ΜΕΘΟΔΟΣ ΝΕΥΤΩΝΕΙΑΣ ΦΑΣΜΑΤΙΚΗΣ ΟΜΑΔΟΠΟΙΗΣΗΣ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ - ΑΞΙΟΛΟΓΗΣΗ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από την

Κυριακή Χριστοδουλίδου

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Ιούλιος 2009

ΑΦΙΕΡΩΣΗ

Στην οικογένειά μου

ΕΥΧΑΡΙΣΤΙΕΣ

Η διατριβή αυτή εκπονήθηκε στο Τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων με επιβλέποντα τον κ. Κωνσταντίνο Μπλέκα.

Θα ήθελα, καταρχήν, να ευχαριστήσω θερμά τον κ. Κωνσταντίνο Μπλέκα, επιβλέποντα της διατριβής μου και καθηγητή μου στο Τμήμα Πληροφορικής, για την υπομονή που επέδειξε και για τη συνεχή και άψογη καθοδήγησή του καθ' όλη τη διάρκεια της συνεργασίας μας.

Ακόμη, αισθάνομαι την ανάγκη να ευχαριστήσω τους γονείς μου και την αδελφή μου, που ήταν πάντα δίπλα μου με την αγάπη, την κατανόηση και την υποστήριξη που μου προσέφεραν, καθώς και με την υπομονή που επιδείκνυαν.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Κυριακή Χριστοδουλίδου του Θεόδωρου και της Χρυσούλας.
MSc, ΤμήμαΠληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιούλιος, 2009.
Νευτώνεια Φασματική Ομαδοποίηση.
Επιβλέπων: Κωνσταντίνος Μπλέκας.

Η διατριβή αυτή περιέχει πέντε κεφάλαια. Στο πρώτο κεφάλαιο αναλύουμε την έννοια της ομαδοποίησης και τη βασική κατηγοριοποίηση των αλγορίθμων ομαδοποίησης. Στο δεύτερο κεφάλαιο παρουσιάζουμε διεξοδικά τη μέθοδο της Φασματικής Ομαδοποίησης. Η παραπάνω μέθοδος χρησιμοποιεί πληροφορία από τα ιδιοδιανύσματα και τις ιδιοτιμές των πινάκων ομοιότητας. Στο τρίτο κεφάλαιο παρουσιάζουμε τη μέθοδο της Νευτώνειας Ομαδοποίησης, που βασίζεται στις εξισώσεις κίνησης του Νεύτωνα. Η εφαρμογή της έχει ως αποτέλεσμα τη συγκέντρωση καθενός από τα σημεία του συνόλου δεδομένων γύρω από το κέντρο της ομάδας στην οποία ανήκει. Στο τέταρτο κεφάλαιο παρουσιάζουμε τη μέθοδο της Νευτώνειας Φασματικής Ομαδοποίησης που προτείνουμε, η οποία συνδυάζει τις δύο προηγούμενες μεθόδους. Η μέθοδός μας, μέσω της πληροφορίας που παρέχει η Νευτώνεια Ομαδοποίηση, οδηγεί στην κατασκευή ενός αραιού πίνακα ομοιότητας που μπορεί να χρησιμοποιηθεί στην επίλυση προβλημάτων Φασματικής Ομαδοποίησης. Στο πέμπτο κεφάλαιο ακολουθεί η πειραματική μελέτη και αξιολόγηση της προτεινόμενης μεθόδου. Τα πειραματικά αποτελέσματα που παρουσιάζονται δείχνουν την υπεροχή της μεθόδου που προτείνουμε σε σχέση με τις κλασικές μεθόδους ομαδοποίησης.

EXTENDED ABSTRACT IN ENGLISH

Christodoulidou Kyriaki, T.
MSc, Computer Science Department, University of Ioannina, Greece. July, 2009.
Newtonian Spectral Clustering
Thesis Supervisor: Konstadinos Blekas.

This dissertation consists of five chapters. In the first chapter, firstly we analyze the concept of clustering. Moreover, we present some of the basic clustering algorithms that have been proposed in the literature. In the second chapter, we describe thoroughly the method of Spectral Clustering. This method uses the information derived from the eigenvectors and the eigenvalues of the similarity matrices in order to discover the data clusters. In the third chapter, we describe the method of Newtonian Clustering. This method is based in the Newton's equations of motion. The application of this method results in each data point shrinking around its center. As a result, the data clusters are more obvious. In the fourth chapter, we describe the method of Newtonian Spectral Clustering, which we propose. This method combines the two methods previously mentioned. It employs the information derived from the Newtonian Clustering and leads in the construction of a sparse similarity matrix. Lastly, in the fifth chapter we present an experimental evaluation of the proposed method. It is experimentally shown that our method outperforms the performance of the classical clustering algorithms.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1. Ομαδοποίηση

Η ανάλυση δεδομένων είναι μια διαδικασία η οποία διέπει πολλές υπολογιστικές εφαρμογές. Βασικό στοιχείο της διαδικασίας αυτής είναι η ομαδοποίηση (*clustering*) ή κατηγοριοποίηση (*classification*) των δεδομένων που μπορεί να προκύψει κατά τη διάρκειά της. Ομαδοποίηση είναι η διαίρεση του συνόλου δεδομένων σε ομάδες όμοιων αντικειμένων. Πιο συγκεκριμένα, μία ομάδα (*cluster*) παρουσιάζει μία δομή στο χώρο. Η παραπάνω δομή προσδίδει εσωτερική συνοχή στην ομάδα σε σχέση με τα άλλα δεδομένα του συνόλου. Επομένως, κάθε ομάδα αποτελείται από δεδομένα που είναι όμοια μεταξύ τους και ανόμοια ως προς τα δεδομένα των άλλων ομάδων.

Είναι σημαντικό να κατανοήσουμε τη διαφορά ανάμεσα στην ομαδοποίηση, που είναι κατηγοριοποίηση χωρίς επίβλεψη, και στην κατηγοριοποίηση με επίβλεψη. Στην περίπτωση της ομαδοποίησης στόχος είναι η διαμέριση μίας συλλογής προτύπων (παρατηρήσεων, δεδομένων-αντικείμενων ή διανυσμάτων χαρακτηριστικών) χωρίς ετικέτα σε ομάδες. Αντίθετα, στην κατηγοριοποίηση με επίβλεψη στόχος είναι, δεδομένης μιας συλλογής από πρότυπα με ετικέτες (σύνολο εκπαίδευσης), να δοθεί ετικέτα σε ένα νέο πρότυπο χωρίς ετικέτα. Οι ετικέτες, σε αντίθεση με τις ομάδες, παρέχονται αποκλειστικά από τα δεδομένα.

Η ομαδοποίηση αποτελεί το αντικείμενο έρευνας πολλών πεδίων, ανάμεσά τους, η εξόρυξη δεδομένων (*data mining*) [25], η στατιστική, η αναγνώριση προτύπων (*pattern recognition*) [11, 13], η μηχανική μάθηση (*machine learning*) [3], η υπολογιστική όραση (*computer vision*) [8, 21, 24], και η ανάλυση εικόνας (*image analysis*) [6].

Η ομαδοποίηση είναι μια από τις πιο ευρέως διαδεδομένες τεχνικές σε εφαρμογές εξόρυξης δεδομένων που κυμαίνονται από την επιστημονική εξερεύνηση δεδομένων, στην ανάκτηση πληροφορίας και στην εξόρυξη κειμένου [10], στις ιατρικές διαγνώσεις, στην ανάλυση DNA στη βιοπληροφορική [2, 15, 22], στην ανάλυση Δικτύου και πολλές άλλες. Επομένως, η εξόρυξη δεδομένων αντιμετωπίζει μεγάλες βάσεις δεδομένων με μεγάλες υπολογιστικές απαιτήσεις.

1.1.1. Κατάρα της διάστασης

Ένα σημαντικό πρόβλημα στη Μηχανική Μάθηση είναι η κατάρα της διάστασης (*curse of dimensionality*). Η κατάρα της διάστασης αναφέρεται στο φαινόμενο σύμφωνα με το οποίο καθώς αυξάνει η διάσταση, τα δεδομένα γίνονται πιο αραιά στο χώρο που καταλαμβάνουν. Επομένως, όταν έχουμε δεδομένα μεγάλης διάστασης, η επίδοση των τεχνικών ομαδοποίησης δεν είναι καλή. Μία μέθοδος για να αντιμετωπιστεί το παραπάνω πρόβλημα είναι να εφαρμόσουμε τεχνικές μείωσης της διάστασης.

Η μείωση της διάστασης έχει και περαιτέρω σημαντικά πλεονεκτήματα. Αρχικά, μπορεί να εξαλείψει μη-σχετικά χαρακτηριστικά των δεδομένων και να μειώσει, κατά ένα μέρος, τον θόρυβο. Επιπλέον, μπορεί να οδηγήσει σε ένα πιο κατανοητό μοντέλο καθώς, πλέον, το μοντέλο θα περιέχει λιγότερα χαρακτηριστικά.

Οι τεχνικές μείωσης της διάστασης, όπως προαναφέραμε, εφαρμόζονται σε σύνολα δεδομένων με μεγάλο αριθμό χαρακτηριστικών. Για παράδειγμα, σε ένα σύνολο κειμένων όπου το κάθε κείμενο αναπαρίσταται από ένα διάστημα τα στοιχεία του οποίου είναι οι συχνότητες εμφάνισης της κάθε λέξης στο κείμενο αυτό. Σε τέτοιες

περιπτώσεις, υπάρχουν χιλιάδες χαρακτηριστικά (ένα για κάθε λέξη που χρησιμοποιείται στο λεξικό).

1.1.2. Βασικά βήματα της ομαδοποίησης

Η διαδικασία της ομαδοποίησης περιλαμβάνει τα ακόλουθα βήματα:

(1). Αναπαράσταση των προτύπων. Στο στάδιο αυτό περιλαμβάνεται και η εξαγωγή χαρακτηριστικών (*feature extraction*) και/ή επιλογή χαρακτηριστικών (*feature selection*).

(2). Επιλογή ενός μέτρου ομοιότητας κατάλληλου για το πεδίο ορισμού των δεδομένων.

(3). Ομαδοποίηση.

(4). Αφαίρεση δεδομένων (data abstraction).

(5). Αξιολόγηση του αποτελέσματος.

Με τον όρο αναπαράσταση των προτύπων αναφερόμαστε στο πλήθος των ομάδων, στο πλήθος των διαθέσιμων προτύπων, καθώς, επίσης, και στο πλήθος, το είδος και την κλίμακα των χαρακτηριστικών τα οποία είναι διαθέσιμα σε έναν αλγόριθμο ομαδοποίησης προς αξιοποίηση, ελεγχόμενα ή μη. Η επιλογή χαρακτηριστικών είναι η διαδικασία της αναγνώρισης του πιο σημαντικού υποσυνόλου χαρακτηριστικών των δεδομένων που μπορούν να χρησιμοποιηθούν κατά την ομαδοποίηση. Η εξαγωγή χαρακτηριστικών είναι η χρήση ενός ή περισσοτέρων μετασχηματισμών χαρακτηριστικών εισόδου προκειμένου να παραχθούν νέα salient χαρακτηριστικά. Οι τεχνικές αυτές χρησιμοποιούνται για την εύρεση ενός κατάλληλου συνόλου χαρακτηριστικών τα οποία μπορούν να χρησιμοποιηθούν κατά την ομαδοποίηση.

Η επιλογή του μέτρου ομοιότητας, όπως θα δούμε αναλυτικότερα και στη συνέχεια, καθορίζει τον τρόπο υπολογισμού της ομοιότητας δύο σημείων.

Η ομαδοποίηση, δηλαδή η διαίρεση του συνόλου δεδομένων σε ομάδες όμοιων αντικειμένων, μπορεί να γίνει με πολλούς τρόπους. Ο τρόπος ο οποίος ακολουθείται εξαρτάται από τον αλγόριθμο ομαδοποίησης που χρησιμοποιείται σε κάθε περίπτωση.

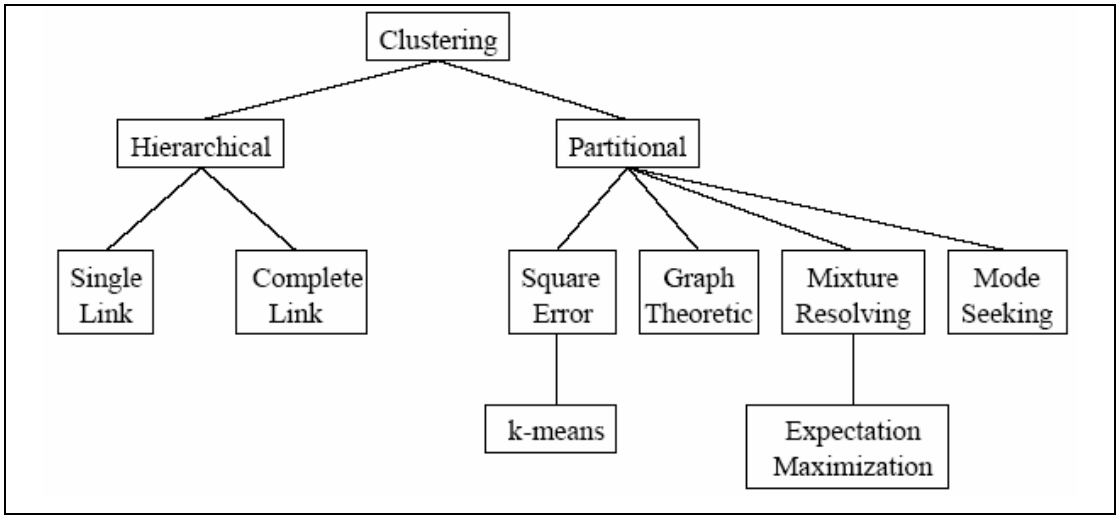
Αφαίρεση δεδομένων είναι η διαδικασία εξαγωγής μιας απλής και συμπαγούς αναπαράστασης του συνόλου δεδομένων.

Η αξιολόγηση του αποτελέσματος έγκειται στην εκτίμηση του αποτελέσματος της ομαδοποίησης, δηλαδή στο κατά πόσο η ομαδοποίηση που προκύπτει είναι καλή ή όχι.

Παρακάτω, αναφέρουμε τις πιο ευρέως διαδεδομένες μεθόδους ομαδοποίησης [18] στο πεδίο της εξόρυξης δεδομένων.

1.1.3. Κατηγοριοποίηση των αλγορίθμων ομαδοποίησης

Στο Σχήμα 1.1 παρουσιάζουμε μία ταξινόμηση των διαφόρων προσεγγίσεων της ομαδοποίησης δεδομένων. Στο κορυφαίο επίπεδο, οι αλγόριθμοι ομαδοποίησης χωρίζονται σε ιεραρχικούς (*hierarchical*) και διαμεριστικούς (*partitional*). Οι ιεραρχικοί αλγόριθμοι παράγουν μία εμφωλευμένη σειρά διαμερίσεων. Αντίθετα, οι διαμεριστικοί αλγόριθμοι παράγουν μία μόνο διαμέριση.



Σχήμα 1.1 Ταξινόμηση των προσεγγίσεων ομαδοποίησης.

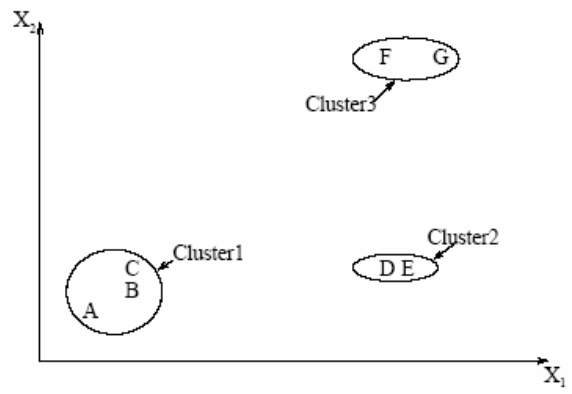
Πριν την ανάλυση της παραπάνω ταξινόμησης, αξίζει να αναφέρουμε την διαφορά μεταξύ αλγορίθμων συσσώρευσης και διαιρετών αλγορίθμων η οποία επηρεάζει τις παραπάνω προσεγγίσεις ανεξάρτητα από τη θέση τους στην ταξινόμηση.

Στους αλγορίθμους συσσώρευσης (*agglomerative*), αρχικά το κάθε στοιχείο αποτελεί ξεχωριστή ομάδα, ενώ στη συνέχεια ακολουθούν συγχωνεύσεις των στοιχείων σε διαδοχικά μεγαλύτερες ομάδες (*bottom-up*). Στους διαιρετούς αλγορίθμους (*divisive*), αρχικά υπάρχει μία ομάδα που αποτελείται από το σύνολο των στοιχείων, ενώ στη συνέχεια ακολουθούν διαχωρισμοί σε διαδοχικά μικρότερες ομάδες (*top-down*). Η διαδικασία συνεχίζεται μέχρι να ισχύσει ένα κριτήριο τερματισμού (συνήθως, ο αριθμός των ομάδων να ισούται με K).

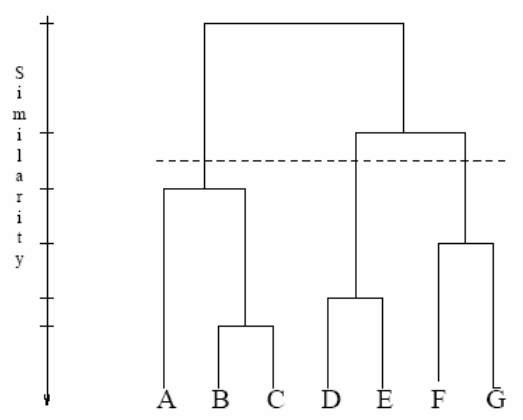
1.1.3.1. Ιεραρχικοί αλγόριθμοι ομαδοποίησης

Ο τρόπος λειτουργίας ενός ιεραρχικού αλγορίθμου ομαδοποίησης μπορεί να παρουσιαστεί χρησιμοποιώντας το παράδειγμα δεδομένων του Σχήματος 1.2. Συγκεκριμένα, παρουσιάζονται 7 πρότυπα με ετικέτες A, B, C, D, E, F και G μέσα σε τρεις ομάδες. Ένας ιεραρχικός αλγόριθμος κατασκευάζει τις ομάδες σταδιακά οργανώνοντας τα δεδομένα σε μία δενδρική δομή.

Στο Σχήμα 1.3 παρουσιάζουμε το δενδρόγραμμα που αντιστοιχεί στα 7 σημεία του Σχήματος 1.2.



Σχήμα 1.2 Εφτά σημεία που ανήκουν σε τρεις ομάδες.



Σχήμα 1.3 Το δενδρόγραμμα που αντιστοιχεί στα σημεία του Σχήματος 1.2.

Οι περισσότεροι ιεραρχικοί αλγόριθμοι διακρίνονται σε *single-link* και *complete-link* αλγορίθμους. Στους single-link αλγορίθμους η απόσταση μεταξύ δύο ομάδων είναι η ελάχιστη μεταξύ των αποστάσεων όλων των ζευγών προτύπων που προκύπτουν από τις δύο ομάδες. Στους complete-link αλγορίθμους η απόσταση μεταξύ δύο ομάδων είναι η μέγιστη όλων των ανά ζεύγη αποστάσεων μεταξύ των προτύπων των δύο ομάδων. Οι complete-link αλγόριθμοι παράγουν ιδιαίτερα συμπαγείς ομάδες. Αντίθετα, οι single-link αλγόριθμοι έχουν την τάση να παράγουν ανάκατες ή διαμήκεις ομάδες.

Ο *agglomerative single-link* αλγόριθμος ομαδοποίησης:
(1). Τοποθετούμε κάθε πρότυπο στην ομάδα του. Κατασκευάζουμε μία λίστα αποστάσεων για όλα τα ζεύγη προτύπων και την ταξινομούμε κατά αύξουσα σειρά.
(2). Κατασκευάζουμε για κάθε τιμή απόστασης d_k ένα γράφο προτύπων όπου τα ζεύγη των προτύπων που απέχουν απόσταση μικρότερη από d_k συνδέονται με ακμή. Αν όλα τα πρότυπα είναι μέλη ενός συνδεδεμένου γράφου προχωράμε στο βήμα (3) αλλιώς, επαναλαμβάνουμε το βήμα (2).
(3). Η έξοδος του αλγορίθμου είναι μία εμφωλευμένη ιεραρχία γράφων. Η παραπάνω ιεραρχία μπορεί να διαχωριστεί σε απλές συνδεδεμένες συνιστώσες του γράφου (ομαδοποίηση).

Ο *agglomerative complete-link* αλγόριθμος ομαδοποίησης:

(1). Τοποθετούμε κάθε πρότυπο στην ομάδα του. Κατασκευάζουμε μία λίστα αποστάσεων για όλα τα ζεύγη προτύπων και την ταξινομούμε κατά αύξουσα σειρά.
(2). Κατασκευάζουμε για κάθε τιμή απόστασης d_k ένα γράφο προτύπων όπου τα ζεύγη των προτύπων που απέχουν απόσταση μικρότερη από d_k συνδέονται με ακμή. Αν όλα τα πρότυπα είναι μέλη ενός πλήρους συνδεδεμένου γράφου προχωράμε στο βήμα (3) αλλιώς, επαναλαμβάνουμε το βήμα (2).
(3). Η έξοδος του αλγορίθμου είναι μία εμφωλευμένη ιεραρχία γράφων. Η παραπάνω ιεραρχία μπορεί να διαχωριστεί σε πλήρεις συνδεδεμένες συνιστώσες του γράφου (ομαδοποίηση).

Οι ιεραρχικοί αλγόριθμοι ομαδοποίησης είναι περισσότερο ευπροσάρμοστοι από τους διαμεριστικούς. Για παράδειγμα, ο *single-link* αλγόριθμος ομαδοποίησης παρουσιάζει καλή επίδοση για σύνολα δεδομένων που ανήκουν σε μη-ισοτροπικές ομάδες (όπως είναι οι καλώς-διαχωρίσιμες ομάδες, οι ομόκεντρες, καθώς και οι ομάδες με τη μορφή αλυσίδας), ενώ ένας διαμεριστικός αλγόριθμος, όπως είναι ο K-means που περιγράφεται στη συνέχεια, έχει καλή επίδοση για δεδομένα που ανήκουν σε ισοτροπικές ομάδες. Αντίθετα, η χρονική και χωρική πολυπλοκότητα των διαμεριστικών αλγορίθμων είναι μικρότερη από αυτή των ιεραρχικών.

Ο ιεραρχικός *agglomerative* αλγόριθμος ομαδοποίησης:
(1). Υπολογίζουμε τον πίνακα ομοιότητας (πίνακας αποστάσεων μεταξύ του κάθε ζεύγους προτύπων). Θεωρούμε ότι κάθε πρότυπο αποτελεί μία ομάδα.
(2). Βρίσκουμε τα πιο όμοια ζεύγη ομάδων (με βάση τον πίνακα ομοιότητας) και τα συγχωνεύουμε σε μία ομάδα. Στη συνέχεια, ενημερώνουμε τον πίνακα ομοιότητας.
(3). Ο αλγόριθμος σταματά αν όλα τα πρότυπα είναι μέλη μίας ομάδας. Αλλιώς, επαναλαμβάνουμε το βήμα (2).

1.1.3.2. Διαμεριστικοί αλγόριθμοι ομαδοποίησης

Οι διαμεριστικοί αλγόριθμοι, σε αντίθεση με τους ιεραρχικούς, καθορίζουν όλες τις ομάδες ταυτόχρονα. Πιο συγκεκριμένα, προσπαθούν να ανακαλύψουν ομάδες βελτιστοποιώντας μία συνάρτηση-κριτήριο ορισμένη είτε τοπικά (σε ένα υποσύνολο προτύπων), είτε καθολικά (ορισμένη πάνω σε όλα τα πρότυπα). Πρακτικά, οι

διαμεριστικοί αλγόριθμοι εκτελούνται επαναληπτικά βελτιώνοντας έτσι σταδιακά τις ομάδες, με αποτέλεσμα τελικά να έχουμε πολύ καλά αποτελέσματα. Παρόλο, βέβαια, που υπερτερούν σε εφαρμογές όπου υπάρχουν μεγάλα σύνολα δεδομένων, ένα μειονέκτημα που συνδέεται με τη χρήση τους είναι ότι πρέπει να επιλεχθεί ο αριθμός των ομάδων.

Οι αλγόριθμοι τετραγωνικού σφάλματος (*squared error*): Το πιο διαδεδομένο κριτήριο στις διαμεριστικές τεχνικές ομαδοποίησης είναι το κριτήριο του τετραγωνικού σφάλματος που τείνει να δίνει καλά αποτελέσματα σε περιπτώσεις δεδομένων με απομονωμένες ή συμπαγείς ομάδες. Το τετραγωνικό σφάλμα της L ομαδοποίησης ενός συνόλου K ομάδων είναι ίσο με:

$$e^2(K,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| x_i^{(j)} - c_j \right\|^2, \qquad \text{Εξ. 1.1}$$

όπου, $x_i^{(j)}$ είναι το i-οστό πρότυπο και c_j το κεντροειδές της j-οστής ομάδας.

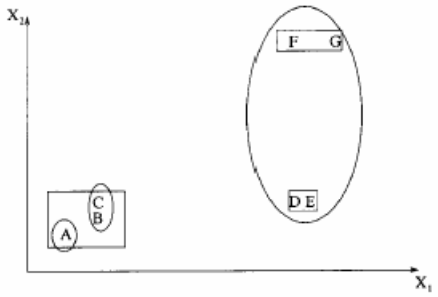
Ο K-means είναι ο πιο διαδεδομένος και πιο απλός αλγόριθμος τετραγωνικού σφάλματος.

