ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων

«Advanced clustering techniques in images and videos»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνος Ε. Σωτηρόπουλος
email:kosotiro@uth.gr

Επιβλέπουσα Καθηγήτρια: Αλεξία Μπριασούλη

Επιτροπή: Ηλίας Χούστης, Αλεξία Μπριασούλη

Βόλος, Ιούλιος 2008

Πίνακας περιεχομένων

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω την καθηγήτριά μου κ. Αλεξία Μπριασούλη για την πολύτιμη συνδρομή της κατά τη διάρκεια εκπόνησης της διπλωματικής μου. Επίσης οι γνώσεις της πάνω στον τομέα της ψηφιακής εικόνας αποτέλεσαν για εμένα ισχυρό κίνητρο για να ασχοληθώ με το συγκεκριμένο θέμα.

Επίσης, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ηλία Χούστη που χωρίς την πολύτιμη συνδρομή του δε θα μπορούσε να ολοκληρωθεί η παρούσα διπλωματική εργασία.

Τέλος, ένα μεγάλο ευχαριστώ στους γονείς μου και στους δικούς μου ανθρώπους που με την αγάπη τους με έμαθαν να ζω ανθρώπινα και να παλεύω με αισιοδοξία για το καλύτερο αποτέλεσμα.

Κεφάλαιο 1

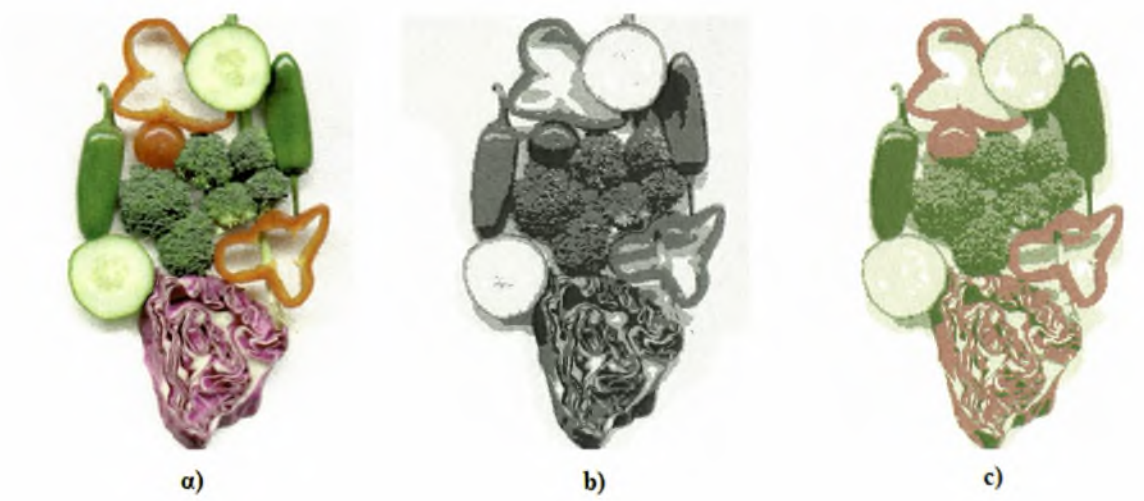
Εισαγωγή

1.1 Γενικά

Η συσταδοποίηση είναι μία από τις πιο διαδεδομένες τεχνικές που χρησιμοποιούνται στην εξερευνητική ανάλυση δεδομένων με αμέτρητες εφαρμογές στη στατιστική, στην επιστήμη των υπολογιστών, στη βιολογία, στις κοινωνικές επιστήμες, στην ψυχολογία και σε άλλους ακόμα τομείς. Σε ένα συγκεκριμένο πρόβλημα συσταδοποίησης μας δίνεται ένα σύνολο: $X = \{\chi_1,\chi_2,\ldots,\chi_n\}$, $\chi_i \in \Re^d$, που αποτελείται από n δεδομένα διάστασης d, όπου σκοπός μας είναι να χωρίσουμε αυτά τα δεδομένα του συνόλου X (που συνήθως αναφέρονται και με τον όρο πρότυπα) σε ένα πλήθος k ομάδων, χαρακτηριζόμενες για την ακρίβειά τους, έτσι ώστε τα πρότυπα που ανήκουν στην ίδια ομάδα να είναι όμοια μεταξύ τους ως προς κάποια ιδιότητα. Για παράδειγμα κατά τη συσταδοποίηση των pixel μιας ψηφιακής εικόνας ως προς την ποικιλία και την φωτεινότητα του κάθε χρώματος που αυτή διαθέτει θα πρέπει να δημιουργηθούν αφενός μεν τόσες ομάδες (συστάδες) όσα τα χρώματα των pixel της εικόνας στο χώρο RGB, που συνοδεύονται ταυτόχρονα και με την ανάλογη φωτεινότητα, αφετέρου να τοποθετηθούν τα pixel και στην ανάλογη συστάδα. Φυσικά σε μια εικόνα μπορεί να γίνει συσταδοποίηση των στοιχείων της (pixel) είτε μόνο ως προς τις διάφορες εκφάνσεις φωτεινότητας αυτών είτε μόνο ως προς τις χρωματική ποικιλία αυτών (Βλέπε Εικόνα 1.1).

Η συσταδοποίηση κάποιων δεδομένων εισόδου, είτε αναφερόμαστε για τα στοιχεία μιας εικόνας (pixels), είτε για τα πλαίσια ενός video (frames) πρέπει να έχει ως έξοδο ακριβείς συστάδες. Τι ακριβώς όμως ορίζουμε με την έννοια ότι οι τελικές συστάδες που θα δημιουργηθούν ύστερα από την εκτέλεση κάποιου αλγορίθμου συσταδοποίησης, είναι ακριβείς; Το πόσο καλή και ακριβής είναι μια συστάδα

επιβεβαιώνεται μέσα από τρία κύρια κριτήρια: 1) Όλα τα στοιχεία της συγκεκριμένης συστάδας θα πρέπει να χαρακτηρίζονται από ένα υψηλό βαθμό ομοιότητας 2) Η *διακύμανση*[1] της εκάστοτε συστάδας θα πρέπει να είναι όσο το δυνατόν μικρότερη 3) Το μέγεθός της θα πρέπει να βρίσκεται και εντός κάποιων λογικών ορίων και για να επιτευχθεί κάτι τέτοιο λαμβάνεται υπόψη και η *μέση διακύμανση*[2].



Εικόνα 1.1

α) Η αρχική εικόνα της οποίας τα στοιχεία συσταδοποιήθηκαν με τον αλγόριθμό k-means ως προς τις ιδιότητες: b) φωτεινότητα χρώματος και c) χρώμα. Σε κάθε στοιχείο της εικόνας που ανήκει σε μια συγκεκριμένη συστάδα έχει δοθεί η μέση τιμή της φωτεινότητας (για τη b) και η μέση τιμή του χρώματος (για τη c) που χαρακτηρίζει τα στοιχεία που ανήκουν στην ίδια συστάδα.

Κάποιοι αλγόριθμοι συσταδοποίησης ενδεχομένως να δίνουν αρκετά καλά αποτελέσματα ακόμα και με την παρουσία θορύβου ενώ κάποιοι άλλοι ενδεχομένως να μην καταφέρνουν να δώσουν κάποια αξιόπιστα αποτελέσματα σε αυτή την περίπτωση. Σε κάθε εφαρμογή πρέπει να επιλέγεται από τον χρήστη ο καταλληλότερος αλγόριθμος που έχει την καλύτερη και πιο αξιόπιστη απόδοση αποτελεσμάτων.

[1] Ορίζεται ως το άθροισμα των τετραγωνικών διαφορών του κάθε στοιχείου από τη μέση τιμή.
[2] Πρόκειται ουσιαστικά για τη συνολική διακύμανση διαιρεμένη με το μέγεθος της συστάδας.

1.2 Κίνητρα και Στόχοι

Στόχος της παρούσας διπλωματικής εργασίας είναι η περιγραφή και η αξιολόγηση προχωρημένων αλγορίθμων συσταδοποίησης, που εξετάζουν το φάσμα ενός πίνακα, και μπορούν να εφαρμοστούν πάνω σε δεδομένα που σχετίζονται τόσο με ψηφιακές εικόνες όσο και ψηφιακά βίντεο. Οι συγκεκριμένες πηγές πληροφορίας χαρακτηρίζονται συνήθως από την παρουσία θορύβου, γεγονός το οποίο επιφέρει επιπλέον φόρτο κατά τη διάρκεια της συσταδοποίησης. Συγκεκριμένα επιβάλλεται μια αρχική επεξεργασία των δεδομένων πριν εφαρμοστεί κάποιος ήδη γνωστός αλγόριθμος συσταδοποίησης, που θα καθορίσει τις τελικές συστάδες και τα περιεχόμενα αυτών.

Κατά καιρούς έχουν προταθεί αρκετοί αλγόριθμοι συσταδοποίησης δεδομένων που όμως δεν μπορούν να εφαρμοστούν απευθείας πάνω σε ορισμένα δεδομένα γιατί παρουσιάζουν τα εξής σημαντικά προβλήματα και περιορισμούς:

- Συχνά δεν δίνουν ικανοποιητικά αποτελέσματα για δεδομένα που ανήκουν σε σχετικά μεγάλες διαστάσεις. Σε αυτή την περίπτωση ενδέχεται να δημιουργηθούν μη ακριβείς (όχι αρκετά καλές) συστάδες ή ακόμα και να αποτύχουν στον προσδιορισμό υποσυστάδων που ενδεχομένως υπάρχουν στα προς εξερεύνηση δεδομένα μας.
- Επίσης αλγόριθμοι όπως ο k-means, ο οποίος περιγράφεται και πιο κάτω, αποτυγχάνει πολύ εύκολα στην περίπτωση που οι συστάδες δεν αντιστοιχίζονται σε κυρτές περιοχές. Το συγκεκριμένο πρόβλημα έρχονται να λύσουν κάποιες νέες τεχνικές που βασίζονται πάνω σε αλγόριθμους φασματικής συσταδοποίησης.
- Δεν παρέχουν πολλές φορές την απαιτούμενη σταθερότητα. Πιο συγκεκριμένα ένας αλγόριθμος φασματικής συσταδοποίησης μας παρέχει τη δυνατότητα, έστω και μετά από μια πειραματική μελέτη, να προσδιορίσουμε τον αριθμό των τελικών συστάδων και μια ικανοποιητική τιμή της παραμέτρου κλιμάκωσης, για την οποία θα γίνει εκτενής αναφορά πιο κάτω.

- Αποτυγχάνουν στον εντοπισμό του πως γίνεται η καθολική κατανομή των δεδομένων που χρησιμοποιούνται για την εκπαίδευση του συστήματός μας.

Αν και οι μέθοδοι φασματικής συσταδοποίησης είναι σχετικά καινούργιες τεχνικές επίλυσης προβλημάτων συσταδοποίησης, εντούτοις καταφέρνουν να λύσουν αρκετά προβλήματα με έναν πιο κομψό, ημι-επιβλεπόμενο και καινοτόμο τρόπο, δίνοντας έτσι πιο αποδοτικά αποτελέσματα.

1.3 Γενικά για τις Μεθόδους, τα Βήματα και τις Ιδιότητές τους

Σε αυτή την ενότητα παρουσιάζουμε κάποιες από τις πιο γνωστές μεθόδους φασματικής συσταδοποίησης που χρησιμοποιούνται για τη συσταδοποίηση των στοιχείων μιας εικόνας.

Γενικά οι περισσότερες μέθοδοι φασματικής συσταδοποίησης εισάγουν δύο παραμέτρους που θα πρέπει είτε να δοθούν ως είσοδος από τον χρήστη είτε να προσδιοριστούν αυτόματα από τον ίδιο τον αλγόριθμο. Οι δύο αυτές παράμετροι είναι ο αριθμός των τελικών συστάδων, που αποτελεί και το μεγαλύτερο πρόβλημα για όλους τους αλγορίθμους συσταδοποίησης, και η τιμή των παραμέτρων καθολικής ή τοπικής κλιμάκωσης που καθορίζουν ποια στοιχεία σχετίζονται μεταξύ τους και ποια όχι.

Η πρώτη μέθοδος που παρουσιάζεται είναι αυτή που προτάθηκε από τους Ng, Jordan και Weiss το 2002 [1]. Ο συγκεκριμένος αλγόριθμος επιτρέπει στο χρήστη να επιλέξει ο ίδιος τόσο τον αριθμό των συστάδων όσο και μια παράμετρο καθολικής κλιμάκωσης, μετά από την επαναληπτική εκτέλεση του ίδιου πειράματος για διαφορετικές τιμές της καθολικής παραμέτρου. Η επιλογή μιας παραμέτρου, που δεν ανταποκρίνεται επαρκώς στη διάταξη και τη σχέση των δεδομένων αφενός μεν επηρεάζει την αξιοπιστία των τελικών αποτελεσμάτων αφετέρου δε και τον αριθμό των συστάδων.

Η μέθοδος της ασυμμετρικής φασματικής συσταδοποίησης είναι αποτέλεσμα της εργασίας των Fischer και Polland που εκδόθηκε το 2005 [6]. Σε αυτή τη μέθοδο προτείνεται η χρήση τοπικών παραμέτρων κλιμάκωσης που υπολογίζονται αυτόματα

από τον ίδιο τον αλγόριθμο, ενώ ο αριθμός των συστάδων καθορίζεται από τον χρήστη. Βέβαια το γεγονός ότι ο χρήστης ασχολείται μόνο με την τιμή της μιας από τις δύο παραμέτρους, αποτελεί για τον ίδιο σχετικά εύκολη υπόθεση ο προσδιορισμός μόνο του αριθμού των τελικών συστάδων.

Η τελευταία μέθοδος φασματικής συσταδοποίησης που παρουσιάζεται, ονομάζεται ZP συσταδοποίηση και δημοσιεύτηκε από τους Zelnik-Manor και Perona το 2004 [4]. Η δουλειά τους βασίστηκε σε αυτή της πρώτης μεθόδου που υλοποίησαν οι Ng, Jordan και Weiss, αλλά με τη μόνη διαφορά ότι κατάφεραν να αυτοματοποιήσουν τόσο τον υπολογισμό των τοπικών παραμέτρων κλιμάκωσης όσο και αυτόν του υπολογισμού των τελικών συστάδων. Ο υπολογισμός του πλήθους των τελικών συστάδων γίνεται μέσα από τον έλεγχο ενός συγκεκριμένου διαστήματος ακεραίων τιμών καθεμία από τις οποίες εξετάζεται για το τελικό αποτέλεσμα που δίνει.

Όλες αυτές οι μέθοδοι φασματικής συσταδοποίησης χρησιμοποιούν στο τελευταίο τους βήμα αλγορίθμους συσταδοποίησης που ήδη υπάρχουν για να συσταδοποιήσουν τα μορφοποιημένα πλέον αρχικά δεδομένα εισόδου. Γι' αυτό στο Κεφάλαιο 2 παρουσιάζονται οι υπάρχουσες κλασσικές κατηγορίες αλγορίθμων συσταδοποίησης, καθώς και κάποιοι από τους αλγορίθμους που ανήκουν σε αυτές τις κατηγορίες.

1.4 Ορισμοί

1.4.1 Γενικά

Σε αυτή την εισαγωγική ενότητα θα δοθούν κάποιοι ορισμοί και κάποιες σημειογραφίες που θα χρησιμοποιηθούν στα παρακάτω κεφάλαια.

Αρχικά θεωρούμε ένα γράφο $G = (V, E)$, όπου V είναι το σύνολο των κορυφών και E το σύνολο των ακμών.

1.4.2 Ορισμός συμπληρώματος ενός συνόλου κορυφών

Για ένα δοθέν σύνολο κορυφών $A \subset V$ συμβολίζουμε το *συμπλήρωμά* του ως $\overline{A}$.

1.4.3 Ορισμός του διανύσματος δείκτη 1_A

Για ένα σύνολο κορυφών $A \subset V$ ορίζουμε το *διάνυσμα δείκτη (indicator vector)* $1_A = (f_1,...,f_n)' \in \Re^n$ με στοιχεία: $f_i = 1$ αν η κορυφή $v_i \in A$ και $f_i = 0$ σε κάθε άλλη περίπτωση.

1.4.4 Ορισμοί υπολογισμού του μεγέθους ενός υποσυνόλου κορυφών

Θεωρούμε δύο διαφορετικούς τρόπους για τον υπολογισμό του μεγέθους ενός υποσυνόλου $A \subset V$: Ο πρώτος τρόπος θεωρεί ότι το μέγεθος του συνόλου A ισούται με τον αριθμό των κορυφών που περιέχει αυτό το σύνολο και συμβολίζεται με $|A|$. Ο δεύτερος τρόπος υπολογίζει το μέγεθος του A λαμβάνοντας υπόψη τα βάρη των ακμών του και ορίζεται ως: $vol(A) := \sum_{i \in A} d_i$.

1.4.5 Ορισμός συνδεδεμένου υποσυνόλου

Ένα υποσύνολο κορυφών $A \subset V$ ενός γράφου είναι συνδεδεμένο αν δύο οποιεσδήποτε κορυφές του που ανήκουν στο A μπορούν να συνδεθούν με ένα μονοπάτι έτσι ώστε όλα τα ενδιάμεσα σημεία να βρίσκονται επίσης μέσα στο A.

1.4.6 Ορισμός συνδεδεμένου στοιχείου

Ένα υποσύνολο κορυφών $A \subset V$ καλείται συνδεδεμένο στοιχείο αν είναι συνδεδεμένο και αν δεν υπάρχουν καθόλου συνδέσεις μεταξύ των κορυφών που ανήκουν στο A και σε αυτές που ανήκουν στο $\overline{A}$.

1.4.7 Ορισμός διαμέρισης ενός γράφου

Λέμε ότι τα σύνολα κορυφών $A_1,\ldots,A_k$ σχηματίζουν μια διαμέριση του γράφου G αν $A_i \cap A_j = \varnothing$ και $A_1 \cup \ldots \cup A_k = V$.

