογή της καταλληλότερης τιμής του k. Στην περίπτωση αυτή ο αλγόριθμος global k-means παρέχει όλες τις ενδιάμεσες λύσεις για το k πρόβλημα ομαδοποίησης και δεν απαιτεί πρόσθετους υπολογισμούς.

Όσον αφορά στον υπολογιστικό κόστος η μέθοδος απαιτεί N εκτελέσεις του k-means αλγορίθμου, μια για κάθε υποψήφιο κέντρο της νέας ομάδας. Ανάλογα με τους διαθέσιμους πόρους και τις ποσότητες N και K, ο αλγόριθμος μπορεί να αποτελέσει μια ελκυστική μέθοδο, αφού όπως δείχνουν τα πειραματικά αποτελέσματα, η απόδοσή της είναι εξαιρετική. Προκειμένου να μειωθεί το κόστος, χωρίς σημαντική φθορά στην ποιότητα λύσης έχει προταθεί ο fast global k-means αλγόριθμος [15]. Ο αλγόριθμος απαιτεί μια μόνο εκτέλεση του k-means για κάθε τιμή k=1,...,K του αριθμού των ομάδων, επιταχύνοντας με αυτόν τον τρόπο τον χρόνο εκτέλεσης του αλγορίθμου, ωστόσο τα αποτελέσματα μπορεί να είναι χειρότερα σε σχέση με τον τυπικό global k-means.

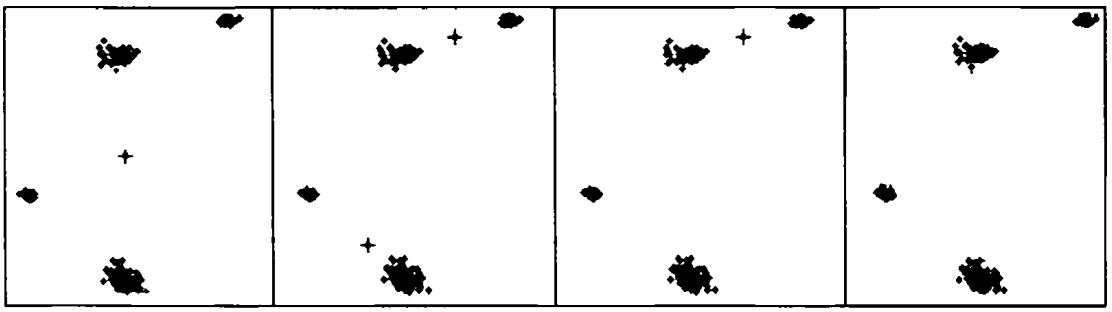
Η μέθοδος global k-means είναι ανεξάρτητη από την αρχικοποίηση των κέντρων των ομάδων και παράγει την ίδια λύση σε κάθε εφαρμογή της. Ο απλός k-means αντίθετα, εξαρτάται από τις αρχικές συνθήκες αρχικοποίησης και επιτυγχάνει καλές λύσεις μετά από πολλαπλές αρχικοποιήσεις. Αξίζει επίσης να αναφέρουμε ότι ο αλγόριθμος global k-means παράγει πολύ καλές λύσεις, και είναι τουλάχιστον όσο αποτελεσματικός είναι ο απλός k-means με πολλαπλές τυχαίες αρχικοποιήσεις

Συνοπτικά, τα βασικά βήματα του αλγορίθμου παρουσιάζονται παρακάτω:

Αλγόριθμος 2: global k-means

1. Για k=1 υπάρχει μια ομάδα με βέλτιστο κέντρο, το κέντρο όλων των δεδομένων.
2. Για να προστεθεί μια νέα ομάδα k, τα k-1 κέντρα αρχικοποιούνται στις βέλτιστες θέσεις τους που προκύπτουν από τη λύση για k-1 ομάδες.
    a. Εκτελείται ο αλγόριθμος k-means N φορές. Κατά την n εκτέλεση το k-οστό κέντρο τοποθετείται στη θέση του στοιχείου x_n (n = 1…N).
    b. Η καλύτερη λύση που προκύπτει μετά από τις N εκτελέσεις του αλγορίθμου k-means είναι και η αποδεκτή για το πρόβλημα με k ομάδες.
3. k= k+1, επανέλαβε το βήμα 2 μέχρι k=K.

Στο σχήμα που ακολουθεί (Σχήμα 3.1) βλέπουμε την ομαδοποίηση σε ένα τεχνητό διδιάστατο σύνολο δεδομένων, μετά από 4 επαναλήψεις του αλγορίθμου global k-means.



Σχήμα 3.1 Συμπεριφορά του global k-means σε 4 επαναλήψεις.

3.3. Μεθοδολογία Πολλαπλής Ομαδοποίησης

Η μεθοδολογία που προτείνουμε για το πρόβλημα της πολλαπλής ομαδοποίησης βασίζεται στον αλγόριθμο global k-means και επιχειρεί να επεκτείνει τη λειτουργία του. Εκμεταλλεύεται το βασικό πλεονέκτημα του αλγορίθμου, ότι παρέχει δηλαδή όλες τις λύσεις για όλα τα k υπό-προβλήματα ομαδοποίησης με $k \leq K$. Όπως περιγράφηκε προηγουμένως, για να προστεθεί μια νέα ομάδα, ο αλγόριθμος k-means εκτελείται N φορές. Μετά τις N εκτελέσεις του αλγορίθμου λαμβάνουμε N λύσεις και

κρατάμε την καλύτερη, αυτή με το μικρότερο σφάλμα ομαδοποίησης. Εκτός λοιπόν από την βέλτιστη λύση εκμεταλλευόμαστε και μερικές λιγότερο «καλές» λύσεις. Για το πρόβλημα ομαδοποίησης κάθε τάξης η μέθοδος αξιοποιεί έναν αριθμό λύσεων που προέρχονται από την τοπική αναζήτηση χρησιμοποιώντας τον παραδοσιακό αλγόριθμο k-means. Οι λύσεις αυτές συντελούν τελικά στην εύρεση πολλαπλών εναλλακτικών ομαδοποιήσεων.

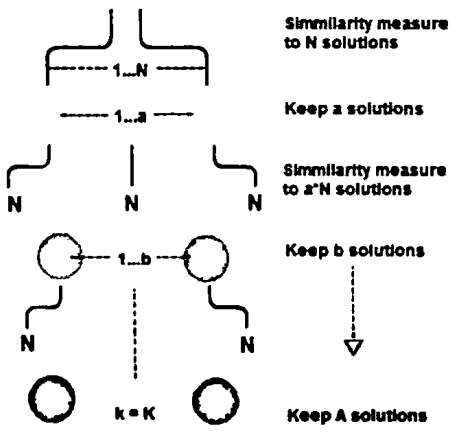
Η προτεινόμενη μεθοδολογία τροποποιεί τον αλγόριθμο global k-means και ακολουθεί την παρακάτω διαδικασία:

Ξεκινώντας με k=1, υπάρχει μόνο μια λύση, η οποία είναι το κέντρο των ομάδων όπως συμβαίνει και στον global k-means αλγόριθμο. Για να λύσει το πρόβλημα με δύο ομάδες, αρχικά ο k-means εκτελείται N φορές. Από τον N αριθμό ομαδοποιήσεων που παράγονται, επιλέγεται ένας αριθμός λύσεων S για k=2, που περιλαμβάνει κάποιες επιπλέον λύσεις εκτός από την βέλτιστη. Η επιλογή των επιπλέον λύσεων πραγματοποιείται χρησιμοποιώντας μια συνάρτηση ομοιότητας ώστε να διατηρούνται ανόμοιες λύσεις. Στη συνέχεια εκτελείται πάλι ο αλγόριθμος k-means N φορές, για κάθε μια από τις επιλεγμένες λύσεις, και τελικά παράγεται S·N αριθμός ομαδοποιήσεων. Εξετάζεται πάλι πόσο διαφέρουν οι νέες λύσεις μεταξύ τους και παράγεται, όπως προηγουμένως, ένας αριθμός αποδεκτών ομαδοποιήσεων A. Η καλύτερη λύση (αυτή με το μικρότερο σφάλμα ομαδοποίησης) καθώς επίσης και οι λιγότερο «καλές» που επιλέχθηκαν εφαρμόζοντας τη συνάρτηση ομοιότητας, αποτελούν τις εναλλακτικές λύσεις για το πρόβλημα των 3 ομάδων.

Χρησιμοποιώντας επαναληπτικά την παραπάνω μέθοδο και εκμεταλλευόμενοι τις λύσεις που παράγονται σε κάθε προσθήκη μιας επιπλέον ομάδας, μπορούμε να υπολογίσουμε πολλαπλές εναλλακτικές ομαδοποιήσεις για κάθε k=2,…,K διαμερισμό. Ο τροποποιημένος αλγόριθμος Multiple Clusters Global k-Means για το πρόβλημα με k ομάδες παρουσιάζεται παρακάτω:

Αλγόριθμος 3: Multiple Clusters Global K-Means (MCGKM)

1. Για k=1 υπάρχει μια ομάδα με βέλτιστο κέντρο, το κέντρο όλων των δεδομένων.
2. Για k=2 εκτελείται ο αλγόριθμος k-means N φορές, δημιουργώντας N λύσεις. Κατά την n εκτέλεση το κέντρο τοποθετείται στη θέση του στοιχείου x_n(n=1…N). Από τις N λύσεις, επιλέγουμε S αριθμό λύσεων εφαρμόζοντας μια συνάρτηση ομοιότητας. Οι λύσεις S είναι οι αποδεκτές εναλλακτικές λύσεις για το πρόβλημα με 2 ομάδες.
3. Για να προστεθεί μια νέα ομάδα k>2, τα k-1 κέντρα αρχικοποιούνται στις βέλτιστες θέσεις τους που προκύπτουν από τη λύση για k-1 ομάδες για κάθε μια από τις S λύσεις.
    a. Εκτελείται ο αλγόριθμος k-means N φορές, για κάθε μια από τις S λύσεις, δημιουργώντας N·|S| αριθμό ομαδοποιήσεων.
    b. Από τον N·|S| αριθμό λύσεων, επιλέγουμε A αριθμό ομαδοποιήσεων εφαρμόζοντας μια συνάρτηση ομοιότητας.
    c. Οι λύσεις A είναι οι αποδεκτές εναλλακτικές λύσεις για το πρόβλημα με k ομάδες.
    d. S = A.
4. k=k+1, επανέλαβε το βήμα 3 μέχρι k=K.