Ο αλγόριθμος K-means: Ο K-means είναι ο πιο ευρέως διαδεδομένος και πιο απλός αλγόριθμος τετραγωνικού σφάλματος. Ο αλγόριθμος ξεκινά με μία τυχαία διαμέριση των προτύπων σε ομάδες και στη συνέχεια, επανατοποθετεί τα πρότυπα στις ομάδες με βάση την ομοιότητα μεταξύ αυτών και των κέντρων των ομάδων. Έπειτα, τα κέντρα κάθε ομάδας ενημερώνονται με βάση τα σημεία που ανατίθενται στη συγκεκριμένη ομάδα. Επαναλαμβάνουμε τις αναθέσεις και τις ενημερώσεις μέχρι να παραμένει αμετάβλητη η ομάδα καθενός σημείου ή μέχρι το τετραγωνικό σφάλμα να μειωθεί σημαντικά. Στο Σχήμα 1.4 παρουσιάζουμε τα βήματα του K-means αλγορίθμου.

Αλγόριθμος K-means
1: επιλέγουμε K κέντρα έτσι ώστε να συμπίπτουν με K τυχαία επιλεγμένα πρότυπα. 2: επαναλαμβάνουμε 3: δημιουργούμε K ομάδες αναθέτοντας κάθε πρότυπο στο κοντινότερό του κέντρο. 4: ενημερώνουμε το κέντρο της κάθε ομάδας. 5: ***μέχρι*** τα κέντρα να μην μεταβάλλονται ή την ελαχιστοποίηση του τετραγωνικού σφάλματος.

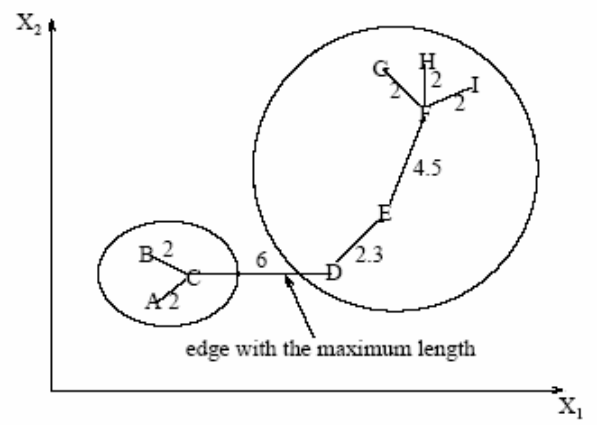
Σχήμα 1.4 Βήματα αλγορίθμου K-means.

Ο αλγόριθμος K-means είναι ιδιαίτερα διαδεδομένος καθώς έχει χρονική πολυπλοκότητα O(n) (όπου n ο αριθμός των προτύπων). Ένα βασικό μειονέκτημα του K-means είναι ότι το αποτέλεσμα της εφαρμογής του εξαρτάται από την αρχική διαμέριση. Επομένως, αν η αρχική διαμέριση δεν έχει επιλεχθεί σωστά μπορεί ο αλγόριθμος να συγκλίνει σε τοπικό ελάχιστο. Στο Σχήμα 1.5 μπορούμε να διακρίνουμε τη σημασία της αρχικής διαμέρισης για τον K-means. Συγκεκριμένα, στ παράδειγμα υπάρχουν 7 πρότυπα. Αν ξεκινήσουμε την εφαρμογή του K-means έχοντας τα πρότυπα A, B και C ως αρχικά κέντρα γύρω από τα οποία θα δημιουργηθούν οι τρεις ομάδες, η τελική μας διαμέριση θα είναι {{A}, {B, C}, {D, E, F, G}}. Όμως, στην παραπάνω περίπτωση, το τετραγωνικό σφάλμα είναι πολύ μεγαλύτερο σε σχέση με αυτό που έχουμε όταν η τελική διαμέριση είναι η βέλτιστη δηλαδή, {{A, B, C}, {D, E}, {F, G}}. Η σωστή τελική διαμέριση παρέχεται όταν για παράδειγμα επιλεχθούν ως αρχικά κέντρα των ομάδων τα πρότυπα A, D και F.



Σχήμα 1.5 Η εξάρτηση του K-means από την αρχική διαμέριση.

Γραφοθεωρητική ομαδοποίηση: Ο πιο γνωστός *divisive* γραφοθεωρητικός αλγόριθμος ομαδοποίησης βασίζεται στην κατασκευή του ελάχιστου σκελετικού δέντρου (*Minimal Spanning Tree* ή *MST*) και στη μετέπειτα διαγραφή των MST ακμών με τα μεγαλύτερα μήκη. Στο Σχήμα 1.6 παρουσιάζουμε το MST που προήλθε από ένα σύνολο 9 δεδομένων. Πιο αναλυτικά, διαγράφεται η ακμή CD μήκους 6 μονάδων και δημιουργούνται δύο νέες ομάδες ({A, B, C} και {D, E, F, G, H, I}).



Σχήμα 1.6 Δημιουργία ομάδων μέσω του ελάχιστου σκελετικού δέντρου.

Οι mixture-resolving και mode-seeking αλγόριθμοι: Στους mixture-resolving αλγορίθμους υποθέτουμε ότι τα πρότυπα που πρόκειται να ομαδοποιηθούν επιλέγονται από μία μεταξύ πολλών άλλων κατανομών. Με άλλα λόγια, οι παραπάνω μέθοδοι υποθέτουν ότι τα δεδομένα δημιουργούνται από μια μίξη κατανομών πιθανότητας στην οποία, κάθε συνιστώσα αντιστοιχεί σε μία διαφορετική ομάδα. Στόχος τους είναι η αναγνώριση των παραμέτρων της κάθε κατανομής και, ίσως, η εύρεση του πλήθους των κατανομών. Συνήθως, υποθέτουμε πως οι ξεχωριστές συνιστώσες είναι Γκαουσιανές κατανομές. Στις προαναφερθείσες μεθόδους, συχνά επιλέγουμε για τον συντονισμό των μικτών παραμέτρων τον αλγόριθμο πρόβλεψης-μεγιστοποίησης (*Expectation Maximization ή EM*). Ο αλγόριθμος ΕΜ είναι μία γενική μέθοδος εύρεσης εκτιμητών μέγιστης πιθανοφάνειας (*maximum likelihood estimators*) των παραμέτρων μιας δοθείσας κατανομής, σε προβλήματα στα οποία κάποιες μεταβλητές δεν έχουν παρατηρηθεί. Στην πράξη είναι πιο βολικό να μεγιστοποιήσουμε τον λογάριθμο της συνάρτησης πιθανοφάνειας.

Ομαδοποίηση μέσω Τεχνητών Νευρωνικών Δικτύων (*ANNs*): Τα σχήματα ομαδοποίησης που βασίζονται σε ANNs αποτελούν νευρωνικές εφαρμογές των αλγορίθμων ομαδοποίησης. Συνήθως χρησιμοποιούνται οι competitive neural networks αλγόριθμοι ομαδοποίησης στους οποίους όμοια πρότυπα ομαδοποιούνται από το δίκτυο και αναπαρίστανται μέσω ενός νευρώνα. Τα πιο γνωστά ANNs έχουν ένα μόνο επίπεδο, τα πρότυπα που παρουσιάζονται στην είσοδο του δικτύου και σχετίζονται με τους κόμβους εξόδου, ενώ τα βάρη μεταξύ των κόμβων εισόδου και εξόδου μεταβάλλονται διαδοχικά, έως ότου πληρείται ένα κριτήριο τερματισμού (*μάθηση*).

Ασαφής ομαδοποίηση (*fuzzy clustering*): Οι παραδοσιακοί αλγόριθμοι ομαδοποίησης παράγουν μία διαμέριση στην οποία κάθε πρότυπο ανήκει σε μία μόνο ομάδα (*hard clustering*). Αντίθετα, οι ασαφείς αλγόριθμοι ομαδοποίησης συνδέουν το κάθε πρότυπο με όλες τις υπάρχουσες ομάδες μέσω ενός βαθμού συμμετοχής. Επομένως, στην τελική ομαδοποίηση υπάρχουν επικαλυπτόμενες ομάδες. Το σημαντικότερο πρόβλημα στην ασαφή ομαδοποίηση είναι ο σχεδιασμός των συναρτήσεων που καθορίζουν το βαθμό συμμετοχής κάθε προτύπου σε κάθε ομάδα. Ο πιο διαδεδομένος ασαφής αλγόριθμος ομαδοποίησης είναι ο *ασαφής c-means* αλγόριθμος (*FCM* αλγόριθμος), ο οποίος είναι καλύτερος από τον παραδοσιακό K-means αλγόριθμο ως προς την αποφυγή τοπικών ελαχίστων. Οι παραπάνω αλγόριθμοι μπορούν να χρησιμοποιηθούν για ανάκτηση κειμένων (*document retrieval*) ή όταν έχουμε μικτούς τύπους δεδομένων. Ένα μειονέκτημα, πάντως, της ασαφούς ομαδοποίησης είναι το πλήθος υπολογισμών που απαιτεί, οπότε και δεν ενδείκνυται για μεγάλα σύνολα δεδομένων.

Εξελικτικές προσεγγίσεις (*evolutionary approaches*) για ομαδοποίηση: Οι προσεγγίσεις αυτές βασίζονται στη θεωρία της φυσικής εξέλιξης και χρησιμοποιούν εξελικτικούς τελεστές (*evolutionary operators*), καθώς και ένα υποψήφιο σύνολο λύσεων, που μπορεί να είναι τυχαίες k-διαμερίσεις ων δεδομένων, έτσι ώστε να επιτύχουμε την ολικά βέλτιστη διαμέριση των δεδομένων. Οι υποψήφιες λύσεις του προβλήματος ομαδοποίησης κωδικοποιούνται ως χρωμοσώματα, ενώ οι εξελικτικοί τελεστές είναι η επιλογή (*selection*), ο ανασυνδυασμός (*recombination*) και η μετάλλαξη (*mutation*). Κάθε ένας από τους παραπάνω τελεστές μετατρέπει ένα

χρωμόσωμα εισόδου σε ένα ή περισσότερα χρωμοσώματα εξόδου, που αποτελούν την επόμενη γενιά υποψήφιων λύσεων, ενώ μία συνάρτηση (*fitness function*) που εφαρμόζεται σε κάθε νέο χρωμόσωμα καθορίζει την πιθανότητα που έχει το χρωμόσωμα αυτό να επιβιώσει μέχρι την επόμενη γενιά (όπως και στην περίπτωση της φυσικής εξέλιξης). Η διαδικασία συνεχίζεται έως ότου πληρείται κάποιο κριτήριο τερματισμού. Οι εξελικτικές προσεγγίσεις περιλαμβάνουν τους γενετικούς αλγορίθμους (*genetic algorithms* ή *GAs*), τις στρατηγικές εξέλιξης (*evolutionary strategies* ή *ESs*) και τον εξελικτικό προγραμματισμό (*evolutionary programming* ή *EP*). Αξίζει να σημειωθεί πως η επίδοση όλων των αλγορίθμων αυτής της κατηγορίας εξαρτάται από την επιλογή των παραμέτρων.

1.1.4. Εφαρμογές των αλγορίθμων ομαδοποίησης

Στη συνέχεια, παρουσιάζουμε συνοπτικά κάποια από τα πεδία στα οποία αξιοποιούνται πρακτικά οι αλγόριθμοι ομαδοποίησης. Στα πεδία αυτά περιλαμβάνονται: (1) η κατάτμηση εικόνας (*image segmentation*), (2) η αναγνώριση αντικειμένων και χαρακτήρων (*object and character recognition*), (3) η ανάκτηση κειμένων (*document retrieval*) και (4) η εξόρυξη δεδομένων (*data mining*).

Στην κατάτμηση εικόνας, επιχειρούμε να διαμερίσουμε μια εικόνα σε περιοχές, κάθε μια από τις οποίες παρουσιάζει μία ομοιογένεια ως προς κάποια ιδιότητα της εικόνας που μας ενδιαφέρει (π.χ. ένταση, χρώμα, υφή). Στην αναγνώριση αντικειμένων, επιχειρούμε να ομαδοποιήσουμε όψεις τρισδιάστατων αντικειμένων, δηλαδή εικόνες αντικειμένων οι οποίες έχουν ληφθεί από διαφορετικό σημείο παρατήρησης, έτσι ώστε να αναγνωρίσουμε τελικά το ίδιο το αντικείμενο. Στην αναγνώριση χαρακτήρων, επιχειρούμε να αναγνωρίσουμε χειρόγραφους χαρακτήρες, ανεξαρτήτως του τρόπου γραφής του κειμένου. Στην ανάκτηση πληροφορίας, επιχειρούμε να κατηγοριοποιήσουμε κείμενα βάσει κάποιων κριτηρίων. Η εξόρυξη δεδομένων, τέλος, επιτρέπει την εξαγωγή χρήσιμων πληροφοριών που μας ενδιαφέρουν από ένα τεράστιο σύνολο δεδομένων. Το πλήθος αυτών των πολλών και διαφορετικών εφαρμογών αναδεικνύει τη μεγάλη χρησιμότητα των αλγορίθμων ομαδοποίησης με τους οποίους ασχολούμαστε στην παρούσα διατριβή.

Στη συνέχεια, περιγράφουμε τους στόχους της διατριβής αυτής, καθώς και τη δομή της.

1.2. Στόχοι της Διατριβής

Στόχος της μελέτης μας είναι, σε πρώτη φάση, η παρουσίαση μιας συστηματικής μεθοδολογίας, της Νευτώνειας Φασματικής Ομαδοποίησης (*NSC*), που βασίζεται στη Φασματική Ομαδοποίηση μέσω Νευτώνειων δυνάμεων. Η μέθοδός μας, μέσω των Νευτώνειων εξισώσεων κίνησης, οδηγεί στην κατασκευή ενός αραιού πίνακα ομοιότητας που μπορεί να χρησιμοποιηθεί στην επίλυση προβλημάτων Φασματικής Ομαδοποίησης.

Στην συνέχεια, αφού τροποποιήσαμε ελαφρά την μέθοδό μας, επεκτείναμε τη μελέτη μας στην αντιμετώπιση του προβλήματος της ομαδοποίησης δεδομένων μεγάλης διάστασης. Συγκεκριμένα, εφαρμόσαμε τη μέθοδό μας σε προβλήματα ομαδοποίησης κειμένων.

Έπειτα, στόχος μας είναι η αποτίμηση της προτεινομένης μεθόδου. Εξετάσαμε τη μέθοδό μας σε ευρέως γνωστές δοκιμασίες επιδόσεως που κυμαίνονται από συνεχή δεδομένα σε κατάτμηση εικόνας και σε προβλήματα ομαδοποίησης κειμένων. Ακόμα, συγκρίναμε τα πειραματικά αποτελέσματα της μεθόδου με αυτά της καθιερωμένης μεθόδου που βασίζεται στη Φασματική Ομαδοποίηση και του αλγορίθμου K-means.

1.3. Δομή της Διατριβής

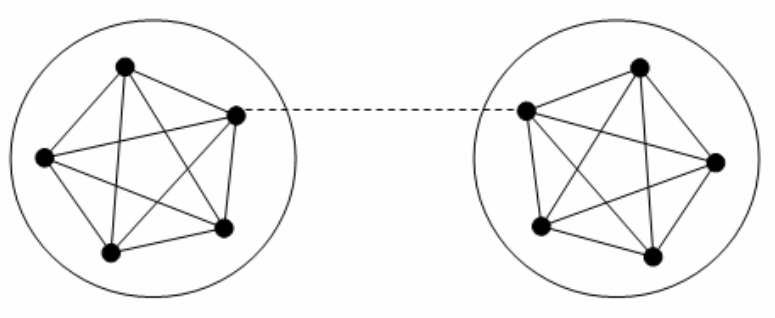
Η διατριβή περιέχει πέντε κεφάλαια. Στο πρώτο κεφάλαιο αναλύουμε την έννοια της ομαδοποίησης. Στο δεύτερο κεφάλαιο παρουσιάζουμε διεξοδικά την μέθοδο της Φασματικής Ομαδοποίησης. Στο τρίτο κεφάλαιο παρουσιάζουμε τη μέθοδο της Νευτώνειας Ομαδοποίησης. Στο τέταρτο κεφάλαιο παρουσιάζουμε και προτείνουμε τη μέθοδο της Νευτώνειας Φασματικής Ομαδοποίησης, που συνδυάζει τις δύο παραπάνω μεθόδους. Στο πέμπτο κεφάλαιο ακολουθεί η πειραματική μελέτη και αξιολόγηση της προτεινόμενης μεθόδου.

ΚΕΦΑΛΑΙΟ 2. ΦΑΣΜΑΤΙΚΗ ΟΜΑΔΟΠΟΙΗΣΗ

2.1. Εισαγωγή

Στο κεφάλαιο αυτό, θα παρουσιάσουμε τη μέθοδο της Φασματικής Ομαδοποίησης [19, 20]. Οι τεχνικές Φασματικής Ομαδοποίησης χρησιμοποιούν το φάσμα του πίνακα ομοιότητας των σημείων με στόχο, την μείωση της διάστασης και την τελική ομαδοποίηση των σημείων σε μικρότερη διάσταση. Συγκεκριμένα, η Φασματική Ομαδοποίηση χρησιμοποιεί πληροφορία από τα ιδιοδιανύσματα και τις ιδιοτιμές των πινάκων ομοιότητας. Η Φασματική Ομαδοποίηση αποτελεί μία προσέγγιση της διαμέρισης γράφου (*graph partitioning*). Η διαμέριση γράφου αποτελεί τον διαχωρισμό του γράφου με τέτοιο τρόπο ώστε οι ακμές μεταξύ των διαφορετικών ομάδων να έχουν μικρό βάρος ενώ οι ακμές μεταξύ σημείων της ίδιας ομάδας να έχουν μεγάλο βάρος. Κάθε δεδομένο αποτελεί και ένα σημείο-κορυφή του αντίστοιχου γράφου. Ακόμα, δύο κορυφές συνδέονται αν η ομοιότητα μεταξύ των αντίστοιχων δεδομένων είναι θετική(ή μεγαλύτερη από ένα ορισμένο κατώφλι). Επιπλέον, με τον όρο βάρος μεταξύ δύο σημείων-κορυφών αναφερόμαστε στην μεταξύ τους ομοιότητα. Στο Σχήμα 2.1 παρουσιάζουμε ένα γράφο στον οποίο διακρίνουμε δύο ομάδες (είναι κυκλωμένες).

Οι διάφορες μέθοδοι ομαδοποίησης χρησιμοποιούν διαφορετικά μέτρα ομοιότητας. Επομένως, ένα σημαντικό βήμα στην ομαδοποίηση είναι η επιλογή ενός μέτρου ομοιότητας, το οποίο καθορίζει τον τρόπο υπολογισμού της ομοιότητας δύο σημείων. Στη συνέχεια, θα αναλύσουμε την έννοια της ομοιότητας και θα παρουσιάσουμε κάποια διαδεδομένα μέτρα ομοιότητας.



Σχήμα 2.1 Γράφος με δύο ομαδοποιημένα σύνολα κορυφών.

Οι Φασματικοί αλγόριθμοι βασίζονται στα ιδιοδιανύσματα των πινάκων *Laplace*. Παρακάτω, θα ορίσουμε τους πίνακες Laplace και τις βασικές τους ιδιότητές τους. Ακόμα, θα παρουσιάσουμε το βασικό σχήμα της μεθόδου Φασματικής Ομαδοποίησης και στη συνέχεια διάφορες παραλλαγές της.

2.2. Μέτρα Ομοιότητας

2.2.1. Ορισμός συνάρτησης ομοιότητας

Αν έχουμε δύο αντικείμενα x, y ενός χώρου, ένα μέτρο ομοιότητας ή μία συνάρτηση ομοιότητας εκφράζει αριθμητικά το ποσό της ομοιότητάς τους. Αντίθετα, η ανομοιότητα μεταξύ δύο αντικειμένων εκφράζει αριθμητικά το κατά πόσο τα δύο αυτά αντικείμενα διαφέρουν. Ένα είδος ανομοιότητας με συγκεκριμένες ιδιότητες, τις οποίες και θα παρουσιάσουμε στη συνέχεια, είναι η απόσταση.

Συχνά, μπορούμε να εφαρμόσουμε κάποιο μετασχηματισμό για να μετατρέψουμε ένα μέτρο ομοιότητας σε ένα μέτρο ανομοιότητας ή το αντίθετο. Επιπλέον, υπάρχουν διάφοροι μετασχηματισμοί έτσι ώστε ένα μέτρο ομοιότητας ή ένα μέτρο ανομοιότητας να κυμαίνεται μεταξύ ενός ορισμένου εύρους τιμών. Συνήθως, η ομοιότητα είναι μη-αρνητική και κυμαίνεται μεταξύ 0 (καθόλου ομοιότητα) και 1 (πλήρης ομοιότητα) ενώ, η ανομοιότητα κυμαίνεται συνήθως στο διάστημα [0, 1] ή στο εύρος από 0 έως ∞.

Συγκεκριμένα, αν w(x,y) είναι η ομοιότητα μεταξύ των σημείων x και y, τότε οι ισχύουν οι παρακάτω ιδιότητες:

1. $w(x, y) = 1 \;\; \mu ονο \;\; \alpha ν \;\; x = y. (0 \le w \le 1)$ Εξ. 2.1

2. $w(x, y) = w(y, x) \;\; \gamma ια \;\; ολα \;\; τα \;\; x \;\; y$ ($συμμετρικη \;\; ιδιοτητα$). Εξ. 2.2

Ακόμα, αν d(x, y) είναι η απόσταση μεταξύ των σημείων x και y, τότε ισχύουν οι παρακάτω ιδιότητες:

1. Ιδιότητα θετικότητας:

$$d(x, y) \ge 0 \;\; \gamma ια \;\; ολα \;\; τα \;\; x \;\; y. \qquad \text{Εξ. 2.3}$$

$$d(x, y) = 0 \;\; \mu ονο \;\; \alpha ν \;\; x = y. \qquad \text{Εξ. 2.4}$$

2. Συμμετρική ιδιότητα:

$$d(x, y) = d(y, x) \;\; \gamma ια \;\; ολα \;\; τα \;\; x \;\; y. \qquad \text{Εξ. 2.5}$$

3. Τριγωνική ανισότητα:

$$d(x, y) \le d(x, y) + d(y, z) \;\; \gamma ια \;\; ολα \;\; τα \;\; x \;\; y \;\; \gamma ια \;\; ολα \;\; τα \;\; x \;\; y \;\; και \;\; z. \quad \text{Εξ. 2.6}$$

2.2.2. Συνημιτονοειδής ομοιότητα (cosine similarity)

Τα κείμενα συνήθως παρουσιάζονται ως διανύσματα όπου, κάθε χαρακτηριστικό εκφράζει τη συχνότητα με την οποία ένας συγκεκριμένος όρος (λέξη) εμφανίζεται στο κείμενο. Ένα από τα πιο δημοφιλή μέτρα ομοιότητας στο χώρο της ομαδοποίησης κειμένων είναι αυτό της συνημιτονοειδούς ομοιότητας. Δεδομένου ότι

x, y είναι δύο διανύσματα κειμένων, η συνημιτονοειδής ομοιότητα υπολογίζεται από την σχέση:

$$\cos\left(\theta\left(x,y\right)\right)=\frac{x\cdot y}{\|x\|_2\cdot\|y\|_2}=\left(\frac{x}{\|x\|_2}\right)\cdot\left(\frac{y}{\|y\|_2}\right), \qquad \text{Εξ. 2.7}$$

όπου,

$$x\cdot y=\sum_{k=1}^{n}x_k y_k \qquad \text{Εξ. 2.8}$$

και ||x||, που είναι το μήκος το διανύσματος x, είναι ίσο με:

$$\|x\|=\sqrt{\sum_{k=1}^{n}x_k^2}=\sqrt{x\cdot x}. \qquad \text{Εξ. 2.9}$$

Εξαιτίας της δεύτερης ισότητας στην Εξίσωση (2.7) η συνημιτονοειδής ομοιότητα δεν λαμβάνει υπόψη τα μεγέθη των x και y καθώς, τα κανονικοποιεί να έχουν μήκος 1.

Επομένως, η συνημιτονοειδής ομοιότητα είναι ένα μέτρο του συνημίτονου της γωνίας μεταξύ των x και y. Άρα, αν η συνημιτονοειδής ομοιότητα είναι ίση με 1, η γωνία μεταξύ των x και y είναι ίση με 0°, και τα x, y είναι ίδια (όλες οι λέξεις τους είναι κοινές, το μήκος τους όμως διαφέρει). Αν η συνημιτονοειδής ομοιότητα είναι ίση με 0, η γωνία μεταξύ των x και y είναι ίση με 90°, και δεν έχουν καμία κοινή λέξη.

2.2.3. Συναρτήσεις πυρήνα (kernel functions)

Οι συναρτήσεις πυρήνα είναι συναρτήσεις ομοιότητας μεταξύ ζευγών αντικειμένων οι οποίες αν και χειρίζονται απευθείας τα πραγματικά αντικείμενα, δίνουν ισοδύναμα αποτελέσματα με το να προβάλλαμε τα αντικείμενα σε έναν άλλο χώρο και ύστερα να υπολογίζαμε την ομοιότητας τους μέσω της πράξης του εσωτερικού γινομένου.

Τώρα, για ένα μοντέλο που βασίζεται σε ένα καθορισμένο, μη γραμμικό χώρο χαρακτηριστικών $\phi(x)$, μία συνάρτηση πυρήνα ορίζεται μέσω της σχέσης:

$$k(x_i,x_j)=\phi(x_i)^T\phi(x_j). \qquad \text{Εξ. 2.10}$$

Παρατηρούμε ότι οι συναρτήσεις πυρήνα είναι συμμετρικές καθώς ισχύει ότι:

$$k(x_i, x_j) = k(x_j, x_i). \qquad \text{Εξ. 2.11}$$

Πρέπει να διασφαλίσουμε ότι η συνάρτηση που θα επιλέξουμε θα είναι έγκυρη, δηλαδή, θα εκφράζει ομοιότητα στο χώρο χαρακτηριστικών $\phi(x)$. Η παραπάνω απαίτηση μπορεί να διασφαλιστεί αν ικανοποιείται το θεώρημα του *Mercer*.