1.4.8 Στροφή Givens

Μία στροφή Givens αναπαριστάται από έναν πίνακα της μορφής:

$$G(i,k,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}$$

όπου $c = \cos(\theta)$ και $s = \sin(\theta)$ και εμφανίζονται στις τομές των $i-$οστών και $k-$οστών γραμμών και στηλών. Ο πίνακας στροφής *Givens* προκύπτει από τον μοναδιαίο πίνακα αντικαθιστώντας τα εξής στοιχεία με τις αντίστοιχες τιμές:

$g_{ii} = c$, $g_{kk} = c$, $g_{ik} = s$, $g_{ki} = -s$. Το γινόμενο $G(i,k,\theta)^T$ **x** αναπαριστά μια στροφή γωνίας θ του διανύσματος **x,** αντίθετη με τη φορά των δεικτών του ρολογιού, πάνω στο (i,k) επίπεδο. Η κύρια χρήση της στροφής Givens είναι η εισαγωγή μηδενικών σε διανύσματα και πίνακες.

1.4.9 Ορισμός ενός πίνακα συνδιακύμανσης Σ

Θεωρούμε το διάνυσμα στήλη $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ όπου τα στοιχεία του είναι τυχαίες μεταβλητές, καθεμία από τις οποίες έχει πεπερασμένη διακύμανση, τότε ο πίνακας συνδιακύμανσης Σ έχει ως στοιχείο του (i, j) τη συνδιακύμανση $\Sigma_{ij} = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$ όπου $\mu_i = E(X_i)$ είναι η μέση τιμή της i-οστής καταχώρησης στο διάνυσμα X. Με άλλα λόγια έχουμε ότι:

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

Κεφάλαιο 2

Κατηγορίες Αλγορίθμων Συσταδοποίησης

2.1 Κατηγοριοποίηση των κλασσικών αλγορίθμων συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης δεδομένων μπορεί να είναι *ιεραρχικοί* ή *διαμεριστικοί*. Οι ιεραρχικοί αλγόριθμοι εντοπίζουν διαδοχικές συστάδες χρησιμοποιώντας τη γνώση από συστάδες που έχουν ήδη υπολογισθεί, ενώ οι διαμεριστικοί αλγόριθμοι προσδιορίζουν όλες τις συστάδες ταυτόχρονα. Οι ιεραρχικοί αλγόριθμοι μπορεί να είναι *συσσωρευτικοί ή διαιρετικοί*. Οι συσσωρευτικοί αλγόριθμοι ξεκινούν με το κάθε στοιχείο να ανήκει σε μια ξεχωριστή συστάδα και τα συγχωνεύει σε διαδοχικά μεγαλύτερες συστάδες. Οι διαιρετικοί αλγόριθμοι ξεκινούν με ολόκληρο το σύνολο των δεδομένων και προχωρούν διαιρώντας το σε διαδοχικά μικρότερες συστάδες. Σε αντίθεση με τους συσσωρευτικούς αλγορίθμους που δημιουργούν τις συστάδες με διαδικασία από *κάτω προς τα πάνω (bottom-up)*, οι διαιρετικοί αλγόριθμοι δημιουργούν τις συστάδες με διαδικασία από *πάνω προς τα κάτω (top-down)*. Μια άλλη ομάδα αλγορίθμων εντάσσονται σε μια κατηγορία συσταδοποίησης που είναι γνωστή με την ονομασία *συσταδοποίηση 2-τρόπων (two-way clustering ή biclustering ή co-clustering)*. Οι αλγόριθμοι που επιτελούν τέτοιου είδους συσταδοποίηση καθορίζουν ότι θα εισέλθουν σε συστάδες όχι μόνο τα αντικείμενα αλλά και τα χαρακτηριστικά αυτών των αντικειμένων. Πέρα όμως από αυτές τις προαναφερθείσες κατηγορίες αλγορίθμων συσταδοποίησης υπάρχουν και αλγόριθμοι που χρησιμοποιούν τις ιδιοτιμές και τα ιδιοδιανύσματα ενός πίνακα για να συσταδοποιήσουν τα αρχικά

δεδομένα. Οι αλγόριθμοι αυτοί ονομάζονται *φασματικοί αλγόριθμοι ή αλγόριθμοι φασματικής συσταδοποίησης.*

Τέλος ένας άλλος τρόπος διάκρισης των αλγορίθμων συσταδοποίησης σχετίζεται με το αν η συσταδοποίηση χρησιμοποιεί συμμετρικές ή μη συμμετρικές αποστάσεις. Μια ιδιότητα του Ευκλείδειου χώρου είναι ότι οι αποστάσεις είναι συμμετρικές (δηλαδή η απόσταση του αντικειμένου Α από το Β είναι ίδια με την απόσταση του Β από το Α). Σε αυτό το σημείο αξίζει να αναφέρουμε και ποια είναι τα συνηθέστερα μέτρα απόστασης που χρησιμοποιούνται από τους αλγορίθμους συσταδοποίησης.

2.1.1 Μέτρα απόστασης των αλγορίθμων συσταδοποίησης

Ένα σημαντικό βήμα σε κάθε συσταδοποίηση είναι η επιλογή του καταλληλότερου μέτρου απόστασης καθώς είναι αυτό που προσδιορίζει πως υπολογίζεται η ομοιότητα μεταξύ δύο στοιχείων. Φυσικά, το μέτρο απόστασης είναι αυτό που επηρεάζει και το μέγεθος των συστάδων, καθώς μερικά στοιχεία μπορεί να βρίσκονται κοντά το ένα με το άλλο σύμφωνα με ένα μέτρο απόστασης και αρκετά μακριά σύμφωνα με κάποιο άλλο. Για παράδειγμα, σε ένα χώρο δύο διαστάσεων, η απόσταση ανάμεσα μεταξύ των σημείων (1,0) και (0,0) είναι πάντοτε 1 σύμφωνα με τις συνήθεις νόρμες, αλλά η απόσταση του σημείου (1,1) από το σημείο (0,0) μπορεί να είναι 2, $\sqrt{2}$ ή 1 ανάλογα με το αν κάποιος πάρει τη νόρμα απόστασης 1, 2 ή αυτή του απείρου. Μερικά από τα πιο κοινά μέτρα απόστασης είναι τα παρακάτω:

- *Η Ευκλείδεια απόσταση.* Ουσιαστικά πρόκειται για την 2-νόρμα. Η ευκλείδεια απόσταση μεταξύ δύο σημείων:
  $P = (p_1, p_2, \ldots, p_n)$, $Q = (q_1, q_2, \ldots, q_n)$ που ανήκουν στο n-διάστατο Ευκλείδειο χώρο ορίζεται από την ακόλουθη σχέση:
  $$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$
- *Η απόσταση Manhattan.* Κατέχει και την ονομασία 1-νόρμα (στην ξενόγλωσση βιβλιογραφία απαντάται αρκετές φορές και με την ονομασία *taxicab norm).* Η απόσταση Manhattan για δύο σημεία που ανήκουν σε έναν

Ευκλείδειο χώρο με καθορισμένο σύστημα καρτεσιανών συντεταγμένων ορίζεται ως το άθροισμα των μηκών των προβολών του ευθύγραμμου τμήματος, που εδράζεται εντός των σημείων, πάνω στους άξονες συντεταγμένων. Θεωρώντας ως παράδειγμα το επίπεδο, η απόσταση Manhattan μεταξύ των σημείων P_1 με συντεταγμένες (χ_1,y_1) και P_2 με συντεταγμένες (χ_2,y_2) είναι ίση με $|\chi_1 - \chi_2| + |y_1 - y_2|$.

- *Η νόρμα μεγίστου.* Για ένα διάνυσμα $\chi = (\chi_1,\ldots,\chi_n)$ που ανήκει σε έναν πεπερασμένης διάστασης χώρο συντεταγμένων ορίζεται από την ακόλουθη σχέση: $\|\chi\|_\infty = \max\{|\chi_1|,\ldots,|\chi_n|\}$. Ο λόγος για τον δείκτη «∞» προκύπτει από το γεγονός ότι:

$$\lim_{p\to\infty} \|f\|_p = \|f\|_\infty,$$

όπου

$$\|f\|_p = \left(\int_D |f|^p \, d\mu\right)^{1/p}$$

όπου D είναι ο χώρος της f και στην περίπτωση που το D είναι ένα διακριτό σύνολο τότε το ολοκλήρωμα γίνεται άθροισμα.

- *Η απόσταση Mahalanobis.* Είναι ένα μέτρο απόστασης που παρέχει ένα χρήσιμο τρόπο προσδιορισμού της *ομοιότητας* ενός αγνώστου δειγματοληπτικού συνόλου με ένα άλλο γνωστό. Αν θέλαμε να ορίσουμε το συγκεκριμένο μέτρο απόστασης για μια ομάδα τιμών με μέση τιμή $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_p)^T$ και πίνακα συνδιακύμανσης Σ για διάνυσμα $x = (x_1, x_2, x_3, \ldots, x_p)^T$ τότε έχουμε την ακόλουθη σχέση:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

Η απόσταση Mahalanobis μπορεί να οριστεί και ως το μέτρο ανομοιότητας μεταξύ δύο τυχαίων διανυσμάτων $\vec{x}$ και $\vec{y}$ που έχουν την ίδια κατανομή και με πίνακα συνδιακύμανσης P, τότε ισχύει:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T P^{-1} (\vec{x}-\vec{y})}.$$

Αν ο πίνακας συνδιακύμανσης είναι ο μοναδιαίος πίνακας τότε η απόσταση Mahalanobis μας δίνει την Ευκλείδεια απόσταση. Αν ο πίνακας

συνδιακύμανσης είναι διαγώνιος τότε το προκύπτον μέτρο απόστασης ονομάζεται *κανονικοποιημένη ευκλείδεια απόσταση* και ισούται με τη σχέση:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}},$$

όπου το σ_i είναι η διασπορά του χ_i πάνω στο σύνολο δείγματος.

- *Η γωνία που σχηματίζεται μεταξύ δύο διανυσμάτων* μπορεί να χρησιμοποιηθεί ως μέτρο απόστασης όταν γίνεται συσταδοποίηση δεδομένων υψηλής διάστασης.
- *Η απόσταση Hamming.* Για ένα συγκεκριμένο μέγεθος n, η απόσταση Hamming είναι ένα μέτρο απόστασης στο χώρο των διανυσμάτων των λέξεων αυτού του μεγέθους που μετράει των ελάχιστο αριθμό αντικαταστάσεων που απαιτούνται για την αλλαγή ενός αντικειμένου σε ένα άλλο. Για παράδειγμα για τις ακόλουθες συμβολοσειρές: "**toned**" και "**roses**" η απόσταση Hamming είναι ίση με 3.

2.2 Ιεραρχική συσταδοποίηση

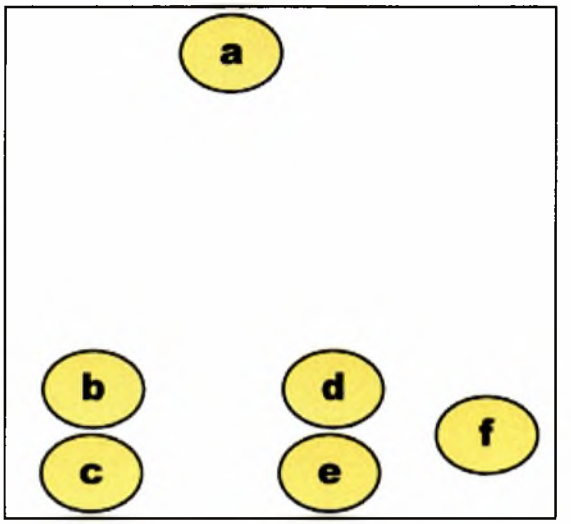
2.2.1 Γενικά

Η ιεραρχική συσταδοποίηση κατασκευάζει (*συσσωρευτικοί αλγόριθμοι*) ή διασπά (*διαιρετικοί αλγόριθμοι*) μια ιεραρχία από συστάδες. Ο συνήθης τρόπος απεικόνισης αυτής της ιεραρχίας είναι μέσω ενός δέντρου που αποκαλείται *δενδρόγραμμα* και το οποίο στη μια του άκρη έχει τα αντικείμενα ως ξεχωριστές οντότητες και στην άλλη, μια απλή συστάδα που περιέχει όλα τα αντικείμενα.

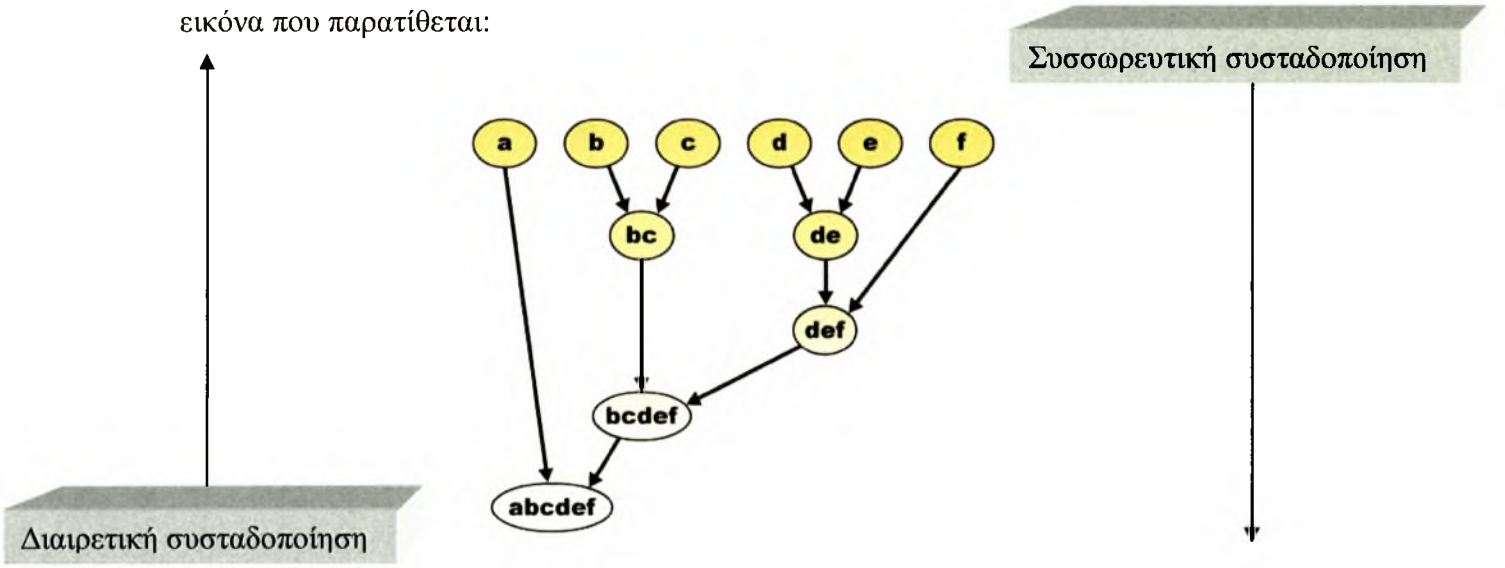
2.2.2 Συσσωρευτική συσταδοποίηση

Οι συσσωρευτικοί αλγόριθμοι ξεκινούν από την κορυφή (το πάνω μέρος) του δέντρου ενώ οι διαιρετικοί από τη ρίζα του. Αν κόψουμε το δέντρο σε ένα συγκεκριμένο ύψος τότε αυτό θα έχει ως αποτέλεσμα την επιστροφή μιας

συσταδοποίησης των αντικειμένων που αντιστοιχεί στην επιλεγόμενη ακρίβεια. Για να γίνει πιο εύκολα αντιληπτό κάτι τέτοιο θα οπτικοποιήσουμε την εφαρμογή ενός συσσωρευτικού αλγορίθμου πάνω σε ένα σύνολο δεδομένων ο οποίος χρησιμοποιεί ως μέτρο απόστασης την Ευκλείδεια απόσταση. Στην ακόλουθη εικόνα παρουσιάζονται τα δεδομένα εισόδου για τον συσσωρευτικό αλγόριθμο συσταδοποίησης:



Το δενδρόγραμμα ιεραρχικής συσταδοποίησης που θα προκύψει απεικονίζεται στην εικόνα που παρατίθεται:



Σύμφωνα λοιπόν με το παραπάνω δενδρόγραμμα αν κόψουμε το δέντρο ακριβώς μετά τη δεύτερη σειρά τότε θα έχουν προκύψει οι εξής συστάδες: {a}, {b c}, {d e}, {f}, ενώ αν το κόψουμε ακριβώς μετά την τρίτη σειρά τότε οι συστάδες που θα έχουν

προκύψει θα είναι οι εξής: {a}, {b c}, {d e f}. Η μέθοδος αυτή συσταδοποίησης κατασκευάζει την ιεραρχία από τα ξεχωριστά στοιχεία μέσω της σταδιακής συγχώνευσης των συστάδων. Στο συγκεκριμένο παράδειγμα έχουμε 6 στοιχεία προς συσταδοποίηση και ως πρώτο βήμα είναι η επιλογή εκείνων των στοιχείων που θα συγχωνευτούν σε μία συστάδα. Συνήθως επιλέγουμε τα δύο εκείνα στοιχεία που βρίσκονται πιο κοντά σύμφωνα με το επιλεγόμενο κάθε φορά μέτρο απόστασης. Μια άλλη εναλλακτική επιλογή θα ήταν η κατασκευή ενός πίνακα απόστασης σε αυτό το πρωταρχικό στάδιο, όπου το στοιχείο που βρίσκεται στη θέση (i,j) του πίνακα δηλώνει την απόσταση μεταξύ των στοιχείων i και j. Έτσι, καθώς η συσταδοποίηση εξελίσσεται, οι γραμμές και οι στήλες του πίνακα συγχωνεύονται εφόσον και οι ίδιες οι συστάδες συγχωνεύονται ενώ ταυτόχρονα γίνεται και κατάλληλη ενημέρωση των αποστάσεων. Αν υποθέσουμε ότι έχουμε συγχωνεύσει τα δύο πιο κοντινά αντικείμενα b και c, τότε προκύπτουν οι ακόλουθες συστάδες: {a}, {b, c}, {d}, {e} και {f}, τις οποίες θέλουμε να συγχωνεύσουμε περαιτέρω. Για να επιτευχθεί κάτι τέτοιο θα πρέπει να λάβουμε υπόψη την απόσταση μεταξύ των συστάδων {a} and {b c}.