Σχήμα 3.2 Συμπεριφορά του αλγορίθμου MCGKM.

Η προτεινόμενη μεθοδολογία είναι ανεξάρτητη των αρχικών συνθηκών, αφού χρησιμοποιεί στην βάση της τον αλγόριθμο global k-means. Το υπολογιστικό κόστος ωστόσο της μεθόδου, τείνει να αυξάνεται κατά πολύ σε σχέση με τον global k-means, αφού για κάθε προσθήκη μιας νέας ομάδας ο αλγόριθμος k-means εκτελείται επιπλέον $N \cdot |S|$ φορές αντί για N φορές.

3.4. Λεπτομέρειες Υλοποίησης

3.4.1. Ο δείκτης Rand Index

Για την αξιολόγηση των ομαδοποιήσεων χρησιμοποιείται ο δείκτης Rand Index[16]. Πιο συγκεκριμένα, έστω ότι θέλουμε να συγκρίνουμε δύο ομαδοποιήσεις $Clust_1$ και $Clust_2$ ενός συνόλου X, που αποτελείται από n αντικείμενα. Για κάθε ζεύγος αντικειμένων ορίζουμε [12][17]:

- a, το πλήθος των ζευγών των αντικειμένων που είναι στην ίδια ομάδα στην $Clust_1$ και στην $Clust_2$.
- b, το πλήθος των ζευγών των αντικειμένων που είναι σε διαφορετική ομάδα και στην $Clust_1$ και στην $Clust_2$.
- c, το πλήθος των ζευγών των αντικειμένων που είναι στην ίδια ομάδα στην $Clust_1$ και σε διαφορετικές ομάδες στην $Clust_2$.
- d, το πλήθος των ζευγών των αντικειμένων που είναι σε διαφορετικές ομάδες στην $Clust_1$ και στην ίδια ομάδα στην $Clust_2$.

Σύμφωνα με τα παραπάνω, η συνάρτηση Rand Index παίρνει την ακόλουθη μορφή:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \qquad \text{Εξ. 3.2}$$

Ο δείκτης παίρνει τιμές στο διάστημα [0,1]. Τιμή 1 του R δηλώνει απόλυτη συμφωνία των δύο ομαδοποιήσεων ενώ τιμή κοντά στο 0 υποδηλώνει ασυμφωνία των ομαδοποιήσεων.

3.4.2. Υπολογισμός Ομοιότητας Λύσεων

Αρχικά, η επιλογή των λύσεων πραγματοποιήθηκε με βάση το τετραγωνικό σφάλμα. Όλες οι λύσεις που συμμετέχουν στο τελικό σύνολο, διαφέρουν μεταξύ τους ως προς το σφάλμα ομαδοποίησης περισσότερο από ένα συγκεκριμένο κατώφλι. Ο διαχωρισμός αυτός, ωστόσο προκαλεί απώλεια πληροφορίας αφού υπάρχουν λύσεις για τις οποίες το σφάλμα κυμαίνεται στα ίδια επίπεδα, αλλά τα κέντρα των αντίστοιχων ομάδων διαφέρουν σε μεγάλο βαθμό.

Για να ξεπεραστεί το συγκεκριμένο πρόβλημα η επιλογή των λύσεων έγινε με βάση τον δείκτη Rand Index, χρησιμοποιώντας παράλληλα την πληροφορία από το τετραγωνικό σφάλμα κάθε λύσης και ένα κατώφλι ομοιότητας. Πιο αναλυτικά εξετάζονται οι $N \cdot |S|$ λύσεις ομαδοποίησης με τη σειρά, ξεκινώντας από εκείνη με το μικρότερο τετραγωνικό σφάλμα. Η λύση προστίθεται στο τελικό αποτέλεσμα. Στη συνέχεια υπολογίζεται ο δείκτης Rand Index για την υπό εξέταση λύση, συγκριτικά με κάθε άλλη από τις λύσεις. Εντοπίζονται οι λύσεις που είναι όμοιες και αφαιρούνται από το υπό εξέταση σύνολο ομαδοποιήσεων. Η διαδικασία συνεχίζεται μέχρι να γίνει έλεγχος όλων των λύσεων. Το κατώφλι ομοιότητας, βρίσκεται στο διάστημα [0,1] όπως αναφέρθηκε. Στην περίπτωση που πάρει την τιμή 1, όλες οι υποψήφιες ομαδοποιήσεις προστίθεται στο τελικό σύνολο, ενώ αντίθετα για τιμή κοντά στο 0 καμία ομαδοποίηση δεν προστίθεται στο τελικό σύνολο.

Υποθέτουμε ότι Clust = $\{Clust_i\}$ το σύνολο των ομαδοποιήσεων που έχουν παραχθεί, $Er(Clust_i)$ το σφάλμα ομαδοποίησης για κάθε μια από τις ομαδοποιήσεις Clust, Simthres το κατώφλι ομοιότητας που έχει ορίσει ο χρήστης, $R(Clust_i,Clust_j)$ η τιμή του δείκτη Rand Index για δύο ομαδοποιήσεις $Clust_i$ και $Clust_j$ και B το τελικό σύνολο ομαδοποιήσεων, ο αλγόριθμος Rand Index Partition παρουσιάζεται παρακάτω.

Αλγόριθμος 4: Rand Index Partition (w.r.t Clustering Error)

U = Clust, B = ø

While U ≠ ø

b = best (U) //Το καλύτερο αντικείμενο $Clust_i$ (με το μικρότερο σφάλμα ομαδοποίησης).

B = B∨{b} //Προσθήκη του b στο σύνολο αποτελεσμάτων.

U_b = {$Clust_i \in U$: R ($Clust_i$, b) > Simthres}//Εύρεση των $Clust_i$ που είναι όμοιες με το best.

U = U - U_b //Αφαίρεση των λύσεων που εντοπίστηκαν στο προηγούμενο βήμα από το σύνολο ομαδοποιήσεων.

End

3.4.3. Τελικός Αλγόριθμος

Στο σημείο αυτό παρουσιάζονται τα βήματα που ακολουθούνται για την προτεινόμενη διαδικασία πολλαπλής ομαδοποίησης.

Αρχικά, για k=1 υπάρχει μόνο μια ομάδα, με βέλτιστο κέντρο το κέντρο όλων των δεδομένων. Στην περίπτωση αυτή η προτεινόμενη μεθοδολογία δεν παράγει κάποια εναλλακτική ομαδοποίηση, αφού πάντα η λύση είναι μία. Για την προσθήκη δεύτερης ομάδας τοποθετείται το κέντρο k=1 στην βέλτιστη θέση του και εκτελείται ο αλγόριθμός k-means N φορές. Με αυτόν τον τρόπο παράγεται N αριθμός λύσεων. Ως κριτήριο χρησιμοποιείται το σφάλμα ομαδοποίησης (Εξ. 3.1).

Γίνεται επιλογή ενός αριθμού λύσεων με τον αλγόριθμο Rand Index Partition. Τελικά διατηρείται ένας αριθμός λύσεων S. Για k=2 οι λύσεις S αποτελούν τις προτεινόμενες εναλλακτικές ομαδοποιήσεις. Στο σύνολο αυτό περιλαμβάνεται και η βέλτιστη λύση που παράγεται από τον απλό global k-means αλγόριθμο. Για κάθε επόμενη προσθήκη ομάδας χρησιμοποιούνται όλες οι εναλλακτικές ομαδοποιήσεις που υπολογίστηκαν στο προηγούμενο βήμα. Αυτό σημαίνει ότι εκτελείται ο αλγόριθμος k-means N·|S| φορές από τις ακόλουθες θέσεις:

για κάθε μία από τις παραχθείσες ομαδοποιήσεις S τοποθετούνται τα (k-1) κέντρα στις βέλτιστες θέσεις τους και εκτελείται ο απλός k-means N φορές.Στις καινούριες λύσεις που παράγονται εφαρμόζεται πάλι η μεθοδολογία που περιγράφηκε προηγουμένως. Ακολούθως αυτές χρησιμοποιούνται σαν αφετηρία για την προσθήκη μιας επιπλέον ομάδας.

Αξίζει να αναφέρουμε, πως ο αριθμός S των λύσεων που επιλέγεται κάθε φορά είναι συνάρτηση του συντελεστή ομοιότητας που χρησιμοποιείται για να ελέγξει το πόσο διαφέρουν οι λύσεις μεταξύ τους, καθώς επίσης και της δομής του συνόλου δεδομένων που εξετάζεται. Στην περίπτωση λοιπόν που για έναν αριθμό ομάδων οι υποψήφιες λύσεις διαφέρουν λιγότερο από όσο ορίζει ο συντελεστής που έχει επιλέξει ο χρήστης, ή η δομή του συνόλου δεδομένων δεν επιτρέπει πολλαπλές διαφορετικές ομαδοποιήσεις, κρατείται μόνο η βέλτιστη και δεν δημιουργούνται εναλλακτικές λύσεις για τη συγκεκριμένη τάξη.

Το κατώφλι ομοιότητας γενικά ορίζεται κοντά στο 0.9 (συνήθως 0.95). Όσο μικρότερο είναι το κατώφλι τόσο λιγότερες είναι και οι εναλλακτικές που επιλέγονται. Στα πειράματα που πραγματοποιήθηκαν σε συνθετικά τεχνητά δεδομένα καθώς επίσης και στην κατάτμηση εικόνας χρησιμοποιείται τιμή 0.95, ενώ το κατώφλι διαφοροποιείται για τα πειράματα που πραγματοποιήθηκαν σε ακολουθίες βίντεο. Τα πειράματα θα παρουσιαστούν στο επόμενο κεφάλαιο.