Θεώρημα 1.1 (*Mercer's Theorem*). Μία συνάρτηση πυρήνα $k(x_i, x_j)$ μπορεί να εκφραστεί μέσω της σχέσης:

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j) \qquad \text{Εξ. 2.12}$$

αν και μόνο αν, για κάθε συνάρτηση g(x_i) τέτοια ώστε το $\int g(x_i)^2 dx_i$ να είναι πεπερασμένο, ισχύει ότι:

$$\int k(x_i, x_j) g(x_i) g(x_j) dx_i dx_j \geq 0. \qquad \text{Εξ. 2.13}$$

Παρακάτω, παρουσιάζουμε κάποια παραδείγματα συναρτήσεων πυρήνα που χρησιμοποιούνται συχνά. Συγκεκριμένα, μία συνάρτηση πυρήνα είναι το εσωτερικό γινόμενο. Συγκριμένα, στην περίπτωση αυτή ισχύει ότι:

$$k(x_i, x_j) = x_i^T x_j \text{ (\textit{linear kernel})}. \qquad \text{Εξ. 2.14}$$

Μία άλλη συνάρτηση πυρήνα είναι η πολυωνυμική συνάρτηση που ακολουθεί:

$$k(x_i, x_j) = \left(x_i^T x_j + c\right)^p, \text{ όπου c>0 (\textit{polynomial kernel})}. \qquad \text{Εξ. 2.15}$$

Μία άλλη συνάρτηση πυρήνα είναι η γκαουσιανή και έχει τη μορφή:

$$k(x_i, x_j) = \exp\left(-\| x_i - x_j \|^2 / 2\sigma^2\right) \text{ (\textit{Gaussian kernel})}. \qquad \text{Εξ. 2.16}$$

Τέλος, ένα άλλο παράδειγμα συνάρτησης πυρήνα είναι η σιγμοειδής και έχει την ακόλουθη μορφή:

$$k(x_i, x_j) = \tanh\left(a x_i^T x_j + b\right) \text{ (\textit{sigmoidal kernel})}. \qquad \text{Εξ. 2.17}$$

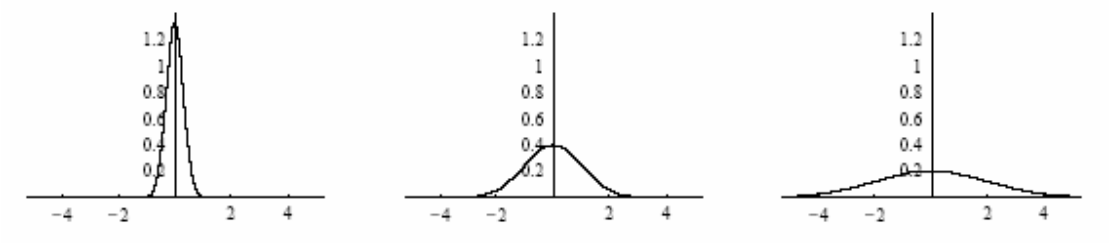
2.2.4. Γκαουσιανή συνάρτηση πυρήνα (Gaussian kernel)

Η Γκαουσιανή συνάρτηση πυρήνα ορίζεται στις δύο διαστάσεις όπως φαίνεται στην Εξίσωση (2.16). Για υψηλές διαστάσεις, έστω N, μπορεί να περιγραφεί ως το γινόμενο των N μονοδιάστατων συναρτήσεων πυρήνα. Η μεταβλητή σ καθορίζει το εύρος της Γκαουσιανής συνάρτησης πυρήνα και μπορεί να έχει μόνο θετική τιμή.

Η κανονικοποιημένη Γκαουσιάνη συνάρτηση πυρήνα έχει τη μορφή που ακολουθεί:

$$\frac{1}{2\pi\sigma^2}\exp\left(-\| x_i - x_j \|^2 / 2\sigma^2\right). \qquad \text{Εξ. 2.18}$$

Ο όρος $\frac{1}{\sqrt{2\pi}\sigma}$ στην Εξίσωση (2.18) αποτελεί την σταθερά κανονικοποίησης. Στην κανονικοποιημένη Γκαουσιανή συνάρτηση πυρήνα ισχύει ότι η περιοχή κάτω από την καμπύλη είναι συνεχής χωρίς αποκλίσεις. Σαν αποτέλεσμα, η αύξηση της μεταβλητής σ συνεπάγεται τη μείωση του πλάτους της συνάρτησης. Για παράδειγμα, στο Σχήμα 2.2 παρουσιάζουμε τις κανονικοποιημένες συναρτήσεις πυρήνα για σ=0.3, σ=1 και σ=2.



Σχήμα 2.2 Η Γκαουσιανή συνάρτηση για σ=0.3, σ=1 και σ=2.

Η συγχώνεψη δύο Γκαουσιανών συναρτήσεων πυρήνα συνεπάγεται μία νέα μεγαλύτερου εύρους Γκαουσιανή συνάρτηση με διακύμανση (σ^2) ίση με το άθροισμα των διακυμάνσεων των συγχωνευμένων συναρτήσεων.

2.2.5. Βασικά μέτρα απόστασης

Το μέτρο της απόστασης *Minkowski* ορίζεται ως εξής:

$$d(x,y)=\left(\sum_{k=1}^{n}|x_k-y_k|^r\right)^{\frac{1}{r}}, \qquad \text{Εξ. 2.19}$$

όπου r είναι μία παράμετρος.

Οι συνηθέστερες επιλογές είναι: α) για r=1, η απόσταση *Manhattan*, ένα παράδειγμα είναι η απόσταση *Hamming* που αντιστοιχεί στον αριθμό των bits στα οποία διαφέρουν δύο αντικείμενα με μόνο δυαδικά χαρακτηριστικά β) για r=2, η *Ευκλείδεια* απόσταση, γ) για r=3, η απόσταση *Tschebyshev*.

Το μέτρο της απόστασης *Mahalanobis* αποτελεί μία γενίκευση της Ευκλείδειας απόστασης και χρησιμοποιείται όταν κάποια από τα χαρακτηριστικά των δεδομένων συσχετίζονται, έχουν διαφορετικό εύρος τιμών και η κατανομή που ακολουθούν προσεγγίζει την γκαουσιανή κατανομή. Συγκεκριμένα, η απόσταση Mahalanobis μεταξύ των σημείων x, y ορίζεται ως εξής:

$$mahalanobis(x,y)=(x-y)\Sigma^{-1}(x-y)^T, \qquad \text{Εξ. 2.20}$$

όπου ο Σ^{-1} είναι ο αντίστροφος του πίνακα συμμεταβλητότητας των δεδομένων. Ο πίνακας Σ είναι συμμετρικός και θετικά ημιορισμένος. Ένας πίνακας καλείται θετικά ημιορισμένος αν όλες οι ιδιοτιμές του είναι μη-αρνητικές (θετικές ή μηδέν).

Ο πίνακα Σ έχει την ακόλουθη μορφή:

$$\Sigma=\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1D}\\ \sigma_{12} & \ddots & & \\ \vdots & & & \\ \sigma_{1D} & \sigma_{2D} & \cdots & \sigma_D^2\end{bmatrix}$$

2.3. Πίνακες Laplace και οι Βασικές τους Ιδιότητες

Ο πίνακας Laplace, τον οποίο συμβολίζουμε με L, αποτελεί αναπαράσταση σε πίνακα ενός γράφου. Παρακάτω, θεωρούμε ένα μη κατευθυνόμενο γράφο G(V, E) με πλήθος κορυφών ||V||=n, και πίνακα βαρών (ή ομοιότητας) $W=(w_{ij})_{i,j=1,\ldots,n}$ όπου $w_{ij}\geq 0$. Αν $w_{ij}=0$, τότε, οι κορυφές i και j δεν συνδέονται. Δύο κορυφές συνδέονται με μία ακμή αν η ομοιότητα μεταξύ των αντίστοιχων σημείων-κορυφών είναι σημαντική(ή

μεγαλύτερη από ένα ορισμένο κατώφλι). Ακόμα, επειδή ο G είναι μη κατευθυνόμενος, ισχύει ότι $w_{ij}=w_{ji}$ (ο G είναι συμμετρικός). Ακόμα, θεωρούμε πως ο βαθμός (degree) ενός σημείου-κορυφής $i \in V$ ορίζεται ως το πλήθος των βαρών των ακμών που καταλήγουν στην κορυφή i δηλαδή,

$$d_i = \sum_{j=1}^{n} w_{ij}. \qquad \text{Εξ. 2.21}$$

Με τον όρο D, συμβολίζουμε τον διαγώνιο $n \times n$ πίνακα με τιμές $d_1,\ldots d_n$ στην κύρια διαγώνιο.

$$d_{i,j} = \begin{cases} d_i, \ \alpha\nu \ \ i = j & \text{Εξ. 2.22} \\ 0, \ \alpha\lambda\lambda\iota\omega\varsigma & \text{Εξ. 2.23} \end{cases}$$

2.3.1. Ιδιοδιανύσματα και ιδιοτιμές

Δοθέντος ενός γραμμικού μετασχηματισμού, που αναπαρίσταται μέσω ενός τετραγωνικού πίνακα A, ένα μη μηδενικό διάνυσμα x ορίζεται ως ιδιοδιάνυσμα του A αν ικανοποιεί την παρακάτω εξίσωση:

$$Ax = \lambda x, \qquad \text{Εξ. 2.24}$$

όπου λ είναι ένα βαθμωτό μέγεθος που καλείται ιδιοτιμή του A και αντιστοιχεί στο ιδιοδιάνυσμα x. Δεδομένου ότι υπάρχει ο αντίστροφος του πίνακα A, μπορούμε να υπολογίσουμε τις ιδιοτιμές του επιλύοντας την παρακάτω εξίσωση:

$$\det(A - \lambda I) = 0, \qquad \text{Εξ. 2.25}$$

όπου I είναι ο ταυτοτικός πίνακας (δηλαδή, ο πίνακας με μονάδες στην κύρια διαγώνιο και μηδενικά στις υπόλοιπες θέσεις).

Τα ιδιοδιανύσματα που αντιστοιχούν σε διαφορετικές ιδιοτιμές είναι γραμμικά ανεξάρτητα, δηλαδή, σε ένα n-διάστατο χώρο ο γραμμικός μετασχηματισμός A δεν μπορεί να έχει παραπάνω από n ιδιοδιανύσματα με διαφορετικές ιδιοτιμές. Η πολλαπλότητα μιας ιδιοτιμής είναι ο αριθμός των γραμμικά ανεξάρτητων ιδιοδιανυσμάτων που έχουν την ίδια ιδιοτιμή.

Ένα ιδιοδιάνυσμα ενός γράφου G ορίζεται ως ένα ιδιοδιάνυσμα του πίνακα ομοιότητας του γράφου ή ως ένα ιδιοδιάνυσμα του πίνακα Laplace του γράφου αυτού.

2.3.2. Ο πίνακας Laplace

Οι Φασματικοί αλγόριθμοι βασίζονται στα ιδιοδιανύσματα των πινάκων Laplace, που αποτελούν ένα συνδυασμό των πινάκων D (πίνακας βαθμών) και W (πίνακας ομοιότητας).

Ο μη-κανονικοποιημένος πίνακας Laplace ορίζεται ως:

$$L = D - W, \qquad \text{Εξ. 2.26}$$

και έχει την ακόλουθη μορφή:

$$L = (l_{i,j})_{n\times n}$$

$$l_{i,j} = \begin{cases} d_i, & \alpha\nu \quad i = j \\ -1, & \alpha\nu \quad i \neq j \quad \kappa\alpha\iota \quad i, j \quad \sigma\upsilon\nu\delta\varepsilon o\nu\tau\alpha\iota \quad \mu\varepsilon \quad \alpha\kappa\mu\eta \\ 0, & \alpha\lambda\lambda\iota\omega\varsigma \end{cases} \qquad \text{Εξ. 2.27}$$

Δοθέντος γράφου G και του πίνακα Laplace του L με ιδιοτιμές $\lambda_0 \le \lambda_1 \le ... \le \lambda_{n-1}$, ισχύουν οι παρακάτω ιδιότητες:

1. ο L είναι πάντα θετικά ημιορισμένος (δηλαδή, όλες οι ιδιοτιμές του είναι θετικές ή ίσες με μηδέν).

2. ο αριθμός των φορών που εμφανίζεται η ιδιοτιμή 0 στο L είναι ίσος με τον αριθμό των συνδεδεμένων συνιστωσών του γράφου. Ο γράφος G είναι συνδεδεμένος αν δύο οποιεσδήποτε κορυφές του μπορούν να συνδεθούν μέσω ενός μονοπατιού όπου όλοι οι ενδιάμεσοι σταθμοί να ανήκουν και αυτοί στον G. Συνδεδεμένη συνιστώσα ονομάζεται ένα υποσύνολο B του G αν είναι συνδεδεμένο και επιπλέον αν δεν υπάρχουν συνδέσεις μεταξύ των σταθμών του B και του $\overline{B}$.

3. η ιδιοτιμή λ_1 είναι μεγαλύτερη από το 0 αν και μόνο αν ο γράφος G είναι συνδεδεμένος.
4. η μικρότερη, μη μηδενική ιδιοτιμή του L ονομάζεται φασματικό κενό (*spectral gap*).

Ο κανονικοποιημένος πίνακας Laplace ορίζεται με δύο διαφορετικούς τρόπους:

$$L_{sym} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = 1 - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \text{ (συμμετρικός πίνακας)} \qquad \text{Εξ. 2.28}$$

$$L_{rw} := D^{-1} L = 1 - D^{-1} W \text{ (τυχαίος περίπατος)} \qquad \text{Εξ. 2.29}$$

και έχει την παρακάτω μορφή:

$$L = (l_{i,j})_{n \times n}$$

$$l_{i,j} = \begin{cases} 1, & \alpha\nu \quad i = j \quad \kappa\alpha\iota \quad d_i \neq 0 \\ -\dfrac{1}{\sqrt{d_i d_j}}, & \alpha\nu \quad i \neq j \quad \kappa\alpha\iota \quad i, j \quad \sigma\upsilon\nu\delta\varepsilon o\nu\tau\alpha\iota \quad \mu\varepsilon \quad \alpha\kappa\mu\eta \\ 0, & \alpha\lambda\lambda\iota\omega\varsigma \end{cases} \qquad \text{Εξ. 2.30}$$

Δοθέντος γράφου G και του πίνακα Laplace του L κάποιες ιδιότητες που ισχύουν είναι οι παρακάτω:
1. οι L_{sym} και L_{rw} είναι πάντα θετικά ημιορισμένοι και έχουν n πραγματικές, μη αρνητικές ιδιοτιμές $0 = \lambda_1 \leq \lambda_1 \leq ... \leq \lambda_n$.
2. ο αριθμός των φορών που εμφανίζεται η ιδιοτιμή 0 στους L_{sym} και L_{rw} στο L είναι ίσος με τον αριθμό των συνδεδεμένων συνιστωσών του γράφου.

2.4. Μέθοδοι Φασματικής Ομαδοποίησης

Οι τεχνικές Φασματικής Ομαδοποίησης χρησιμοποιούν το φάσμα του πίνακα ομοιότητας των σημείων με στόχο, την μείωση της διάστασης και την τελική ομαδοποίηση των σημείων σε μικρότερη διάσταση. Με τον όρο φάσμα ενός πίνακα αναφερόμαστε στις ιδιοτιμές, στα ιδιοδιανύσματα και στα ιδιοδιαστήματά του. Δοθέντος ενός γραμμικού μετασχηματισμού, ένα ιδιοδιάστημα για μία συγκεκριμένη

ιδιοτιμή είναι το σύνολο των ιδιοδιανυσμάτων που συνδέονται με την ιδιοτιμή αυτή μαζί με το μηδενικό διάνυσμα.

Υπάρχουν διάφοροι αλγόριθμοι Φασματικής Ομαδοποίησης που διαφέρουν κυρίως ως προς τον αριθμό των ιδιοδιανυσμάτων που χρησιμοποιούν για τον διαχωρισμό. Συγκεκριμένα, οι αλγόριθμοι Φασματικής Ομαδοποίησης διακρίνονται σε *recursive* και *multiway*. Οι recursive αλγόριθμοι χρησιμοποιούν ένα μόνο ιδιοδιάνυσμα επαναληπτικά για τους διαχωρισμούς ενώ, οι multiway αλγόριθμοι χρησιμοποιούν πολλά ιδιοδιανύσματα και υπολογίζουν απευθείας τον διαχωρισμό των δεδομένων.

Μία τεχνική Φασματικής Ομαδοποίησης είναι ο αλγόριθμος των Shi και Malik που διαχωρίζει τα δεδομένα σε δύο σύνολα με βάση το ιδιοδιάνυσμα που αντιστοιχεί στην δεύτερη μικρότερη ιδιοτιμή του πίνακα Laplace.

Εμείς χρησιμοποιούμε τον αλγόριθμο Φασματικής Ομαδοποίησης των Ng, Jordan και Weiss (NJW). Ο NJW αλγόριθμος εμπεριέχει την επιλογή των κορυφαίων ιδιοδιανυσμάτων του πίνακα ομοιότητας και στην συνέχεια την χρησιμοποίησή τους για την ομαδοποίηση των διαφόρων σημείων.

2.4.1. Αλγόριθμος των Ng, Jordan και Weiss

Στον αλγόριθμο NJW, συμβολίζουμε με N είναι το σύνολο των δεδομένων, και με |N| το πλήθος των σημείων του, θεωρούμε πώς |N|=n. Ακόμα, συμβολίζουμε με $C = \{C_1, C_2, ...C_K\}$, όπου K το πλήθος των ομάδων, μία διαμέριση του N σε μη-άδεια και μη-συνδεδεμένα υποσύνολα $C_1, C_2, ...C_K$.

Σύμφωνα με τον αλγόριθμό NJW, αρχικά κατασκευάζουμε τον πίνακα ομοιότητας μεταξύ των σημείων και μηδενίζουμε τα διαγώνια στοιχεία του. Στη συνέχεια, με βάση τον πίνακα ομοιότητας, υπολογίζουμε τον πίνακα Laplace όπως αυτός ορίζεται στην Εξίσωση 2.28. Έπειτα, κατασκευάζουμε ένα πίνακα U, μεγέθους $(n \times K)$, που

έχει ως στήλες τα K κανονικοποιημένα ιδιοδιανύσματα που αντιστοιχούν στις K μεγαλύτερες ιδιοτιμές του πίνακα Laplace. Ακολούθως, κατασκευάζουμε ένα πίνακα Y, μεγέθους $(n \times K)$, μέσω του πίνακα U. Πιο αναλυτικά, ο Y είναι αποτέλεσμα της κανονικοποίησης των αθροισμάτων των γραμμών του U σε μοναδιαίο μήκος. Στο τελικό βήμα του αλγορίθμου NJW ομαδοποιούμε τις γραμμές του Y μέσω του K-means αλγορίθμου ομαδοποίησης σε K ομάδες. Μπορούμε να παρατηρήσουμε αναλυτικά τα βήματα του NJW αλγορίθμου στο Σχήμα 2.3.

Αλγόριθμος των Ng, Jordan και Weiss (NJW)
Βήμα 1ο: Κατασκευάζουμε τον πίνακα ομοιότητας W, ο αριθμός των ομάδων είναι K. Στη συνέχεια, θέτουμε τα διαγώνια στοιχεία του (w_{ii}) ίσα με μηδέν.
Βήμα 2ο: Υπολογίζουμε τον πίνακα $$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}},$$ όπου L είναι ο πίνακας Laplace του W.
Βήμα 3ο: Έστω ότι είναι $1 = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K$ οι K μεγαλύτερες ιδιοτιμές του και $u_1, u_2, ..., u_K$ είναι τα αντίστοιχα ιδιοδιανύσματα. Όλα τα ιδιοδιανύσματα είναι κανονικοποιημένα ώστε να έχουν μοναδιαίο μήκος. Κατασκευάζουμε τον πίνακα $U \in R^{n \times K}$ που έχει ως στήλες τα ιδιοδιανύσματα $u_1, u_2, ..., u_K$.
Βήμα 4ο: Κατασκευάζουμε τον πίνακα $Y \in R^{n \times k}$ από τον U κανονικοποιώντας τα αθροίσματα των γραμμών σε μοναδιαίο μήκος. Επομένως, κάθε στοιχείο του πίνακα Y έχει τη μορφή $$y_{ij} = u_{ij} / \sqrt{\sum_{j=1} u_{ij}^{\ 2}},$$ για κάθε i=1, …,n έστω ότι $y_i \in R^K$ είναι το διάνυσμα που αντιστοιχεί στην i-οστή γραμμή του U.
Βήμα 5ο: Ομαδοποιούμε τα σημεία $(y_i)_{i=1,...n}$ στο R^K, μέσω του αλγορίθμου K-means, στις ομάδες $C_1, ..., C_k$.

Σχήμα 2.3 Βήματα αλγορίθμου NJW.

2.4.2. Διαμέριση γράφου (Graph Cut)

Όπως προαναφέραμε, στόχος της Φασματικής Ομαδοποίησης είναι ο διαχωρισμός των σημείων σε K ομάδες με βάση τον πίνακα ομοιότητας. Μπορούμε να

αναδιατυπώσουμε το παραπάνω ως εξής: στόχος της ομαδοποίησης είναι η διαμέριση του γράφου έτσι ώστε οι ακμές μεταξύ διαφορετικών ομάδων να έχουν ιδιαίτερα μικρό βάρος και οι ακμές μέσα σε μία ομάδα να έχουν μεγάλο βάρος. Επομένως, στη Φασματική Ομαδοποίηση θέλουμε να ελαχιστοποιήσουμε την ομοιότητα μεταξύ των διαφόρων ομάδων και να μεγιστοποιήσουμε την ομοιότητα μεταξύ των δεδομένων της ίδιας ομάδας.

Στη συνέχεια, συμβολίζουμε το πλήθος των κορυφών του A με |A|. Ακόμα, συμβολίζουμε με $vol(A)$ το άθροισμα των βαθμών των κορυφών που ανήκουν στο A και το ορίζουμε ως:

$$vol(A) := \sum_{i \in A} d_i, \qquad \text{Εξ. 2.31}$$

Το σύνολο των ακμών μεταξύ δύο μη συνδεδεμένων συνόλων $A, B \subseteq N$ ονομάζεται edge cut ή απλά cut μεταξύ των A, B. Ακόμα, ορίζουμε το συμπλήρωμα του υποσυνόλου $A \subset V$ ως $\overline{A}$. Ακόμα, θεωρούμε ότι τα σύνολα $A_1,\ldots A_k$ αποτελούν μία διαμέριση του γράφου αν είναι ανά δύο ξένα μεταξύ τους (δηλαδή $A_i \cap A_j = \varnothing$) και αν η ένωση τους δίνει το γράφο V (δηλαδή $A_1 \cup ... \cup A_k = V$).

Υποθέτουμε ότι έχουμε δύο μη συνδεδεμένα υποσύνολα A, B τότε, ισχύει ότι $A \cup B = V, A \cap B = \varnothing,$ απλά αφαιρώντας τις ακμές που συνδέουν τα δύο αυτά υποσύνολα. Μπορούμε να υπολογίσουμε τον βαθμό ανομοιότητας, μεταξύ των δύο υποσυνόλων, ως το συνολικό βάρος των ακμών που έχουμε αφαιρέσει. Επομένως, θα έχουμε:

$$cut(A, B) = \sum_{i \in A,\, j \in B} w_{ij} \qquad \text{Εξ. 2.32}$$

Ο πιο απλός και ευθύς τρόπος να κατασκευάσουμε μία διαμέριση του γράφου είναι μέσω της επίλυσης του προβλήματος ελάχιστης κατάτμησης. Συγκεκριμένα, επιλέγουμε την διαμέριση $A_1,...A_k$ που ελαχιστοποιεί το

$$cut(A_1,...,A_k) := \sum_{i=1}^{k} cut(A_i, \overline{A}_i) \qquad \text{Εξ. 2.33.}$$

Παρόλα αυτά, η επίλυση του προβλήματος ελάχιστης κατάτμησης, σε πολλές περιπτώσεις οδηγεί στον διαχωρισμό μίας ξεχωριστής κορυφής από το υπόλοιπο γράφο. Το παραπάνω είναι ένα μη επιθυμητό αποτέλεσμα καθώς στόχος της ομαδοποίησης είναι αρκετά μεγάλες ομάδες. Ένας τρόπος να αντιμετωπίσουμε το παραπάνω πρόβλημα είναι να απαιτούμε τα σύνολα $A_1,...,A_k$ να είναι αρκετά μεγάλα, θα αναφέρουμε τις δύο πιο κοινές συναρτήσεις που επιλύουν το πρόβλημα με τον τρόπο αυτό.

Η συνάρτηση RatioCut παρουσιάστηκε πρώτη φορά το 1992 από τους Hagen και Kahng και ορίζεται ως:

$$RatioCut(A_1,...,A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A}_i)}{|A_i|}. \qquad \text{Εξ. 2.34}$$

Η συνάρτηση Ncut (Normalized cut) [24] παρουσιάστηκε πρώτη φορά το 2002 από τους Shi και Malik και ορίζεται ως:

$$Ncut(A_1,...,A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A}_i)}{vol(A_i)}. \qquad \text{Εξ. 2.35}$$

Σημειώνουμε ότι και οι δύο συναρτήσεις αποκτούν μικρή τιμή αν οι ομάδες A_i δεν είναι πολύ μικρές. Συγκεκριμένα, η συνάρτηση $\sum_{i=1}^{k}(1/|A_i|)$ ελαχιστοποιείται αν όλα τα $|A_i|$ συμπίπτουν και η συνάρτηση $\sum_{i=1}^{k}(1/vol(A_i))$ ελαχιστοποιείται αν όλα τα $vol(A_i)$ συμπίπτουν.