2.2.3 Απόσταση μεταξύ δύο συστάδων

Η απόσταση μεταξύ δύο συστάδων $\mathcal{A}$ και $\mathcal{B}$ μπορεί να υπολογιστεί με έναν από τους ακόλουθους ορισμούς αποστάσεων που σχετίζουν τα στοιχεία των δύο συστάδων μεταξύ τους:

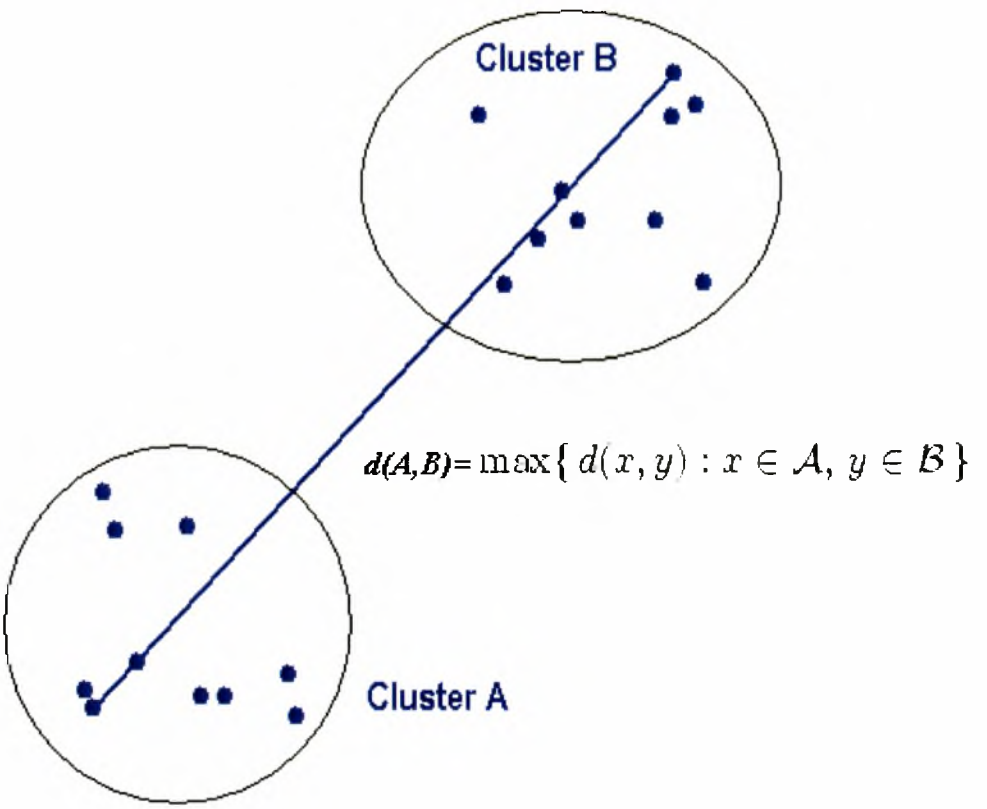
- *Η μέγιστη απόσταση μεταξύ στοιχείων της κάθε συστάδας (αναφέρεται και ως συσταδοποίηση πλήρους συνδέσμου):*
  $$\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \,\}.$$
- *Η ελάχιστη απόσταση μεταξύ στοιχείων της κάθε συστάδας (αναφέρεται και ως συσταδοποίηση απλού συνδέσμου):*
  $$\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \,\}.$$

- *Η μέση απόσταση μεταξύ στοιχείων της κάθε συστάδας (αναφέρεται και ως συσταδοποίηση μέσου συνδέσμου):*

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

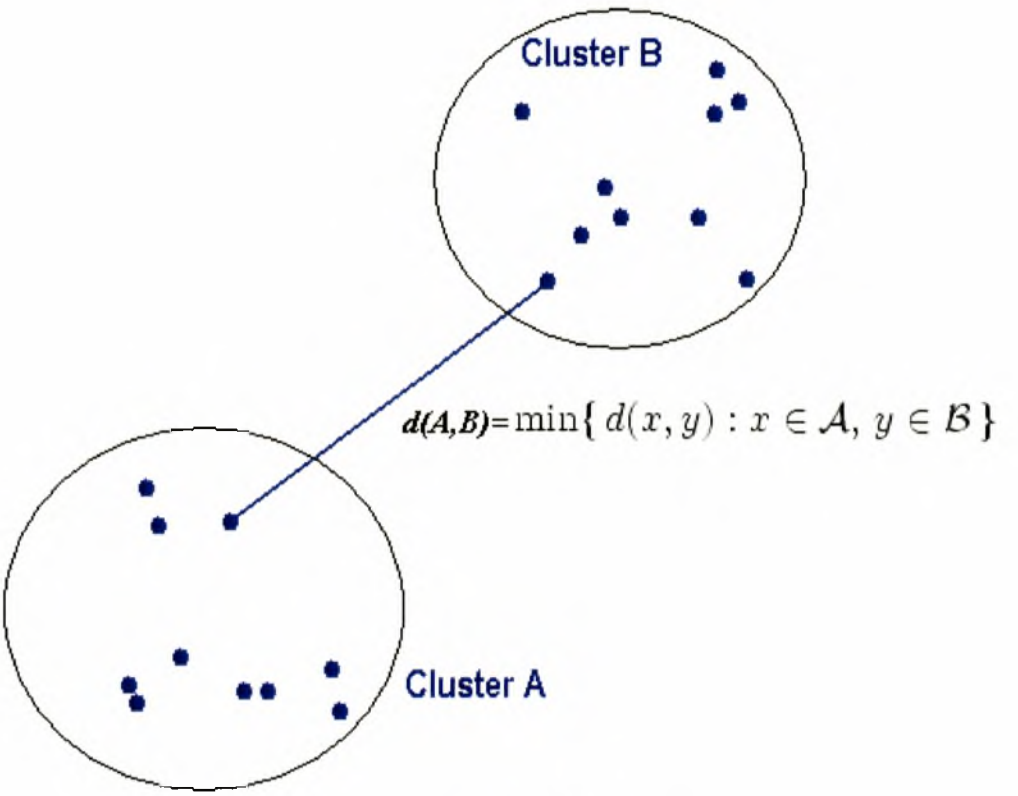
- *Το άθροισμα της διακύμανσης στο εσωτερικό μιας συστάδας.*
- *Η αύξηση που παρατηρείται στη διακύμανση για τη συστάδα που συγχωνεύεται.*

Στην *Εικόνα 2.2.3 (α)* παρουσιάζεται ο υπολογισμός της απόστασης μεταξύ δύο συστάδων με βάση τη συσταδοποίηση πλήρους συνδέσμου, όπου, όπως φαίνεται η απόσταση μεταξύ των δύο συστάδων είναι η απόσταση που ορίζεται μεταξύ των δύο πιο απομακρυσμένων στοιχείων της κάθε συστάδας. Η συγκεκριμένη μέθοδος δεν συνίσταται για δεδομένα στα οποία μπορεί να υφίσταται αρκετός θόρυβος. Δύο από τα βασικά της πλεονεκτήματα είναι ότι δημιουργεί συμπαγείς συστάδες και θεωρείται ιδιαίτερα χρήσιμη αν είναι αναμενόμενο ότι οντότητες της ίδιας συστάδας βρίσκονται σε μεγάλη μεταξύ τους απόσταση στον πολυδιάστατο χώρο (εφόσον φυσικά δεν υπάρχει θόρυβος).



Εικόνα 2.2.3 (α)

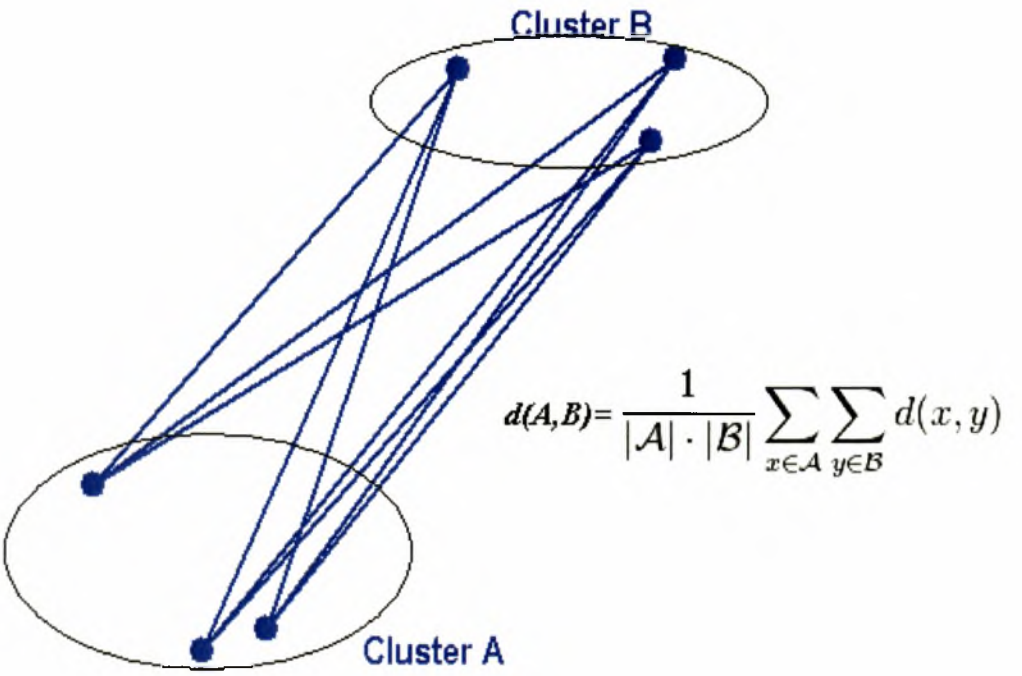
Στην *Εικόνα 2.2.3 (b)* παρουσιάζεται ο υπολογισμός της απόστασης μεταξύ δύο συστάδων με βάση τη συσταδοποίηση απλού συνδέσμου, όπου όπως μπορούμε να διακρίνουμε η απόσταση ισούται με την απόσταση των δύο πιο κοντινών στοιχείων της κάθε συστάδας. Η χρήση αυτής της μεθόδου δημιουργεί *το φαινόμενο της αλυσίδας (chaining phenomenon)* καθώς μπορεί να επιβάλλεται η συγχώνευση δύο συστάδων λόγω της ύπαρξης οντοτήτων που είναι αρκετά κοντά η μία με την άλλη χωρίς να ασχολείται με τις θέσεις των άλλων οντοτήτων μέσα στη συστάδα.



Εικόνα 2.2.3 (b)

Τέλος η απόσταση που υπολογίζεται με βάση τη συσταδοποίηση του μέσου συνδέσμου απεικονίζεται σχηματικά στην *Εικόνα 2.2.3 (c)*. Είναι αυτή που απαιτεί το μεγαλύτερο κόστος σε υπολογισμούς καθώς υπολογίζει τη μέση απόσταση μεταξύ όλων των πιθανών ζευγών στοιχείων από τις δύο συστάδες που διερευνούνται. Υπάρχουν αρκετές παραλλαγές τις συγκεκριμένης μεθόδου αλλά αυτό που είναι αρκετά σημαντικό είναι ότι η χρήση της δεν δημιουργεί το φαινόμενο της αλυσίδας ενώ τα απομακρυσμένα στοιχεία (outliers) δε χρίζουν ιδιαίτερης εύνοιας κατά την

απόφαση της δημιουργίας των συστάδων. Ως άμεση συνέπεια αυτού αποτελεί το γεγονός ότι η συγκεκριμένη μέθοδος είναι πιο δημοφιλής από τις άλλες δύο.



Εικόνα 2.2.3 (c)

Επίσης κάτι που αξίζει να αναφερθεί είναι ότι η κάθε συσσώρευση συμβαίνει σε απόσταση συστάδων μεγαλύτερη από αυτή του προηγούμενου σταδίου συσσώρευσης, δίνοντας έτσι τη δυνατότητα τερματισμού της συσταδοποίησης είτε όταν οι συστάδες βρίσκονται πολύ μακριά για να συγχωνευτούν (κριτήριο της απόστασης) είτε όταν υπάρχει ικανοποιητικά μικρός αριθμός υπολογισμένων συστάδων (κριτήριο του αριθμού των συστάδων).

2.2.4 Διαιρετική συσταδοποίηση

Οι διαιρετικοί αλγόριθμοι όπως αναφέρθηκε και προηγουμένως ξεκινούν με όλα τα στοιχεία να εμπεριέχονται μέσα σε μία συστάδα και σταδιακά τη διαμερίζουν σε μικρότερες συστάδες έως ότου να ικανοποιηθεί η συνθήκη τερματισμού. Η βασική ιδέα είναι ότι μια συστάδα διασπάται όταν κάποια από τα στοιχεία της δεν βρίσκονται αρκετά κοντά στα υπόλοιπα στοιχεία της. Κάποιες φορές μπορεί να περιλαμβάνουν

τεχνικές κλαδέματος και συγχώνευσης ούτως ώστε να επιτευχθεί ένα πιο βελτιωμένο τελικό αποτέλεσμα. Αξίζει να αναφερθεί ότι συγκριτικά με τους συσσωρευτικούς αλγορίθμους είναι λιγότερο αποδοτικοί καθώς παρουσιάζονται υπολογιστικές δυσκολίες λόγω της θεώρησης ότι το σύνολο δεδομένων κάθε φορά διαιρείται σε δύο ομάδες.

2.3 Διαμεριστική συσταδοποίηση

2.3.1 Γενικά

Οι πιο γνωστοί αλγόριθμοι διαμεριστικής συσταδοποίησης που θα περιγραφούν σε αυτή την ενότητα, είναι οι εξής: ο fuzzy c-means και ο k-means αλγόριθμος συσταδοποίησης.

2.3.2 Fuzzy c-means συσταδοποίηση

Τεωρούμε το αρχικό σύνολο X των δεδομένων προς επεξεργασία, που αποτελείται από n διανύσματα χ_k, με p στοιχεία στο καθένα, δηλαδή: X = $\{\chi_1,\chi_2,\ldots,\chi_n\}$, $\chi_k \in \Re^p$.

Σε αυτή τη μέθοδο συσταδοποίησης κάθε στοιχείο μπορεί να ανήκει σε περισσότερες από μία συστάδες και κάθε σημείο x συνδέεται με έναν συντελεστή που εκφράζει το κατά πόσο ανήκει στην k-οστή συστάδα. Ο συντελεστής αυτός ονομάζεται *βαθμός* του εκάστοτε σημείου και συμβολίζεται με $u_k(x)$, ενώ το άθροισμα όλων των συντελεστών για το δεδομένο σημείο ισούται με 1:

$\sum_{i=1}^{c} u_{ik} = 1$, $\forall$ k και όπου c είναι ο αριθμός των συστάδων. Η τιμή του κάθε συντελεστή ισούται με:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d(c_i,k)}{d(c_j,k)} \right)^{\frac{2}{m-1}}} .$$

Σύμφωνα με αυτόν τον αλγόριθμο, το κεντροειδές μιας συστάδας είναι ο μέσος όλων των σημείων σταθμισμένος ως προς το βαθμό τους να ανήκουν στη συστάδα, δηλαδή:

$$c_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}}$$, όπου m είναι μία παράμετρος για την οποία ισχύει m=1 και ονομάζεται εκθέτης στάθμισης (weighting degree) ή βαθμός ασάφειας (fuzziness degree).

Αλγόριθμος fuzzy c-means

Είσοδος: X = $\{\chi_1, \chi_2, ..., \chi_n\}$ // Το σύνολο των στοιχείων

c //ο αριθμός των επιθυμητών συστάδων

Έξοδος: Οι c συστάδες που έχουν σχηματιστεί

Βήμα 1: Επιλογή μια αρχικής διαμέρισης των n αντικειμένων σε c συστάδες επιλέγοντας ένα πίνακα γειτνίασης U μεγέθους n×c. Η τιμή u_{ij} όπως ήδη αναφέρθηκε καθορίζει το βαθμό συμμετοχής του αντικειμένου χ_i στη συστάδα j και υπολογίζεται σύμφωνα με τη σχέση που έχει αναφερθεί πιο πάνω.

Βήμα 2: Υπολογισμός ενός fuzzy κριτηρίου με τη βοήθεια του πίνακα U που δημιουργήθηκε:

$$E^2(\mathcal{X}, \mathbf{U}) = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} \|\mathbf{x}_i - \mathbf{c_k}\|^2,$$

όπου c_k είναι το κεντροειδές της συστάδας k που δίνεται από τη φόρμουλα που προαναφέρθηκε.

Βήμα 3: Επανυπολογισμός των κέντρων των συστάδων για την ελαχιστοποίηση του πάνω κριτηρίου. Πήγαινε στο Βήμα 2.

2.3.3 k-means συσταδοποίηση

2.3.3.1 Ο αλγόριθμος k-means

Ο αλγόριθμος k-means είναι ένας επαναληπτικός αλγόριθμος ο οποίος καλείται να τοποθετήσει σε k συστάδες ένα σύνολο n στοιχείων που διαθέτουν κάποια γνωρίσματα και για τον οποίο ισχύει ότι k < n. Βασικά θεωρεί ότι τα γνωρίσματα των στοιχείων σχηματίζουν ένα διανυσματικό χώρο, ενώ η αντικειμενική συνάρτηση που προσπαθεί να ελαχιστοποιήσει είναι η συνολική διακύμανση στο εσωτερικό μιας συστάδας ή η συνάρτηση τετραγωνικού σφάλματος:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

,

όπου k είναι ο αριθμός των συστάδων και μ_i είναι το κεντροειδές όλων των στοιχείων $\chi_j \in S_i$. Σε αυτό το σημείο παρουσιάζεται ο αλγόριθμος k-means:

Αλγόριθμος k-means

Είσοδος: $X = \{\chi_1, \chi_2, ..., \chi_n\}$ // Το σύνολο των στοιχείων

k //ο αριθμός των επιθυμητών συστάδων

Έξοδος: Οι k συστάδες που έχουν σχηματιστεί

Βήμα 1: Ανάθεση των αρχικών κεντροειδών μ_i, $i = 1,2,...,k$ για το σύνολο των k συστάδων.