ΚΕΦΑΛΑΙΟ 4. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Το παρόν κεφάλαιο αποτελείται από τρεις επιμέρους ενότητες. Αρχικά παραθέτουμε κάποια πειράματα που έγιναν σε συνθετικά δεδομένα και στη συνέχεια ακολουθούν τα αποτελέσματα από την εφαρμογή του προτεινόμενου αλγορίθμου πολλαπλής ομαδοποίησης σε κατάτμηση εικόνας καθώς και μια εφαρμογή εξαγωγής χαρακτηριστικών εικονοπλαισιών από ακολουθίες βίντεο.

4.1. Συνθετικά Δεδομένα

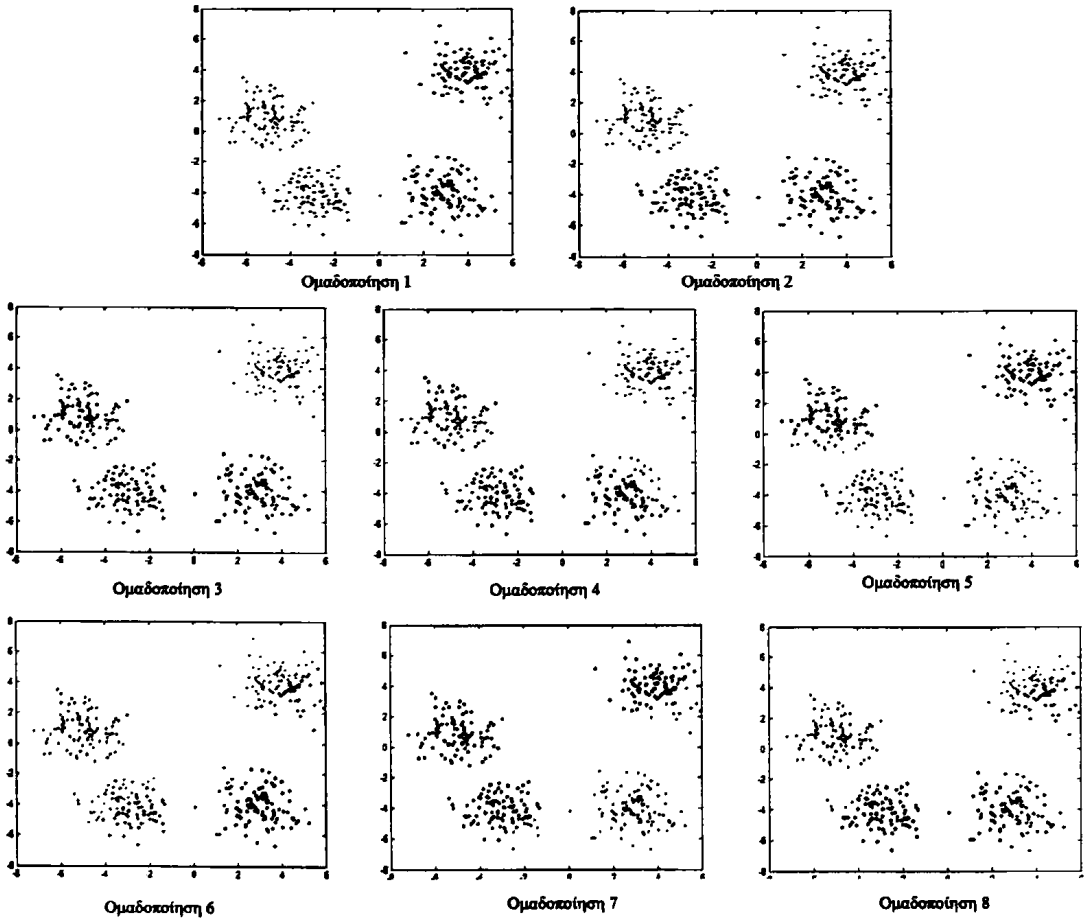
Σε πρώτη φάση κρίθηκε σκόπιμο να εφαρμόσουμε τον αλγόριθμο σε συνθετικά σύνολα δεδομένων, προκειμένου να διασφαλίσουμε ότι η μέθοδος που αναπτύχθηκε δουλεύει σωστά. Θελήσαμε να μελετήσουμε τη δυνατότητα εύρεσης εναλλακτικών λύσεων πέραν της βέλτιστης που πιθανώς να περιγράφουν καλύτερα τη δομή του συνόλου. Για τα πειράματα χρησιμοποιήθηκαν δύο διδιάστατα τεχνητά σύνολα δεδομένων αποτελούμενα από 202 και 400 σημεία αντίστοιχα.

4.1.1. Πειραματικά Αποτελέσματα

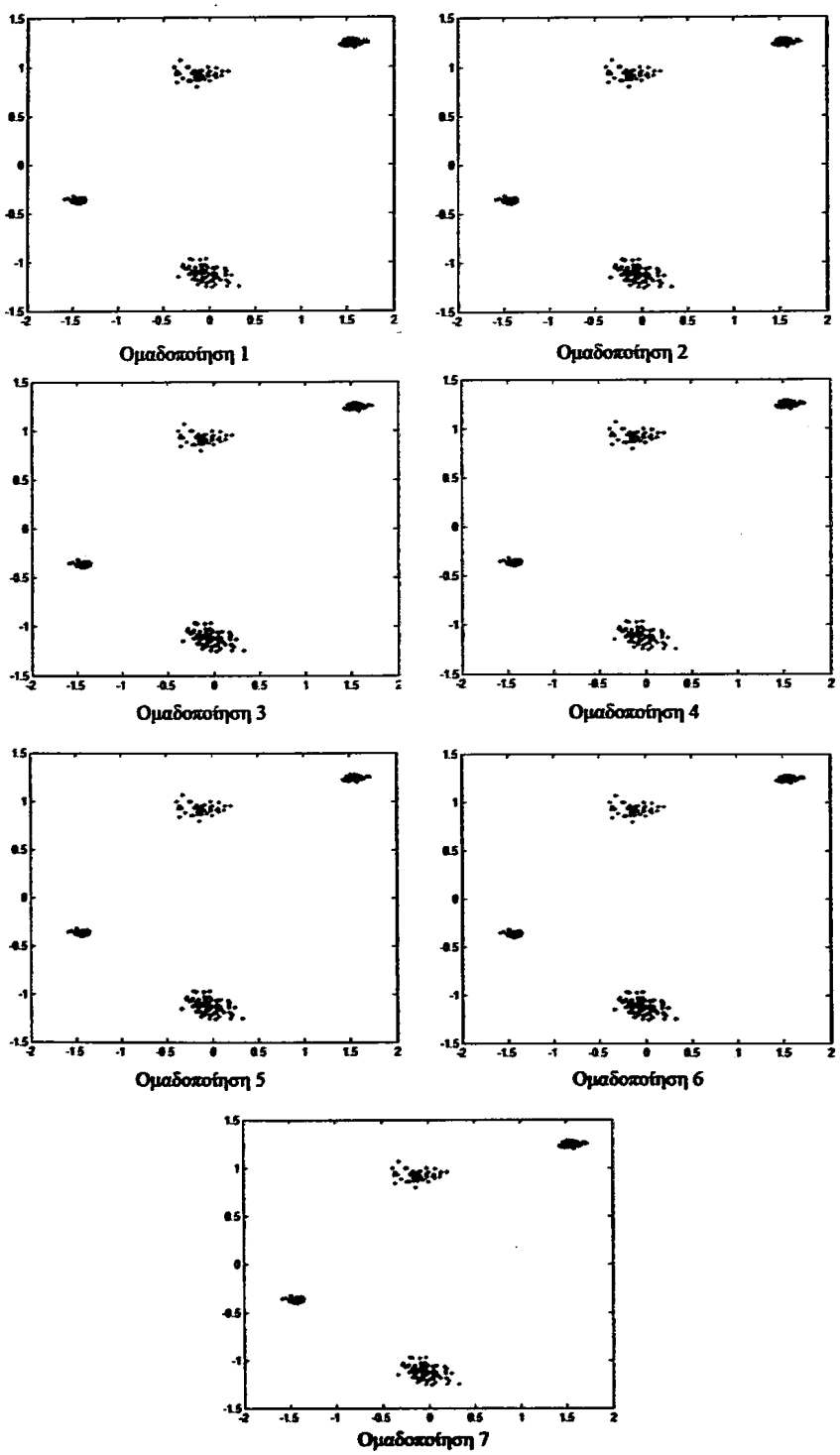
Δημιουργήθηκε ένα σύνολο συνθετικών δεδομένων με 400 στοιχεία, που περιέχει τέσσερις ομάδες. Αρχικά επιλέγεται η ομαδοποίηση των δεδομένων σε δύο ομάδες (k=2) αφού με αυτόν τον τρόπο τα αποτελέσματα έχουν περισσότερο νόημα για το σύνολο δεδομένων. Η τιμή κατωφλίου για τον υπολογισμό της ομοιότητας των λύσεων τέθηκε ίση με 0.95. Στο σχήμα 4.1 μπορούμε να δούμε τις 8 εναλλακτικές

λύσεις ομαδοποίησης που παράχθηκαν μετά την εφαρμογή του αλγορίθμου σε αυτό το τεχνητό σύνολο δεδομένων. Στη συνέχεια στο σχήμα 4.2 παρουσιάζονται οι 7 εναλλακτικές ομαδοποιήσεις που βρέθηκαν μετά την εφαρμογή της μεθόδου σε σύνολο δεδομένων με 202 στοιχεία για πρόβλημα τεσσάρων ομάδων (k=4).

Στο σχήμα 4.1 παρατηρούμε ότι η Ομαδοποίηση 1 και Ομαδοποίηση 2 χαρακτηρίζονται από υψηλή ποιότητα, χαρακτηρίζουν από διαφορετικές οπτικές τα δεδομένα και περιγράφουν εναλλακτικά τη δομή του συνόλου. Οι υπόλοιπες εναλλακτικές ομαδοποιήσεις που βρέθηκαν παρουσιάζουν μικρότερη ποιότητα στη λύση.