2.4.3. Χρήση τυχαίων περιπάτων (Random Walks)

Ένας άλλος τρόπος να εξηγήσουμε τη Φασματική Ομαδοποίηση βασίζεται στους τυχαίους περιπάτους στο γράφο ομοιότητας [14]. Ο τυχαίος περίπατος σε ένα γράφο είναι μία στοχαστική διαδικασία που με τυχαίο τρόπο μεταβαίνει από την μία κορυφή

στην άλλη. Μπορούμε να αναδιατυπώσουμε το πρόβλημα ως εξής: στόχος της ομαδοποίησης είναι η διαμέριση του γράφου έτσι ώστε ο τυχαίος περίπατος να παραμένει επί μακρών μέσα σε μία ομάδα και σπάνια να μεταβαίνει σε κάποια άλλη.

Συγκεκριμένα, η πιθανότητα μετάβασης σε ένα βήμα από την κορυφή i στην κορυφή j είναι ανάλογη του βάρους w_{ij} και ίση με $p_{ij} := w_{ij} / d_i$. Επομένως, ο πίνακας μετάβασης $P = (p_{ij})_{i,j=1,\ldots,n}$ του τυχαίου περιπάτου ορίζεται ως:

$$P = D^{-1}W . \qquad \text{Εξ. 2.36}$$

Παρατηρούμε, από τον παραπάνω τύπο, ότι υπάρχει στενή σχέση ανάμεσα στον πίνακα μετάβασης και στον πίνακα Laplace.

ΚΕΦΑΛΑΙΟ 3. ΝΕΥΤΩΝΕΙΑ ΟΜΑΔΟΠΟΙΗΣΗ

3.1. Εισαγωγή

Στο κεφάλαιο αυτό, θα παρουσιάσουμε μία δυναμική μέθοδο ομαδοποίησης, την Νευτώνεια Ομαδοποίηση [4], που βασίζεται στις εξισώσεις κίνησης του Νεύτωνα. Σύμφωνα με τη μέθοδο της Νευτώνειας Ομαδοποίησης εφαρμόζεται αρχικά μία δυναμική διαδικασία στα σημεία του συνόλου δεδομένων, χρησιμοποιώντας την πρώτη εξίσωση κίνησης του Νεύτωνα. Αυτό έχει ως αποτέλεσμα τη μετακίνηση του κάθε σημείου προς το κέντρο της ομάδας στην οποία ανήκει. Στο κεφάλαιο αυτό, θα ορίσουμε τις εξισώσεις κίνησης του Νεύτωνα και θα αναφέρουμε κάποιες έννοιες και ορισμούς. Έπειτα, θα αναλύσουμε τη φάση της συρρίκνωσης των ομάδων δεδομένων και θα αναφέρουμε τις λοιπές φάσεις που συνιστούν την μέθοδο της Νευτώνειας Ομαδοποίησης. Τέλος, θα παρουσιάσουμε μία μέθοδο υπολογισμού της παραμέτρου σ. Η παράμετρος αυτή προσδιορίζει το εύρος του δυναμικού και είναι ιδιαίτερα σημαντική, αφού επηρεάζει τη δυναμική διεργασία σμίκρυνσης των ομάδων, καθώς και την επίδοση της επακόλουθης ομαδοποίησης.

3.2. Εξισώσεις Κίνησης του Νεύτωνα

Οι τρεις εξισώσεις κίνησης του Νεύτωνα λαμβάνουν υπόψη την Αρχή της Σχετικότητας του Γαλιλαίου η οποία προβλέπει ότι, η ελεύθερη κίνηση των αντικειμένων είναι σε ευθεία γραμμή και έχει σταθερή ταχύτητα. Παρακάτω, αρχικά θα υπενθυμίσουμε το πώς ορίζονται η ταχύτητα και η επιτάχυνση και στη συνέχεια θα παρουσιάσουμε τις προαναφερθείσες εξισώσεις.

Όπως γνωρίζουμε, η ταχύτητα είναι ο ρυθμός αλλαγής της μετατόπισης. Με άλλα λόγια, είναι η διανυόμενη απόσταση ανά χρονική μονάδα. Επομένως, αν s είναι η διανυόμενη απόσταση από ένα σώμα σε χρόνο t, η ταχύτητα v του σώματος είναι ίση με:

$$v = s / t\,. \qquad \text{Εξ. 3.1}$$

Στη συνέχεια, η επιτάχυνση είναι ο ρυθμός αλλαγής της ταχύτητας. Με άλλα λόγια, είναι η αλλαγή στην ταχύτητα ανά χρονική μονάδα. Επομένως, αν u είναι η αρχική ταχύτητα (η ταχύτητα όταν t=0) ενός κινούμενου σώματος και v είναι η τελική ταχύτητα του σώματος σε χρονικό διάστημα t, τότε η επιτάχυνση του σώματος σε χρόνο t είναι ίση με:

$$a = (v - u) / t. \qquad \text{Εξ. 3.2}$$

Παραπάνω, θεωρούμε ότι η επιτάχυνση του σώματος παραμένει σταθερή σε όλο το χρονικό διάστημα t. Από την Εξίσωση (3.2) εύκολα προκύπτει η πρώτη εξίσωση κίνησης του Νεύτωνα:

$$\boxed{v = u + at.} \qquad \text{Εξ. 3.3}$$

Έστω ότι ένα σώμα κινείται με σταθερή ταχύτητα v τότε από την Εξίσωση (3.1) ισχύει ότι $s = vt$. Έστω τώρα ότι ένα σώμα κινείται με σταθερή επιτάχυνση, αν u η αρχική του ταχύτητα και v η ταχύτητά του μετά από χρόνο t τότε, η μέση ταχύτητα του σώματος είναι ίση με:

$$(u + v) / 2. \qquad \text{Εξ. 3.4}$$

Άρα, από τις Εξισώσεις (3.1) και (3.4) προκύπτει ότι:

$$s = [(u+v)/2]t. \qquad \text{Εξ. 3.5}$$

Από την Εξίσωση (3.3) (πρώτη εξίσωση του Νεύτωνα) και την Εξίσωση (3.5) προκύπτει η δεύτερη εξίσωση του Νεύτωνα:

$$\boxed{s = ut + \left(\frac{1}{2}\right)at^2.} \qquad \text{Εξ. 3.6}$$

Τώρα, από την Εξίσωση (3.3) έχουμε ότι:

$$v^2 = u^2 + 2a\left(ut + \left(\frac{1}{2}\right)at^2\right) \qquad \text{Εξ. 3.7}$$

και μέσω της Εξίσωσης (3.6) (δεύτερη εξίσωση του Νεύτωνα), προκύπτει η τρίτη εξίσωση κίνησης του Νεύτωνα:

$$\boxed{v^2 = u^2 + 2as.} \qquad \text{Εξ. 3.8}$$

Η μέθοδος της Νευτώνειας Ομαδοποίησης, που θα παρουσιάσουμε στη συνέχεια, βασίζεται στην πρώτη εξίσωση κίνησης του Νεύτωνα.

3.3. Νευτώνεια Ομαδοποίηση

Η μέθοδος της Νευτώνειας Ομαδοποίησης απαριθμεί και εντοπίζει της ομάδες που περιέχονται σε ένα σύνολο δεδομένων. Συγκεκριμένα, αποτελείται από δύο φάσεις.

Στην πρώτη φάση, προσδιορίζεται ο αριθμός Κ των ομάδων και προσεγγίζεται η θέση των κέντρων τους. Συγκεκριμένα, εφαρμόζεται μία δυναμική διαδικασία στα σημεία του συνόλου δεδομένων με αποτέλεσμα να συγκεντρώσει, το καθένα από αυτά, γύρω από το κέντρο της ομάδας στην οποία ανήκει. Το μονοπάτι που ακολουθεί το κάθε σημείο καθορίζει το εύρος μίας σχετικής συνάρτησης πυκνότητας πιθανότητας (pdf). Το πλήθος των παραπάνω συναρτήσεων πυκνότητας πιθανότητας συντελεί σε μία συνάρτηση πολυτροπικού μοντέλου, όπου, κάθε μέγιστο αντιστοιχεί σε διαφορετική ομάδα με κέντρο που προσεγγιστικά τοποθετείται στη θέση αιχμής. Το πλήθος των σημείων αιχμής αποτελεί τον αριθμό των ομάδων του συνόλου δεδομένων ενώ, οι

θέσεις δίνουν μία προσέγγιση της τοποθέτησης των κέντρων των ομάδων. Για την εύρεση του αριθμού των ομάδων και των κέντρων τους χρησιμοποιούνται μέθοδοι στοχαστικής και καθολικής βελτιστοποίησης (global optimization) που ανακαλύπτουν τα τοπικά μέγιστα [1, 7, 16, 17, 26, 27]. Το πλήθος των τοπικών μέγιστων της συνάρτησης αντιστοιχεί στον αριθμό Κ των ομάδων που υπάρχουν στο σύνολο δεδομένων.

Στη δεύτερη φάση, χρησιμοποιείται μία τεχνική για την τοπική ομαδοποίηση. Στη συνέχεια, θα περιγράψουμε την διαδικασία συγκέντρωσης των δεδομένων γύρω από τα κέντρα των ομάδων στις οποίες ανήκουν και τον υπολογισμό της παραμέτρου σ. Τα υπόλοιπα στάδια της μεθόδου δεν θα αναλυθούν καθώς υπερβαίνουν τους σκοπούς της διατριβής.

3.3.1. Συρρίκνωση των ομάδων

Συμβολίζουμε με $X = \{x_1, ...x_N\}$ ένα σύνολο δεδομένων. Ακόμα, θεωρούμε ότι ο αριθμός των ομάδων, στις οποίες θέλουμε να διαμερίσουμε το σύνολο δεδομένων, είναι γνωστός και ίσος με Κ. Ακόμα, συμβολίζουμε με V_{ij} το δυναμικό που αναπτύσσεται μεταξύ των σωματιδίων-σημείων του συνόλου που είναι τοποθετημένα στις θέσεις x_i και x_j. Υποθέτουμε ότι η V_{ij} δίνεται από μία συνάρτηση του Gauss (δηλ. Γκαουσιανό δυναμικό) την οποία και παραθέτουμε:

$$V_{ij} = -\exp(-\| x_i - x_j \|^2 / 2\sigma^2), \qquad \text{Εξ. 3.9}$$

όπου η κλιμακωτή παράμετρος (σ) προσδιορίζει το εύρος του δυναμικού. Εκτός της γνωστής Γκαουσιανής συνάρτησης, θα μπορούσαμε να είχαμε χρησιμοποιήσει οποιαδήποτε άλλη συνάρτηση πυρήνα ανάλογα με των τύπο των δεδομένων και το πρόβλημα που εξετάζουμε. Στη συνέχεια της διατριβής, θα δούμε πως μπορούμε να εφαρμόσουμε την Νευτώνεια εξίσωση κίνησης σε δεδομένα μεγάλης διάστασης και συγκεκριμένα, στο πρόβλημα της ομαδοποίησης κειμένων.

Η παράμετρος σ είναι ιδιαίτερα σημαντική καθώς επηρεάζει τη δυναμική διαδικασία συρρίκνωσης των ομάδων και άρα αναπόφευκτα την επίδοση της ομαδοποίησης. Το

σ εξαρτάται από το σύνολο δεδομένων. Πιο αναλυτικά, αραιά σύνολα δεδομένων απαιτούν μεγαλύτερου εύρους δυναμικό σε σχέση με αυτό που απαιτούν πυκνότερα σύνολα. Συγκεκριμένα, αν το σ που θα επιλέξουμε είναι ιδιαίτερα μεγάλο θα έχει ως αποτέλεσμα Γκαουσιανό δυναμικό μεγάλου εύρους και επομένως, εξαιτίας του μεγάλου αριθμού επικαλύψεων, τελικά θα καταλήγουμε σε μία μόνο ομάδα. Αντίθετα, αν το σ που θα επιλέξουμε οδηγεί σε ιδιαίτερα "στενό" Γκαουσιανό δυναμικό, εξαιτίας του ότι ακόμα και κοντινά σημεία θα καταλήγουν σε διαφορετικές ομάδες, στην τελική ομαδοποίηση θα έχουμε μεγάλο αριθμό από ομάδες.

Η κίνηση των δεδομένων μας επηρεάζεται από την ύπαρξη μίας δύναμης. Τα δεδομένα που κινούνται προς διαφορετικές ομάδες είτε απωθούνται μεταξύ τους, είτε είναι πολύ μακριά για να υπάρξει αλληλεπίδραση μεταξύ τους. Μετά από ένα επαρκή αριθμό βημάτων, τα σημεία που ανήκουν στην ίδια ομάδα θα μετακινηθούν πιο κοντά το ένα στο άλλο και επομένως, η ομάδα θα συρρικνωθεί. Η δυναμική διαδικασία βασίζεται στις εξισώσεις κίνησης του Νεύτωνα που ακολουθούν, όπου και θεωρούμε ότι η επιτάχυνση είναι σταθερή:

$$d^2 x_i(t)/dt^2 = -\nabla_i \sum_{\substack{j=1 \\ j\neq i}}^{N} V_{ij} \equiv F_i, \qquad \forall i = 1,2,...,N \qquad \text{Εξ. 3.10}$$

Θεωρούμε ότι οι αρχικές θέσεις είναι τα αρχικά μας δεδομένα, π.χ. $x_i(t=0) = x_i$ $(\forall i = 1,...,N)$ και ότι οι αρχικές ταχύτητες $(u_i \equiv dx_i/dt)$ είναι ίσες με μηδέν. Ενσωματώνουμε τις εξισώσεις κίνησης σε μικρά χρονικά βήματα, δt, όπου θεωρείται ότι οι δυνάμεις F_i παραμένουν σταθερές καθ' όλη τη διάρκεια του παραπάνω μικρού χρονικού διαστήματος. Σε κάθε βήμα, επαναφέρουμε τις ταχύτητες στο μηδέν για να αποφύγουμε τεχνουργήματα εξαιτίας της θερμικής ενέργειας. Τελικά, έχουμε το ακόλουθο σχήμα κίνησης:

$$x_i(t+\delta t) = x_i(t) + \frac{1}{2}\delta t^2 F_i. \qquad \text{Εξ. 3.11}$$

Η μεταβλητή δt είναι ιδιαίτερα σημαντική. Συγκεκριμένα, αν τα σημεία του συνόλου δεδομένων έχουν μικρές τιμές το δt πρέπει να έχει πάρα πολύ μικρή τιμή, αντίθετα,

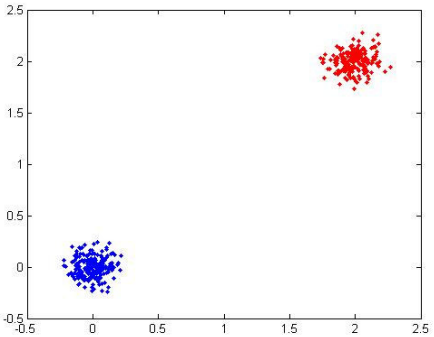
αν τα σημεία του συνόλου δεδομένων έχουν μεγάλες τιμές το δt πρέπει να έχει μικρή τιμή όμως μεγαλύτερη σε σχέση με πριν. Στόχος μας είναι το χρονικό βήμα των σημείων να προκαλεί πολύ μικρή αλλαγή της θέσης τους σε σχέση με την προηγούμενη τους θέση.

Επειδή η αλληλεπίδραση είναι ελκτική, μετά από μία χρονική περίοδο Τ τα σωματίδια που ανήκουν στην ίδια γειτονιά-ομάδα θα συγκεντρωθούν γύρω από το κέντρο της. Επομένως, μία αρχικά διάπλατη ομάδα στη συνέχεια μαζεύεται σε μέγεθος. Μετά από καθορισμένο αριθμό βημάτων ή όταν τα βήματα καταστούν τόσο μικρά ώστε οι επιπλέον επαναλήψεις να μην προκαλούν κάποια ιδιαίτερη διαφοροποίηση, η προσομοίωσή ολοκληρώνεται. Προσοχή, αν ο αριθμός των βημάτων είναι μεγαλύτερος από το προβλεπόμενο, τα σημεία του συνόλου δεδομένων θα τείνουν να συγκεντρωθούν γύρω από μοναδικό κέντρο. Συγκεκριμένα, χρησιμοποιούμε το ακόλουθο κριτήριο:

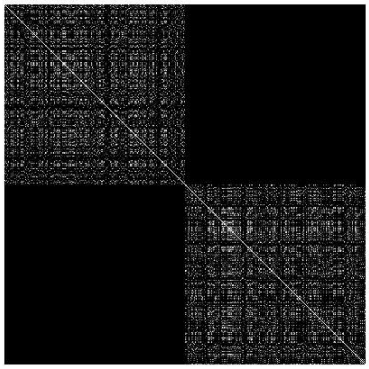
$$\frac{\sum_{i=1}^{N} | x_i(t+\delta t) - x_i(t) |}{\sum_{i=1}^{N} | x_i(t+\delta t) - x_i |} < \eta, \qquad \text{Εξ. 3.12}$$

όπου η είναι μία μικρή θετική τιμή (όπως π.χ. 10^{-3}).

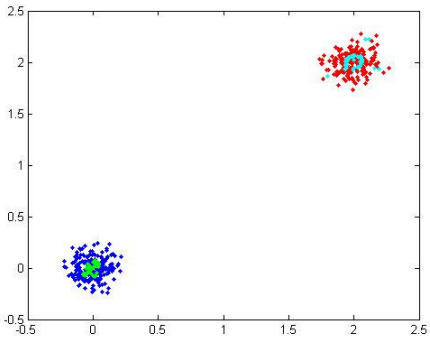
Στο Σχήματα 3.1, 3.2, 3.3 και 3.4 παραθέτουμε κάποια παραδείγματα συρρίκνωσης ομάδων. Συγκεκριμένα, δημιουργήσαμε με τυχαία δειγματοληψία από την κανονική κατανομή δύο σύνολα δεδομένων (Σχήματα 3.1 και 3.2) με 200 σημεία το καθένα, δt ίσο με 10^{-4} και σ περίπου ίσο με 0.03. Με τον ίδιο τρόπο δημιουργήσαμε ακόμα δύο σύνολα δεδομένων (Σχήματα 3.3 και 3.4) με 1000 σημεία το καθένα, δt ίσο με 10^{-4} και σ περίπου ίσο με 0.3. Για κάθε σύνολο δεδομένων εφαρμόσαμε τη μέθοδο για αυξανόμενο κάθε φορά αριθμό Νευτώνειων βημάτων. Στα σχήματα 3.2 και 3.5 παρατηρούμε ότι είναι περισσότερο ευδιάκριτες οι ομάδες που υπάρχουν στα σύνολα δεδομένων μετά την εφαρμογή της συρρίκνωσης (δηλαδή μετά από κάποιο αριθμό Νευτώνειων βημάτων) σε σχέση με το πόσο ευδιάκριτες είναι στα αρχικά (μη συρρικνωμένα) σύνολα.



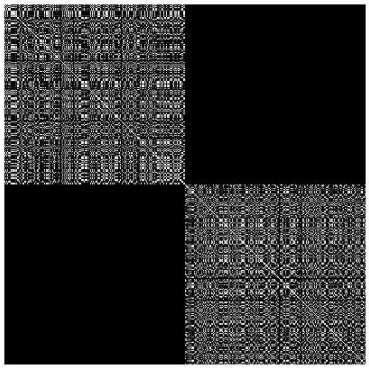
α) αρχικό σύνολο δεδομένων



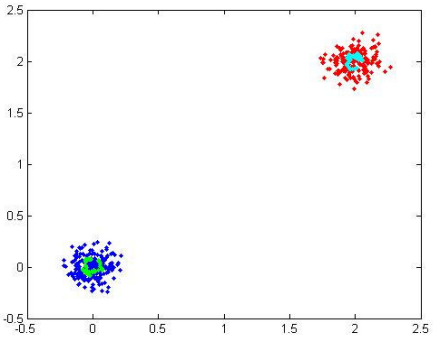
α) πίνακας ομοιότητας του αρχικού δεδομένων



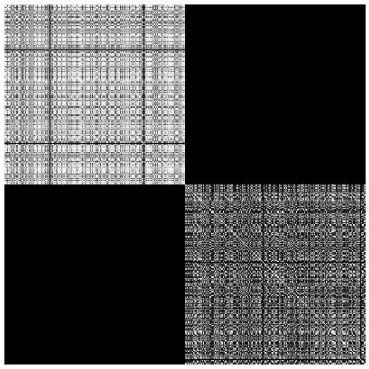
β) σύνολο δεδομένων μετά από 50 Νευτώνεια βήματα



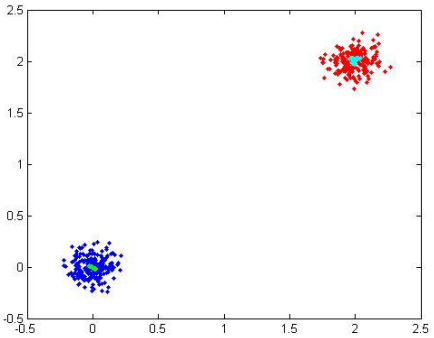
β) πίνακας ομοιότητας του συνόλου δεδομένων μετά από 50 Νευτώνεια βήματα



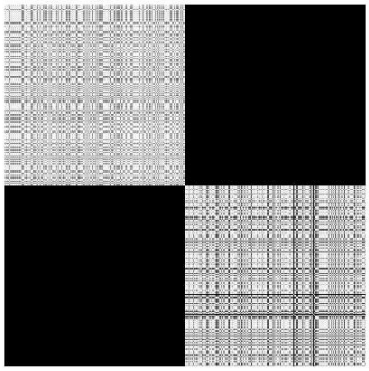
γ) σύνολο δεδομένων μετά από 100 Νευτώνεια βήματα



γ) πίνακας ομοιότητας του συνόλου δεδομένων μετά από 100 Νευτώνεια βήματα

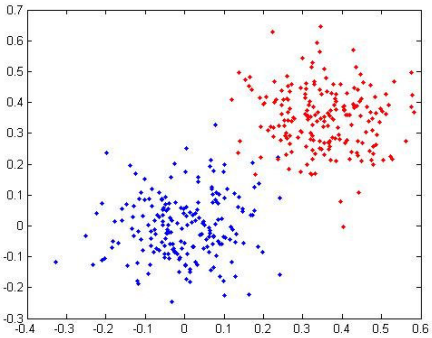


δ) σύνολο δεδομένων μετά από 150 Νευτώνεια βήματα

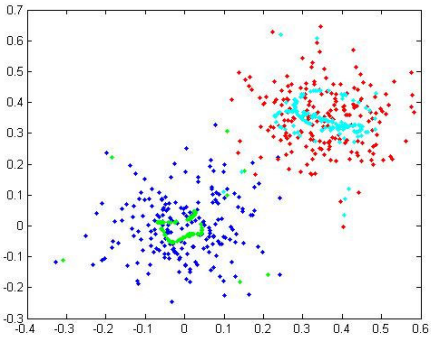


δ) πίνακας ομοιότητας του συνόλου δεδομένων μετά από 150 Νευτώνεια βήματα

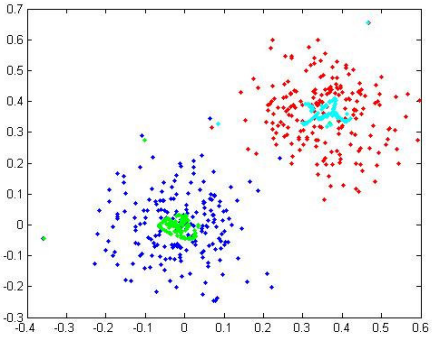
Σχήμα 3.1 Πρώτο σύνολο δεδομένων (K=2, N=200).



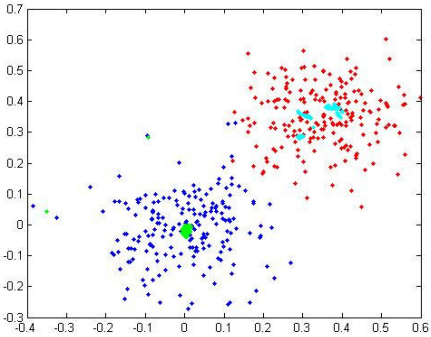
α) αρχικό σύνολο δεδομένων



β) σύνολο δεδομένων μετά από 25 Νευτώνεια βήματα

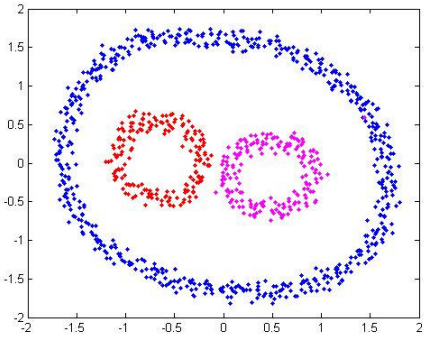


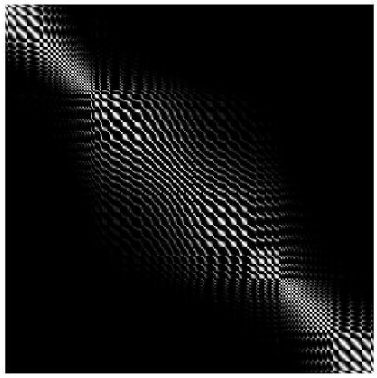
γ) σύνολο δεδομένων μετά από 75 Νευτώνεια βήματα



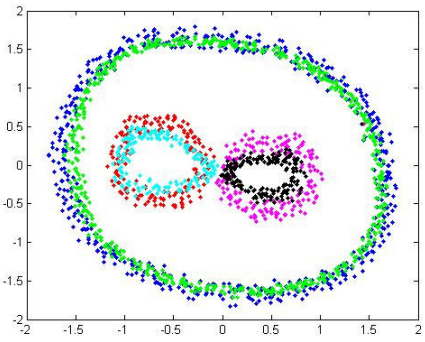
δ) σύνολο δεδομένων μετά από 200 Νευτώνεια βήματα

Σχήμα 3.2 Δεύτερο σύνολο δεδομένων (K=2, N=200).