Για κάθε επανάληψη $r = 1,2,...,r_{max}$

Βήμα 2: Υπολογισμός της απόστασης κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε συστάδας: $d_{ji} = (\chi_j - \mu_i)^2$, j = 1,2,…,n και i=1,2,…,k

Βήμα 3: Κάθε στοιχείο χ_j αντιστοιχίζεται στη συστάδα για την οποία ισχύει: $\min_{j,i}(d_{ij}), \forall i, j.$

Βήμα 4: Υπολογισμός των νέων κεντροειδών των συστάδων ως εξής: $m_i^{(r+1)} = \frac{\sum_{j=1}^{n_i} \chi_j}{n_i}$, όπου

n_i ο αριθμός των στοιχείων που ανήκουν στη συστάδα i μέχρι στιγμής.

Βήμα 5: Αν $\left\| m_i^r - m_i^{(r+1)} \right\| < \varepsilon$ τότε σταμάτησε

αλλιώς

$r = r + 1$, πήγαινε στο Βήμα 2.

2.3.3.2 Ιδιότητες του αλγορίθμου k-means

Πολλές φορές ο k-means αλγόριθμος θεωρείται αρκετά ικανοποιητικός για κάποιες εφαρμογές ενώ σε ορισμένες περιπτώσεις δεν αποτελεί την καλύτερη επιλογή αλγορίθμου συσταδοποίησης δεδομένων. Πριν γίνει η επιλογή του αλγορίθμου k-means για κάποια εφαρμογή θα πρέπει να ληφθούν υπόψη κάποιες ιδιότητες που χαρακτηρίζουν το συγκεκριμένο αλγόριθμο. Οι ιδιότητες αυτές είναι:

- *Επιτυγχάνει σύγκλιση σε τοπικό βέλτιστο.* Ο αλγόριθμος k-means δεν μπορεί να εγγυηθεί την εύρεση ενός καθολικού μεγίστου ενώ το τελικό αποτέλεσμα της συσταδοποίησης μπορεί να επηρεαστεί σημαντικά από την αρχική επιλογή των κεντροειδών.
- *Δυνατότητα τυχαίας επιλογής των αρχικών κεντροειδών.* Η συσταδοποίηση μπορεί να επαναληφθεί αρκετές φορές, με διαφορετικά κάθε φορά κεντροειδή, ούτως ώστε να υπολογιστεί κάποια στιγμή μια ένθεση των δεδομένων σε συστάδες που θα βρίσκεται πιο κοντά στη βέλτιστη λύση.
- *Μεγάλη ταχύτητα:* Πρόκειται για μια σχετικά γρήγορη μέθοδο συσταδοποίησης με την οποία μπορεί επιτευχθεί σύγκλιση μέσα σε σύντομο χρονικό διάστημα.
- *Έχει την τάση να δημιουργεί σφαιρικές και ίσου μεγέθους συστάδες.*
- *Χρήση διαφορετικών κριτηρίων τερματισμού εκτέλεσής του.* Μερικά από τα πιο γνωστά κριτήρια τερματισμού εκτέλεσης του k-means είναι τα ακόλουθα:

  - ✓ *Όταν δε συμβαίνουν κάποιες αλλαγές στην ανάθεση του συνόλου δείγματος.*
  - ✓ *Όταν ο αριθμός των επαναλήψεων ξεπεράσει τον μέγιστο αριθμό επαναλήψεων που έχει οριστεί.*
  - ✓ *Όταν η μεταβολή στην συνολική παραμόρφωση* D[3] *πέφτει κάτω από ένα κατώφλι T, δηλαδή:* $1-\frac{D(n+1)}{D(n)}<T$.

- *Επιτυγχάνεται μείωση της παραμόρφωσης* D *καθώς γίνεται μετάβαση από το n-οστό βήμα στο n+1 βήμα του αλγορίθμου, δηλαδή:* $D(n+1)<D(n)$.

Είναι προφανές ότι οι ιδιότητες του αλγορίθμου k-means υποδεικνύουν κάποιες αδυναμίες που δεν τον καθιστούν την ιδανικότερη επιλογή για κάποιες εφαρμογές. Οι σημαντικότερες από αυτές τις αδυναμίες είναι:

- *Δυσκολία προσδιορισμού των πραγματικών συστάδων.* Αν τα στοιχεία ενός συγκεκριμένου συνόλου δεδομένων εισαχθούν με διαφορετική σειρά στον αλγόριθμο k-means, τότε το αποτέλεσμα της συσταδοποίησης μπορεί να είναι

---

[3] $D=\sum_{i=1}^{k}\sum_{\chi\in C_i}\|\chi-\mu_i\|^2$

εντελώς διαφορετικό από αυτό που θα υπολογιστεί από την εισαγωγή των στοιχείων με άλλη σειρά. Βέβαια αυτό εξαρτάται κατά πολύ και από το πλήθος των δεδομένων που διαθέτουμε. Αν ο αριθμός των δεδομένων δεν είναι ικανοποιητικός τότε υπάρχει μεγαλύτερη πιθανότητα δημιουργίας ενός νέου αποτελέσματος συσταδοποίησης κάθε φορά που αλλάζουμε τη σειρά εισαγωγής στον αλγόριθμο.

- *Μεγάλη ευαισθησία στην αρχική ανάθεση κεντροειδών.* Ο αλγόριθμος k-means εξαρτάται κατά πολύ από την αρχική επιλογή των κεντροειδών καθώς διαφορετικές αναθέσεις ενδέχεται να δημιουργήσουν διαφορετικά αποτελέσματα συσταδοποίησης ή ακόμα και να εγκλωβίσουν τον αλγόριθμο σε κάποιο τοπικό βέλτιστο.
- *Το μέτρο της μέσης τιμής επηρεάζεται αρκετά από τα απομακρυσμένα στοιχεία των δεδομένων (outliers).* Αν στο σύνολο δεδομένων υπάρχουν στοιχεία που εντοπίζονται αρκετά μακριά από το κεντροειδές τότε ίσως να οδηγήσουν στην απομάκρυνση του κεντροειδούς από την πραγματική του θέση. Μια λύση για το συγκεκριμένο πρόβλημα αποτελεί η χρήση του μέσου ως μέτρο υπολογισμού των κεντροειδών.

2.4 Συσταδοποίηση 2-τρόπων

2.4.1 Γενικά

Πρόκειται για μια μέθοδο συσταδοποίησης που επιτρέπει την ταυτόχρονη συσταδοποίηση των γραμμών και των στηλών ενός πίνακα. Ο όρος εισήχθη για πρώτη φορά από τον Mirkin, αν και η συγκεκριμένη τεχνική συσταδοποίησης παρουσιάστηκε πολύ πιο νωρίς από τον Hartigan. Οι περισσότεροι από αυτούς τους αλγορίθμους χρησιμοποιούνται σε εφαρμογές της βιοπληροφορικής και στη συγκεκριμένη ενότητα θα αναφέρουμε ονομαστικά κάποιους αλγορίθμους της συγκεκριμένης κατηγορίας.

2.4.2 Δημιουργία συστάδων

Δοθέντος ενός πίνακα μεγέθους $m \times n$ ο αλγόριθμος συσταδοποίησης 2-τρόπων δημιουργεί συστάδες με τον εξής τρόπο: ένα υποσύνολο γραμμών του πίνακα που παρουσιάζει παρόμοια συμπεριφορά κατά πλάτος ενός υποσυνόλου στηλών ή και το αντίθετο θα τοποθετηθεί σε μια νέα συστάδα.

2.4.3 Είδη αλγορίθμων συσταδοποίησης 2-τρόπων

Υπάρχουν διάφορα είδη συσταδοποίησης 2-τρόπων τα οποία παρουσιάζονται οπτικά και μέσω των σχημάτων που ακολουθούν. Συγκεκριμένα τα είδη αυτών των αλγορίθμων είναι τα εξής:

1. Συσταδοποίηση 2-τρόπων με σταθερές τιμές
2. Συσταδοποίηση 2-τρόπων με σταθερές τιμές σε γραμμές και στήλες
3. Συσταδοποίηση 2-τρόπων με συνεκτικές τιμές

2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0
2.0	2.0	2.0	2.0	2.0

(a)

1.0	1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0	4.0
5.0	5.0	5.0	5.0	5.0

(b)

1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0
1.0	2.0	3.0	4.0	5.0

(c)

1.0	4.0	5.0	0.0	1.5
4.0	7.0	8.0	3.0	4.5
3.0	6.0	7.0	2.0	3.5
5.0	8.0	9.0	4.0	5.5
2.0	5.0	6.0	1.0	7.5

(d)

1.0	0.5	2.0	0.2	0.8
2.0	1.0	4.0	0.4	1.6
3.0	1.5	6.0	0.6	2.4
4.0	2.0	8.0	0.8	3.2
5.0	2.5	10	1.0	4.0

(e)

a) Συσταδοποίηση 2-τρόπων σταθερής τιμής b) συσταδοποίηση 2-τρόπων σταθερής σειράς c) συσταδοποίηση 2-τρόπων σταθερής στήλης d) συσταδοποίηση 2-τρόπων συνεκτικής τιμής (προσθετική) e) συσταδοποίηση 2-τρόπων συνεκτικής τιμής (μοντέλο)

2.4.4 Αλγόριθμοι συσταδοποίησης 2-τρόπων

Το εύρος των αλγορίθμων αυτής της κατηγορίας που χρησιμοποιούνται σε εφαρμογές της βιοπληροφορικής είναι αρκετά μεγάλο. Μερικά παραδείγματα τέτοιων αλγορίθμων είναι οι ακόλουθοι: ο αλγόριθμος μπλοκ συσταδοποίησης, ο CTWC, ITWC, ο δ-bicluster, d-pCluster, d-pattern, ο FLOC, ο OPC, ο Plaid Model, ο OPSMs, ο Gibbs, ο SAMBA, ο RoBA (Robust Biclustering Algorithm), ο Crossing Minimization, ο cMonkey , ο PRMs και ο DCC.

Ένα μεγάλο πρόβλημα που προκύπτει με όλους τους αλγορίθμους που ανήκουν σε αυτή την κατηγορία είναι ο βαθμός αξιοπιστίας των αποτελεσμάτων που δίνουν καθώς η συσταδοποίηση 2-τρόπων επιτρέπει την επικάλυψη μεταξύ διαφορετικών συστάδων. Το γεγονός ότι κάποιοι από αυτούς τους αλγορίθμους δεν είναι ντετερμινιστικοί και συνάμα ότι το πρόβλημα εμπεριέχει μη επιβλεπόμενη κατηγοριοποίηση, δυσκολεύουν τον εντοπισμό σφαλμάτων στα αποτελέσματα. Μια προσέγγιση για την επίλυση του συγκεκριμένου προβλήματος είναι η χρήση πολλών αλγορίθμων συσταδοποίησης 2-τρόπων από τους οποίους, η πλειονότητα θα ψηφίζει για να αποφασιστεί ποιο είναι το καλύτερο αποτέλεσμα συσταδοποίησης.

Παρόλα αυτά οι συγκεκριμένοι αλγόριθμοι βρίσκουν αρκετά καλή εφαρμογή στο χώρο της βιολογίας και γενικότερα της βιοπληροφορικής.

2.5 Φασματική συσταδοποίηση

2.5.1 Γενικά

Τα τελευταία χρόνια η φασματική συσταδοποίηση έγινε μία από τις πιο δημοφιλείς και μοντέρνες τεχνικές συσταδοποίησης. Η υλοποίησή της θεωρείται αρκετά εύκολη με τη χρήση ενός λογισμικού γραμμικής άλγεβρας, ενώ αρκετές φορές τα αποτελέσματα που δίνει υπερέχουν αυτών που δίνουν αλγόριθμοι όπως ο k-means.

2.5.2 Αρχική οργάνωση των δεδομένων

Οι αλγόριθμοι φασματικής συσταδοποίησης αντιμετωπίζουν τη συσταδοποίηση ενός συνόλου δεδομένων $X = \{\chi_1,\chi_2,\ldots,\chi_n\}$, $\chi_i \in \Re^d$, το οποίο συνοδεύεται και από κάποια λογική ομοιότητας $s_{ij} \geq 0$, ως μία διαίρεση σε ικανό αριθμό ομάδων, στο εσωτερικό των οποίων παρουσιάζεται μεγάλη ομοιότητα μεταξύ των εντεθειμένων δεδομένων. Αν η μόνη πληροφορία που είναι διαθέσιμη αφορά τις ομοιότητες που μπορεί να εμφανίζουν τα δεδομένα μεταξύ τους τότε ένας καλός τρόπος αναπαράστασης αυτών είναι με τη μορφή ενός γράφου ομοιότητας $G = (V, E)$. Κάθε κορυφή v_i του γράφου αντιστοιχίζεται με κάποιο από τα δεδομένα χ_i. Δύο κορυφές συνδέονται μεταξύ τους αν η ομοιότητα s_{ij} μεταξύ των αντίστοιχων σημείων χ_i και χ_j είναι θετική (ή μεγαλύτερη από ένα συγκεκριμένο κατώφλι) με την ακμή που συνδέει τα δύο σημεία να βαρύνεται με s_{ij}. Έτσι βάσει των ομοιοτήτων s_{ij} μπορεί να κατασκευαστεί ο πίνακας ομοιότητας S. Είναι προφανές τώρα ότι το πρόβλημα της συσταδοποίησης μπορεί να αλλάξει μορφή και να βασιστεί στη χρήση του γράφου ομοιότητας. Στην περίπτωση του γράφου ομοιότητας επιθυμούμε να βρούμε μία διαμέριση του γράφου τέτοια ώστε οι ακμές μεταξύ των διαφορετικών ομάδων να έχουν πολύ μικρό βάρος (δηλαδή τα σημεία που ανήκουν σε διαφορετικές ομάδες είναι ανόμοια μεταξύ τους) και οι ακμές εντός μιας ομάδας να έχουν μεγάλο βάρος (δηλαδή τα σημεία που βρίσκονται εντός της ίδιας συστάδας είναι όμοια μεταξύ τους).

2.5.3 Γράφοι ομοιότητας

2.5.3.1 Γενικά

Υπάρχουν πολλές δημοφιλείς κατασκευές που μετασχηματίζουν ένα δοθέν σύνολο σημείων $X = \{\chi_1,\chi_2,\ldots,\chi_n\}$, $\chi_i \in \Re^d$ με ομοιότητες και αποστάσεις ανά ζεύγη s_{ij} και d_{ij} αντίστοιχα, σε ένα γράφο. Όταν κατασκευάζουμε γράφους ομοιότητας στοχεύουμε στη μοντελοποίηση των σχέσεων μεταξύ των σημείων όπως αυτές εμφανίζονται σε μια τοπική γειτονιά. Επιπλέον, οι περισσότερες από τις κατασκευές που θα παρουσιαστούν παρακάτω οδηγούν σε μια αραιή αναπαράσταση των δεδομένων κάτι που είναι αρκετά ευνοϊκό κατά τη διενέργεια υπολογισμών, αφού γίνονται σαφώς πιο γρήγορα.

2.5.3.2 Είδη γράφων ομοιότητας

2.5.3.2.1 Ο γράφος της ε-γειτονιάς

Σε έναν γράφο ε-γειτονιάς συνδέουμε όλα εκείνα τα σημεία των οποίων οι αποστάσεις ανά ζεύγη σημείων είναι μικρότερες του ε. Καθώς οι αποστάσεις μεταξύ όλων των συνδεδεμένων σημείων είναι το πολύ κλίμακας ε, η τοποθέτηση βαρών στις ακμές του γράφου δε θα προσθέσει επιπλέον πληροφορία για τα δεδομένα πάνω στο γράφο. Παρόλα ταύτα αν μας δοθούν οι ανά ζεύγη αποστάσεις των δεδομένων μπορούμε να μετατρέψουμε αυτές τις αποστάσεις σε ομοιότητες και να χρησιμοποιήσουμε τις ομοιότητες ως βάρη. Αυτός είναι και ο λόγος που ο γράφος της e-γειτονιάς συνήθως θεωρείται ως γράφος χωρίς βάρη.

2.5.3.2.2 Ο γράφος των k-κοντινότερων γειτόνων

Σε αυτούς τους γράφους μια κορυφή v_i συνδέεται με μια άλλη κορυφή v_j αν η v_j είναι μεταξύ των k κοντινότερων γειτόνων της v_i. Ωστόσο, ένας τέτοιος ορισμός οδηγεί σε έναν κατευθυνόμενο γράφο, καθώς η σχέση που περιγράφει την έννοια της

γειτονιάς δεν είναι συμμετρική. Παρόλα ταύτα υπάρχουν δύο τρόποι για την μετατροπή αυτού του κατευθυνόμενου γράφου σε μη κατευθυνόμενο. Ο πρώτος τρόπος επιτυγχάνεται με το να αγνοήσουμε απλά τις κατευθύνσεις των ακμών και να συνδέσουμε μια κορυφή v_i και μια κορυφή v_j με μια μη κατευθυνόμενη ακμή αν η v_i είναι μεταξύ των k κοντινότερων γειτόνων της v_j *ή* αν η v_j είναι μεταξύ των k κοντινότερων γειτόνων της v_i. Ο γράφος που προκύπτει συνήθως αποκαλείται *γράφος των k κοντινότερων γειτόνων ή συμμετρικός γράφος των k κοντινότερων γειτόνων*. Ο δεύτερος τρόπος είναι να συνδέσουμε τις κορυφές v_i και v_j αν η v_i είναι μεταξύ των k κοντινότερων γειτόνων της v_j *και* αν η v_j είναι μεταξύ των k κοντινότερων γειτόνων της v_i. Ο γράφος που προκύπτει ονομάζεται *αμοιβαίος γράφος των k κοντινότερων γειτόνων*. Και στις δύο περιπτώσεις, αφού γίνει η σύνδεση των κατάλληλων κορυφών έπειτα βαραίνουμε τις ακμές με την ομοιότητα των γειτονικών σημείων.