Σχήμα 4.1 Ομαδοποίηση με k=2 ομάδες χρησιμοποιώντας τον MCGKM αλγόριθμο



Σχήμα 4.2 Ομαδοποίηση με k=4 ομάδες χρησιμοποιώντας τον MCGKM αλγόριθμο

4.2. Κατάτμηση Εικόνας

Η κατάτμηση εικόνας στην υπολογιστική όραση αναφέρεται στη διαδικασία του διαχωρισμού μιας εικόνας σε πολλαπλές ομογενείς περιοχές[18]. Χρησιμοποιείται συνήθως για να εντοπίσουμε αντικείμενα όπως πρόσωπα και όρια περιοχών όπως γραμμές, γωνίες και καμπύλες σε εικόνες. Κατά την κατάτμηση πραγματοποιείται ανάθεση μιας «ετικέτας» σε κάθε εικονοστοιχείο (pixel) μιας εικόνας, έτσι ώστε τα εικονοστοιχεία με την ίδια ετικέτα να έχουν συγκεκριμένα κοινά οπτικά χαρακτηριστικά. Το αποτέλεσμα της κατάτμησης είναι ένα σύνολο από τμήματα που συλλογικά καλύπτουν ολόκληρη την εικόνα ή ένα σύνολο περιγραμμάτων που εξάγονται από τα αντικείμενα της εικόνας. Τελικά, εικονοστοιχεία που ανήκουν στην ίδια περιοχή είναι παρόμοια με βάση κάποιο χαρακτηριστικό, όπως το χρώμα, η φωτεινότητα και η υφή και αντίστοιχα διαφορετικά τμήματα εικοστοιχείων διαφέρουν μεταξύ τους ως προς τα ίδια χαρακτηριστικά.

Υπάρχουν πολλοί αλγόριθμοι και μέθοδοι κατάτμησης εικόνας. Ανάμεσα σε αυτές που χρησιμοποιούνται συχνότερα είναι οι αλγόριθμοι κατωφλίωσης (Thresholding methods), αλγόριθμοι ομαδοποίησης, τεχνικές που βασίζονται στο ιστόγραμμα (Histogram-based methods), αλγόριθμοι ανίχνευσης ακμών (edge detection) κ.λπ.

Ο τυπικός αλγόριθμος k-means χρησιμοποιείται στην κατάτμηση της εικόνας λόγω της απλότητάς του και της ικανότητας να ομαδοποιεί αποτελεσματικά τα εικονοστοιχεία μιας εικόνας. Αρχικά επιλέγονται τα διανύσματα χαρακτηριστικών για κάθε εικονοστοιχείο (όπως χρώμα, υφή, φωτεινότητα). Στόχος είναι να χωριστούν τα διανύσματα σε Κ ομάδες. Για κάθε ομαδοποίηση εξετάζεται κάθε εικονοστοιχείο στην εικόνα και ανατίθεται στην ομάδα με την μικρότερη απόσταση μεταξύ του εικονοστοιχείου και του μέσου της ομάδας. Τελικά η εικόνα χωρίζεται σε Κ ομάδες.

Ο αλγόριθμος global k-means χρησιμοποιεί τον τυπικό k-means σαν εσωτερική διαδικασία. Επιλέγονται τα διανύσματα χαρακτηριστικών, παρόμοια με τον τυπικό k-means. Για k=1 υπάρχει μια ομάδα, με κέντρο το κέντρο όλων των εικονοστοιχείων της εικόνας. Για να προστεθεί μια νέα ομάδα k, τα k-1 κέντρα είναι τοποθετημένα στις βέλτιστες θέσεις τους, που προέκυψαν λύνοντας το k-1 πρόβλημα ομαδοποίησης.

Το k-οστό κέντρο τοποθετείται στην κατάλληλη θέση μετά από N εκτελέσεις του k-means (όπου N ο αριθμός των εικονοστοιχείων της εικόνας).

Ο αλγόριθμος MCGKM χρησιμοποιείται για την κατάτμηση της εικόνας σε K ομάδες και την εξαγωγή πολλαπλών αποτελεσμάτων κατάτμησης. Στόχος είναι να βρεθούν K χρωματικά κέντρα στην εικόνα. Τα εικονοστοιχεία είναι διανύσματα που περιλαμβάνουν τιμές στον RGB χώρο. Εφαρμόζοντας την προτεινόμενη μεθοδολογία πραγματοποιείται κατάτμηση με βάση το χρώμα. Προκειμένου να μειωθεί το υπολογιστικό κόστος, επιλέχθηκε να μην εξετάζονται όλα τα εικονοστοιχεία, δηλαδή να μην εκτελείται ο k-means N φορές, για κάθε μια από τις S λύσεις. Αντί αυτού, ο αλγόριθμος δέχεται μια παράμετρο βήματος (step) η οποία ορίζει τον αριθμό εκτελέσεων του αλγορίθμου N/step. Με αυτόν τον τρόπο το υπολογιστικό κόστος μειώνεται.

Επειδή η εφαρμογή του αλγορίθμου σε μια εικόνα μεγάλων διαστάσεων είναι χρονοβόρα, η μέθοδος αρχικά χρησιμοποιήθηκε σε εικόνα μικρότερων διαστάσεων, εφαρμόζοντας έναν συντελεστή μείωσης ύψους και πλάτους. Αφού υπολογιστούν τα κέντρα των ομάδων στην μικρή εικόνα για κάθε μια από τις πολλαπλές ομαδοποιήσεις, αυτά χρησιμοποιούνται στη συνέχεια για να αντικαταστήσουν τα εικονοστοιχεία στην αρχική εικόνα.

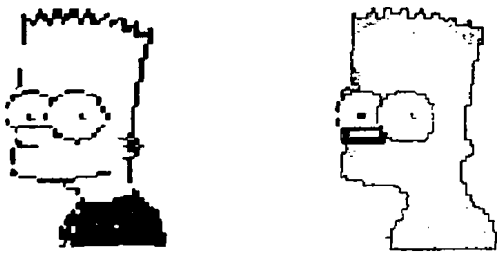
4.2.1. Πειραματικά Αποτελέσματα

Στο σχήμα 4.3 αριστερά βλέπουμε την αρχική εικόνα 100 x 100 του φανταστικού χαρακτήρα Bart Simpson. Η εικόνα κατατμήθηκε σε τρεις ομάδες, παράγοντας τρεις εναλλακτικές ομαδοποιήσεις. Την ομαδοποίηση 1 στο κέντρο, η οποία αποτελεί και τη λύση για τον global k-means αλγόριθμο, την ομαδοποίηση 2 και 3. Τρία χρώματα χρησιμοποιήθηκαν για να δείξουν τρεις διαφορετικές ομάδες και στις δύο λύσεις. Όλα τα εικονοστοιχεία σε μια ομάδα έχουν το ίδιο χρώμα. Στην ομαδοποίηση 1 τα μάτια αποτελούν μια ξεχωριστή ομάδα, ενώ στην ομαδοποίηση 2 σχηματίζονται τα χαρακτηριστικά του προσώπου, τα οποία μαζί με την μπλούζα ανήκουν στην ίδια ομάδα και είναι καλύτερα ορατά και τέλος στην ομαδοποίηση 3 βλέπουμε την λεπτομέρεια της σκιάς. Με έναν συνδυασμό, των λύσεων μπορούμε να διακρίνουμε

όλα τα βασικά χαρακτηριστικά της εικόνας, που δεν ήταν ορατά αρχικά με τον τυπικό global k-means.

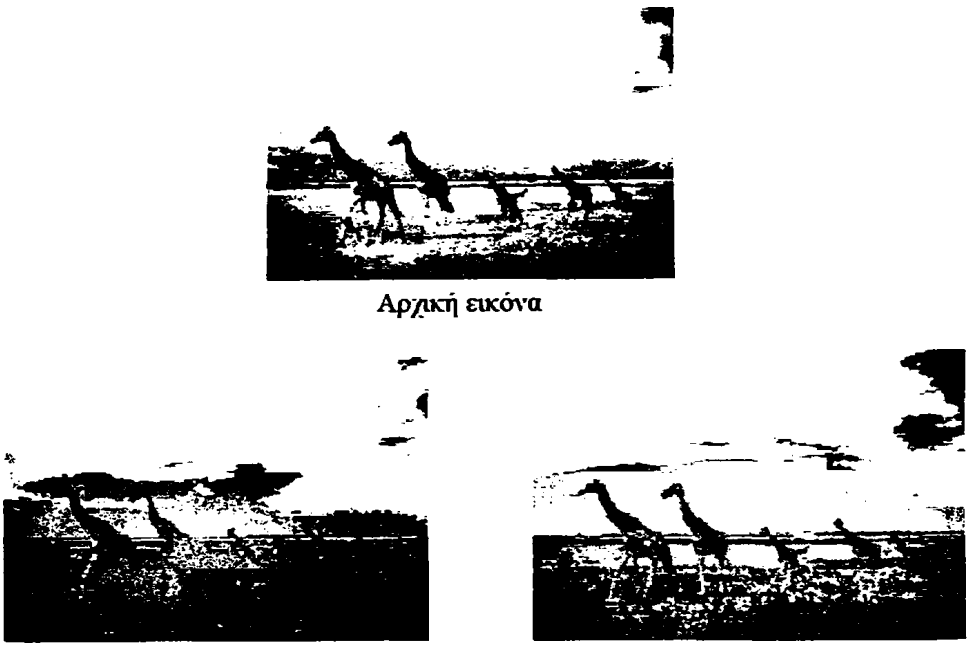
Στο σχήμα 4.4 φαίνεται η εικόνα "Λεοπαρδάλεις" διάστασης 250 x 170 που περιέχει πέντε λεοπαρδάλεις με φόντο που διαχωρίζεται σε δύο επίπεδα, ουρανό και πράσινο. Η εικόνα κατατμήθηκε σε τέσσερις ομάδες δημιουργώντας δύο εναλλακτικές λύσεις ομαδοποίησης. Τέσσερα χρώματα χρησιμοποιήθηκαν για να αναδείξουν τις ξεχωριστές ομάδες σε κάθε ομαδοποίηση. Η ομαδοποίηση 1, η οποία αποτελεί και την λύση του global k-means δεν επιτυγχάνει τον σαφή διαχωρισμό των λεοπαρδάλεων από το φόντο, κυρίως κάτω από το επίπεδο του ουρανού. Η εναλλακτική ομαδοποίηση 2 φαίνεται να διαχωρίζει τα αντικείμενα από το φόντο καλύτερα, με ελαφρά απώλεια πληροφορίας στην περιοχή του ουρανού.