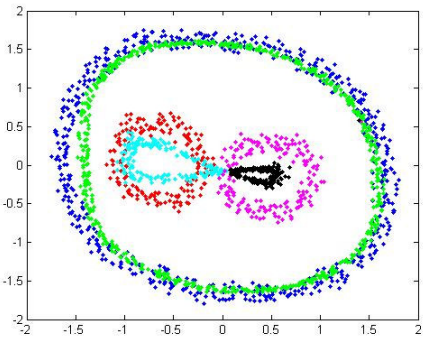




α) αρχικό σύνολο δεδομένων

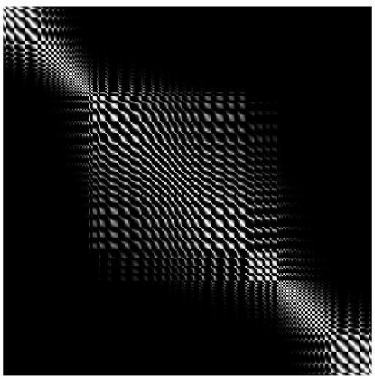


β) σύνολο δεδομένων μετά από 20 Νευτώνεια βήματα

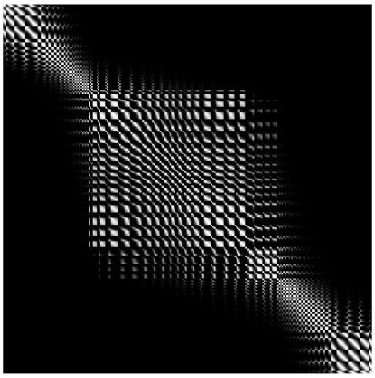


γ) σύνολο δεδομένων μετά από 30 Νευτώνεια βήματα

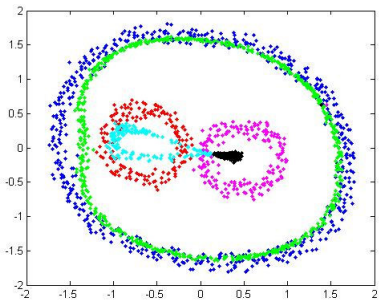
α) πίνακας ομοιότητας του αρχικού συνόλου δεδομένων



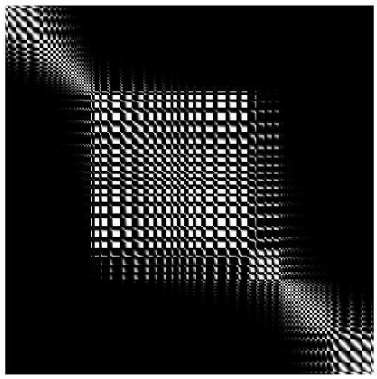
β) πίνακας ομοιότητας του συνόλου δεδομένων μετά από 20 Νευτώνεια βήματα



γ) πίνακας ομοιότητας του συνόλου δεδομένων μετά από 30 Νευτώνεια βήματα

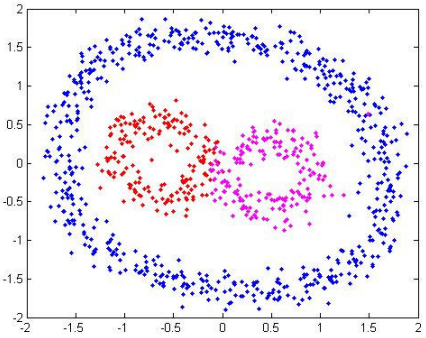


δ) σύνολο δεδομένων μετά από 40 Νευτώνεια βήματα

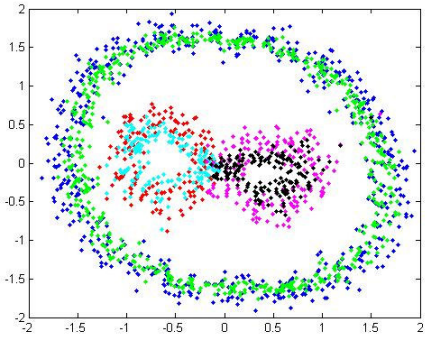


δ) πίνακας ομοιότητας του συνόλου δεδομένων μετά από 40 Νευτώνεια βήματα

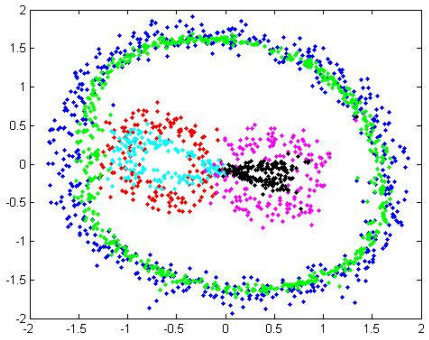
Σχήμα 3.3 Τρίτο σύνολο δεδομένων με (K=3, N=1000).



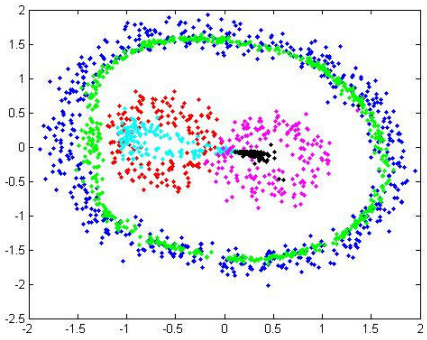
α) αρχικό σύνολο δεδομένων



β) σύνολο δεδομένων μετά από 20 Νευτώνεια βήματα



γ) σύνολο δεδομένων μετά από 30 Νευτώνεια βήματα



δ) σύνολο δεδομένων μετά από 40 Νευτώνεια βήματα

Σχήμα 3.4 Τέταρτο σύνολο δεδομένων με (K=3, N=1000).

3.4. Καθορισμός του Εύρους του Δυναμικού (σ)

Όπως προαναφέραμε, ο υπολογισμός της παραμέτρου σ έχει ιδιαίτερη σημασία. Συγκεκριμένα, η τιμή του σ καθορίζει το εύρος του Γκαουσιανού δυναμικού και επομένως το πλήθος των ομάδων της τελικής ομαδοποίησης. Στην εργασία [20] έχει προταθεί μία απλή μέθοδος εκτίμησης της τιμής του σ μέσω επαναληπτικής εκτέλεσης του αλγορίθμου ομαδοποίησης για διάφορες τιμές του σ με ένα βήμα διακριτοποίησης στο $[\sigma_{\min}, \sigma_{\max}]$. Τελικά, η μέθοδος επιλέγει εκείνο το σ που οδηγεί στη λύση με την καλύτερη ομαδοποίηση.

Παρακάτω, παρουσιάζουμε μία περισσότερο συστηματική μεθοδολογία. Τα βασικά στοιχεία που χρησιμοποιούμε στην ανάλυση μας είναι η μέση απόσταση του κοντινότερου γείτονα και η θεωρία των διατεταγμένων παρατηρήσεων (order statistics). Συγκεκριμένα, συμβολίζουμε με $d_j^{(i)}$ την απόσταση μεταξύ του σημείου x_i και του j-οστού κοντινότερού του γείτονα. Επομένως, έχουμε την ακόλουθη διατεταγμένη σειρά αποστάσεων (κατά αύξουσα σειρά):

$$d_1^{(i)} < d_2^{(i)} ... < d_{N-1}^{(i)} \qquad \forall i = 1, 2, ..., N-1 \qquad \text{Εξ. 3.13}$$

Η μέση δειγματική απόσταση της μεταβλητής d_j από τον j-οστό κοντινότερο γείτονα σε ένα σύνολο N σημείων δίνεται από:

$$\left\langle d_j \right\rangle = \frac{1}{N}\sum_{i=1}^{N} d_j^{(i)} \qquad \forall j = 1,2,...,N-1 \qquad \text{Εξ. 3.14}$$

Ακόμα, η μέση τετραγωνισμένη δειγματική απόσταση της μεταβλητής d_j από τον j-οστό κοντινότερο γείτονα σε ένα σύνολο N σημείων δίνεται από:

$$\left\langle d_j^2 \right\rangle = \frac{1}{N}\sum_{i=1}^{N} (d_j^{(i)})^2 \qquad \forall j = 1,2,...,N-1 \qquad \text{Εξ. 3.15}$$

Επομένως, η δειγματική διακύμανση της μεταβλητής d_j μπορεί να υπολογιστεί ως εξής:

$$\tilde{\sigma}_{N,m}^2 = \frac{1}{m}\sum_{k=1}^{m} (\left\langle d_k^2 \right\rangle - \left\langle d_k \right\rangle^2), \quad \forall m = 1,2,...,N-1 \qquad \text{Εξ. 3.16}$$

όπου m είναι το πλήθος των κοντινότερων γειτόνων. Στην εργασία [4] έχει αποδειχθεί ότι, στην περίπτωση που υπάρχει μία μόνο ομάδα, ισχύει ότι:

$$\tilde{\sigma}_{N,m}^2 = \alpha(N)(m+1)^2 + \beta(N)(m+1), \qquad \text{Εξ. 3.17}$$

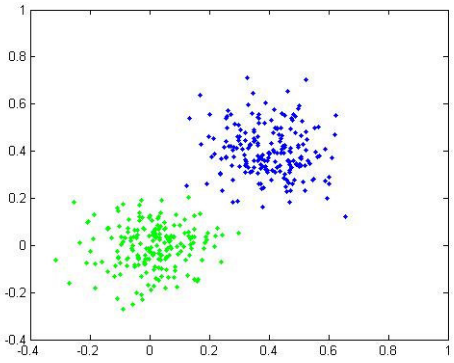
όπου $a(N)$ και $\beta(N)$ δύο συντελεστές. Δηλαδή, ότι υπάρχει μία δευτέρου βαθμού σχέση που διέπει το σ ως συνάρτηση του m. Όταν υπάρχουν περισσότερες από μία ομάδες, η απόσταση $\left\langle d_j \right\rangle$ αποκτά ασυνέχειες και το $\tilde{\sigma}_{N,m}^2$ προκύπτει από μία εναπόθεση από πάνω των μετασχηματισμένων τετραγώνων. Στη συνέχεια, στόχος μας είναι να εκτιμήσουμε το εύρος του δυναμικού. Επομένως, προσπαθούμε να εντοπίσουμε την μικρότερη τιμή που μπορεί να έχει το m, την οποία συμβολίζουμε με m^*, και για την οποία ισχύει ότι η δεύτερη παράγωγος του $\tilde{\sigma}_{N,m}^2/(m+1)$,ως προς m, μηδενίζεται. Επειδή, η ποσότητα $\frac{\tilde{\sigma}_{N,m}^2}{m+1}$ είναι γραμμική ως προς το m, η δεύτερη παράγωγος μηδενίζεται. Οπότε, με m^* θα συμβολίζουμε τη μικρότερη τιμή που μπορεί να έχει το m και για την οποία ισχύει ότι η δεύτερη παράγωγος του $\tilde{\sigma}_{N,m}^2/(m+1)$,ως προς m, μηδενίζεται.

Πρακτικά, αναζητούμε το ελάχιστο m για το οποίο ισχύει ότι:

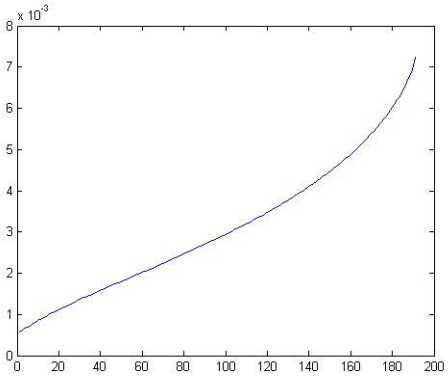
$$\left| \frac{\tilde{\sigma}_{N,m+1}^2}{m+2} + \frac{\tilde{\sigma}_{N,m-1}^2}{m} - \frac{2\tilde{\sigma}_{N,m}^2}{m+1} \right| < \varepsilon \left| \frac{\tilde{\sigma}_{N,m}^2}{m+1} \right|. \qquad \text{Εξ. 3.18}$$

Από τα πειράματα που πραγματοποιήσαμε, παρατηρούμε ότι, γύρω από το m^* υπάρχει μία ευρεία και σταθερή περιοχή για να υπολογίσουμε τη τιμή της διακύμανσης $\left(\hat{\sigma}^2 = \sigma^2_{m^*}\right)$, όπου και η επίδοση της προσέγγισής μας ήταν πανομοιότυπη. Με άλλα λόγια, οποιαδήποτε τιμή του σ ανήκει στο προαναφερθέν σταθερό διάστημα οδηγεί στον ίδιο αριθμό ομάδων.

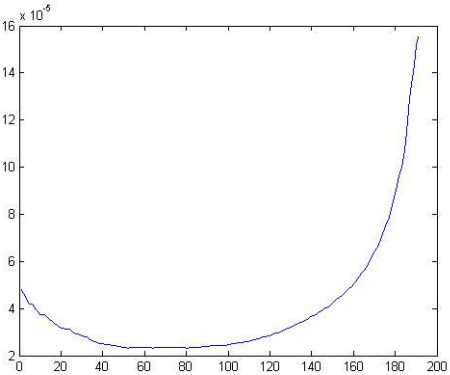
Στα Σχήματα 3.8 και 3.9 παραθέτουμε δύο παραδείγματα συνόλων δεδομένων με δύο ομάδες (K=2) και τρεις ομάδες (K=3) αντίστοιχα. Ακόμα, παραθέτουμε το διάγραμμα της πρώτης παραγώγου του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$ ως προς m και το διάγραμμα της δεύτερης παραγώγου του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$ ως προς m. Σημειώνουμε το σημείο m*, που προέκυψε μέσω της Εξίσωσης (3.18), με ένα βέλος.



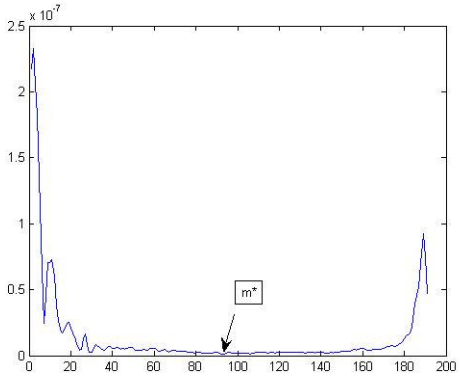
α) σύνολο δεδομένων με K=2



β) $\tilde{\sigma}^2_{N,m}$

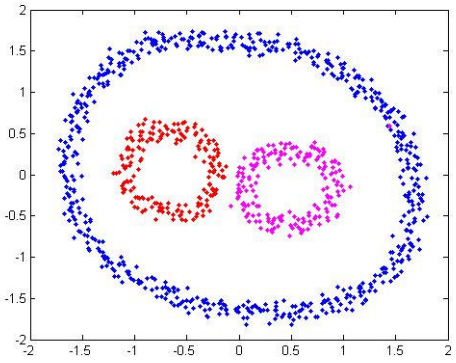


γ) 1η παράγωγος του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$

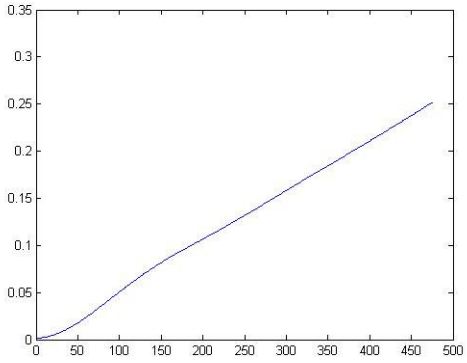


δ) 2η παράγωγος του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$

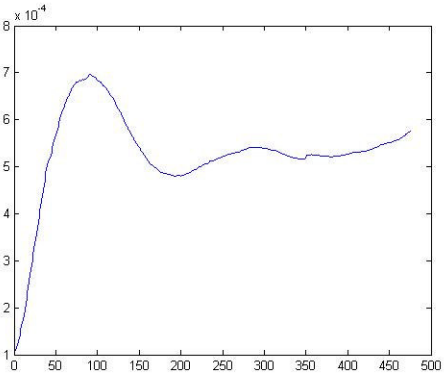
Σχήμα 3.5 Πρώτο σύνολο δεδομένων με διαγράμματα του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$.



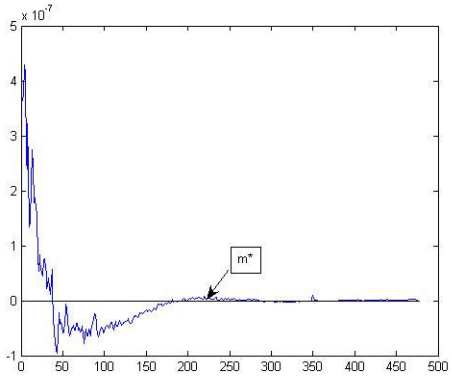
α) σύνολο δεδομένων με K=3



β) $\tilde{\sigma}^2_{N,m}$



γ) 1η παράγωγος του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$



δ) 2η παράγωγος του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$

Σχήμα 3.6 Δεύτερο σύνολο δεδομένων με διαγράμματα του $\frac{\tilde{\sigma}^2_{N,m}}{m+1}$.

ΚΕΦΑΛΑΙΟ 4. ΝΕΥΤΩΝΕΙΑ ΦΑΣΜΑΤΙΚΗ ΟΜΑΔΟΠΟΙΗΣΗ

4.1. Εισαγωγή

Σε αυτό το κεφάλαιο, παρουσιάζουμε και αναλύουμε τη Νευτώνεια Φασματική Ομαδοποίηση [5] η οποία συνδυάζει τα χαρακτηριστικά της Φασματικής και της Νευτώνειας Ομαδοποίησης. Η μέθοδός μας, μέσω των Νευτώνειων εξισώσεων κίνησης, οδηγεί στην κατασκευή ενός αραιού (sparse) πίνακα ομοιότητας που μπορεί να χρησιμοποιηθεί στην επίλυση προβλημάτων Φασματικής Ομαδοποίησης. Η επίδοση που εμφανίζει είναι καλύτερη σε σχέση με αυτήν της Φασματικής Ομαδοποίησης. Στη συνέχεια, περιγράφουμε πώς τροποποιήσαμε τη μέθοδό μας ώστε να μπορούμε να αντιμετωπίσουμε προβλήματα μεγάλης διάστασης όπως αυτά της ομαδοποίησης κειμένων (document clustering). Συγκεκριμένα, επιλέξαμε μία διαφορετική δυναμική συνάρτηση και επομένως μία ελαφρώς τροποποιημένη εξίσωση κίνησης.

4.2. Προτεινομένη Μέθοδος

Κεντρικό ζήτημα της προτεινόμενης μεθοδολογίας είναι η κατασκευή ενός αραιού πίνακα ομοιότητας και ο εμπλουτισμός του με μεγαλύτερη πληροφορία.

Δανειζόμενη στοιχεία από την Νευτώνεια Ομαδοποίηση, η μέθοδος χρησιμοποιεί τις εξισώσεις κίνησης του Νεύτωνα ώστε να μετακινήσει τα δεδομένα. Έτσι, συρρικνώνει τα γειτονικά σημεία έλκοντάς τα προς το κέντρο της γειτονιάς, και απωθεί τα απομακρυσμένα σημεία καθώς, δεν ανήκουν στην ίδια ομάδα. Με τον τρόπο αυτό, "ανταμείβει" τα γειτονικά σημεία ενισχύοντας το ποσό της μεταξύ τους ομοιότητας και "τιμωρεί" τα ανόμοια σημεία μειώνοντας ή μηδενίζοντας το ποσό ομοιότητάς τους. Το παραπάνω οφείλεται στη μείωση της Ευκλείδειας απόστασης μεταξύ των γειτονικών σημείων και στην περεταίρω αύξησή της όταν τα σημεία είναι μεταξύ τους ανόμοια. Τελικά, η διαδικασία αυτή επιτρέπει τον εμπλουτισμό του πίνακα ομοιότητας με μεγαλύτερη πληροφορία και την μετατροπή του σε αραιό πίνακα με σημαντικά πλεονεκτήματα.

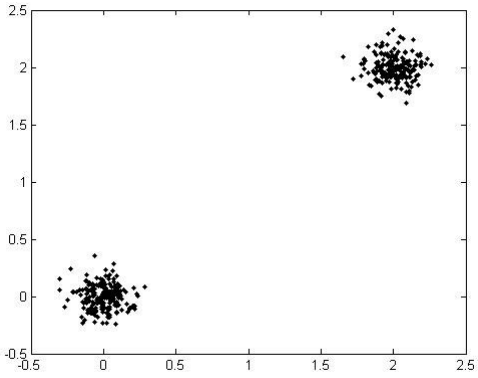
Στη συνέχεια, έχοντας υπολογίσει το πίνακα ομοιότητας W, εφαρμόζουμε τα βήματα του NJW αλγορίθμου όπως παρουσιάζονται στο Σχήμα 2.3. Δηλαδή, περιληπτικά, υπολογίζουμε τον πίνακα Laplace L του W. Έπειτα, υπολογίζουμε τα K κανονικά ιδιοδιανύσματα $u_1,u_2,\ldots,u_K$ που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές. Τελικά, καταλήγουμε στην τελική ομαδοποίηση με τη βοήθεια του αλγορίθμου K-means. Στο Σχήμα 4.1 παραθέτουμε τον ψευδοκώδικα της προτεινομένης μεθόδου.

Ψευδοκώδικας της Νευτώνειας Φασματικής Ομαδοποίησης
Είσοδος: $X = \{x_1, ..., x_N\}$, όπου N το πλήθος των παρατηρήσεων **Βήμα 1^{ο}:** Υπολογισμός του σ και ανάθεση τιμής στη μεταβλητή δt (χρονικό βήμα) (κάποια ενδεικτική τιμή είναι το 10^{-4}). **Βήμα 2^{ο}:** Εφαρμογή της Νευτώνειας Ομαδοποίησης $\rightarrow$ συρρίκνωση των ομάδων Αν $x_i(0), \ \forall i = 1, ..., N$: η αρχική θέση των δεδομένων τότε $x_i(\mathrm{T}) \rightarrow$ η τελική τους θέση μετά από χρονικό βήμα *T*. **Βήμα 3^{ο}:** Υπολογισμός του αραιού πίνακα ομοιότητας W: υπολογισμός των Ευκλείδειων αποστάσεων όλων των σημείων του αρχικού συνόλου δεδομένων ανά ζεύγη $\rightarrow dist_{ij}(0)$. υπολογισμός των Ευκλείδειων αποστάσεων όλων των σημείων όπως προέκυψαν μετά τη συρρίκνωση του 2^{ου} βήματος $\rightarrow dist_{ij}(T)$. Αν $dist_{ij}(0) - dist_{ij}(T) < 0$ τότε $W_{ij} = 0$ (μηδενισμός των θέσεων του W που αντιστοιχούν σε σημεία που μετά την τελική τους μετακίνηση αποκλίνουν). **Βήμα 4^{ο}:** Εφαρμογή του NJW αλγορίθμου.

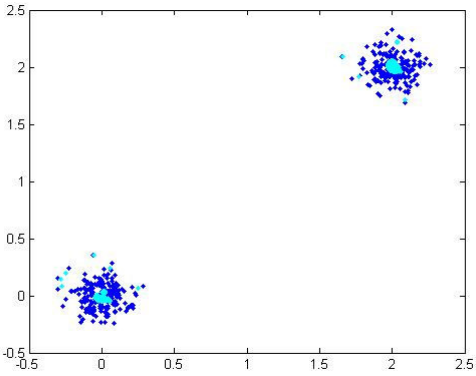
Σχήμα 4.1 Ψευδοκώδικας προτεινόμενης μεθόδου.

Στα σχήματα που ακολουθούν παρουσιάζουμε σταδιακά όλες τις φάσεις της μεθόδου της Νευτώνειας Φασματικής Ομαδοποίησης σε δύο διαφορετικά σύνολα δεδομένων. Συγκεκριμένα, δημιουργήσαμε τα σύνολα δεδομένων (Σχήματα 4.2 και 4.4) με τυχαία δειγματοληψία από την κανονική κατανομή. Κατά την διάρκεια των πειραμάτων, ο

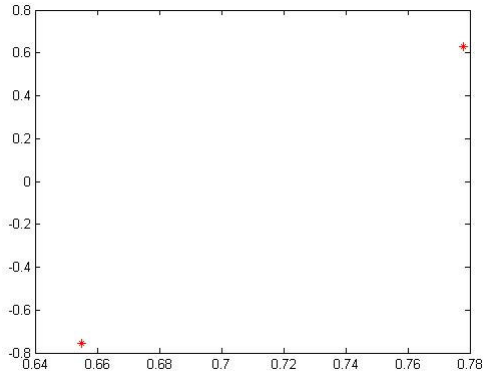
αριθμός των Νευτώνειων βημάτων (T) ήταν ίσος με 50, το χρονικό βήμα (δt) ήταν ίσο με 10^{-4} και το σ περίπου ίσο με 0.05 και 0.3 αντίστοιχα. Αρχικά, παρουσιάζουμε για κάθε σύνολο δεδομένων τα σημεία όπως έχουν μετακινηθεί προς το κέντρο της ομάδας στην οποία ανήκει το καθένα. Ακόμα, παρουσιάζουμε το διάγραμμα των K βασικών ιδιοδιανυσμάτων. Επιπλέον, παρουσιάζουμε τα K ιδιοδιανύσματα μετά την τελική ομαδοποίηση μέσω του αλγορίθμου K-means. Τέλος, παρουσιάζουμε τα σύνολα δεδομένων μετά την τελική ομαδοποίηση.