2.5.3.2.3 Ο πλήρως συνδεδεμένος γράφος

Εδώ δημιουργούμε συνδέσεις μεταξύ όλων των σημείων που έχουν θετική ομοιότητα και βαραίνουμε όλες τις ακμές με s_{ij}. Καθώς ο γράφος θα πρέπει να μοντελοποιεί τις σχέσεις που υπάρχουν μέσα στις τοπικές γειτονιές, η κατασκευή αυτή συνήθως επιλέγεται μόνο αν η συνάρτηση ομοιότητας κωδικοποιεί ήδη από μόνη της τις τοπικές γειτονιές. Ένα παράδειγμα μιας τέτοιας συνάρτησης ομοιότητας είναι η περίπτωση της Γκαουσιανής συνάρτησης ομοιότητας: $s(\chi_i, \chi_j) = \exp\left(-\frac{\|\chi_i - \chi_j\|^2}{2\sigma^2}\right)$. Εδώ η παράμετρος s καθορίζει το πλάτος των γειτονιών όπως κάνει και η παράμετρος ε στην περίπτωση των γράφων της ε-γειτονιάς.

Να σημειωθεί ότι στο κεφάλαιο των πειραματικών αποτελεσμάτων γίνεται απεικόνιση των προηγούμενων γράφων ομοιότητας για ένα σύνολο δεδομένων.

2.5.4 Πίνακες των αλγορίθμων φασματικής συσταδοποίησης

2.5.4.1 Γενικά

Οι περισσότεροι αλγόριθμοι φασματικής συσταδοποίησης χρησιμοποιούν κάποια είδη πινάκων που αποτελούν ένα κοινό στοιχείο γι' αυτούς. Ένας από αυτούς, ο πίνακας ομοιότητας, περιγράφηκε ήδη, ενώ ο πίνακας γειτνίασης, ο πίνακας βαθμών και ο πίνακας Laplace αποτελούν τις άλλες τρεις πιο συνηθισμένες μορφές πινάκων για την εκτέλεση της φασματικής συσταδοποίησης.

2.5.4.2 Πίνακας γειτνίασης

Ο συγκεκριμένος πίνακας συνήθως κατασκευάζεται στο ίδιο βήμα με τον πίνακα ομοιότητας και αρκετές φορές στη βιβλιογραφία ταυτίζονται. Θεωρούμε ότι κάθε ακμή του γράφου G βαρύνεται με κάποια τιμή $w_{ij} \geq 0$. Ο πίνακας βαρών W του γράφου που κατασκευάζεται βασιζόμενος στον πίνακα ομοιότητας S ονομάζεται και *πίνακας γειτνίασης (adjacency matrix)* και έχει ως στοιχεία του τα μη αρνητικά βάρη w_{ij} με $i, j = 1,...,n$. Αν για κάποιο από τα βάρη ισχύει $w_{ij} = 0$ τότε οι δύο κορυφές δε συνδέονται μεταξύ τους και δεδομένου ότι ο γράφος G είναι μη κατευθυνόμενος ισχύει ότι: $w_{ij} = w_{ji}$.

2.5.4.3 Πίνακας βαθμών

Για τον βαθμό d_i μιας κορυφής $v_i \in V$ έχουμε ότι: $d_i = \sum_{j=1}^{n} w_{ij}$. Ο πίνακας βαθμών D ορίζεται ως ο διαγώνιος πίνακας που έχει ως στοιχεία διαγωνίου τους βαθμούς $d_1,...,d_n$.

2.5.4.4 Πίνακες Laplace

2.5.4.4.1 Γενικές Θεωρήσεις

Θεωρούμε πάντα τον ίδιο γράφο G της υπόθεσης ο οποίος είναι μη κατευθυνόμενος και βεβαρημένος με πίνακα βαρών W για τα στοιχεία του οποίου ισχύει $w_{ij} = w_{ji} \geq 0.$ Επίσης όταν αναφερόμαστε σε ιδιοδιανύσματα ενός πίνακα δε θεωρούμε απαραίτητα ότι είναι κανονικοποιημένα στη μονάδα. Για παράδειγμα, το σταθερό διάνυσμα **1** και το διάνυσμα $a\mathbf{1}$ για κάποιο $a\neq 0$, θεωρούνται δύο ίδια διανύσματα.

2.5.4.4.2 Μη κανονικοποιημένος πίνακας Laplace

Ο μη κανονικοποιημένος πίνακας Laplace L ενός γράφου ορίζεται ως $L = D - W$. Μερικές από τις πιο σημαντικές ιδιότητες του μη κανονικοποιημένου πίνακα Laplace, που θεωρούνται χρήσιμες στη φασματική συσταδοποίηση και μπορούν να περιγράψουν πολλές από τις ιδιότητες των γράφων, είναι οι ακόλουθες [2]:

1. *Για κάθε διάνυσμα* $f \in \Re^n$ *ισχύει:* $f^{\wedge} Lf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(f_i - f_j\right)^2$.
2. *Είναι συμμετρικός και θετικά ημι-ορισμένος.*
3. *Η μικρότερη ιδιοτιμή του είναι το 0 στην οποία αντιστοιχεί το ιδιοδιάνυσμα* ***1.***
4. *Ο L έχει n μη αρνητικές, πραγματικές ιδιοτιμές:* $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.

Για ένα μη κατευθυνόμενο γράφο G που δεν έχει αρνητικά βάρη, η πολλαπλότητα k της ιδιοτιμής 0 του πίνακα Laplace του ισούται με τον αριθμό των συνδεδεμένων στοιχείων $A_1,..., A_k$ *που εντοπίζονται μέσα σε αυτό το γράφο .Ο ιδιοχώρος της ιδιοτιμής 0 επικαλύπτεται από τα διανύσματα δείκτες* $1_{A_1},...,1_{A_k}$ *των στοιχείων* $A_1,..., A_k$.

2.5.4.4.3 Κανονικοποιημένοι πίνακες Laplace

Υπάρχουν δύο πίνακες που ονομάζονται κανονικοποιημένοι πίνακες Laplace και εμφανίζονται συχνά στη βιβλιογραφία. Οι δύο αυτές μορφές είναι οι ακόλουθες:

$$L_{sym} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$L_{rw} := D^{-1} L = I - D^{-1} W .$$

Ο πρώτος πίνακας συμβολίζεται με L_{sym} καθώς ο πίνακας είναι συμμετρικός, ενώ ο δεύτερος συμβολίζεται με L_{rw} επειδή συνδέεται άμεσα με έναν τυχαίο περίπατο.

Οι κανονικοποιημένοι πίνακες Laplace ικανοποιούν τις ακόλουθες ιδιότητες [2]:

1. *Για κάθε* $f \in \Re^n$ *ισχύει:* $f' L_{sym} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 .$
2. *Το* λ *αποτελεί ιδιοτιμή του πίνακα* L_{rw} *που αντιστοιχεί σε ένα ιδιοδιάνυσμα* v *αν και μόνο αν η* λ *είναι μια ιδιοτιμή του πίνακα* L_{sym} *που αντιστοιχεί σε ένα ιδιοδιάνυσμα* $w = D^{\frac{1}{2}} v$.
3. *Το* λ *αποτελεί ιδιοτιμή του* L_{rw} *που αντιστοιχεί σε ένα ιδιοδιάνυσμα* v *αν και μόνο αν η* λ *και το* v *επιλύουν το γενικευμένο ιδιοπρόβλημα* $Lv = \lambda Dv$.
4. *Το 0 αποτελεί ιδιοτιμή του* L_{rw} *που αντιστοιχεί στο μοναδιαίο σταθερό ιδιοδιάνυσμα* ***1****. Το 0 αποτελεί ιδιοτιμή του* L_{sym} *που αντιστοιχεί στο ιδιοδιάνυσμα* $D^{\frac{1}{2}}$ ***1****.*

5. *Οι πίνακες* L_{rw} *και* L_{sym} *είναι θετικά ημι-ορισμένοι και έχουν n μη αρνητικές πραγματικές ιδιοτιμές* $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.

Για έναν μη κατευθυνόμενο γράφο G που δεν έχει αρνητικά βάρη, η πολλαπλότητα k της ιδιοτιμής 0 τόσο του πίνακα L_{rw} *όσο και του πίνακα* L_{sym} *ισούται με τον αριθμό των συνδεδεμένων στοιχείων* $A_1,...,A_k$ *που εντοπίζονται μέσα στο γράφο. Για τον πίνακα* L_{rw}*, ο ιδιοχώρος του 0 επικαλύπτεται από τα διανύσματα δείκτες* $1_{A_1},...,1_{A_k}$ *των στοιχείων* $A_1,...,A_k$*. Για τον πίνακα* L_{sym}*, ο ιδιοχώρος του 0 επικαλύπτεται από τα διανύσματα* $D^{\frac{1}{2}}1_{A_i}$.

Κεφάλαιο 3

Αλγόριθμος F ασματικής Συσταδοποίησης των Ng, Jordan και Weiss

3.1 Γενικά

Ο Αλγόριθμος φασματικής συσταδοποίησης που πρότειναν οι Ng, Jordan και Weiss πρόκειται για έναν από τους πιο γνωστούς αλγορίθμους φασματικής συσταδοποίησης. Αρκετοί ερευνητές βασίστηκαν στο συγκεκριμένο αλγόριθμο ώστε να παρέχουν περαιτέρω βελτιώσεις του. Οι βελτιώσεις αφορούσαν τον αυτόματο υπολογισμό του πλήθους των συστάδων για ένα σύνολο δεδομένων καθώς και των παραμέτρων τοπικής ή καθολικής κλιμάκωσης. Για τις συγκεκριμένες παραμέτρους γίνεται αναφορά σε επόμενες υποενότητες.

3.2 Περιγραφή του αλγορίθμου

3.2.1 Γενικά

Σε αυτή την ενότητα θα παρουσιαστούν ο αλγόριθμος NJW και κάποια στοιχεία που τον χαρακτηρίζουν και σχετίζονται κυρίως με τους πίνακες και την καθολική παράμετρο κλιμάκωσης που χρησιμοποιεί κατά την εκτέλεσή του.

3.2.2 Ο αλγόριθμος των NJW

Αλγόριθμος των NJW

Είσοδος: X // πίνακας δεδομένων εισόδου μεγέθους N×d

(N είναι το πλήθος των σημείων διάστασης d)

c // το πλήθος των επιθυμητών συστάδων

σ // παράμετρος που εκφράζει το πόσο γρήγορα «πέφτει» η απόσταση στη θέση A_{ij} του πίνακα A. Για τη παράμετρο σ και για τον πίνακα A θα γίνει εκτενής αναφορά πιο κάτω.

Έξοδος: Ομαδοποίηση των N σημείων σε c συστάδες

Βήμα 1: Κατασκευή του πίνακα A μεγέθους N×N τα στοιχεία του οποίου ορίζονται με τον ακόλουθο τρόπο: $A_{ij} = \exp(-\frac{\|\chi_i - \chi_j\|^2}{2\sigma^2})$ για i ≠ j και $A_{ij} = 0$ για i = j.

Βήμα 2: Ορισμός και κατασκευή του διαγώνιου πίνακα D μεγέθους N×N με στοιχεία διαγωνίου D_{ii} της μορφής: $D_{ii} = \sum_{j=1}^{n} A_{ij}$. Κατασκευή του κανονικοποιημένου πίνακα γειτνίασης L ως εξής: $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$.

Βήμα 3: Υπολογισμός των c μεγαλύτερων ιδιοδιανυσμάτων: $v_1, v_2, \ldots, v_c$ του πίνακα L και κατασκευή του πίνακα $V = [v_1\ v_2 \ldots v_c]$ που έχει μέγεθος N×c.

Βήμα 4: Κανονικοποίηση των γραμμών του πίνακα V ώστε να προκύψει ο πίνακας Y του οποίου οι γραμμές έχουν μοναδιαίο μήκος. Πιο συγκεκριμένα τα στοιχεία πίνακα Y υπολογίζονται ως εξής: $Y_{ij} = \frac{V_{ij}}{\left(\sum_j V_{ij}^2\right)^{\frac{1}{2}}}$.

Βήμα 5: Θεώρηση της κάθε γραμμής του Y ως ένα σημείο στο χώρο $\Re^c$ και εφαρμογή του αλγορίθμου k-means πάνω σε αυτά τα δεδομένα των N γραμμών.

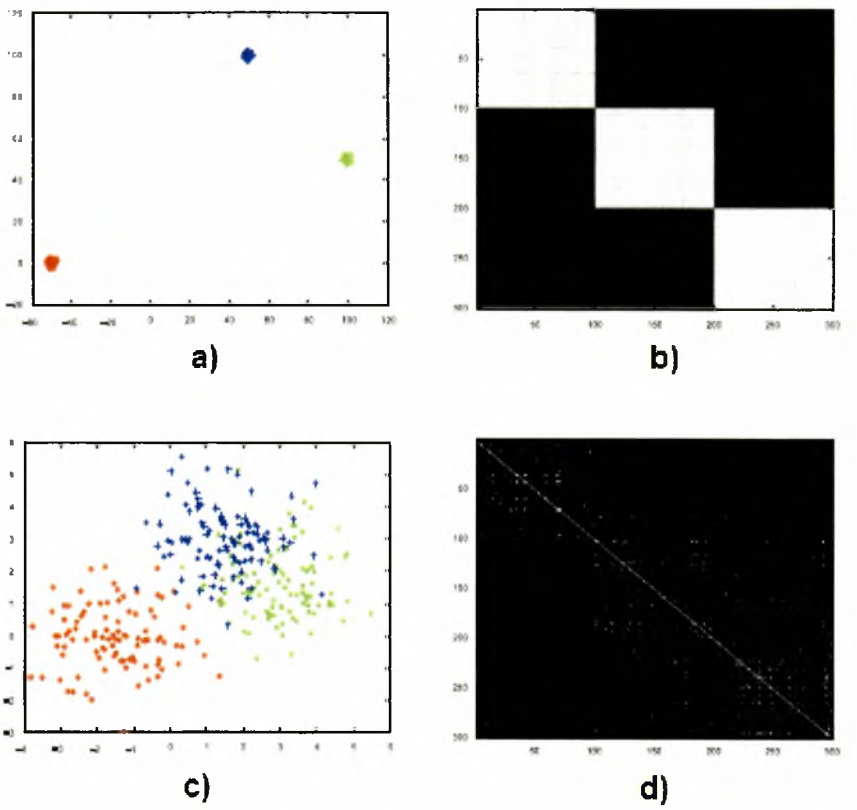
Βήμα 6: Το αρχικό σημείο χ_i θα ανήκει στη συστάδα c αν και μόνο αν η γραμμή i του πίνακα Y ανατέθηκε στη συστάδα c.

3.2.3 Στοιχεία του αλγορίθμου

3.2.3.1 Ο πίνακας γειτνίασης A

Ο πίνακας A του αλγορίθμου ουσιαστικά προκύπτει από την έννοια του γράφου ομοιότητας (similarity graph). Όταν δημιουργούμε ένα γράφο ομοιότητας προσπαθούμε ουσιαστικά να μοντελοποιήσουμε τις τοπικές σχέσεις των στοιχείων που ανήκουν σε μια συγκεκριμένη γειτονιά. Στην περίπτωση του πίνακα A και δεδομένου ενός συνόλου σημείων $\chi_1, \chi_2, ..., \chi_n$ χρησιμοποιούμε τη συνάρτηση ομοιότητας $s(\chi_i, \chi_j) = \exp\left(-\frac{\|\chi_i - \chi_j\|^2}{2\sigma^2}\right)$ για $i \neq j$ και φυσικά όπου αναφερόμαστε για ένα σημείο με τον εαυτό του $(i = j)$ θεωρούμε $s(\chi_i, \chi_j) = 0$. Η παράμετρος σ καθορίζει το πλάτος που θα έχει η κάθε γειτονιά σημείων. Συνεπώς, έτσι καταφέρνουμε να συνδέσουμε ένα σημείο με όλα τα υπόλοιπα και να αναθέσουμε στην ακμή (i, j) του γράφου το βάρος $s_{ij} = s(\chi_i, \chi_j)$. Εδώ ταυτίζουμε το βάρος και την ομοιότητα αν και θα μπορούσαμε να θεωρήσουμε ότι ομοιότητα για δύο σημεία αποτελεί η τετραγωνική Ευκλείδεια απόσταση που υπολογίζεται μεταξύ αυτών. Όσο πιο μεγάλη είναι αυτή η απόσταση τόσο πιο ανόμοια θεωρούνται τα δύο αυτά σημεία μεταξύ τους.

Ο πίνακας γειτνίασης που θα δημιουργηθεί επιθυμούμε να είναι όσο το δυνατόν πιο κοντά στη μορφή ενός διαγωνίου μπλοκ πίνακα (block-diagonal matrix) γιατί τότε η φασματική συσταδοποίηση παρέχει τα πιο καλά αποτελέσματα. Η μορφή όμως ενός διαγωνίου μπλοκ πίνακα εμφανίζεται μόνο όταν οι συστάδες που εμπεριέχουν τα δεδομένα εισόδου είναι καλά σχηματισμένες και διαχωρισμένες εξαρχής (*Εικόνα 3.2.3.1)*. Κάτι τέτοιο όμως είναι αρκετά σπάνιο στην πράξη για την πλειονότητα των δεδομένων που θα εισαχθούν σε έναν αλγόριθμο συσταδοποίησης, και πόσο μάλλον για δεδομένα που αφορούν τα pixel μιας εικόνας, που ως γνωστόν μπορεί να παρουσιάζουν μια έντονη ποικιλομορφία εντός αυτής.