Αρχική εικόνα Ομαδοποίηση 1



Ομαδοποίηση 2 Ομαδοποίηση 3

Σχήμα 4.3 Αριστερά πάνω αρχική εικόνα 100 x 100 του φανταστικού χαρακτήρα Bart Simpson. Στη συνέχεια τρεις διαφορετικές ομαδοποιήσεις με τρεις ομάδες k=3 τον αλγόριθμο MCGKM.



Αρχική εικόνα

Ομαδοποίηση 1

Ομαδοποίηση 2

Σχήμα 4.4 Αρχική εικόνα 250 x 170 "Λεοπαρδάλεις". Ομαδοποίηση με τέσσερις ομάδες k=4 με δύο διαφορετικές εναλλακτικές ομαδοποιήσεις.

Στο σχήμα 4.5 βλέπουμε την αρχική εικόνα ανάλυσης 250 x 166 "Πυροσβέστες" που περιλαμβάνει δύο άτομα μπροστά από ένα κόκκινο αυτοκίνητο. Στην εικόνα εφαρμόστηκε ο αλγόριθμος με τέσσερις ομάδες δημιουργώντας δύο εναλλακτικές λύσεις ομαδοποίησης. Τέσσερα χρώματα χρησιμοποιήθηκαν για την ομαδοποίηση των εικονοστοιχείων της ίδιας ομάδας. Παρόμοια με το προηγούμενο παράδειγμα η πρώτη ομαδοποίηση δεν διαχωρίζει ολόκληρες τις μορφές από το φόντο σωστά, κάτι που επιτυγχάνεται με τη δεύτερη λύση. Σ' αυτήν φαίνονται πιο ξεκάθαρα οι φιγούρες από το κόκκινο φόντο, παρουσιάζοντας μια ελαφρά απώλεια στα χαρακτηριστικά του προσώπου. Έτσι λοιπόν, η εναλλακτική ομαδοποίηση που παράχθηκε με την προτεινόμενη μεθοδολογία διαχωρίζει τις φιγούρες από το φόντο καλύτερα απ' ότι ο παραδοσιακός global-k means αλγόριθμος.

Αρχική εικόνα

Ομαδοποίηση 1

Ομαδοποίηση 2

Σχήμα 4.5 Πάνω αρχική εικόνα 250 x 166 "Πυροσβέστες". Ομαδοποίηση της εικόνας με τέσσερις ομάδες (k=4) και παρουσίαση δύο εναλλακτικών λύσεων.

4.3. Εξαγωγή Χαρακτηριστικών Εικονοπλαισιών από Ακολουθίες Βίντεο

Μια ακολουθία βίντεο συνήθως περιέχει ένα μεγάλο αριθμό από εικονοπλαίσια (frames) ο οποίος είναι απαγορευτικός για πολλές εφαρμογές. Γι αυτό το λόγο είναι απαραίτητη η ύπαρξη μιας μεθόδου που θα επιτρέπει στον χρήστη να αποκτά γρήγορα άποψη για το περιεχόμενο ενός βίντεο χωρίς να είναι απαραίτητο να το παρακολουθήσει όλο. Κάτι τέτοιο μπορεί να καταστεί δυνατό δημιουργώντας μια περίληψη του βίντεο.

Οι τεχνικές περίληψης που έχουν δημιουργηθεί έχουν σαν βασικό στόχο να διευκολύνουν την αναζήτηση ενός βίντεο σε μια βάση δεδομένων. Προκειμένου να είναι αποτελεσματική είναι απαραίτητο ο αριθμός των εικονοπλαισιών να είναι σχετικά μικρός και η περίληψη να είναι όσο το δυνατόν περισσότερο αντιπροσωπευτική του αρχικού βίντεο. Επίσης η περίληψη ενός βίντεο δίνει στον χρήστη καλύτερη άποψη και βοηθά στην περιήγηση ενός βίντεο. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο σε εφαρμογές επεξεργασίας βίντεο, όπως το μοντάζ. Αξίζει

να αναφέρουμε ότι οι περιλήψεις μειώνουν σημαντικά το υπολογιστικό κόστος της ανάλυσης και ανάκτησης του περιεχομένου ενός βίντεο.

Στην παρούσα εργασία ασχοληθήκαμε με τμήματα ενός βίντεο που το περιεχόμενό τους αντιστοιχεί σε ένα πλάνο. Ένα πλάνο ορίζεται ως μια συνεχής ακολουθία εικονοπλαισιών που έχουν καταγραφεί από μόνο μια κάμερα. Η πιο διαδεδομένη προσέγγιση αναπαράστασης ενός πλάνου είναι τα χαρακτηριστικά εικονοπλαίσια (keyframes) ή αλλιώς στατικές περιλήψεις[19]. Πρόκειται για ένα σύνολο από τις πιο βασικές εικόνες που περιέχουν όσο δυνατόν περισσότερη πληροφορία από την ακολουθία βίντεο.

Για την εξαγωγή χαρακτηριστικών εικονοπλαισίων χρησιμοποιούνται περιγραφείς τους. Υπάρχουν αρκετοί περιγραφείς, όπως τα ιστογράμματα χρώματος, οι περιγραφείς CENTRIST, WAVELET και SIFT. Στην εργασία χρησιμοποιήθηκαν τα ιστογράμματα χρώματος. Τα ιστογράμματα χρώματος είναι τα πιο διαδεδομένα χαρακτηριστικά στο πρόβλημα της περίληψης ενός βίντεο. Είναι η αναπαράσταση της κατανομής των χρωμάτων σε μια εικόνα και χρησιμοποιούνται συχνότερα για την ανίχνευση ορίων σε ένα πλάνο. Το σύνολο των χρωμάτων χωρίζεται σε κάδους (bins) που περιέχουν ένα προκαθορισμένο εύρος χρωμάτων. Ένα ιστόγραμμα χρώματος αναπαριστά την κατανομή των εικονοστοιχείων στους κάδους αυτούς.

Ένα ιστόγραμμα χρώματος μπορεί να δημιουργηθεί για οποιονδήποτε χώρο χρωμάτων, αλλά συνήθως χρησιμοποιείται στον τρισδιάστατο χώρο, όπως είναι το RGB ή HSV. Στην εργασία χρησιμοποιήθηκαν κανονικοποιημένα ιστογράμματα στον χώρο χρώματος HSV. Για κάθε εικονοπλαίσιο υπολογίζεται ένα κανονικοποιημένο ιστόγραμμα με 8 κάδους για την Απόχρωση H (Hue) και 4 κάδους για τον Κορεσμό S (Saturation) και Αξία V (Value). Τα τρία αυτά ιστογράμματα ενώνονται και σχηματίζουν ένα διάνυσμα διάστασης 16 (8+4+4). Χρησιμοποιείται η τρισδιάστατη αναπαράστασή τους με 128 κάδους. Αξίζει να αναφέρουμε ότι το βασικότερο μειονέκτημα των ιστογραμμάτων χρώματος είναι ότι αποτελούν μια περιγραφή μόνο του χρώματος του αντικειμένου, ενώ αγνοούνται το σχήμα και η επιφάνειά του. Με αυτόν τον τρόπο είναι δυνατό να δημιουργηθούν ακριβώς ίδια ιστογράμματα χρώματος δύο αντικειμένων επειδή έχουν το ίδιο χρώμα.

Η ομαδοποίηση είναι η ευρύτερα χρησιμοποιούμενη μέθοδος για την εξαγωγή χαρακτηριστικών εικονοπλαισίων από βίντεο. Στα πειράματα που πραγματοποιήθηκαν χρησιμοποιήθηκε ο προτεινόμενος αλγόριθμος ομαδοποίησης MCGKM για να πραγματοποιηθεί η εξαγωγή των keyframes, ομαδοποιώντας τα εικονοπλαίσια. Στη συνέχεια ως keyframe χαρακτηρίζεται το medoid κάθε ομάδας, δηλαδή εκείνο το εικονοπλαίσιο της ομάδας το οποίο έχει τη μεγαλύτερη μέση ομοιότητα με όλα τα εικονοπλαίσια της ομάδας[19]. Τελικά παράγεται η περίληψη ενός πλάνου και συγκρίνεται με το αποτέλεσμα του παραδοσιακού global k-means αλγορίθμου.

4.3.1. Πειραματικά Αποτελέσματα

Κατά τη διάρκεια της πειραματικής διαδικασίας εξετάστηκαν 27 διαφορετικά πλάνα, ανάλυσης 320 x 240. Τα πλάνα αυτά περιλαμβάνουν εναλλαγή σχημάτων, κίνηση σε αυτοκίνητο και εναλλαγή κίνησης μέσα σε συγκεκριμένο χώρο. Η τιμή του κατωφλιού ομοιότητας για τον υπολογισμό των εναλλακτικών λύσεων έλαβε την τιμή 0.85. Παρακάτω φαίνονται τα αποτελέσματα από πέντε διαφορετικά πλάνα, χρησιμοποιώντας αρχικά τον MCGKM και στη συνέχεια παρουσιάζεται η αντίστοιχη λύση εφαρμόζοντας τον παραδοσιακό global k-means αλγόριθμο.