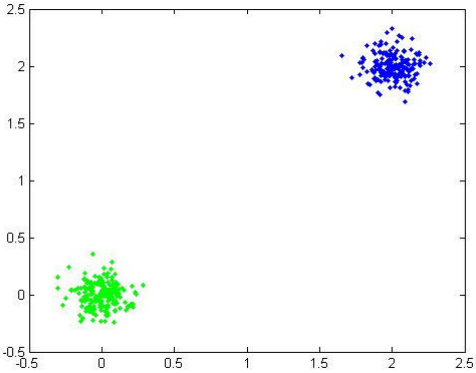
Σχήμα 4.2 Δεύτερο σύνολο δεδομένων.



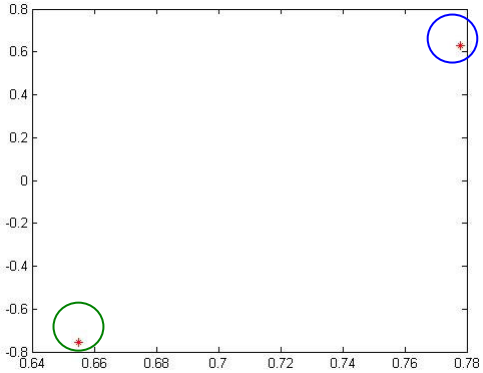
α) Σύνολο δεδομένων μετά από 50 Νευτώνεια βήματα



β) Διάγραμμα των δύο κύριων ιδιοδιανυσμάτων.

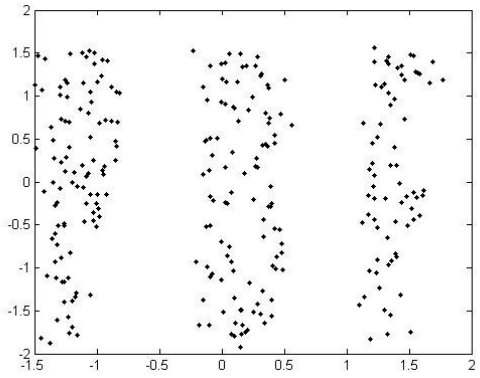


γ) σύνολο των δεδομένων μετά την τελική ομαδοποίηση.

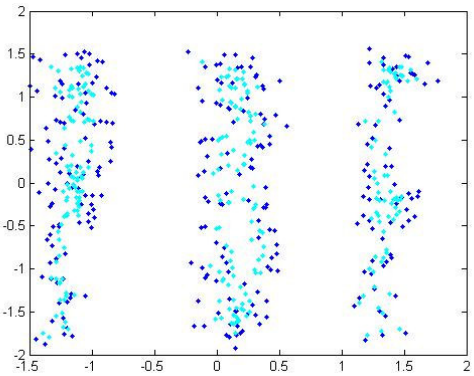


δ) μετά την εκτέλεση του K-means.

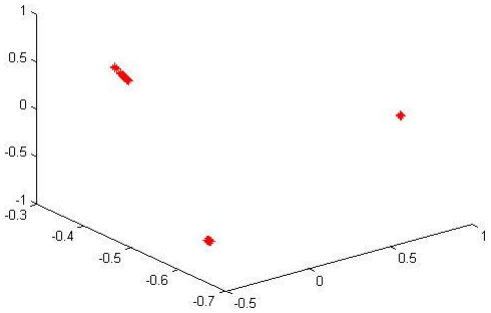
Σχήμα 4.3 Τα βήματα της Νευτώνειας Φασματικής ομαδοποίησης στο σύνολο του Σχήματος 4.2.



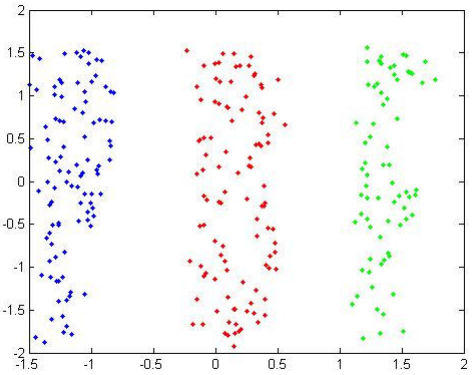
Σχήμα 4.4 Τρίτο σύνολο δεδομένων.



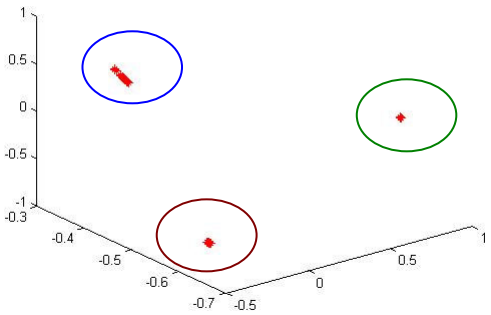
α) Σύνολο δεδομένων μετά από 50 Νευτώνεια βήματα



β) Διάγραμμα των τριών κύριων ιδιοδιανυσμάτων



γ) σύνολο των δεδομένων μετά την τελική ομαδοποίηση



δ) μετά την εκτέλεση του K-means.

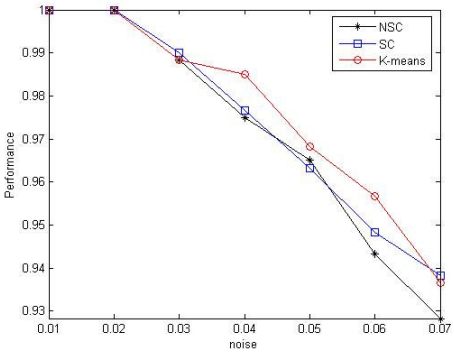
Σχήμα 4.5 Τα βήματα της Νευτώνειας Φασματικής ομαδοποίησης στο σύνολο του Σχήματος 4.4.

Στα Σχήματα 4.6 και 4.7, για να τονίσουμε τη σημασία του χρονικού βήματος (μεταβλητή δt), παρουσιάζουμε την εφαρμογή της μεθόδου μας στα σύνολα δεδομένων που παρουσιάζονται στα Σχήματα 4.2 και 4.4. Συγκεκριμένα, συγκρίναμε την επίδοση και την αποτελεσματικότητα της μεθόδου μας με τα αποτελέσματα της κλασικής μεθόδου της Φασματικής Ομαδοποίησης και του παραδοσιακού K-means αλγορίθμου. Εφαρμόσαμε την μέθοδό μας σε κάθε συνόλου δεδομένων, με στόχο τη συγκέντρωσή των σημείων του γύρω από το κέντρο της ομάδας στην οποία ανήκουν.

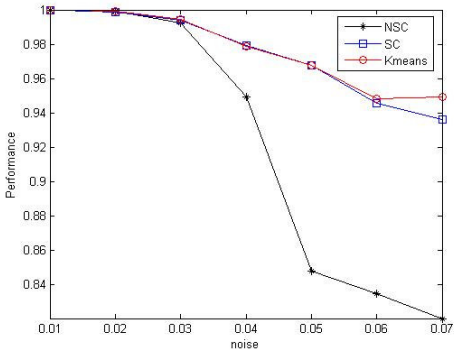
Επιπλέον, για την διεξαγωγή των πειραμάτων μας, συνεχώς μεταβάλλαμε το εύρος των ομάδων του κάθε συνόλου δεδομένων έτσι ώστε σε κάθε βήμα οι ομάδες δεδομένων να "πλησιάζουν" όλο και περισσότερο και να γίνονται με αυτόν τον τρόπο λιγότερο ευδιάκριτες. Με άλλα λόγια, προσθέσαμε θόρυβο στα δεδομένα. Συγκεκριμένα, διεξήγαμε 40 πειράματα για κάθε διαφορετική τιμή θορύβου. Στα Σχήματα 4.6(α) και 4.7(α) το χρονικό βήμα ήταν ίσο με 10^{-4}. Ενώ στα 4.6(β), 4.7(β) και 4.6(γ), 4.7(γ) το χρονικό βήμα ήταν ίσο με 10^{-6} (μικρό σε σχέση με τις τιμές των σημείων του συνόλου δεδομένων) και 10^{-2} (μεγάλο σε σχέση με τις τιμές των σημείων του συνόλου δεδομένων) αντίστοιχα.

Στο Σχήμα 4.7, για δt ίσο με 10^{-4}, μπορούμε να παρατηρήσουμε ότι οι επιδόσεις των όλων των μεθόδων δίνουν εξίσου καλή επίδοση. Για τις υπόλοιπες τιμές του δt παρατηρούμε ότι η επίδοση της μεθόδου μας αν και είναι αρκετά ικανοποιητική δεν υπερτερεί ως προς τις υπόλοιπες χρησιμοποιούμενες μεθόδους.

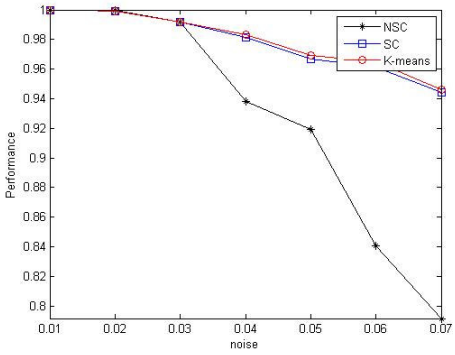
Στο Σχήμα 4.8, για δt ίσο με 10^{-4}, μπορούμε εύκολα να παρατηρήσουμε ότι η επίδοση της προτεινομένης μεθόδου (NSC) είναι καλύτερη σε σχέση με τη επίδοση της κλασικής μεθόδου Φασματικής Ομαδοποίησης ιδιαίτερα σε περιβάλλον υψηλού θορύβου. Ακόμα, ο παραδοσιακός K-means αλγόριθμος αποτυγχάνει στην εύρεση των ομάδων στα δοθέντα σύνολά δεδομένων καθώς το σχήμα των ομάδων δεν είναι σφαιρικό. Για δt ίσο με 10^{-6}, μπορούμε να παρατηρήσουμε ότι η επίδοση της μεθόδου μας δεν ήταν αρκετά ικανοποιητική αν και παρέμεινε καλύτερη σε σχέση με τις επιδόσεις της κλασικής μεθόδου Φασματικής Ομαδοποίησης και του K-means αλγορίθμου. Για δt ίσο με 10^{-2}, παρατηρούμε ότι η επίδοση της μεθόδου μας ούτε είναι αρκετά ικανοποιητική ούτε είναι καλύτερη σε σχέση με τις επιδόσεις της κλασικής μεθόδου Φασματικής Ομαδοποίησης και του K-means αλγορίθμου. Το παραπάνω συμβαίνει καθώς δεν είναι επιθυμητό το βήμα κίνησης των σημείων να προκαλεί πολύ μεγάλη αλλαγή της θέσης τους σε σχέση με την προηγούμενη τους θέση.



α) $\delta t=10^{-4}$

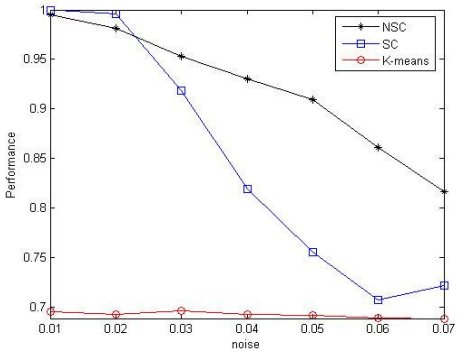


β) $\delta t=10^{-6}$

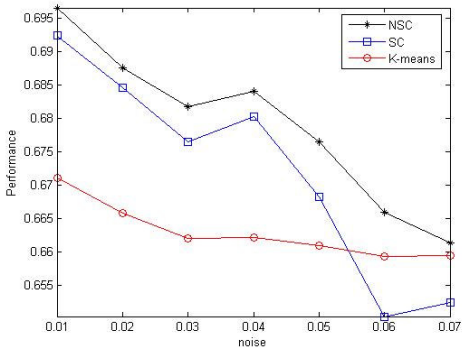


γ) $\delta t=10^{-2}$

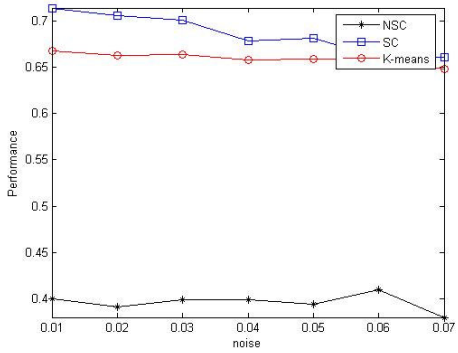
Σχήμα 4.6 Συγκριτικά αποτελέσματα για διαφορετικές τιμές του δt στο σύνολο δεδομένων του Σχήματος 4.2.



α) $\delta t=10^{-4}$



β) $\delta t=10^{-6}$



γ) $\delta t=10^{-2}$

Σχήμα 4.7 Συγκριτικά αποτελέσματα για διαφορετικές τιμές του δt στο σύνολο δεδομένων του Σχήματος 4.4.

Παρακάτω, θα αναφέρουμε μία τροποποίηση της μεθόδου μας ώστε να μπορούμε να αντιμετωπίσουμε προβλήματα μεγάλης διάστασης όπως αυτά της ομαδοποίησης κειμένων. Το πρόβλημα της ομαδοποίησης κειμένων αφορά την διαίρεση μίας συλλογής κειμένων σε ομάδες με βάση την ομοιότητά τους.

4.3. Επέκταση σε Ομαδοποίηση Κειμένων

Συγκεκριμένα, στη μελέτη μας, μετατρέπουμε το κείμενο εισόδου σε ένα διάνυσμα χαρακτηριστικών $x_i \in R^M$, όπου το M είναι το μέγεθος του λεξικού και είναι τέτοιο ώστε το κάθε χαρακτηριστικό να αποτελεί το βάρος του όρου στον οποίο αντιστοιχεί. Για το καθορισμό του βάρους, χρησιμοποιήσαμε το TF-IDF (συχνότητα όρου,

αντίστροφος συχνότητας κειμένου) σχήμα. Επιπλέον, για να υπολογίσουμε την εγγύτητα μεταξύ κάθε ζεύγους κειμένων, χρησιμοποιήσαμε την συνημιτονοειδή ομοιότητα. Ακόμα, καθώς τα κείμενα είναι κανονικά διανύσματα, μπορούμε, με απλό τρόπο, να υπολογίσουμε την ομοιότητα με τον ακόλουθο τύπο:

$$V_{ij} = x_i^T x_j . \qquad \text{Εξ. 4.1}$$

Ο παραπάνω τύπος αποτελεί την νέα δυναμική συνάρτηση, οπότε, θα επιλέξουμε και μία ελαφρώς τροποποιημένη εξίσωση κίνησης.

Αρχικά, θεωρούσαμε ότι τα δεδομένα μας αντιστοιχούν σε σωματίδια που αλληλεπιδρούν διαμέσου ενός, εφαρμόσιμου μεταξύ δύο σωμάτων, ελκτικού δυναμικού. Πλέον, η αλληλεπίδραση μεταξύ των σωματιδίων δεν είναι πάντοτε ελκτική. Συγκεκριμένα, κάθε σωματίδιο επηρεάζεται θετικά από παρόμοια κείμενα (δηλαδή, κείμενα της ίδιας ομάδας) και η μεταξύ τους αλληλεπίδραση είναι ελκτική (εφαρμόζεται θετική δύναμη). Αντίθετα, η αλληλεπίδραση μεταξύ διαφορετικών κειμένων είναι απωστική (εφαρμόζεται αρνητική δύναμη). Τελικά, υπολογίζουμε τη δύναμη F_i με βάση τον τύπο που ακολουθεί

$$F_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} c_{ij}(t) x_j \text{ , όπου } c_{ij}(t) = \begin{cases} +1, \alpha \nu V_{ij} > \overline{V}_{ij} \\ -1, \alpha \lambda \lambda \iota \omega \varsigma \end{cases} \qquad \text{Εξ. 4.2}$$

Στην πραγματικότητα, η ποσότητα $\overline{V}_{ij}$ δρα σαν κατώφλι ομοιότητας για την διάκριση μεταξύ των όμοιων και των ανόμοιων κειμένων.

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ

5.1. Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζουμε τα πειραματικά αποτελέσματα της μεθόδου της Νευτώνειας Φασματικής Ομαδοποίησης για διάφορα σύνολα δεδομένων. Ειδικότερα, ο προσανατολισμός και στόχος των πειραμάτων αυτών είναι η σύγκριση της επίδοσης και της αποτελεσματικότητας της προτεινόμενης μεθόδου με τα αποτελέσματα της κλασικής μεθόδου της Φασματικής Ομαδοποίησης και του παραδοσιακού K-means αλγορίθμου. Συγκεκριμένα, εξετάσαμε τη μέθοδό μας σε ευρέως γνωστές δοκιμασίες επιδόσεως που κυμάνθηκαν από αριθμητικά δεδομένα σε κατάτμηση εικόνας και σε προβλήματα ομαδοποίησης κειμένων.

5.2. Πειραματική Μελέτη σε Δεδομένα με Αριθμητικά Χαρακτηριστικά

Κατά την διάρκεια όλων των πειραμάτων σε δεδομένα με αριθμητικά χαρακτηριστικά, ήταν γνωστός ο αριθμός των ομάδων καθώς και η πραγματική

ομάδα στην οποία ήταν τοποθετημένο το κάθε δεδομένο. Ακόμα, σε κάθε σειρά πειραμάτων ήταν σταθερός ο αριθμός του Νευτώνειου βήματος καθώς και η τιμή του χρονικού βήματος δt. Επιπλέον, και στις δύο προσεγγίσεις, NSC και SC, χρησιμοποιήσαμε την ίδια τιμή για το σ (με τον τρόπο που παρουσιάσαμε στο κεφάλαιο 3).

5.2.1. Τρόπος διεξαγωγής πειραμάτων

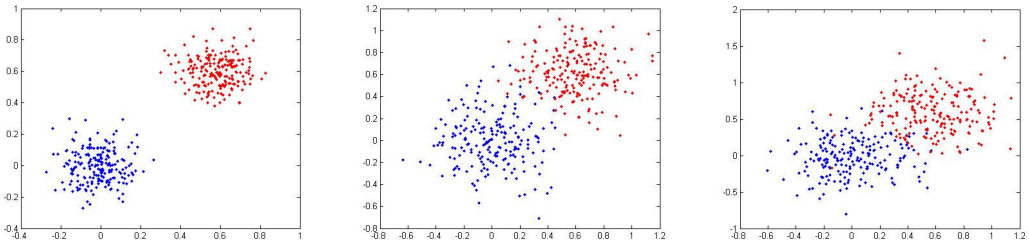
Αρχικά, δημιουργήσαμε με τυχαία δειγματοληψία από την κανονική κατανομή δύο σύνολα δεδομένων (K=2) (Σχήματα 5.1 και 5.2) με 200 σημεία το καθένα. Κατά την διάρκεια των πειραμάτων, ο αριθμός των Νευτώνειων βημάτων (T) ήταν ίσος με 50, το χρονικό βήμα (δt) ήταν ίσο με 10^{-4} και το σ περίπου ίσο με 0.17 και 0.14 αντίστοιχα.

Στη συνέχεια δημιουργήσαμε με το σπρέι της ζωγραφικής των Windows τέσσερα σύνολα δεδομένων. Συγκεκριμένα, κάθε σύνολο αποτελείται από κάποια γράμματα τα οποία και αποτυπώσαμε μέσω κουκίδων. Πιο αναλυτικά, στα σύνολα δεδομένων CAT, KYR, PDF και UOI (Σχήματα 5.3, 5.4, 5.5 και 5.6) έχουμε τρεις ομάδες δεδομένων (K=3) (αφού τρία είναι τα γράμματα που αποτελούν το κάθε σύνολο), ενώ στο σύνολο δεδομένων BIRD (Σχήμα 5.7) έχουμε τέσσερις ομάδες (K=4) (4 γράμματα). Κατά την διάρκεια των πειραμάτων, ο αριθμός των Νευτώνειων βημάτων ήταν ίσος με 50, το δt ίσο με 10^{-4} και το σ περίπου ίσο με 0.3 (για τα σύνολα CAT και PDF) και 0.2 (για τα σύνολα KYR, UOI και BIRD). Εφαρμόσαμε την μέθοδό μας στα σημεία του κάθε συνόλου δεδομένων με στόχο τη συγκέντρωσή τους γύρω από το κέντρο της ομάδας στην οποία ανήκουν. Επιπλέον, για την διεξαγωγή των πειραμάτων μας, συνεχώς μεταβάλλαμε το εύρος των ομάδων του κάθε συνόλου δεδομένων έτσι ώστε σε κάθε βήμα οι ομάδες δεδομένων να "πλησιάζουν" όλο και περισσότερο και να γίνονται με αυτόν τον τρόπο λιγότερο ευδιάκριτες. Με άλλα λόγια, προσθέσαμε θόρυβο στα δεδομένα. Συγκεκριμένα, "μετακινήσαμε" τα δεδομένα μας μέσω μίας δειγματοληπτικά τυχαίας κανονικής κατανομής. Συγκεκριμένα, διεξήγαμε 40 πειράματα για κάθε διαφορετική τιμή θορύβου.

5.2.2. Πειραματικά αποτελέσματα – αξιολόγηση

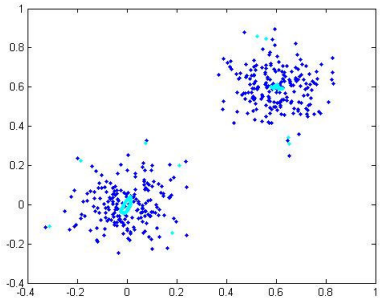
Στα Σχήματα 5.1(α), 5.2(α), 5.3(α), 5.4(α), 5.5(α), 5.6(α) και 5.7(α) παρουσιάζουμε τα σημεία του κάθε συνόλου δεδομένων δοθείσης της ύπαρξης ελάχιστης, μέτριας και μεγάλης τιμής θορύβου.

Στα Σχήματα 5.1(β), 5.2(β), 5.3(β), 5.4(β), 5.5(β), 5.6(β) και 5.7(β) παρουσιάζουμε τη διαδικασία συρρίκνωσης των ομάδων. Τελικά, παρουσιάζουμε την μέση τιμή των παραπάνω πειραμάτων για κάθε μέθοδο και τα συγκριτικά αποτελέσματα στα Σχήματα 5.1(γ), 5.2(γ), 5.3(γ), 5.4(γ), 5.5(γ), 5.6(γ) και 5.7(γ).

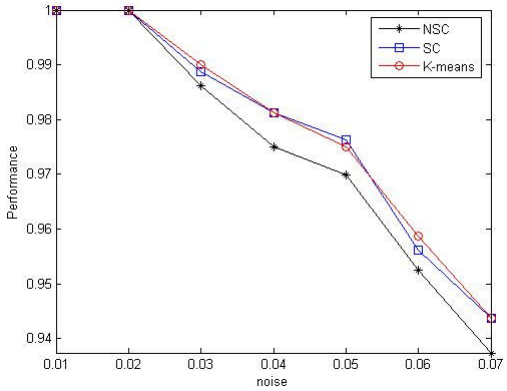


a) b) c)

α)

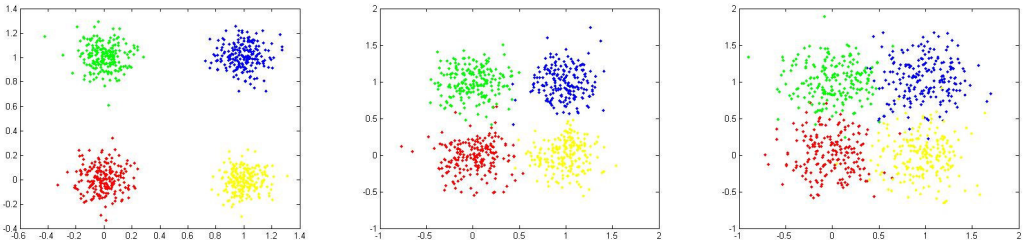


β)



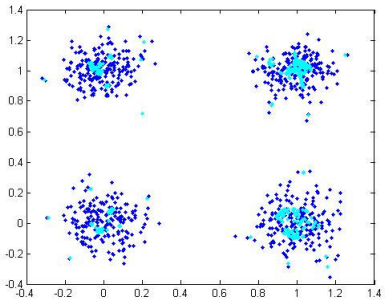
γ)

Σχήμα 5.1 Σύνολο δεδομένων 2-holes.

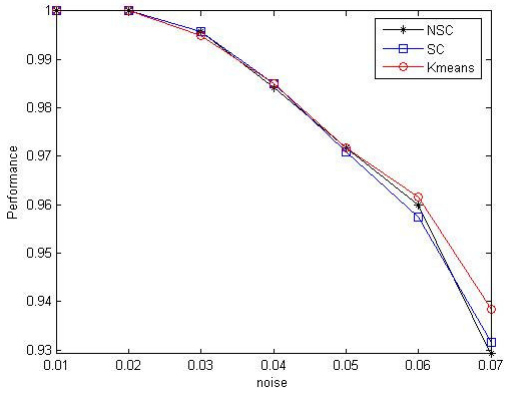


a) b) c)

α)

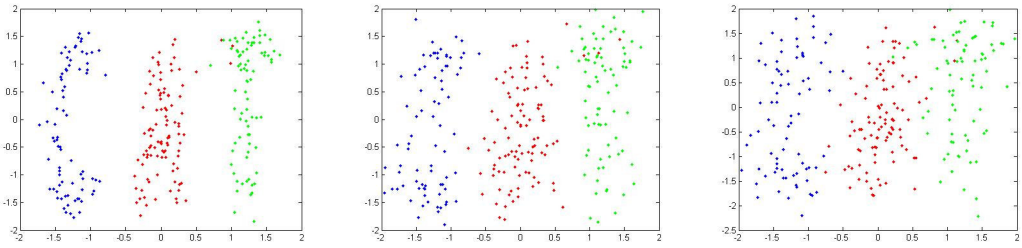


β)



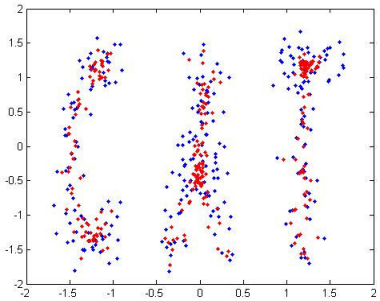
γ)

Σχήμα 5.2 Σύνολο δεδομένων 4-holes.