Εικόνα 3.2.3.1

a) Ένα καλά διαχωρισμένο σύνολο δεδομένων στον δισδιάστατο χώρο που αποτελείται από τρεις ευδιάκριτες συστάδες, b) Ο πίνακας γειτνίασης για το συγκεκριμένο σύνολο δεδομένων, όπου όπως μπορεί εύκολα να παρατηρηθεί έχει τη καθαρή μορφή ενός διαγωνίου μπλοκ πίνακα υποδεικνύοντας ταυτόχρονα ότι και ο αριθμός των συστάδων είναι πράγματι τρεις, c) Ένα σύνολο δεδομένων όπου ο αριθμός των συστάδων δεν είναι ευδιάκριτος καθώς δεν είναι καλώς διαχωρισμένες και σχηματισμένες, d) Ο πίνακας γειτνίασης για το σύνολο δεδομένων της (c). Είναι προφανές ότι ο συγκεκριμένος πίνακας δεν είναι μπλοκ διαγώνιος λόγω της φύσης των δεδομένων εισόδου.

3.2.3.2 Ο κανονικοποιημένος πίνακας γειτνίασης L

Ο πίνακας L ουσιαστικά δεν πρόκειται για τον πίνακα Laplace αλλά για έναν πίνακα που έχει ακριβώς τα ίδια ιδιοδιανύσματα με τον πίνακα Laplace (I – L) αλλά διαφορετικές ιδιοτιμές. Συγκεκριμένα ο πίνακας Laplace έχει ιδιοτιμές της μορφής

$1-\lambda_i$ ενώ ο πίνακας L έχει ιδιοτιμές λ_i. Ο πίνακας αυτός ονομάζεται *κανονικοποιημένος ή σταθμισμένος πίνακας γειτνίασης (scaled adjacency matrix).*

3.2.3.3 Ο υπολογισμός των ιδιοδιανυσμάτων του πίνακα L

Για την υλοποίηση της φασματικής συσταδοποίησης πρέπει να υπολογιστούν τα c μεγαλύτερα ιδιοδιανύσματα του πίνακα L. Τα c μεγαλύτερα ιδιοδιανύσματα είναι αυτά που δίνουν πληροφορία σχετικά με το πως μπορούν να διαχωριστούν τα δεδομένα εισόδου μεταξύ τους. Το μέγεθος του πίνακα L είναι τις περισσότερες φορές σχετικά μεγάλο, γεγονός το οποίο, καθιστά τους υπολογισμούς δύσκολους και χρονοβόρους. Παρόλα αυτά, αν χρησιμοποιηθεί *ο γράφος του k πλησιέστερου γείτονα* ή *ο γράφος της ε-γειτονιάς* τότε ο πίνακας L που προκύπτει είναι αραιός. Επιπλέον έχουν ήδη αναπτυχθεί ικανοποιητικές μέθοδοι για τον υπολογισμό των μεγαλύτερων ιδιοδιανυσμάτων ενός αραιού πίνακα, με τις πιο αντιπροσωπευτικές και πιο γνωστές να είναι αυτή της εκθετικής μεθόδου και οι μέθοδοι του υποχώρου Krylov (Krylov subspace methods) όπως για παράδειγμα η μέθοδος του Lanczos (μια μέθοδος που έχει υιοθετηθεί και από την εφαρμογή Matlab μέσω της συνάρτησης *eigs* που την χρησιμοποιεί). Η ταχύτητα σύγκλισης αυτών των αλγορίθμων εξαρτάται από το μέγεθος του *«φασματικού ιδιοκενού»* όπως συνηθίζεται να αποκαλείται και ορίζεται ως: $\gamma_k = |\lambda_k - \lambda_{k+1}|$. Όσο μεγαλύτερο είναι το φασματικό ιδιοκενό τόσο πιο γρήγορα συγκλίνουν οι αλγόριθμοι κατά τον υπολογισμό των c μεγαλύτερων ιδιοδιανυσμάτων. Παρατηρώντας την πολλαπλότητα που μπορεί να έχουν οι ιδιοτιμές του πίνακα μπορούμε να πούμε ότι προκύπτει ένα πρόβλημα στην περίπτωση που κάποια από τις ιδιοτιμές παρουσιάζει πολλαπλότητα μεγαλύτερη του ένα. Για να κατανοηθεί καλύτερα το πρόβλημα που εμφανίζεται, θεωρούμε την ιδανική περίπτωση των k ασύνδετων συστάδων όπου η ιδιοτιμή 0 έχει πολλαπλότητα k. Σε αυτή την περίπτωση ο ιδιοχώρος καλύπτεται από k διανύσματα που υποδεικνύουν τις συστάδες και στα οποία δεν συγκλίνουν κατ' ανάγκη οι αλγόριθμοι εύρεσης ιδιοδιανυσμάτων και ιδιοτιμών. Στην πραγματικότητα επιτυγχάνουν σύγκλιση σε κάποια ορθοκανονική βάση του ιδιοχώρου που συνήθως εξαρτάται από τις λεπτομέρειες υλοποίησης του αλγορίθμου σύγκλισης. Επιπλέον όλα τα διανύσματα που εντοπίζονται στο χώρο,

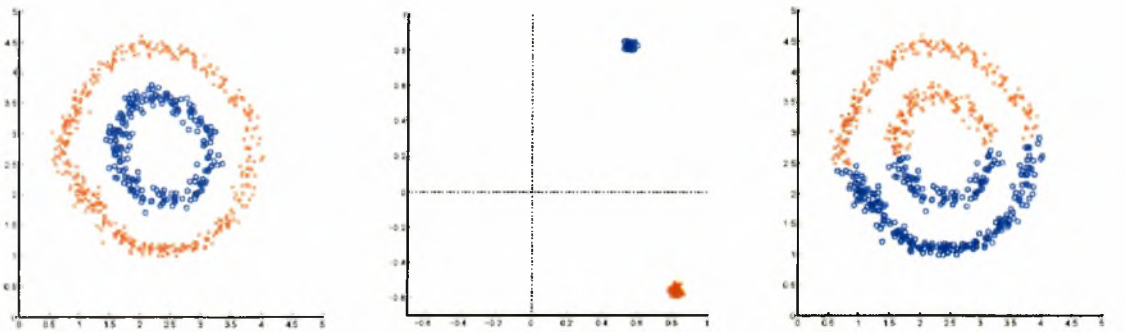
επικαλύπτονται από τα διανύσματα 1_{A_i} που υποδεικνύουν τις συστάδες, έχουν την μορφή: $\upsilon = \sum_{i=1}^{k} a_i 1_{A_i}$, για κάποιους συντελεστές α_i , τα οποία είναι τμηματικά σταθερά εντός των συστάδων. Συνεπώς τα διανύσματα που επιστρέφουν οι αλγόριθμοι εμπεριέχουν την πληροφορία για τις συστάδες η οποία μπορεί έπειτα να χρησιμοποιηθεί από τον k-means αλγόριθμο για την ανακατασκευή των συστάδων.

3.2.3.4 Η παράμετρος σ

3.2.3.4.1 Καθολική κλιμάκωση

Στον συγκεκριμένο αλγόριθμο η παράμετρος κλιμάκωσης *s* ορίζεται ως καθολική για το εκάστοτε σύνολο δεδομένων. Η τιμή της παίζει ιδιαίτερο ρόλο για την εξαγωγή καλών τελικών συστάδων αφού ενισχύει τα σημεία που παρουσιάζουν μεγάλη ομοιότητα ενώ η τιμή της κάθε φορά εξαρτάται από την εφαρμογή που καλούμαστε να μοντελοποιήσουμε. Πιο συγκεκριμένα επαναλαμβάνουμε το πείραμα όσες φορές χρειαστεί ώστε να πάρουμε μια καλή εκτίμηση της παραμέτρου. Είναι προφανές ότι κάτι τέτοιο αφενός μεν αυξάνει τον υπολογιστικό χρόνο που απαιτείται αφετέρου de μια καθολική τιμή της συγκεκριμένης παραμέτρου, όσο καλή και αν μπορεί να θεωρηθεί, κάποιες φορές αποτυγχάνει πάνω σε δεδομένα που έχουν σχετικά μεγάλη μεταβλητότητα. Σε αυτή την περίπτωση μπορούν να οριστούν τοπικές παράμετροι κλιμάκωσης για κάθε γειτονιά στοιχείων που υπάρχει στο σύνολο των δεδομένων. Έχουν προταθεί αρκετές σχετικά μέθοδοι για το πώς μπορούμε να υπολογίσουμε αυτές τις παραμέτρους. Μερικές από τις πιο γνωστές θα παρουσιαστούν πιο κάτω.

Στο παρακάτω σχήμα δίνουμε μια εξήγηση για το ότι πολλές φορές συμβατικοί αλγόριθμοι, όπως ο k-means, δε δίνουν αποδοτικά αποτελέσματα για κάποιες περιπτώσεις δεδομένων εισόδου.



α) *β)* *γ)*

α) *Δεδομένα εισόδου υπό την μορφή 2 κύκλων. Προφανώς ο αριθμός των ομάδων (συστάδων) που θα πρέπει να δημιουργηθούν είναι 2. Η μία ομάδα θα περιλαμβάνει τα σημεία του εξωτερικού κύκλου και η άλλη αυτά του εσωτερικού.* ***β)*** *Επιτυχής συσταδοποίηση που πετυχαίνεται με φασματική συσταδοποίηση μέσω του αλγορίθμου NJW.* ***γ)*** *Μη επιτυχής συσταδοποίηση που προέκυψε ύστερα από την εφαρμογή του αλγορίθμου k-means.*

3.2.3.4.2 Τοπική κλιμάκωση

Παρόλο που ο αλγόριθμος φασματικής συσταδοποίησης των Ng-Jordan-Weiss θεωρεί μια καθολική τιμή για την παράμετρο κλιμάκωσης s, εντούτοις μια τέτοια θεώρηση δε δίνει πάντα τα επιθυμητά αποτελέσματα σε ένα σύνολο δεδομένων με πολλαπλές κλιμακώσεις. Για να λύσουμε λοιπόν το συγκεκριμένο πρόβλημα χρησιμοποιούμε τις παραμέτρους τοπικής κλιμάκωσης. Στη συνέχεια παρουσιάζονται τρεις τρόποι υπολογισμού των τοπικών παραμέτρων κλιμάκωσης.

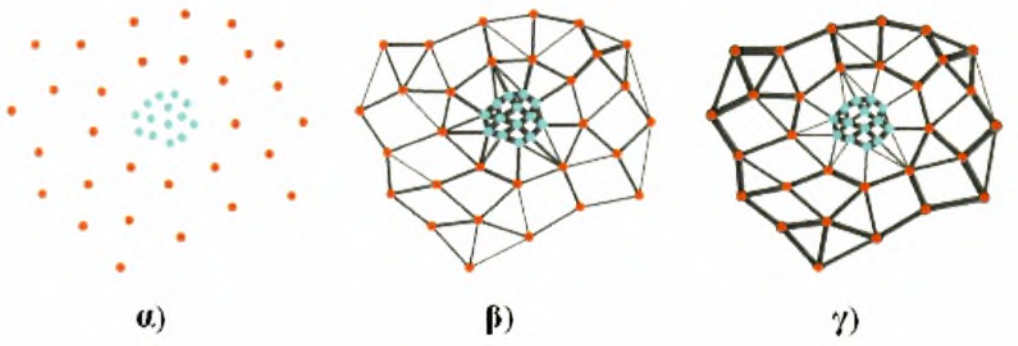
a. Ένας απλός τρόπος υπολογισμού μπορεί να θεωρηθεί ο ακόλουθος:
   Ξεκινώντας το συλλογισμό μας μπορούμε να πούμε ότι η απόσταση του χ_i από το χ_j όπως αυτή «φαίνεται» από το σημείο χ_i, είναι $\frac{d(\chi_i,\chi_j)}{\sigma_i}$ και συμμετρικά ισχύει ότι $\frac{d(\chi_j,\chi_i)}{\sigma_j}$. Τότε για την τετραγωνική απόσταση d^2

προκύπτει $\frac{d^2(\chi_i,\chi_j)}{\sigma_i\sigma_j}$. Συνεπώς τώρα προκύπτει ένας νέος πίνακας $\hat{A}$ κάθε στοιχείο του οποίου παίρνει την μορφή $\hat{A}_{ij} = \exp\left(-\frac{d^2(\chi_i,\chi_j)}{\sigma_i\sigma_j}\right)$ για $i \neq j$ και $\hat{A}_{ii} = 0$. Η τιμή του σ_i καθορίζεται από την ακόλουθη σχέση $\sigma_i = d(\chi_i,\chi_k)$, όπου χ_k ο k-οστός γείτονας του σημείου χ_i. Η τιμή του k εξαρτάται αποκλειστικά από την διάσταση των δεδομένων και τις περισσότερες φορές θεωρείται ίση με 7 [4]. Παρόλα αυτά μπορούμε να θεωρήσουμε ότι k = 1 + 2×D, όπου D είναι η διάσταση των δεδομένων μας. Η συγκεκριμένη φόρμουλα μπορεί να δικαιολογηθεί ως εξής: κάθε σημείο είναι γείτονας με τον εαυτό του και ταυτόχρονα θεωρούμε ότι έχει και 2 γείτονες για κάθε διάσταση. Έτσι για παράδειγμα αν θεωρήσουμε ότι D = 3 στον RGB χώρο τότε προκύπτει ότι k = 1 + 2×3 = 7.

b. Θεωρώντας ότι κάθε στοιχείο $\tilde{A}_{ij}$ ενός πίνακα $\tilde{A}$, έχει την εξής τιμή $\tilde{A}_{ij} = \exp\left(-\frac{\|\chi_i - \chi_j\|^2}{2\sigma_i^2}\right)$, εισάγουμε την ακόλουθη συνθήκη για να προσδιορίσουμε την τιμή του σ_i: $\sum_{j=1}^{n} \exp\left(-\frac{\|\chi_i - \chi_j\|^2}{2\sigma_i^2}\right) = \tau$, για κάθε σειρά i, όπου 1= i = n και τ είναι μια θετική σταθερά. Με αυτό τον τρόπο επιλέγουμε το σ_i ούτως ώστε το άθροισμα της κάθε σειράς του $\tilde{A}$ να ισούται με μια δεδομένη τιμή του τ. Το τ μπορεί να ερμηνευτεί ως ένα καθορισμένο *μέγεθος γειτονιάς*. Η επιλογή του τ είναι πιο εύκολη από αυτή του σ_i, αφού το τ είναι ανεξάρτητο κλιμάκωσης. Ο προσδιορισμός της τιμής της πρέπει να γίνει με τέτοιο τρόπο ώστε να «συνδέονται» μόνο τα άμεσα γειτονικά σημεία των δεδομένων εισόδου. Το τ μπορεί να επιλεχθεί πάλι με τον ίδιο κανόνα όπως επιλέχθηκε και το k στον προηγούμενο τρόπο επιλογής της τιμής της παραμέτρου. Η επίλυση της εξίσωσης που μας δίνει την τιμή του s_i μπορεί να γίνει μέσω της επαναληπτικής μεθόδου της διχοτόμησης. Να σημειωθεί ότι δεν απαιτείται μεγάλος αριθμός βημάτων καθώς δε χρειάζεται

ιδιαίτερη ακρίβεια στον υπολογισμό της τιμής της τοπικής παραμέτρου κλιμάκωσης. Τελικά ο πίνακας γειτνίασης A θα περιέχει τα στοιχεία $A_{ij} = \min\left\{\widetilde{A}_{ij}, \widetilde{A}_{ji}\right\}$.

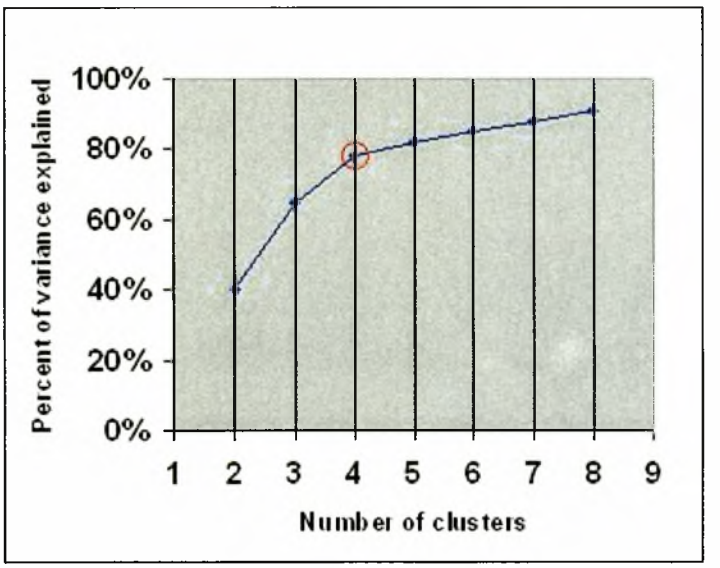
c. Μια άλλη προσέγγιση είναι η εξής: $\sigma_i = \sum_k d^2(\chi_i, \chi_k)$. Ουσιαστικά υπολογίζουμε το σ_i ως το άθροισμα των τετραγώνων των αποστάσεων του σημείου χ_i από τους k κοντινότερους γείτονές του (δηλαδή αυτοί που απέχουν λιγότερο από το συγκεκριμένο σημείο). Η τιμή του k μπορεί πάλι να υπολογιστεί με βάση τη θεώρηση που έγινε και προηγουμένως. Μια παραλλαγή αυτής της προσέγγισης είναι η ακόλουθη: $\sigma_i = \frac{1}{n_i}\sum_k d(\chi_i, \chi_k)$, όπου n_i είναι το σύνολο των γειτόνων που θα επιλέξουμε για το σημείο x_i.