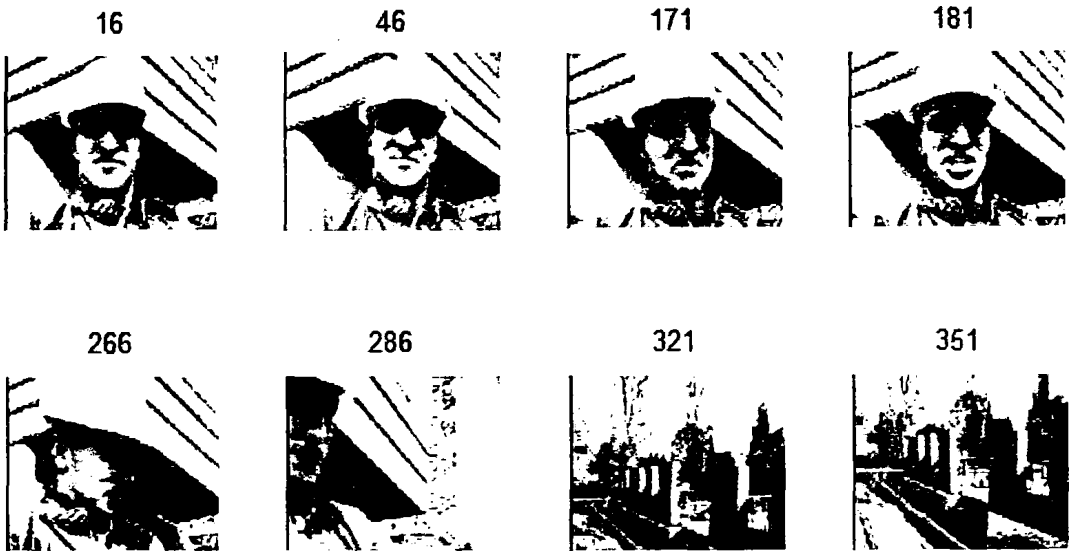
Η προτεινόμενη μεθοδολογία συνενώνει τις πολλαπλές εναλλακτικές λύσεις για έναν αριθμό ομάδων, δημιουργώντας ένα σύνολο εικονοπλαισίων μεγαλύτερης τάξης, ενώ η εφαρμογή του απλού global k-means εξάγει αριθμό εικονοπλαισίων ίσο με την τάξη του προβλήματος που λύνει. Κάθε σύνολο λύσεων του MCGKM με τέσσερις ομάδες για ένα πλάνο συγκρίνεται με το αποτέλεσμα του global k-means, για το αντίστοιχο πρόβλημα υψηλότερης τάξης. Το σύνολο λύσεων του MCGKM ισούται με την τάξη που εξετάζεται στον δεύτερο αλγόριθμο. Από το σύνολο των εικονοπλαισίων που παράγονται με την προτεινόμενη μέθοδο, αφαιρούνται τα επαναλαμβανόμενα στις διαφορετικές λύσεις, καθώς επίσης και εικονοπλαίσια τα οποία είναι όμοια μεταξύ τους. Τα εικονοπλαίσια συγκρίνονται χρησιμοποιώντας τα ιστογράμματά τους.

Στο 1° πλάνο εμφανίζεται η γνωστή ακολουθία του Foreman. Στο σχήμα 4.6 τα αποτελέσματα που παρατίθεται για τον MCGKM αποτελούν έναν συνδυασμό των

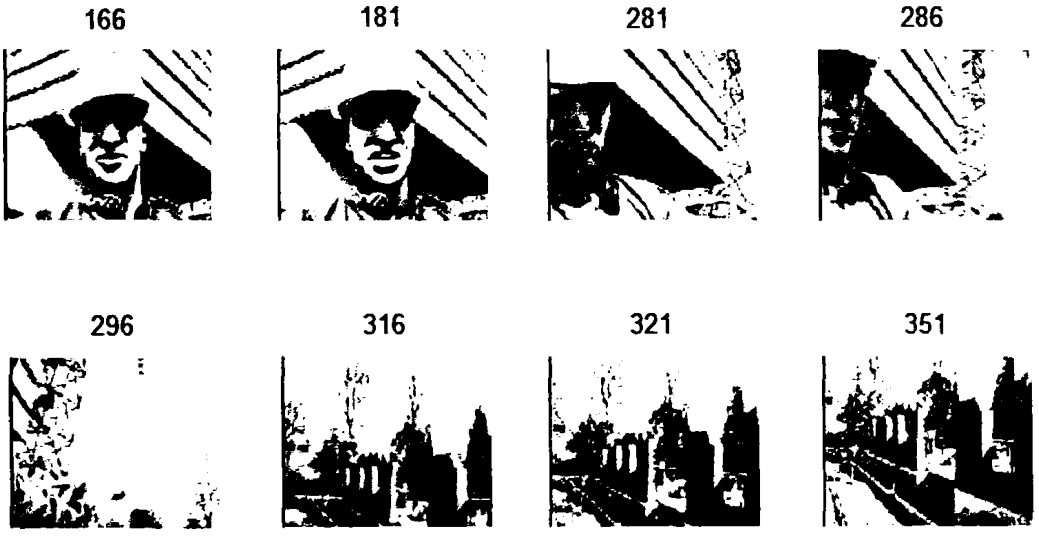
πολλαπλών εναλλακτικών ομαδοποιήσεων για τέσσερις ομάδες (k=4), αποδίδοντας στο αποτέλεσμα οκτώ διαφορετικά keyframes. Όσον αφορά στον απλό global k-means στο σχήμα 4.7, η λύση που παρουσιάζεται αντιπροσωπεύει την ομαδοποίηση με οκτώ ομάδες (k=8), αποδίδοντας επίσης στο αποτέλεσμα οκτώ keyframes. Μπορούμε να παρατηρήσουμε ότι το αποτέλεσμα του MCGKM εστιάζει περισσότερο στην αλλαγή των εκφράσεων του προσώπου.

Στο 2° πλάνο παρουσιάζεται η εναλλαγή ορθογώνιων σχημάτων του ίδιου χρώματος, ξεκινώντας από το μικρότερο και καταλήγοντας στο μεγαλύτερο. Στα σχήματα 4.8 και 4.9 φαίνονται τα αποτελέσματα του MCGKM και του global k-means αντίστοιχα με εφτά εικονοπλαίσια. Παρατηρούμε ότι η προτεινόμενη μεθοδολογία, αν και περιέχει δύο επαναλήψεις εικονοπλαισίων που περιέχουν το ίδιο σχήμα, περιγράφει καλύτερα το πλάνο από την αρχή μέχρι το τέλος σε σχέση με τον απλό global k-means αλγόριθμο.

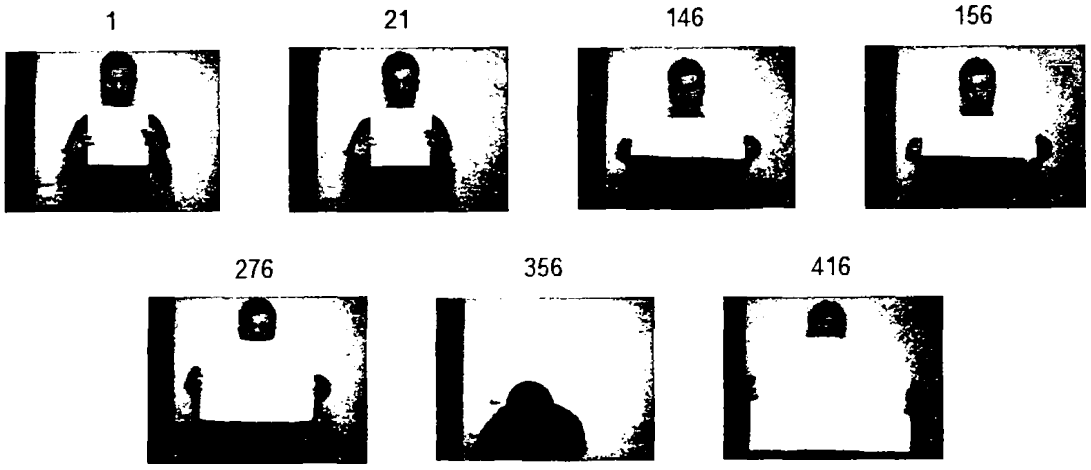
Παρόμοια είναι τα αποτελέσματα και στο 3° πλάνο, το οποίο παρουσιάζει επίσης διάφορα σχήματα του ίδιου χρώματος. Ο αλγόριθμος MCGKM καταφέρνει να περιγράψει το πλάνο πιο συνολικά, σε αντίθεση με τον global k-means ο οποίος «χάνει» το τριγωνικό σχήμα από την περιγραφή του (σχήμα 4.10-4.11). Μπορούμε ωστόσο να παρατηρήσουμε ότι και στις δύο μεθόδους βλέπουμε μερικά εικονοπλαίσια να επαναλαμβάνονται. Στο 4° πλάνο παρουσιάζεται η εναλλαγή κίνησης σε έναν δρόμο. Ο αλγόριθμος MCGKM με τέσσερις ομάδες περιγράφει το πλάνο εξίσου καλά με τον παραδοσιακό global k-means με εννιά ομάδες, αλλά σε μικρότερο χρόνο εκτέλεσης. Ανάλογα συμπεράσματα προκύπτουν και για τα υπόλοιπα πλάνα που εξετάστηκαν.