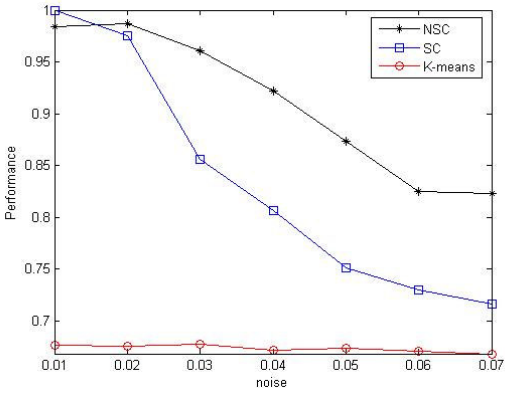


a) b) c)

α)

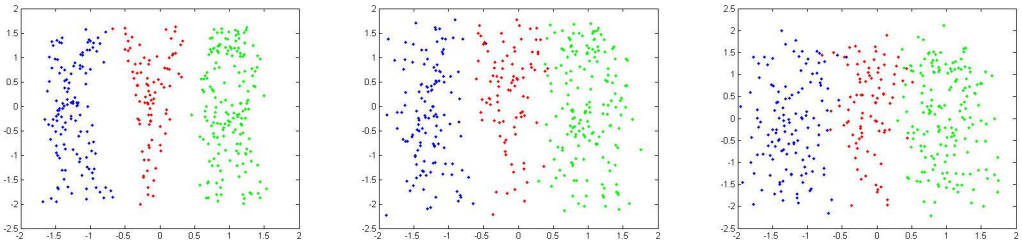


β)



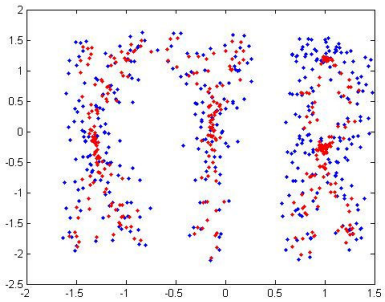
γ)

Σχήμα 5.3 Σύνολο δεδομένων CAT.

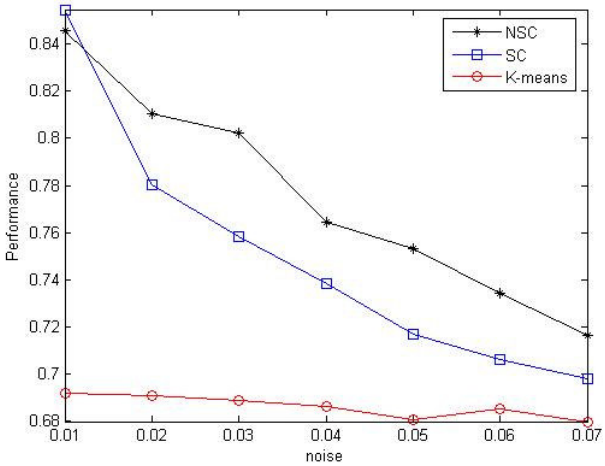


a) b) c)

α)

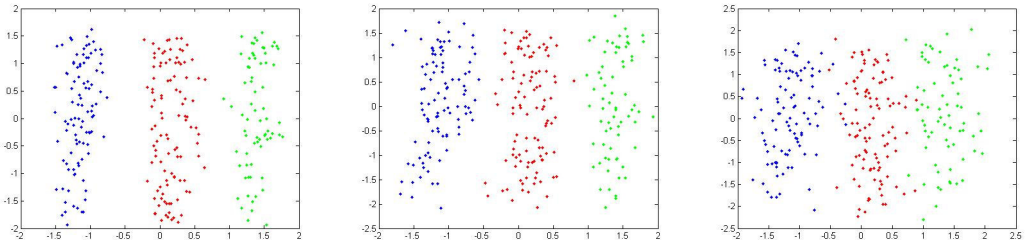


β)



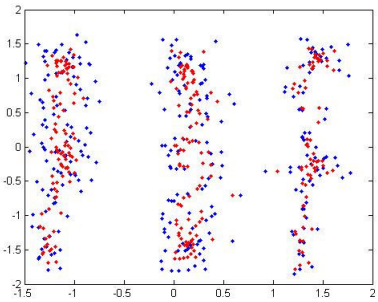
γ)

Σχήμα 5.4 Σύνολο δεδομένων KYR.

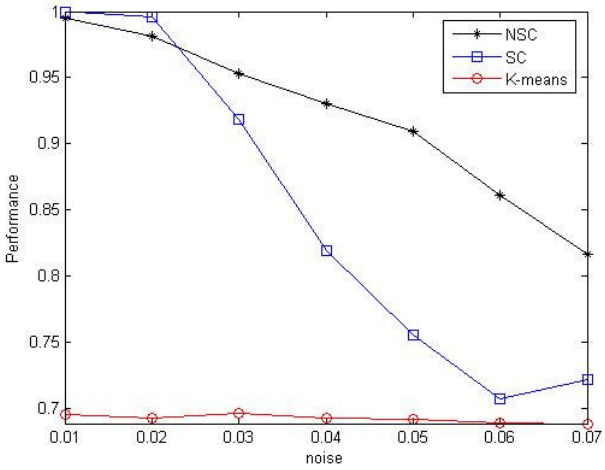


a) b) c)

α)

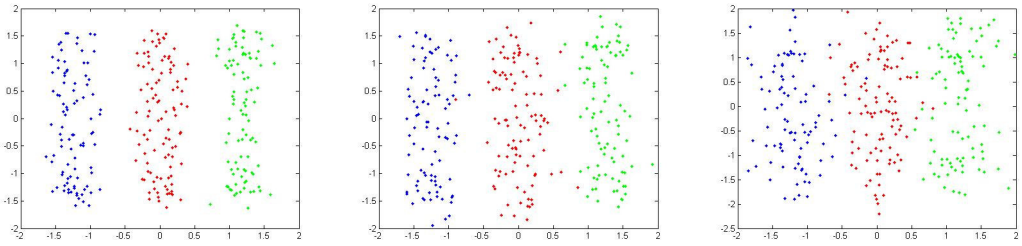


β)



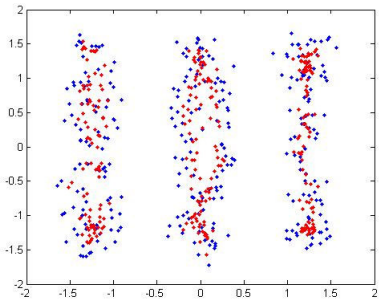
γ)

Σχήμα 5.5 Σύνολο δεδομένων PDF.

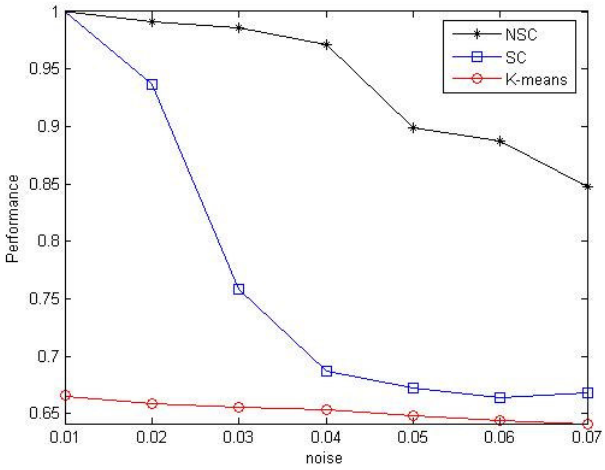


a) b) c)

α)

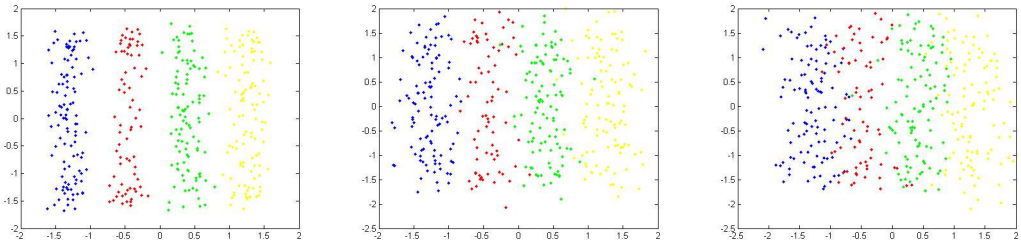


β)



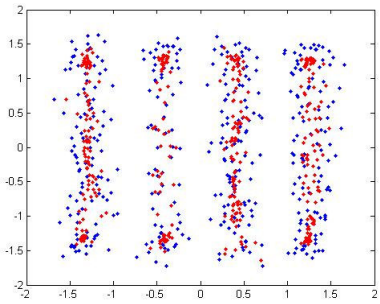
γ)

Σχήμα 5.6 Σύνολο δεδομένων UOI.

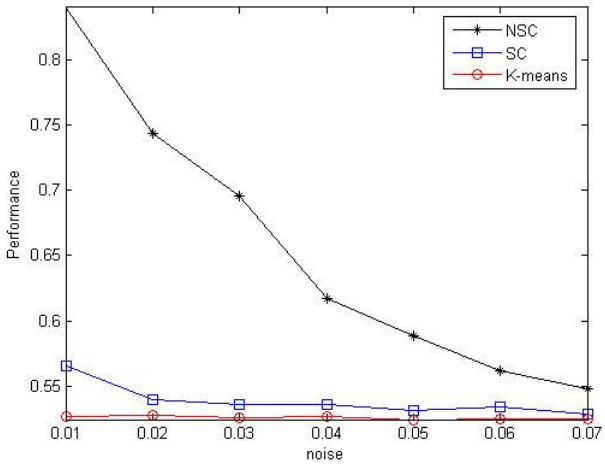


a) b) c)

α)



β)



γ)

Σχήμα 5.7 Σύνολο δεδομένων BIRD.

5.3. Πειραματική Μελέτη σε Γνωστά Δεδομένα με Αριθμητικά Χαρακτηριστικά

Έπειτα, εφαρμόσαμε τη μέθοδό μας σε προβλήματα ομαδοποίησης δεδομένων με αριθμητικά χαρακτηριστικά. Για το σκοπό αυτό, επιλέξαμε κάποιες ευρέως γνωστές

δοκιμασίες επιδόσεως. Ακόμα, σε κάθε σειρά πειραμάτων ήταν σταθερός ο αριθμός του Νευτώνειου βήματος καθώς και η τιμή του χρονικού βήματος δt.

5.3.1. Τρόπος διεξαγωγής πειραμάτων

Αρχικά, επιλέξαμε το σύνολο δεδομένων moon & sun. Το παραπάνω, όπως φαίνεται στο Σχήμα 5.8(α), είναι ένα πρόβλημα δύο κατηγοριών (K=2) όπου η μία έχει σχήμα φεγγαριού και η άλλη έχει σχήμα ήλιου.

Ακόμα, επιλέξαμε το σύνολο δεδομένων CRAB του Ripley [23] που περιέχει 200 δεδομένα τα οποία ανήκουν σε 4 κατηγορίες (K=4) και παρουσιάζεται στο Σχήμα 5.8(β). Το αρχικό σύνολο δεδομένων CRAB είναι 5 διαστάσεων. Εμείς, επιλέξαμε τις προβολές τους στο επίπεδο όπως αυτό ορίζεται από την δεύτερη και την τρίτη βασική συνιστώσα.

Επιπλέον, μελετήσαμε το σύνολο δεδομένων Fisher-IRIS που αποτελείται από 150 σημεία τα οποία ανήκουν σε 3 κατηγορίες (K=3) και παρουσιάζεται στο Σχήμα 5.8(γ). Το αρχικό σύνολο δεδομένων IRIS είναι 4 διαστάσεων. Εμείς, επιλέξαμε τις προβολές τους στο επίπεδο όπως αυτό ορίζεται από την πρώτη και τη δεύτερη βασική συνιστώσα.

Ακόμα, μελετήσαμε το σύνολο δεδομένων wine που αποτελείται από 178 σημεία τα οποία ανήκουν σε 3 κατηγορίες (K=3) με 13 χαρακτηριστικά. Στο παραπάνω σύνολο εφαρμόσαμε zero mean κανονικοποίηση.

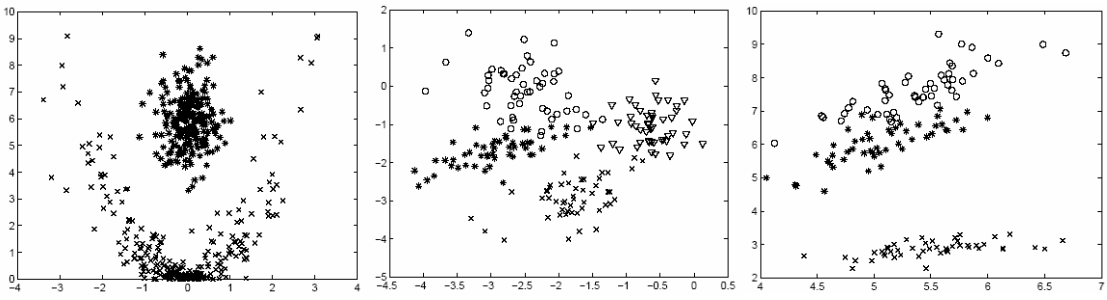
Κατά την διάρκεια των πειραμάτων, ο αριθμός των Νευτώνειων βημάτων ήταν ίσος με 50 ενώ, το χρονικό βήμα ήταν ίσο με 10^{-4}. Επιπλέον, και στις δύο προσεγγίσεις, NSC και SC, χρησιμοποιήσαμε την ίδια τιμή για το σ (με τον τρόπο που παρουσιάσαμε στο κεφάλαιο 3).

Στη συνέχεια, εφαρμόσαμε την προτεινόμενη μέθοδο για την επίλυση του προβλήματος pendigits. Συγκεκριμένα, θεωρήσαμε πως έχουμε ένα σύνολο προσώπων τα οποία γράφουν με το χέρι τους τα αριθμητικά ψηφία από 0 έως 9. Είναι φανερό πως κάθε ένα από τα πρόσωπα αυτά μπορεί να γράφει τα ψηφία αυτά με διαφορετικό τρόπο. Με άλλα λόγια, κάθε ψηφίο x μπορεί να γράφεται με διαφορετικό τρόπο από κάθε πρόσωπο. Η εύρεση ποιου ψηφίου έχει γραφεί με το χέρι από ένα πρόσωπο αποτελεί το πρόβλημα pendigits. Με άλλα λόγια, προσπαθούμε να ξεχωρίσουμε ποιο ψηφίο έχει γράψει κάθε φορά ένα πρόσωπο χωρίς να θεωρούμε ότι όλα τα πρόσωπα γράφουν κάθε συγκεκριμένο ψηφίο με το ίδιο τρόπο.

Το σύνολο των δεδομένων που χρησιμοποιήσαμε αποτελείται από ένα πλήθος χειρόγραφων αριθμητικών ψηφίων από πολλούς διαφορετικούς ανθρώπους. Πιο συγκεκριμένα, ζητήθηκε από ένα σύνολο προσώπων να γράψουν με το χέρι 250 ψηφία με τυχαία σειρά μέσα σε περιοχές με δεδομένη ανάλυση. Στα πρόσωπα που συμμετείχαν στη δειγματοληψία προβάλλονταν τα ψηφία που έπρεπε να γράψουν κάθε φορά με το χέρι. Σημειώνεται πως η γραφή τους δεν έπρεπε να βγαίνει εκτός της οριζόμενης περιοχής. Στη συνέχεια, αναγάγαμε τις οριζόμενες περιοχές σε περιοχές με εύρος [0, 100] σε κάθε μία από τις δύο διαστάσεις. Κάθε ψηφίο, λοιπόν, θεωρούμε ότι βρίσκεται γραμμένο σε μία δισδιάστατη περιοχή με εύρος [0, 100] για κάθε διάσταση. Για την πραγματοποίηση της διαδικασίας ομαδοποίησης, κάθε ψηφίο αναπαρίσταται ως ένα διάνυσμα χαρακτηριστικών σταθερού μήκους. Το διάνυσμα περιέχει τις συντεταγμένες 8 διαδοχικών σημείων, ακολουθούμενων από έναν επιπλέον αριθμό. Αυτά τα 8 σημεία επιλέγονται με τυχαίο τρόπο από την περιοχή όπου έχει γραφεί το ψηφίο, δηλαδή από το σύνολο των σημείων όπου έχει αποτυπωθεί η γραφή του ψηφίου. Προφανώς, οι συντεταγμένες ανήκουν στο εύρος [0, 100]. Αν, λοιπόν, ενώσουμε τα σημεία με τη σειρά την οποία ακολουθούν στο διάνυσμα, έχουμε μια αποτύπωση του τρόπου γραφής του ψηφίου στην περιοχή. Το διάνυσμα αυτό συμπληρώνεται από έναν ακόμη αριθμό ο οποίος δείχνει ποιο ψηφίο αναπαρίσταται στην πραγματικότητα από το υπόλοιπο διάνυσμα. Για παράδειγμα, το διάνυσμα [47 100 27 81 57 37 26 0 0 23 56 53 100 90 40 98 8] αναπαριστά το ψηφίο 8 που έχει γραφεί από κάποιο από τα πρόσωπα που συμμετείχαν στη

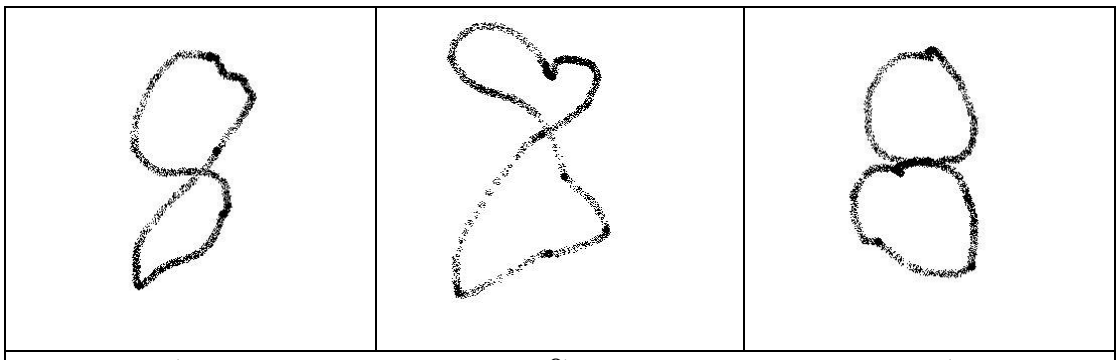
δειγματοληψία, όπως μας δείχνει η $17^{η}$ συνιστώσα του. Οι 16 πρώτες συνιστώσες αποτελούν τα σημεία που έχουν επιλεγεί τυχαία από την περιοχή στην οποία έχει αποτυπωθεί η γραφή του ψηφίου και χρησιμοποιούνται για την ομαδοποίηση. Έτσι, παρατηρούμε πως αν ενώσουμε τα σημεία (47, 100), (27, 81), (57, 37), (26, 0), (0, 23), (56, 53), (100, 90) και (40, 98), τα οποία αντιστοιχούν στις πρώτες συνιστώσες του διανύσματος, θα δούμε τον τρόπο με τον οποίο έχει αποτυπωθεί το ψηφίο 8 στην περιοχή από το πρόσωπο που συμμετείχε στη δειγματοληψία. Στο Σχήμα 5.9 παρουσιάζουμε τρεις διαφορετικούς τρόπους γραφής του ψηφίου 8.

Για να αντιμετωπίσουμε το παραπάνω πρόβλημα, εφαρμόσαμε την τροποποιημένη μας μέθοδο μας που χρησιμοποιεί το δυναμικό (όπως παρουσιάστηκε στο 4.3). Κατά την διάρκεια των πειραμάτων, ο αριθμός των Νευτώνειων βημάτων (T) ήταν ίσος με 50 ενώ, το χρονικό βήμα (δt) ήταν ίσο με 10^{-6}. Κατά τη διάρκεια των πειραμάτων, επειδή το αρχικό σύνολο δεδομένων, δηλαδή το πλήθος των ψηφίων που λάβαμε κατά τη δειγματοληψία, ήταν πολύ μεγάλο, χρησιμοποιήσαμε υποσύνολα αυτού στα οποία, όμως, η αναλογία εμφάνισης των ψηφίων ήταν ίδια με αυτή στο αρχικό σύνολο. Διεξάγαμε πειράματα σε ένα συνεχώς αυξανόμενο δείγμα από τα αρχικά δεδομένα. Για κάθε συγκεκριμένο δείγμα εκτελέσαμε το πείραμα 20 φορές.



α) β) γ)

Σχήμα 5.8 Σύνολα δεδομένων που χρησιμοποιήσαμε α) moon & sun, β) CRAB, γ) IRIS.



α) β) γ)

Σχήμα 5.9 Τρεις διαφορετικοί τρόποι γραφής του ψηφίου 8.

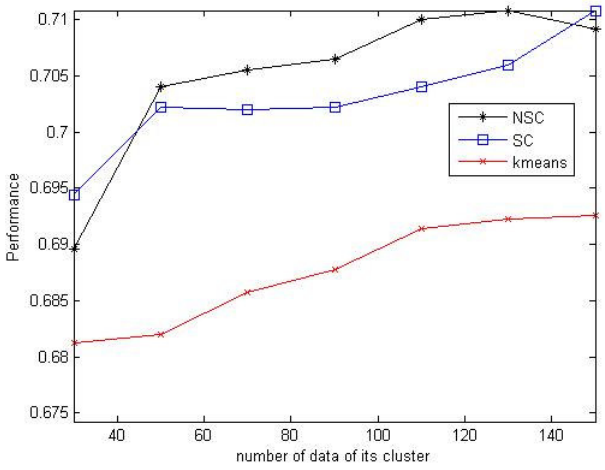
5.3.2. Πειραματικά αποτελέσματα-αξιολόγηση

Στον Πίνακα 5.1, παρουσιάζουμε τις επιδόσεις της προτεινομένης μεθόδου, της κλασικής μεθόδου Φασματικής Ομαδοποίησης και του αλγορίθμου K-means. Παρατηρούμε ότι οι επιδόσεις όλων των μεθόδων είναι ισοδύναμες.

Πειράματα		Επίδοση των		
Σύνολα δεδομένων		NSC	SC	K-means
moon & sun	K=2	0.94	0.94	0.92
crabs	N=200, K=4	0.94	0.92	0.92
iris	N=150, K=3	0.89	0.88	0.88
wine	N=178, K=3	0.95	0.95	0.94

Πίνακας 5.1 Πειραματικά αποτελέσματα.

Στο Σχήμα 5.8 παρουσιάζουμε την μέση τιμή των πειραμάτων στο pendigits για κάθε μέθοδο. Μπορούμε να παρατηρήσουμε ότι η επίδοση της προτεινομένης μεθόδου (NSC) είναι καλή σε σχέση με τη επίδοση της κλασικής μεθόδου Φασματικής Ομαδοποίησης και του K-means αλγορίθμου.



Σχήμα 5.10 Συγκριτικά αποτελέσματα στο σύνολο δεδομένων pendigits.

5.4. Πειραματική Μελέτη σε Ομαδοποίηση Κειμένων

Στη συνέχεια, εξετάσαμε την επίδοση της προτεινομένης μεθόδου σε προβλήματα ομαδοποίησης δεδομένων υψηλής διάστασης. Για το σκοπό αυτό, επιλέξαμε κάποια σύνολα κειμένων. Κατά την διάρκεια όλων των πειραμάτων σε κείμενα, ήταν γνωστός ο αριθμός των κειμένων του κάθε συνόλου, ο αριθμός των ομάδων, το μέγεθος του λεξικού καθώς και η πραγματική κατηγορία στην οποία ανήκε το κάθε κείμενο. Ακόμα, σε κάθε σειρά πειραμάτων ήταν σταθερός ο αριθμός του Νευτώνειου βήματος καθώς και η τιμή του χρονικού βήματος δt.

5.4.1. Τρόπος διεξαγωγής πειραμάτων

Συγκεκριμένα, επιλέξαμε τρία υποσύνολα του δημοφιλούς 2-Newsgroup collection. Το σύνολο Science-400 αποτελείται από 400 επιστημονικά κείμενα τεσσάρων κατηγοριών και το πλήθος των διαφορετικών λέξεων που υπάρχουν σε αυτό (δηλαδή το μήκος του λεξικού) είναι 4855. Το σύνολο Multi-3 αποτελείται από 300 κείμενα τριών κατηγοριών και το πλήθος των διαφορετικών λέξεων που υπάρχουν σε αυτό είναι 4515. Το σύνολο Multi-5 αποτελείται από 250 κείμενα πέντε κατηγοριών και το πλήθος των διαφορετικών λέξεων που υπάρχουν σε αυτό είναι 5589.

Ακόμα, κατά την διάρκεια των πειραμάτων, ο αριθμός των Νευτώνειων βημάτων ήταν ίσος με 50 ενώ, το χρονικό βήμα ήταν ίσο με 10^{-6}.

5.4.2. Πειραματικά αποτελέσματα-αξιολόγηση

Στον Πίνακα 5.2, παρουσιάζουμε τα χαρακτηριστικά των επιλεγμένων υποσυνόλων καθώς και τις επιδόσεις της προτεινομένης μεθόδου καθώς και της κλασικής μεθόδου Φασματικής Ομαδοποίησης. Παρατηρούμε ότι η επίδοση της μεθόδου NSC είναι παρόμοια με αυτή της κλασικής SC.