Η επίδραση που έχει η τοπική κλιμάκωση. α) *Τα δεδομένα εισόδου. Παρατηρούμε ότι στο κέντρο εντοπίζεται μία ευδιάκριτη συστάδα που περιβάλλεται από μια άλλη εξωτερικά αυτής.* **ß)** *Η συνάφεια (affinity ή adjacency) μεταξύ κάθε σημείου με τους γείτονές του υποδεικνύεται από το πάχος της γραμμής που τα συνδέει. Οι συνάφειες κατά πλάτος των συστάδων είναι μεγαλύτερες από αυτές που εντοπίζονται εντός της περιβάλλουσας (της εξωτερικής) αυτής.* **γ)** *Οι προκύπτουσες συνάφειες ύστερα από την τοπική κλιμάκωση. Οι συνάφειες των σημείων κατά πλάτος των συστάδων τώρα είναι σημαντικά πιο ισχνές από αυτές που εντοπίζονται μεταξύ των σημείων εντός οποιασδήποτε συστάδας.*

3.2.3.5 Ο αριθμός των συστάδων

Η επιλογή του αριθμού των συστάδων αποτελεί ένα γενικό πρόβλημα για όλους τους αλγορίθμους συσταδοποίησης, αν και κατά καιρούς έχουν προταθεί διαφορετικοί τρόποι, άλλοι περισσότερο αποδοτικοί και άλλοι όχι τόσο, για την αυτόματη επιλογή του αριθμού τους. Το πιο γνωστό κριτήριο που χρησιμοποιείται στους αλγορίθμους συσταδοποίησης για την επιλογή του αριθμού των συστάδων είναι το *κριτήριο της καμπής* (elbow criterion). Σύμφωνα με αυτό το κριτήριο ο αριθμός των συστάδων που θα επιλεγεί θα πρέπει να είναι τέτοιος ώστε η εισαγωγή μιας επιπλέον συστάδας να μην προσθέτει κάποια σημαντική ποσότητα πληροφορίας στη συσταδοποίηση. Πιο συγκεκριμένα, αν γίνει γραφική παράσταση του ποσοστού της ποικιλομορφίας που εξηγείται από τον αριθμό των συστάδων συναρτήσει του αριθμού των συστάδων, γίνεται αντιληπτό ότι οι πρώτες συστάδες θα προσθέσουν αρκετή ποσότητα πληροφορίας (που αντιστοιχεί σε μεγάλη ποικιλομορφία) έως ότου το περιθώριο κέρδος να πέσει μετά από κάποιο σημείο, σχηματίζοντας στο γράφο μία γωνία (την καμπή) (*Εικόνα 3.2.3.5)*. Το σημείο όπου παρουσιάζεται αυτή η καμπή δεν μπορεί να προσδιοριστεί πάντα με μοναδικό τρόπο.



Εικόνα 3.2.3.5

Οπτική αναπαράσταση του κριτηρίου καμπής για ένα παράδειγμα συνόλου δεδομένων, όπου ο αριθμός των συστάδων που υπολογίστηκαν είναι 4 και υποδεικνύονται από το σημείο όπου εμφανίζεται η καμπή.

Η μέθοδος όμως που ανταποκρίνεται περισσότερο στη φασματική συσταδοποίηση είναι η ευρετική μέθοδος του *φασματικού ιδιοκενού*, που μπορεί να χρησιμοποιηθεί και στους τρεις πίνακες Laplace και ορίζεται ως $\Delta_\kappa = |\lambda_\kappa - \lambda_{\kappa+1}|$. Η βασική ιδέα στην οποία βασίζεται η μέθοδος είναι να επιλεγεί εκείνος ο ακέραιος αριθμός k συστάδων ώστε οι ιδιοτιμές $\lambda_1,...,\lambda_k$ να είναι πολύ μικρές, αλλά η ιδιοτιμή λ_{k+1} να είναι σχετικά μεγάλη. Μία άτυπη απόδειξη που δικαιολογεί την ορθότητα της συγκεκριμένης μεθόδου προκύπτει αν αναλογιστούμε την ιδανική περίπτωση των k μη συνδεδεμένων συστάδων, όπου η ιδιοτιμή 0 έχει πολλαπλότητα k, και έπειτα υπάρχει ένα κενό μέχρι την k+1 ιδιοτιμή για την οποία ισχύει $\lambda_{k+1} > 0$. Επίσης ένας άλλος τρόπος εκτίμησης του αριθμού των συστάδων, που βασίζεται στο γεγονός ότι ο κανονικοποιημένος πίνακας γειτνίασης L σχετίζεται με τις ιδιοτιμές λ_i του πίνακα Laplace με τη σχέση $1-\lambda_i$, είναι ότι ο αριθμός των ομάδων όπου μπορεί να χωριστεί το σύνολο δεδομένων μας ισούται με την πολλαπλότητα της ιδιοτιμής 1, αφού αυτή είναι η μεγαλύτερη ιδιοτιμή του πίνακα L. Δυστυχώς όμως το συγκεκριμένο κριτήριο λειτουργεί μόνο όταν οι συστάδες που υπάρχουν μέσα στα δεδομένα μας είναι πλήρως διαχωρισμένες μεταξύ τους.

3.2.4 Συμπέρασμα

Το μειονέκτημα που παρουσιάζει ο συγκεκριμένος αλγόριθμος είναι ότι ο χρήστης ευθύνεται για την επιλογή κατάλληλης και αξιόπιστης τιμής για την καθολική παράμετρο σ, κάτι που τις περισσότερες φορές δεν θεωρείται ιδιαίτερα εύκολη διαδικασία. Επίσης το γεγονός ότι πρέπει να επιλέξει και τον αριθμό των τελικών συστάδων δυσχεραίνει ακόμα περισσότερο το ήδη υπάρχον πρόβλημα.

Το πρόβλημα επιλογής του κατάλληλου σ έρχεται να επιλύσει ο αλγόριθμος του επόμενου κεφαλαίου με την αυτοματοποίηση του υπολογισμού του, αφήνοντας στην διακριτική ευχέρεια του χρήστη μόνο την επιλογή του αριθμού των συστάδων.

Κεφάλαιο 4

Αλγόριθμοι Φασματικής Συσταδοποίησης των Fischer και Poland

4.1 Γενικά

Ο αλγόριθμος των Ng, Jordan και Weiss δίνει αρκετά καλά αποτελέσματα σε περιπτώσεις όπου ο πίνακας γειτνίασης έχει τη μορφή διαγωνίου μπλοκ πίνακα, αλλά η θεωρία δεν εγγυάται την εύρεση καλής λύσης στην περίπτωση που ο πίνακας είναι τύπου block-band (ο πίνακας αποτελείται από δύο διαγώνιους, υπό μορφή ομάδας, πίνακες). Ωστόσο όταν ο πίνακας γειτνίασης είναι σε αυτή τη μορφή η φασματική συσταδοποίηση με το συγκεκριμένο αλγόριθμο μπορεί να έχει καλά αποτελέσματα αρκεί να επιλεγεί κατάλληλη τιμή για την καθολική παράμετρο s του αλγορίθμου. Η επιλογή ενός κατάλληλου s θεωρείται αρκετά δύσκολη σε αυτή την περίπτωση καθώς η τιμή της πρέπει να είναι ικανοποιητικά μικρή, ώστε να καταλήγουμε σε ένα πίνακα γειτνίασης όπου οι ομοιότητες μόνο των άμεσα γειτονικών σημείων είναι υψηλές. Παρόλα αυτά, η τιμή του s δεν πρέπει να είναι πολύ μικρή διότι σε μερικές περιπτώσεις το διάστημα των αποδεκτών τιμών για το s μπορεί να είναι αρκετά περιορισμένο. Συνεπώς όλες αυτές οι θεωρήσεις επιβάλλουν μια αλλαγή στο μέτρο ομοιότητας.

Συγκεκριμένα, αυτό που επιθυμούμε είναι να ενισχύσουμε τις ασθενείς ομοιότητες που υπάρχουν μέσα στον πίνακα γειτνίασης A. Έτσι δύο σημεία πλέον θεωρούνται όμοια, και τους αναθέτουμε μια μεγάλη τιμή ομοιότητας, αν η συνολική αγωγιμότητα[4] μεταξύ τους, εντός του γράφου ομοιότητας, είναι μεγάλη. Ο αλγόριθμος που θα αναφερθεί προτάθηκε από τους Fischer και Poland και στη βιβλιογραφία αναφέρεται συνήθως με τον όρο ασυμμετρική συσταδοποίηση με

[4] Η αγωγιμότητα δύο σημείων εξαρτάται από όλα τα μονοπάτια του γράφου που υπάρχουν μεταξύ τους και ορίζεται όπως ακριβώς και στη θεωρία των ηλεκτρικών κυκλωμάτων.

ενίσχυση της μορφής μπλοκ του πίνακα (context dependent clustering with block amplification). Πριν γίνει περιγραφή του αλγορίθμου θα οριστεί η κατασκευή κάποιων πινάκων που χρησιμοποιούνται σε κάποια από τα βήματα του αλγορίθμου καθώς επίσης θα γίνει και περιγραφή του αλγορίθμου k-lines, που χρησιμοποιείται στο τελευταίο βήμα του αλγορίθμου για τη συσταδοποίηση των *φασματικών εικόνων*. Η έννοια της φασματικής εικόνας θα αναφερθεί πιο κάτω ούτως ώστε να είναι περισσότερο αντιληπτή. Οι πίνακες που χρησιμοποιούνται στον αλγόριθμο είναι ο πίνακας αγωγιμότητας C και ο πίνακας γειτνίασης A. Βέβαια ο πίνακας αγωγιμότητας C δεν είναι εντελώς ανεξάρτητος από τον πίνακα γειτνίασης A, αφού ο δεύτερος χρησιμοποιείται ούτως ώστε να κατασκευαστεί ο πρώτος.

4.2 Πίνακας αγωγιμότητας

4.2.1 Τρόπος κατασκευής

Η αγωγιμότητα για οποιαδήποτε δύο σημεία χ_i και χ_j υπολογίζεται με τον ακόλουθο τρόπο. Αρχικά επιλύουμε το σύστημα γραμμικών εξισώσεων:

$G \cdot \varphi = \eta_{ij}$ (1), όπου G είναι ένας πίνακας μεγέθους $N \times N$ (αφού θεωρούμε ότι έχουμε ένα σύνολο σημείων εισόδου ίσο με N) που κατασκευάζεται με βάση τον αρχικό πίνακα γειτνίασης A ως εξής:

$$G(p,q) = \begin{cases} \text{if } p = 1: & \begin{cases} 1 & \text{if } q = 1 \\ 0 & \text{else} \end{cases} \\ \text{else}: & \begin{cases} \sum_{k \neq p} A(p,k) & \text{if } p = q \\ -A(p,q) & \text{else} \end{cases} \end{cases} \quad (2)$$

και όπου η_{ij} είναι ένα διάνυσμα δείκτης μεγέθους N, το οποίο υποδεικνύει τα ζεύγη σημείων χ_i και χ_j για τα οποία θέλουμε να υπολογίσουμε την αγωγιμότητα. Τα στοιχεία του διανύσματος η_{ij} υπολογίζονται ως εξής:

$$\eta_{ij}(k) = \begin{cases} -1 & \text{for } k = i \text{ and } i > 1 \\ 1 & \text{for } k = j \\ 0 & \text{otherwise} \end{cases} \quad (3).$$

Με την επίλυση της εξίσωσης (1) υπολογίζουμε το διάνυσμα φ. Η αγωγιμότητα μεταξύ των σημείων χ_i και χ_j, όπου $i < j$, δίνεται από τη σχέση:

$C(i,j) = \frac{1}{\varphi(j) - \varphi(i)}$ (4), η οποία δεδομένου ότι το διάνυσμα η_{ij} είναι αρκετά αραιό, μπορεί να απλοποιηθεί ως εξής: $C(i,j) = \frac{1}{G^{-1}(i,i) + G^{-1}(j,j) - G^{-1}(i,j) - G^{-1}(j,i)}$ (5). Λόγω συμμετρίας ισχύει ότι: $C(i,j) = C(j,i)$ ενώ τα διαγώνια στοιχεία $C(i,i)$ υπολογίζονται ως εξής: $C(i,i) = \max_{i,j} C(i,j)$. Συνεπώς αρκεί να γίνει ο υπολογισμός του πίνακα G^{-1} μόνο μια φορά και σε χρόνο $O(N^3)$, ενώ για να επιτευχθεί ο υπολογισμός του πίνακα αγωγιμότητας C απαιτείται χρόνος $O(N^2)$.

4.2.2 Προέλευση του πίνακα αγωγιμότητας

Η ιδέα της παραπάνω μεθόδου προέρχεται από τη μέθοδο ανάλυσης κόμβων που χρησιμοποιείται στα ηλεκτρικά κυκλώματα. Αρχίζουμε το συλλογισμό μας θεωρώντας ένα κύκλωμα αντιστάσεων όπου η αγωγιμότητα μεταξύ δύο κόμβων i και j, συμβολίζεται με G_{ij}. Η τιμή της αγωγιμότητας G_{ij} μεταξύ των κόμβων i και j, όπου εφαρμόζεται τάση V_{ij} και διέρχονται από ρεύμα I, δίνεται από το νόμο του Ohm: $G_{ij} = \frac{I}{V_{ij}}$. Η τάση V_{ij} ορίζεται ως η διαφορά των δυναμικών μεταξύ των κόμβων i και j: $V_{ij} = \varphi_j - \varphi_i$, όπου τα δυναμικά φ_i και φ_j υπολογίζονται από το νόμο του Kirchhoff σύμφωνα με τον οποίο όλα τα ρεύματα που εισέρχονται σε έναν κόμβο i πρέπει να εξέρχονται και από αυτόν, δηλαδή $\sum_{j \neq i} I_{ij} = 0$. Χρησιμοποιώντας το νόμο του Ohm, τα ρεύματα μπορούν να εκφραστούν συναρτήσει των τάσεων και των αγωγιμοτήτων, οπότε η προηγούμενη εξίσωση

μπορεί να γραφεί πιο αναλυτικά ως εξής: $\sum_{j\neq i} G_{ij}V_{ij} = \sum_{j\neq i} G_{ij}(\varphi_j - \varphi_i) = 0$.

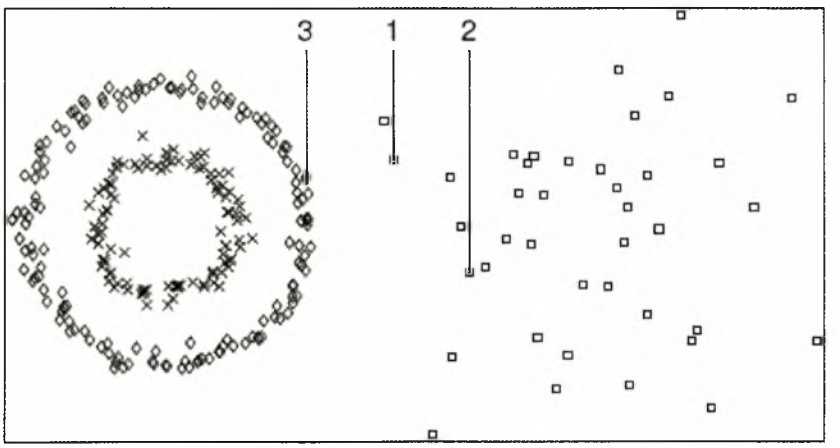
Ομαδοποιώντας τις αγωγιμότητες με τα αντίστοιχα δυναμικά και σχηματίζοντας τις εξισώσεις για όλους τους κόμβους καταλήγουμε στην εξίσωση (1). Το διάνυσμα η_{ij} αντιπροσωπεύει τα γνωστά ρεύματα I που έχουν μεταφερθεί στα δεξιά της εξίσωσης. Σύμφωνα με τη θεωρία των γράφων ο πίνακας γειτνίασης ενός συνδεδεμένου γράφου με *N* κόμβους είναι τάξης *N-1*. Στην περίπτωση δηλαδή που θα κατασκευάζαμε τον πίνακα G βασιζόμενοι μόνο στο νόμο του Kirchhoff και του Ohm τότε οι σειρές του θα άθροιζαν σε 0 και το σύστημα θα ήταν αόριστο. Υπό μια φυσική λογική τα ρεύματα που εισέρχονται και εξέρχονται από N-1 κόμβους προσδιορίζουν επίσης και τα ρεύματα στον *N*-οστό κόμβο, αφού δεν μπορούν να πάνε κάπου αλλού. Για να καταλήξουμε σε ένα σύστημα που έχει λύση, πρέπει να επιλεγεί ένας κόμβος στον οποίο θα ανατεθεί ένα γνωστό δυναμικό ώστε να γίνει κόμβος αναφοράς. Στη μέθοδό που περιγράφεται το δυναμικό του πρώτου κόμβου τίθεται στην τιμή 0 $(\varphi(1) = 0)$. Επίσης με μια πρώτη ματιά, φαίνεται ότι απαιτείται η χρήση ενός διαφορετικού η_{ij} και η επίλυση των εξισώσεων εκ νέου για όλα τα $\binom{N}{2}$ ζευγάρια των κόμβων. Κάτι τέτοιο θα ισοδυναμούσε με τη σύνδεση της πηγής ρεύματος μεταξύ κάθε ζεύγους κόμβων και μέτρηση της αντίστοιχης τάσης.