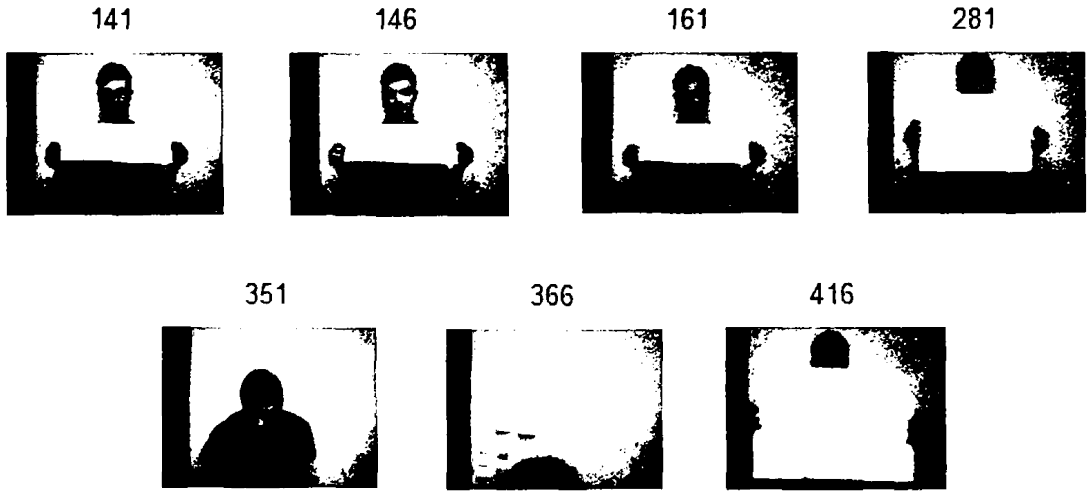
Σχήμα 4.6 Συνδυασμός οκτώ εικονοπλαισίων της 1ης ακολουθίας για k=4 μετά την εφαρμογή του MCGKM



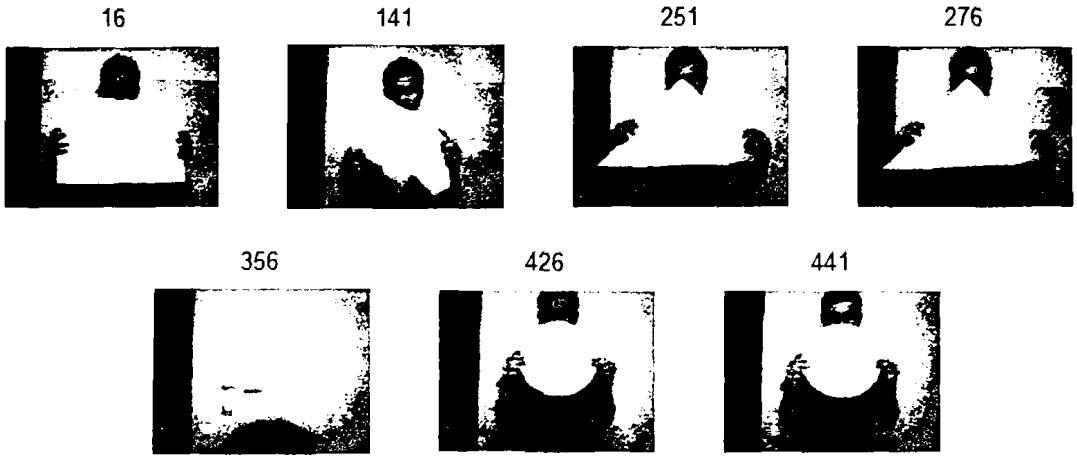
Σχήμα 4.7 Εικονοπλαίσια της 1ης ακολουθίας μετά την εφαρμογή του global k-means για k=8 ομάδες



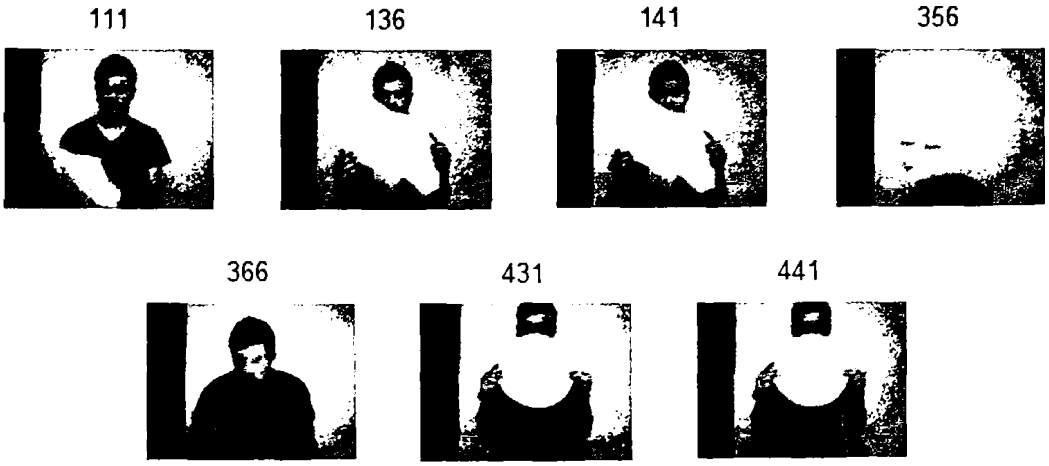
Σχήμα 4.8 Συνδυασμός εφτά εικονοπλαισίων της 2ης ακολουθίας για k=4 μετά την εφαρμογή του MCGKM



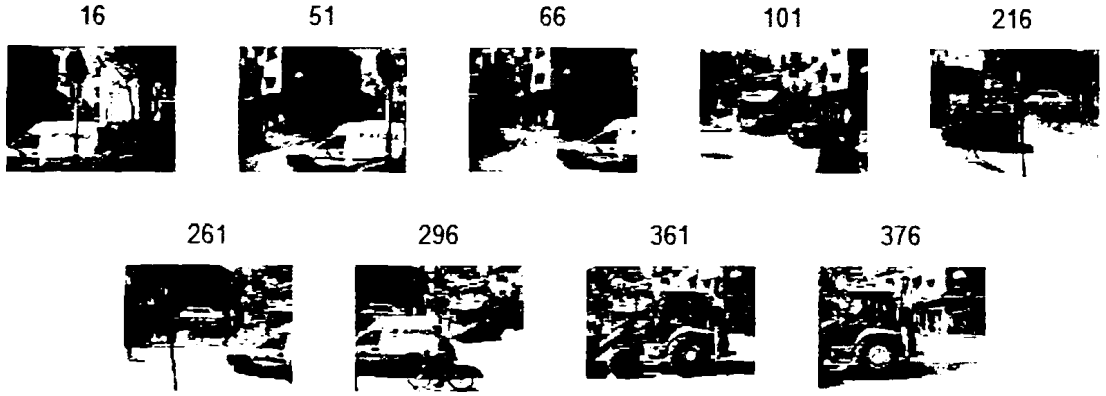
Σχήμα 4.9 Εικονοπλαίσια της 2ης ακολουθίας μετά την εφαρμογή του global k-means για k=7 ομάδες



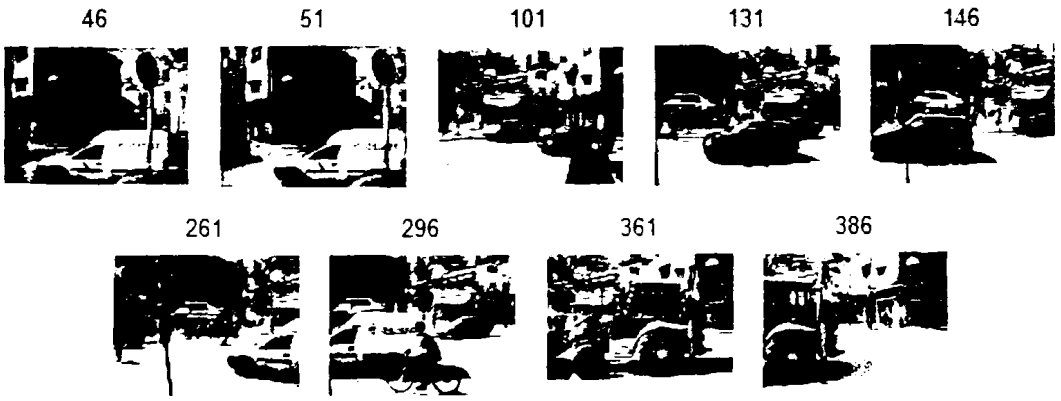
Σχήμα 4.10 Συνδυασμός εφτά εικονοπλαισίων της 3ης ακολουθίας για k=4 μετά την εφαρμογή του MCGKM



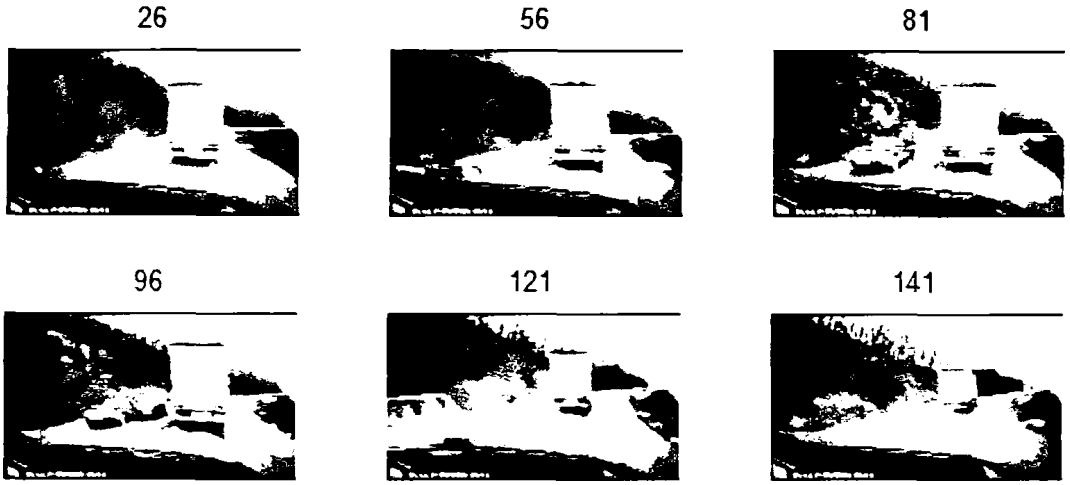
Σχήμα 4.11 Εικονοπλαίσια της 3ης ακολουθίας μετά την εφαρμογή του global k-means για k=7 ομάδες