Σύνολο δεδομένων κειμένων			Επίδοση των	
όνομα	περιγραφή		NSC	SC
Science-400	M=400, K=4, T=4855		0.62	0.61
Multi-3	M=300, K=3, T=4515		0.73	0.72
Multi-5	M=250, K=5, T=5589		0.74	0.64

Πίνακας 5.2 Πειραματικά αποτελέσματα των NSC και SC σε ομαδοποίηση κειμένων.

5.5. Πειραματική Μελέτη σε Προβλήματα Κατάτμηση Εικόνας

Τέλος, εφαρμόσαμε τη μέθοδό μας σε προβλήματα κατάτμησης εικόνας. Κατά την διάρκεια όλων των πειραμάτων στις εικόνες, ήταν γνωστός ο αριθμός των ομάδων καθώς και η πραγματική ομάδα στην οποία ήταν τοποθετημένο το κάθε pixel. Ακόμα, στα πειράματά μας ήταν σταθερός ο αριθμός του Νευτώνειου βήματος καθώς και η τιμή του χρονικού βήματος δt. Επιπλέον, και στις δύο προσεγγίσεις, NSC και SC, χρησιμοποιήσαμε την ίδια τιμή για το σ (με τον τρόπο που παρουσιάσαμε στο κεφάλαιο 3).

5.5.1. Τρόπος διεξαγωγής πειραμάτων

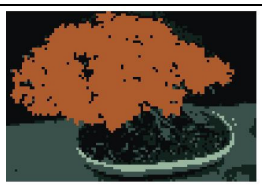
Συγκεκριμένα, επιλέξαμε εφτά έγχρωμες εικόνες από το Berkeley segmentation database. Κατά την διάρκεια των πειραμάτων, ο αριθμός των Νευτώνειων βημάτων ήταν ίσος με 15 ενώ, το χρονικό βήμα ήταν ίσο με 10^{-7}.

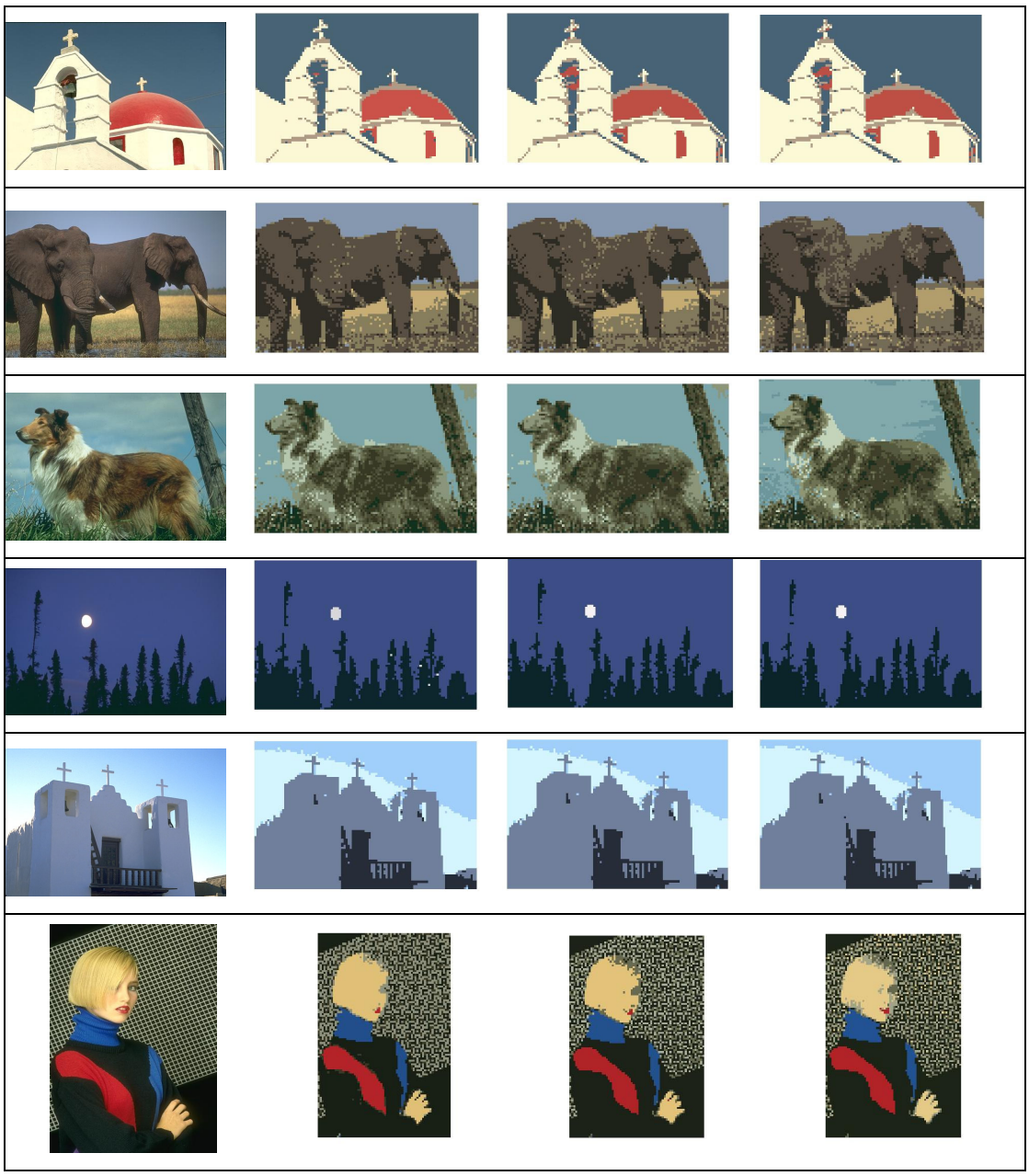
Πρέπει να τονίσουμε ότι, χρησιμοποιήσαμε τη μέθοδο *Nyström* [12] για να διεξάγουμε τη σειρά πειραμάτων, στις εικόνες που επιλέξαμε, εξαιτίας του μεγάλου αριθμού των δεδομένων εισόδου. Περισσότερες πληροφορίες για τη μέθοδο *Nyström* παρατίθενται στο Παράρτημα.

Η μέθοδος *Nyström* αποτελεί μία τεχνική εύρεσης των ιδιοδιανυσμάτων μέσω δειγματοληψίας. Συγκεκριμένα, η προσέγγιση λειτουργεί σε δύο βήματα. Αρχικά, επιλύει το πρόβλημα της ομαδοποίησης σε ένα μικρό και τυχαία επιλεγμένο υποσύνολο από pixels. Στη συνέχεια, συνάγει συμπερασματικά την τελική ομαδοποίηση χρησιμοποιώντας μόνο ένα μικρό αριθμό δειγμάτων από το σύνολο των pixels της εικόνας. Εμείς, στη σειρά πειραμάτων που πραγματοποιήσαμε χρησιμοποιήσαμε 600 δείγματα (από ένα σύνολο 7500 pixels περίπου). Στα αποτελέσματα των πειραμάτων μας, τα pixels κάθε επαναδομημένης εικόνας έχουν την ένταση του κέντρου της ομάδας στην οποία ανήκουν.

5.5.2. Πειραματικά αποτελέσματα-αξιολόγηση

Παρουσιάζουμε τα συγκριτικά αποτελέσματα των πειραμάτων μας στο Σχήμα 5.12. Παρατηρούμε ότι, στην μέθοδο NSC έχουμε ομαλότερη μετάβαση από την μία χρωματική περιοχή στην άλλη σε σχέση με αυτήν που συναντούμε στην μέθοδο SC.

Αρχικές Εικόνες	NCS	SC	K-means
			



Σχήμα 5.11 Συγκριτικά αποτελέσματα τριών μεθόδων ομαδοποίησης.

ΚΕΦΑΛΑΙΟ 6. ΕΠΙΛΟΓΟΣ

6.1 Αξιολόγηση Προτεινόμενης Μεθοδολογίας – Πλεονεκτήματα και Μειονεκτήματα
6.2 Επεκτάσεις

Στη διατριβή αυτή παρουσιάσαμε μία συστηματική μεθοδολογία, τη Νευτώνεια Φασματική Ομαδοποίηση (*NSC*), η οποία συνδυάζει στοιχεία της Νευτώνειας κίνησης και της Φασματικής Ομαδοποίησης. Αρχικά, θεωρώντας τα δεδομένα μας ως σωματίδια και εφαρμόζοντας τις εξισώσεις κίνησης του Νεύτωνα, επιτυγχάνεται η μετακίνησή τους και η συρρίκνωσή τους προς το κέντρο της ομάδας που ανήκουν. Η διαδικασία αυτή επιτρέπει τον εμπλουτισμό του πίνακα ομοιότητας με μεγαλύτερη πληροφορία και τη μετατροπή του σε αραιό (sparse) πίνακα με σημαντικά πλεονεκτήματα. Στη συνέχεια, χρησιμοποιώντας το βασικό σχήμα των μεθόδων φασματικής ομαδοποίησης, καταλήγουμε στην τελική ομαδοποίηση του συνόλου δεδομένων.

Επιπλέον, παρουσιάσαμε μία επέκταση της παραπάνω μελέτης. Πιο συγκεκριμένα, τροποποιήσαμε ελαφρά την μέθοδό μας έτσι ώστε να επιτύχουμε καλή ομαδοποίηση σε δεδομένα μεγάλης διάστασης. Εφαρμόσαμε τη μέθοδό μας σε προβλήματα ομαδοποίησης κειμένων (document clustering).

Έπειτα, μελετήσαμε πειραματικά τις προτεινόμενες μεθόδους και παρουσιάσαμε συγκριτικά αποτελέσματα των μεθόδων σε σχέση με αυτά της καθιερωμένης μεθόδου που βασίζεται στη Φασματική Ομαδοποίηση και του αλγορίθμου K-means.

6.1. Αξιολόγηση Προτεινόμενης Μεθοδολογίας – Πλεονεκτήματα και Μειονεκτήματα

Στην μεθοδολογία της Νευτώνειας Φασματικής Ομαδοποίησης, ιδιαίτερα σημαντικό είναι το γεγονός της κατασκευής ενός αραιού και πλουσιότερου σε πληροφορία πίνακα ομοιότητας. Πιο συγκεκριμένα, η μέθοδος βασίζεται στη Νευτώνεια Ομαδοποίηση της οποίας σημαντικό βήμα είναι η συρρίκνωση των δεδομένων που είναι όμοια μεταξύ τους και η απώθηση των απομονωμένων σημείων. Ως αποτέλεσμα, στον πίνακα ομοιότητας που προκύπτει μετά τη διαδικασία συρρίκνωσης, κάποια ποσά ομοιότητας εμφανίζονται αυξημένα (των γειτονικών σημείων), ενώ κάποια άλλα μηδενίζονται (των απομακρυσμένων). Ακόμη, εκτός της γνωστής Γκαουσιανής συνάρτησης και του δυναμικού που εφαρμόσαμε, μπορούμε να χρησιμοποιήσουμε οποιαδήποτε άλλη συνάρτηση πυρήνα (*kernel function*), ανάλογα με τον τύπο των δεδομένων που θέλουμε να ομαδοποιήσουμε και το αντίστοιχο πρόβλημα που εξετάζουμε, γεγονός που αποτελεί σημαντικό πλεονέκτημα. Επίσης, έχει ιδιαίτερη σημασία το γεγονός πως η προτεινόμενη μέθοδος περιλαμβάνει και μία συστηματική μεθοδολογία για τον καθορισμό του εύρους δυναμικού (σ). Με τον τρόπο αυτό, μπορούμε να υπολογίσουμε το σ εκείνο που θα μας οδηγήσει στη λύση με την καλύτερη ομαδοποίηση. Σημειώνεται ότι η τιμή της παραμέτρου σ εξαρτάται άμεσα από το σύνολο δεδομένων (το εύρος του συνόλου). Τέλος, πρέπει να τονιστεί πως από την εκτεταμένη πειραματική μελέτη που πραγματοποιήσαμε καταλήξαμε στο συμπέρασμα πως η επίδοση της προτεινόμενης μεθοδολογίας είναι ισοδύναμη ή υπερέχει της παραδοσιακής μεθόδου Φασματικής Ομαδοποίησης, ενώ σχεδόν πάντοτε υπερέχει της επίδοσης του αλγορίθμου K-means.

Παρόλα αυτά, η μεθοδολογία μας εμφανίζει και ορισμένα μειονεκτήματα. Ένα σημαντικό ζήτημα είναι ο καθορισμός της τιμής του χρονικού βήματος (δt). Εμείς, κατά τη διάρκεια των πειραμάτων, θεωρούσαμε συνήθως την ενδεικτική τιμή $\delta t=10^{-4}$, όμως η συγκεκριμένη τιμή δεν είναι η καλύτερη δυνατή για όλα τα σύνολα δεδομένων. Για παράδειγμα, αν τα σημεία του συνόλου δεδομένων έχουν μικρές τιμές, η τιμή του δt πρέπει να είναι πολύ μικρή. Αν, όμως, τα σημεία του συνόλου δεδομένων έχουν μεγάλες τιμές, τότε το δt θα έχει και πάλι μικρή τιμή, ίσως όμως

μεγαλύτερη σε σχέση με την προηγούμενη περίπτωση. Ένα ακόμη σημαντικό ζήτημα είναι ο καθορισμός των νευτώνειων βημάτων (T), καθώς, αν το πλήθος τους είναι μεγαλύτερο από κάποιο άνω όριο (το οποίο εξαρτάται από τον τύπο των δεδομένων), τότε μπορεί τα σημεία του συνόλου δεδομένων να τείνουν να συγκεντρωθούν γύρω από ένα μόνο σημείο. Έτσι, η τελική ομαδοποίηση δε θα είναι η επιθυμητή. Τέλος, κατά την πειραματική μελέτη προβλημάτων κατάτμησης εικόνας, εξαιτίας του μεγάλου πλήθους των δεδομένων εισόδου, χρησιμοποιήσαμε τη μέθοδο *Nyström* η οποία αποτελεί μία τεχνική εύρεσης των ιδιοδιανυσμάτων μέσω, όμως, δειγματοληψίας, που δε χρησιμοποιεί όλα τα δεδομένα εισόδου.

6.2. Επεκτάσεις

Μελλοντικό μας στόχο αποτελεί η εύρεση μιας μεθοδολογίας καθορισμού του χρονικού βήματος (δt), του οποίου η τιμή επηρεάζει σημαντικά την επίδοση της προτεινόμενης μεθοδολογίας. Επιπλέον, μελλοντικά θα μπορούσαμε να χρησιμοποιήσουμε και άλλες συναρτήσεις πυρήνα για την αντιμετώπιση διαφορετικών συνόλων δεδομένων, ανάλογες του τύπου των δεδομένων. Τέλος, επέκταση της παραπάνω μεθόδου για την αντιμετώπιση προβλημάτων με περίπλοκο τύπο δεδομένων, όπως είναι οι χρονοσειρές (*time-series*), τα δεδομένα πολυμέσων, κλπ.

ΑΝΑΦΟΡΕΣ

[1] M. M. Ali and C. Storey. “Topological multilevel single linkage”, J. Global Optim. 5, 349-358, 1994.

[2] P. Baldi and S. Brunak. “Bioinformatics: The machine Learning Approach”, MIT Press, Cambridge, 1998.

[3] C. M. Bishop. Neural Networks for Pattern Recognition, Oxford University Press Inc., New York, 1995.

[4] K. Blekas, I.E. Lagaris. “Newtonian clustering: an approach based on molecular dynamics and global optimization”, Pattern Recognition, 40(6), 1734-1744, 2007.

[5] K. Blekas, K. Christodoulidou and I. E. Lagaris. “Newtonian spectral clustering”, 2009.

[6] K. Blekas, A. Likas, N. P. Galatsanos and I.E. Lagaris. “A spatially-constrained mixture model for image segmentation”, IEEE Trans. Neural Networks 16(2), 494-498, 2005.

[7] C. G. E. Boender, A. H. G. R. Kan, G.T. Timmer and L. Stougie. “A stochastic method for global optimization”, Math. Programming 22, 125-140, 1982.

[8] H. Chang and D. Yeung. “Robust path-based spectral clustering with application to image segmentation”, Proc. Intern. Conf. On Computer Vision 278-285, 2005.

[9] W. Chen, Y. Song, H. Bai, C. Lin and E. Y Chang. “Parallel spectral clustering”, ECML/PCDD, Lecture Notes in Science, 5212, 274-389, 2008.

[10] I. S. Dhillon. “Co-clustering documents and words using bipartite spectral graph partitioning”, Proc. seventh ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data mining (KDD), 269-274, 2001.

[11] R. O. Duda, P. E. Hart and D. G. Stork. “Pattern Classification”, Wiley-Interscience, New York, 2001.

[12] C. Fowlkes, S. Belongie, F. Chung and J. Malik. “Spectral grouping using the Nystrom method”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(2), 101-112, February 2004.

[13] K. Fukunaga. “Introduction to Statistical Pattern Recognition”, Academic Press, San Diego, 1990.

[14] D. Harel and Y. Koren. “On clustering using random walks”, FSTTCS 2001, LNCS, 18-41, 2001.

[15] D. J. Higham, G. Kalna and M. Kibble. “Spectral clustering, and its use in bioinformatics”, Journal of Computational and Applied Mathematics 204, 25-37, 2007.

[16] A. H. G. R. Kan and G.T. Timmer. “Stochastic global optimization methods. Part I: clustering methods”, Math. Programming 39, 27-56, 1987.

[17] A. H. G. R. Kan and G.T. Timmer. “Stochastic global optimization methods. Part II: multi level methods”, Math. Programming 39, 57-78, 1987.

[18] A. K. Jain, M. N. Murty and P. J. Flynn. “Data clustering: a review”, ACM Computing Surveys, 31(3), September 1999.

[19] U. Luxburg. “A tutorial on spectral clustering”, Statistics and Computing, 17(4), 395-416, 2007.

[20] A. Ng, M. I. Jordan and Y. Weiss. “On spectral clustering: Analysis and an algorithm”, In advances in Neural Information Processing Systems 14, 849-864, 2001.

[21] J. Park, H. Zha and R. Kasturi. “Spectral clustering for robust motion segmentation”, 8th European Conf. on Computer Vision, 390-401, 2004.

[22] W. Pentney and M. Meila. “Spectral clustering for Biological sequence data”, Proc. of the 25th Annual Conference of AAAI, 845-850, 2005.

[23] B. D. Ripley. Pattern Recognition and Neural Networks, Cambridge Univ. Press Inc., Cambridge, UK, 1996.

[24] J. Shi and J. Malik. “Normalized cuts and image segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888-905, August 2000.

[25] P. N. Tan, M. Steinbach and V. Kumar. “Data mining”, Pearson Education, Inc. 2006.

[26] F. V. Theos, I. E. Lagaris and D. G. Papageorgiou. “PANMIN: sequential and parallel global optimization procedures with a variety of options for the local search strategy”, Comput. Phys. Commun. 159, 63-69, 2004.

[27] A. Torn and S. Viitanen. “Topological global optimization using presampled points”, J. Global Optim. 5, 267-276, 1994.

ΠΑΡΑΡΤΗΜΑ

Η μέθοδος *Nyström*

Με την μέθοδο *Nyström* αποφεύγουμε τη σύγκριση όλων των pixels μίας εικόνας με το σύνολο των κέντρων των ομάδων που υπάρχουν σε αυτήν. Αντίθετα, συγκρίνουμε τα pixels μίας εικόνας με ένα μικρότερο σύνολο δειγμάτων επιλεγμένων με τυχαίο τρόπο. Η παραπάνω προσέγγιση είναι απλή και επιπλέον, η πολυπλοκότητά της αυξάνει γραμμικά ως προς την ανάλυση της εικόνας. Εν προκειμένω εκμεταλλευόμαστε το ότι γενικά ο αριθμός των ομάδων σε μία εικόνα είναι πολύ μικρότερος από τον αριθμό των pixels σε αυτή.

Πιο αναλυτικά, η μέθοδος *Nyström* είναι μία τεχνική εύρεσης αριθμητικών προσεγγίσεων στα προβλήματα ιδιοδιανυσμάτων της μορφής:

$$\int_a^b w(x,y)\phi(y)dy = \lambda\phi(x). \qquad \text{Εξ. Π.1}$$

Μπορούμε να προσεγγίσουμε την παραπάνω εξίσωση εκτιμώντας την σε ένα σύνολο από σημεία ξ_1, ξ_2, … ξ_n ομοίως τοποθετημένων σε ένα διάστημα [a, b] για τα οποία ισχύει ότι:

$$\frac{(b-a)}{n}\sum_{j=1}^{n} w(x,\xi_j)\hat{\phi}(\xi_j) = \lambda\hat{\phi}(x), \qquad \text{Εξ. Π.2}$$

όπου το $\hat{\phi}(x)$ είναι μία προσέγγιση το πραγματικού $\phi(x)$. Για την επίλυση της Εξίσωσης (Π.2) θέτουμε $x = \xi_i$ και επιλύουμε το ακόλουθο σύστημα εξισώσεων:

$$\frac{(b-a)}{n}\sum_{j=1}^{n} w(\xi_i,\xi_j)\hat{\phi}(\xi_j) = \lambda\hat{\phi}(\xi_i) \quad \forall i \in \{1...n\}. \qquad \text{Εξ. Π.3}$$

Χωρίς απώλεια της γενικότητας, θεωρούμε ότι το διάστημα [a, b] αντιστοιχεί στο διάστημα [0, 1] οπότε έχουμε:

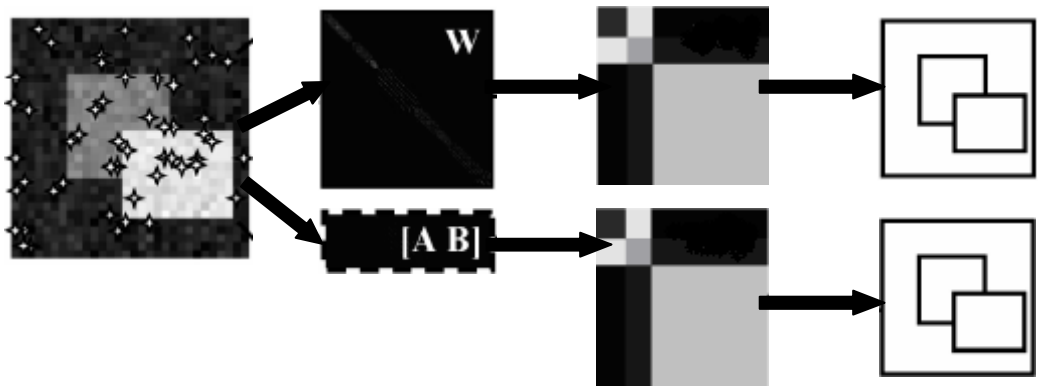
$$A\hat{\Phi} = n\hat{\Phi}\Lambda, \qquad \text{Εξ. Π.4}$$

όπου $A_{ij} = w(\xi_i, \xi_j)$ και $\Phi = [\phi_1\phi_2...\phi_n]$ είναι τα n ιδιοδιανύσματα του A που αντιστοιχούν στις ιδιοτιμές $\lambda_1, \lambda_2, ...\lambda_n$. Μέσω της Εξίσωσης (Π.2) προκύπτει ότι για κάθε $\hat{\phi}_i$ ισχύει ότι:

$$\hat{\phi}_i(x)\frac{1}{n\lambda_i}\sum_{j=1}^{n} w(x, \xi_j)\hat{\phi}_i(\xi_j). \qquad \text{Εξ. Π.5}$$

Η παραπάνω έκφραση μας επιτρέπει να επεκτείνουμε ένα ιδιοδιάνυσμα υπολογισμένο για ένα σύνολο επιλεγμένων σημείων (δείγματα) ως ιδιοδιάνυσμα για ένα τυχαίο σημείο x.

Στο Σχήμα Π.1, στα αριστερά, παρουσιάζουμε μία εικόνα που αποτελείται από τρεις περιοχές (K=3). Ακόμα, στο πρώτο μονοπάτι φαίνεται ο πίνακας ομοιότητας W από τον οποίο μπορούν να εξαχθούν με υπολογιστική ακρίβεια τα τρία κορυφαία ιδιοδιανύσματα του πίνακα Laplace και στη συνέχεια η τελική κατάτμηση των pixels της εικόνας μέσω του K-means. Στο δεύτερο μονοπάτι παρουσιάζεται η εφαρμογή της μεθόδου *Nyström*. Χρησιμοποιούνται μόνο τα pixels της εικόνας που είναι σημειωμένα με αστερίσκο (δείγματα) και στη συνέχεια υπολογίζεται μόνο ένα μέρος του πίνακα W. Κάθε γραμμή του παραπάνω πίνακα περιέχει τις σχέσεις ομοιότητας μεταξύ ενός δείγματος και όλης της εικόνας. Στη συνέχεια εξάγονται κατά προσέγγιση τα κορυφαία ιδιοδιανύσματα και η τελική κατάτμηση των pixels της εικόνας μέσω του K-meansαπεικονίζουμε την εφαρμογή της μεθόδου *Nyström*.



Σχήμα Π.1 Εφαρμογή της μεθόδου *Nyström*.

ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΥΓΓΡΑΦΕΑ

[1] K. Blekas, K. Christodoulidou and I. E. Lagaris. “Newtonian spectral clustering”, 2009.

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Η Χριστοδουλίδου Κυριακή εισήχθη στο Τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων το 2003 και έλαβε το Πτυχίο Πληροφορικής το 2007. Από το Σεπτέμβριο του 2007 είναι Μεταπτυχιακή φοιτήτρια στο Τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων.

Στα επιστημονικά της ενδιαφέροντα συμπεριλαμβάνονται η Εξόρυξη Δεδομένων, η Βιοπληροφορική και η Αναγνώριση Προτύπων.