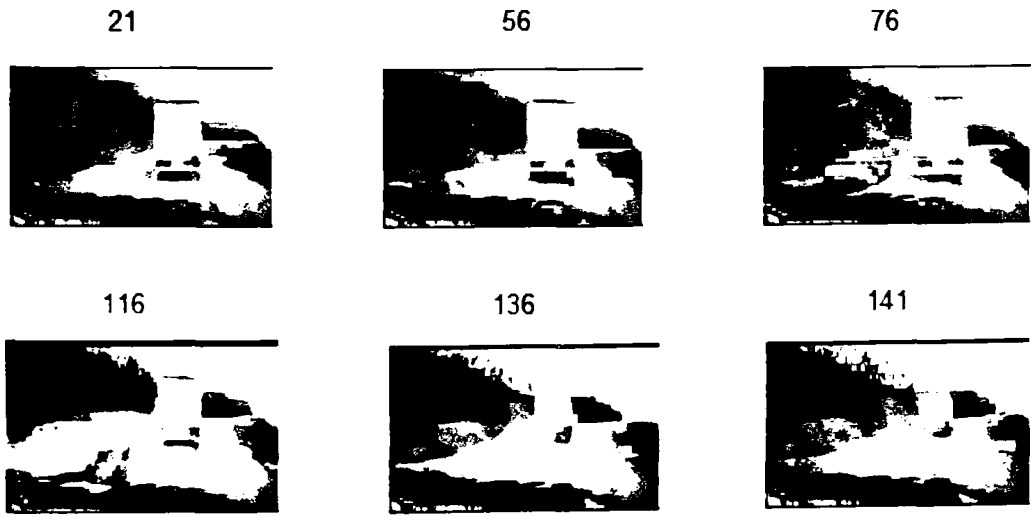
Σχήμα 4.12 Συνδυασμός εννιά εικονοπλαισίων της 4ης ακολουθίας για k=4 μετά την εφαρμογή του MCGKM



Σχήμα 4.13 Εικονοπλαίσια της 4ης ακολουθίας μετά την εφαρμογή του global k-means για k= 9 ομάδες



Σχήμα 4.14 Συνδυασμός έξι εικονοπλαισίων της 5ης ακολουθίας για k=4 μετά την εφαρμογή του MCGKM



Σχήμα 4.15 Εικονοπλαίσια της 5ης ακολουθίας μετά την εφαρμογή του global k-means για k=6 ομάδες

ΚΕΦΑΛΑΙΟ 5. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

5.1. Συμπεράσματα

Στην εργασία αυτή μελετήθηκε το πρόβλημα της πολλαπλής ομαδοποίησης. Αναλύθηκε η ανάγκη ύπαρξης πολλών ομαδοποιήσεων που περιγράφουν τα δεδομένα εναλλακτικά, και όχι μόνο μιας και μοναδικής λύσης. Η αναζήτηση μίας και μόνο λύσης για κάθε πρόβλημα ομαδοποίησης μπορεί να είναι ακατάλληλη, αφού διαφορετικές λύσεις μπορεί να είναι αποδεκτές για διαφορετικές χρήσεις των δεδομένων. Παρουσιάστηκαν μερικές από τις πιο αντιπροσωπευτικές προσεγγίσεις πολλαπλής ομαδοποίησης στον αρχικό χώρο δεδομένων. Αναφέρθηκαν τα βασικά σημεία της κάθε μεθόδου και αναδείχθηκε ο τρόπος με τον οποίο αντιμετωπίζει η κάθε μια το πρόβλημα της πολλαπλής ομαδοποίησης. Έγινε αναφορά στον τρόπο με τον οποίο ταξινομούνται καθώς επίσης και την πρότερη γνώση που μπορεί να είναι απαραίτητη στους υπολογισμούς νέων λύσεων.

Βασικός στόχος της εργασίας ήταν να προσεγγίσει το πρόβλημα της εύρεσης πολλαπλών λύσεων με διαφορετικό τρόπο και να ορίσει μια νέα μεθοδολογία, η οποία αποτελεί επέκταση του παραδοσιακού αλγόριθμου global k-means. Παρουσιάστηκε ένας νέος αλγόριθμος, ο Multiple Clusters Global K-Means και αναλύθηκαν οι πτυχές και τα χαρακτηριστικά του. Επίσης, πραγματοποιήθηκαν πειράματα σε συνθετικά δεδομένα με στόχο να γίνει έλεγχος της προτεινόμενης μεθόδου. Παρουσιάστηκαν στη συνέχεια τα αποτελέσματα που είχε η εφαρμογή του

αλγορίθμου σε κατάτμηση εικόνας. Τέλος, η μέθοδος χρησιμοποιήθηκε για την εξαγωγή πληροφορίας από ακολουθίες βίντεο.

Από τη σειρά πειραμάτων που πραγματοποιήθηκε για την κατάτμηση εικόνων, η μέθοδος αποδείχτηκε τουλάχιστον εξίσου καλή με τον παραδοσιακό αλγόριθμο global k-means. Σε μερικές περιπτώσεις οι εναλλακτικές λύσεις ομαδοποίησης είναι δυνατό να λειτουργήσουν αθροιστικά, κάνοντας ορατά όλα τα βασικά χαρακτηριστικά της εικόνας. Επίσης οι λύσεις της προτεινόμενης μεθοδολογίας συνετέλεσαν, πολλές φορές, στην ανάδειξη χαρακτηριστικών που δεν μπόρεσαν να ανακαλυφθούν από τον απλό global k-means αλγόριθμο.

Όσον αφορά στα πειράματα που πραγματοποιήθηκαν για την εξαγωγή περίληψης από ακολουθίες βίντεο παρατηρήθηκε ότι η προτεινόμενη μεθοδολογία αντιμετωπίζει το πρόβλημα με διαφορετικό τρόπο σε σχέση με τον παραδοσιακό global k–means. Ο MCGKM συνενώνει τις πολλαπλές λύσεις που παράγονται σε κάθε βήμα, παράγοντας λύση σε πρόβλημα υψηλότερης τάξης. Ο χρόνος εκτέλεσης δεν μειώνεται ωστόσο, εξαιτίας του μεγάλου υπολογιστικού φόρτου που συνοδεύει κάθε επόμενη προσθήκη ομάδας.

5.2. Προτάσεις για Μελλοντική Έρευνα

Η μέθοδος που προτείνεται στην παρούσα εργασία θα μπορούσε να υποστεί βελτιώσεις. Ένα ζήτημα που παρουσιάζει ενδιαφέρον σχετίζεται με τη χρήση της παράλληλης επεξεργασίας, προκειμένου να μειωθεί ο χρόνος εκτέλεσης και να υπάρχει δυνατότητα επεξεργασίας μεγαλύτερου όγκου δεδομένων. Επιπλέον, η προτεινόμενη μεθοδολογία θα μπορούσε να αξιολογηθεί σε ένα μεγαλύτερο πλήθος εφαρμογών, έτσι ώστε τα αποτελέσματα να είναι πιο αξιόπιστα.

ΑΝΑΦΟΡΕΣ

[1] J. A. Hartigan, "Statistical theory in clustering", Journal of Classification, 2:63 76, 1985.

[2] D. Niu, "Multiple Alternative Clustering and Dimensionality Reduction", Dissertation, Northeastern University Boston, Massachusetts, pp. 17-18, October 2012.

[3] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, "A survey of kernel and spectral methods for clustering", Pattern Recognition (PR), vol. 41, no. 1, pp. 176–190, 2008.

[4] P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining", pp.491-492, 497, 2005.

[5] A. Jain , M. Murty, P. Flynn,"Data Clustering: A Review", Association for Computing Machinery (ACM) Computing Survey, vol. 31, no. 3, pp. 264–323, 1999.

[6] J. Bailey, "Alternative Clusterings: Current Progress and Open Challenges", MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings, July, 2010.

[7] D. Gondek, T. Hofmann, "Non-redundant Data Clustering", Knowl Inf. Syst., 12(1):1–24, 2007.

[8] E. Bae, J. Bailey, "COALA: A Novel Approach for the Extraction of an Alternate Clustering of High Quality and High Dissimilarity", ICDM '06, pp. 53–62, 2006.

[9] P. Jain, R. Meka, I. S. Dhillon, "Simultaneous Unsupervised Learning of Disparate Clusterings", SDM '08, pp. 858–869, 2008.

[10] R. Caruana, M. Elhawary, N. Nguyen, C. Smith, "Meta Clustering", ICDM '06, pp. 107–118, 2006.

[11] M. S. Hossain, N. Ramakrishnan, I. Davidson, L. T. Watson, "How to Alternatize a clustering algorithm", DMKD, August 2012.

[12] E. Müller, S. Günnemann, I. Färber , T. Seidl, "Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data", 2010.

[13] A. Patrikainen, M. Meila, "Comparing subspace clusterings", IEEE Transactions on Knowledge and Data Engineering, 18(7):902–916, 2006.

[14] J.C. Dunn, "A fuzzy relative of the isodata process and its use in detectingcompact well-separated clusters", Cybernetics and Systems, 3(3):32–57, 1973 .

[15] A. Likas, N. Vlassis, J. J. Verbeek, "The global K-means clustering algorithm" Pattern Recognition, 36(2):451-461, 2003.

[16] W. M. Rand, "Objective criteria for the evaluation of clustering methods" journal of the American statistical association 66 (336): 846–850, 1971.

[17] L.Hubert, P. Arabie, "Comparing partitions", Journal of Classification 2 (1): 193–218, 1985.

[18] L. Shapiro, G. Stockman, "Computer Vision", Prentice-Hall, pp 279-325, 2001.

[19] V. Chasanis, A. Likas, N. Galatsanos, "Efficient Video Shot Summarization Using an Enhanced Spectral Clustering Approach", Proc. International Conference on Artificial Neural Networks (ICANN '08) , Berlin, Heidelberg, Springer-Verlag, Part I, pp. 847-856,2008.

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Η Βασιλεία Σολομωνίδου γεννήθηκε στα Ιωάννινα το 1986. Το 2004 εισήχθη στο τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας στη Θεσσαλονίκη από το οποίο αποφοίτησε με βαθμό 7.44 το 2009. Παρακολούθησε το Πρόγραμμα Μεταπτυχιακών Σπουδών του Τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων από τον Οκτώβριο του 2010 και αποφοίτησε τον Ιούλιο του 2013 αποκτώντας Μεταπτυχιακό Δίπλωμα με ειδίκευση στις «Τεχνολογίες-Εφαρμογές»