Παρόλα αυτά η συγκεκριμένη θεώρηση δεν ισχύει για τη μέθοδο που περιγράφεται για τους εξής λόγους: Πρώτον, αφού οι άμεσες αγωγιμότητες μεταξύ των κόμβων δεν αλλάζουν, αρκεί να αντιστρέψουμε τον πίνακα G μόνο μια φορά. Δεύτερον, για τον υπολογισμό της συνολικής αγωγιμότητας μεταξύ δύο κόμβων, de χρειαζόμαστε όλες τις τάσεις του κυκλώματος, αλλά αρκεί η τάση μεταξύ αυτών των δύο συγκεκριμένων κόμβων. Αυτό μας επιτρέπει να χρησιμοποιούμε μόνο δύο σειρές από τον πίνακα G^{-1}. Επιπλέον το γεγονός ότι για κάθε διάνυσμα η_{ij}, όλα τα στοιχεία εκτός από δύο είναι ίσα με το 0 (αφού η εξωτερική πηγή ρεύματος εφαρμόζεται μόνο σε δύο κόμβους), μας επιτρέπει να θεωρήσουμε δύο στήλες στον πίνακα G^{-1}. Προφανώς η αγωγιμότητα μεταξύ δύο οποιονδήποτε κόμβων μπορεί να υπολογιστεί μόνο από τέσσερα στοιχεία του πίνακα G^{-1}. Κάτι τέτοιο φαίνεται και από την εξίσωση (5).

4.3 Πίνακας γειτνίασης A

Ο πίνακας αγωγιμότητας C προκύπτει από τον πίνακα A, αλλά με τη μόνη διαφορά ότι απεικονίζει μια ενισχυμένη μορφή μπλοκ. Σε αυτό το σημείο πρέπει να αναφέρουμε ότι ο πίνακας A υπολογίζεται με βάση τη (b) μέθοδο όπου έχει αναφερθεί για την εύρεση τοπικών παραμέτρων σ_i στο προηγούμενο κεφάλαιο. Παρατηρούμε ότι τα στοιχεία του τελικού πίνακα A δίνονται από τη σχέση $A_{ij} = \min\left\{\widetilde{A}_{ij}, \widetilde{A}_{ji}\right\}$ γιατί επιθυμούμε να καταλήξουμε σε έναν συμμετρικό πίνακα A με βάση τον αρχικό μη συμμετρικό $\widetilde{A}$. Ισχύει πάλι η υπόθεση που ίσχυε και για τον πίνακα γειτνίασης A του αλγορίθμου NJW, ότι στοιχεία με υψηλές τιμές μέσα στον πίνακα θα υπάρχουν μόνο μεταξύ των άμεσα γειτονικών σημείων. Αυτός είναι και ένας επιπλέον λόγος για τον οποίο θεωρούμε ότι τα στοιχεία A_{ij} του πίνακα δίνονται από τη σχέση $A_{ij} = \min\left\{\widetilde{A}_{ij}, \widetilde{A}_{ji}\right\}$.

Για παράδειγμα στην *Εικόνα 4.3*, τα σημεία 1 και 3 αποκτούν μικρότερη τιμή ομοιότητας από ότι τα σημεία 1 και 2. Αυτό ισχύει επειδή το σημείο 1, παρόλο που είναι πιο κοντά στο σημείο 3 συγκριτικά με την απόσταση του από το σημείο 2, εντούτοις στην πραγματικότητα βρίσκεται στη σχετική γειτονιά του σημείου 2.



Εικόνα 4.3

Δεδομένα σε μορφή δύο δαχτυλιδιών με μικρή διασκόρπιση και μια Γκαουσιανή συστάδα με μεγάλη διασκόρπιση.

4.4 Περιγραφή του αλγορίθμου assυμμετρικής συσταδοποίησης των Fischer και Poland

4.4.1 Γενικά

Δοθέντος του πίνακα αγωγιμότητας C (η του πίνακα γειτνίασης A), υπολογίζουμε τον πίνακα $V = [v_1 \dots v_k] \in \Re^{N \times k}$, που περιέχει τα k ιδιοδιανύσματα και τα οποία αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές. Συμβολίζουμε με $[\psi_1 \dots \psi_N] = Y = V' \in \Re^{K \times N}$ τις *φασματικές εικόνες* των αρχικών δεδομένων $\chi_1, \dots, \chi_N$. Για τον υπολογισμό των τελικών συστάδων με βάση τις φασματικές εικόνες μπορούμε να χρησιμοποιήσουμε διάφορους τρόπους. Οι εικόνες των σημείων που ανήκουν στην ίδια συστάδα παρατηρήθηκε ότι είναι κατανεμημένες υπό γωνία γύρω από ευθείες γραμμές και όσο πιο συμπαγής είναι μια συστάδα τόσο πιο καλή είναι η συσταδοποίηση των φασματικών εικόνων. Η χρήση του αλγορίθμου k-means δεν ενδείκνυται για αυτή την περίπτωση, γιατί ως γνωστόν, ο συγκεκριμένος αλγόριθμος δημιουργεί σφαιρικές συστάδες. Το συγκεκριμένο πρόβλημα προσπάθησαν να το λύσουν οι Ng, Jordan και Weiss με την απεικόνιση των σημείων στη μοναδιαία σφαίρα και τη μετέπειτα εφαρμογή του αλγορίθμου k-means. Ωστόσο αυτή η μετατροπή των δεδομένων έχει ως μειονέκτημα την απώλεια πληροφορίας γι' αυτό και στο συγκεκριμένο αλγόριθμο φασματικής συσταδοποίησης χρησιμοποιούμε τον αλγόριθμο k-lines για τη συσταδοποίηση των φασματικών εικόνων.

4.4.2 Ο αλγόριθμος k-lines

Ο αλγόριθμος k-lines που περιγράφεται στη συνέχεια, ανταποκρίνεται στην προσέγγιση που χρησιμοποιήθηκε από τους Fischer και Poland για την υλοποίηση του αλγορίθμου συσταδοποίησης που οι ίδιο πρότειναν. Στον αλγόριθμο k-lines τα σημεία συσταδοποιούνται γύρω από μία γραμμή και όχι γύρω από κάποια σημεία.

Κάθε γραμμή αναπαρίσταται από ένα διάνυσμα μοναδιαίου μήκους $m_i \in \Re^k$, όπου $1 \le i \le k$.

Αλγόριθμος k-lines

Είσοδος: X // πίνακας δεδομένων εισόδου μεγέθους N×d
(N είναι το πλήθος των σημείων διάστασης d)
k // το πλήθος των επιθυμητών συστάδων

Έξοδος: Ομαδοποίηση των N σημείων σε k συστάδες

Βήμα 1: Αρχικοποίηση των διανυσμάτων $m_1 \ldots m_k$ είτε με τυχαίο τρόπο είτε με τα μεγαλύτερα ιδιοδιανύσματα των φασματικών δεδομένων (ή εικόνων) ψ_i.

Βήμα 2: Για $i = 1 \ldots k$ εκτέλεσε τα Βήματα 3 και 4

Βήμα 3: Δημιουργία του πίνακα $M_i = [\psi_i]_{i \in N_i}$ του οποίου οι στήλες είναι τα σημεία ψ_i που βρίσκονται πιο κοντά στη γραμμή m_i (ελέγχουμε ποια γραμμή της περιοχής έχει το μικρότερο άθροισμα τετραγωνικής απόστασης από τα σημεία), σχηματίζοντας έτσι τη γειτονιά N_i.

Βήμα 4: Υπολογισμός της νέας γραμμής m_i' ως το μεγαλύτερο ιδιοδιάνυσμα του πίνακα $M_i M_i^T$.

Βήμα 5: Εκτέλεση του Βήματος 2 έως ότου επιτευχθεί σύγκλιση.

4.4.3 Ο αλγόριθμος ασυμμετρικής συσταδοποίησης των Fischer και Poland

Αλγόριθμος ασυμμετρικής συσταδοποίησης των Fischer και Poland

Είσοδος: X // πίνακας δεδομένων εισόδου μεγέθους d×N

(N είναι το πλήθος των σημείων διάστασης d)

k // το πλήθος των επιθυμητών συστάδων

τ // το μέγεθος της γειτονιάς (ο χρήστης μπορεί να ορίσει τη δική του τιμή για την παράμετρο ή αλλιώς να θεωρηθεί η εξ' ορισμού 2×d + 1)

Έξοδος: Ομαδοποίηση των N σημείων σε k συστάδες

Βήμα 1: Υπολογισμός των πινάκων γειτνίασης $\tilde{A}$ και A.

Βήμα 2: Κατασκευή του πίνακα αγωγιμότητας C σύμφωνα με την εξίσωση (5).

Βήμα 3: Προσδιορισμός των k μεγαλύτερων ιδιοτιμών του πίνακα αγωγιμότητας C, στις οποίες αντιστοιχούν τα k μεγαλύτερα ιδιοδιανύσματα $v_1 \dots v_k$, όπου

$v_i \in \Re^N$ και $V = [v_1 \dots v_k] \in \Re^{N \times k}$ και $Y = V' \in \Re^{k \times N}$ με τις στήλες $\psi_1 \dots \psi_N$, να αποτελούν τις φασματικές εικόνες.

Βήμα 4: Συσταδοποίηση των φασματικών εικόνων $\psi_1 \dots \psi_N$ με τον αλγόριθμο k-lines.

Το βασικό πλεονέκτημα αυτής της μεθόδου συσταδοποίησης είναι ότι η παράμετρος κλιμάκωσης σ δεν δίνεται από τον χρήστη αλλά υπολογίζεται από την ίδια την μέθοδο. Συνεπώς η μόνη παράμετρος που καθορίζεται από τον χρήστη είναι ο αριθμός των συστάδων.

4.5 Περιγραφή του αλγορίθμου φασματικής συσταδοποίησης με χρήση πίνακα αγωγιμότητας

Εκτός όμως από τη χρήση του πίνακα αγωγιμότητας σε αυτό το νέο αλγόριθμο φασματικής συσταδοποίησης της Ενότητας 4.4.3, ο συγκεκριμένος πίνακας θα μπορούσε να χρησιμοποιηθεί και στον απλό αλγόριθμο φασματικής συσταδοποίησης όπου ο πίνακας γειτνίασης χτίζεται με βάση τη συνάρτηση Γκαουσιανού πυρήνα.

Αλγόριθμος φασματικής συσταδοποίησης με πίνακα αγωγιμότητας C

Είσοδος: X // ο πίνακας δεδομένων εισόδου μεγέθους d×N

(N είναι το πλήθος των σημείων διάστασης d)

c // το πλήθος των επιθυμητών συστάδων

σ // η καθολική παράμετρος κλιμάκωσης

Έξοδος: Ομαδοποίηση των N σημείων σε c συστάδες

Βήμα 1: Κατασκευή του πίνακα A μεγέθους N×N τα στοιχεία του οποίου ορίζονται με τον ακόλουθο τρόπο: $A_{ij} = \exp(-\frac{\|\chi_i - \chi_j\|^2}{2\sigma^2})$ για i ≠ j και $A_{ij} = 0$ για i = j.

Βήμα 2: Κατασκευή του πίνακα αγωγιμότητας C βάσει του πίνακα A που υπολογίστηκε στο προηγούμενο βήμα.

Βήμα 3: Προσδιορισμός των k μεγαλύτερων ιδιοτιμών του πίνακα αγωγιμότητας C, στις οποίες αντιστοιχούν τα k μεγαλύτερα ιδιοδιανύσματα $v_1 \dots v_k$, όπου $v_i \in \Re^N$ και $V = [v_1 \dots v_k] \in \Re^{N \times k}$ και $Y = V' \in \Re^{k \times N}$ με τις στήλες $\psi_1 \dots \psi_N$, να αποτελούν τις φασματικές εικόνες.

Βήμα 4: Συσταδοποίηση των φασματικών εικόνων $\psi_1 \dots \psi_N$ με τον αλγόριθμο k-lines.

4.6 Συμπέρασμα

Οι Fischer και Poland κατάφεραν με τον αλγόριθμο ασυμμετρικής συσταδοποίησης να εισάγουν μια νέα παράμετρο τ, που καθορίζει το μέγεθος της γειτονιάς των σημείων, ενώ παράλληλα πέτυχαν να αποδεσμεύσουν τον αλγόριθμο φασματικής συσταδοποίησης από την ευαισθησία που εισήγαγε η παράμετρο κλιμάκωσης s. Παρόλα αυτά, μια παράμετρος που δεν καθορίστηκε με σαφή τρόπο είναι ο αριθμός των συστάδων που θα πρέπει τελικά να δημιουργηθούν, αλλά θεωρούν ότι είναι μια γνωστή εκ των προτέρων παράμετρος. Ένας αλγόριθμος που υπολογίζει την τιμή της τοπικής παραμέτρου σ_i και τον αριθμό των συστάδων είναι ο αλγόριθμος φασματικής συσταδοποίησης ZP που περιγράφεται στο επόμενο κεφάλαιο.

Κεφάλαιο 5

Ο Αλγόριθμος Φασματικής Συσταδοποίησης ZP

5.1 Γενικά

Ο αλγόριθμος φασματικής συσταδοποίησης ZP προτάθηκε από τους Lihi Zelnik-Manor και Pietro Perona και χρησιμοποίησε ως βάση τον αλγόριθμο των Ng, Jordan και Weiss για να επιλύσει τα προβλήματα τοπικής κλιμάκωσης και αυτόματου προσδιορισμού των συστάδων ενός συνόλου δεδομένων.

5.2 Θεωρήσεις του αλγορίθμου

Σε αυτή την υποενότητα θα παρουσιάσουμε τις θεωρήσεις που κάνει ο αλγόριθμος ZP για την τοπική κλιμάκωση και για τον αυτόματο προσδιορισμό των συστάδων.

5.2.1 Πίνακας γειτνίασης τοπικής κλιμάκωσης

Τα στοιχεία του πίνακα γειτνίασης κατασκευάζονται με βάση τη μέθοδο (α) που περιγράφτηκε στο Κεφάλαιο 2 και αφορά την τοπική κλιμάκωση. Συγκεκριμένα κάθε στοιχείο $\hat{A}_{ij}$ του πίνακα ισούται με $\hat{A}_{ij} = \exp\left(\frac{-d^2(\chi_i, \chi_j)}{\sigma_i \sigma_j}\right)$ (1) με $\sigma_i = d(\chi_i, \chi_k)$ (2) και k ο δείκτης που υποδεικνύει τον k-οστό πλησιέστερο γείτονα του σημείου χ_i.

5.2.2 Ο αριθμός των συστάδων

Ο υπολογισμός του αριθμού των συστάδων είναι ένα από τα προβλήματα που κανένας από τους προηγούμενους αλγορίθμους δε κατάφερε να λύσει επιτυχώς και ειδικά για δεδομένα που εμπεριείχαν κάποιο ποσοστό θορύβου. Συνήθως ο υπολογισμός του αριθμού των συστάδων γίνεται είτε με τη κλασσική μέθοδο της ανάλυσης των ιδιοτιμών του κανονικοποιημένου πίνακα γειτνίασης είτε με την ανάλυση των ιδιοδιανυσμάτων αυτού.

5.2.2.1 Ανάλυση των ιδιοτιμών

Όπως έχει ήδη αναφερθεί ήδη από τον αλγόριθμο των Ng, Jordan και Weiss, ο κοινός τρόπος εντοπισμού του αριθμού των συστάδων είναι με την ανάλυση των ιδιοτιμών του κανονικοποιημένου πίνακα γειτνίασης L. Συγκεκριμένα η πολλαπλότητα της ιδιοτιμής που ισούται με 1 αντιστοιχεί και στον αριθμό των συστάδων που εμπεριέχονται σε ένα σύνολο δεδομένων στα οποία όμως απουσιάζει ο θόρυβος. Στην περίπτωση παρουσίας θορύβου η συγκεκριμένη τεχνική δεν μπορεί να δώσει ικανοποιητικά αποτελέσματα. Ένας άλλος τρόπος υπολογισμού του πλήθους των συστάδων είναι και το κριτήριο του *φασματικού ιδιοκενού* που έχει ήδη αναφερθεί, αλλά που επίσης δε δίνει κάποια αξιόπιστη λύση σε αυτό το πρόβλημα. Η λύση που προτείνεται από το συγκεκριμένο αλγόριθμο είναι μέσω της ανάλυσης των ιδιοδιανυσμάτων.

5.2.2.2 Ανάλυση των ιδιοδιανυσμάτων

Οι ιδιοτιμές του κανονικοποιημένου πίνακα γειτνίασης L είναι το αποτέλεσμα της ένωσης των ιδιοτιμών των υποπινάκων που αντιστοιχούν σε κάθε συστάδα και επομένως εξαρτώνται από τη δομή των διαφορετικών συστάδων που εντοπίζονται μέσα στα δεδομένα. Θεωρώντας την ιδανική περίπτωση που ο πίνακας L είναι αυστηρά διαγώνιος μπλοκ πίνακας με πίνακες μπλοκ $L^{(c)}, c = 1,...,C$, τότε θα ισχύει ότι οι ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα L είναι η ένωση των ιδιοτιμών και των ιδιοδιανυσμάτων των μπλοκ πινάκων γεμίζοντας κατάλληλο αριθμό στοιχείων με

μηδενικά. Εφόσον οι ιδιοτιμές των μπλοκ πινάκων είναι διαφορετικές μεταξύ τους κάθε ιδιοδιάνυσμα θα έχει μη μηδενικές τιμές μόνο σε στοιχεία που αντιστοιχούν σε ένα μπλοκ ή αλλιώς συστάδα, δηλαδή:

$$\hat{X} = \begin{bmatrix} \chi^{(1)} & \vec{0} & \vec{0} \\ \vec{0} & \dots & \vec{0} \\ \vec{0} & \vec{0} & \chi^{(C)} \end{bmatrix}_{n \times C}$$ όπου $\chi^{(C)}$ είναι ένα ιδιοδιάνυσμα του υποπίνακα $L^{(c)}$

που αντιστοιχεί στη συστάδα c. Όπως ήδη αναφέρθηκε η ιδιοτιμή 1 είναι εκ φύσεως μια ιδιοτιμή πολλαπλότητας C, όπου C είναι το πλήθος των ομάδων. Συνεπώς θα μπορούσε πολύ εύκολα να επιλεγεί και κάποιο άλλο σύνολο ορθογώνιων διανυσμάτων που επικαλύπτουν τον ίδιο υποχώρο με τις στήλες του $\hat{X}$. Έτσι ο $\hat{X}$ θα μπορούσε να αντικατασταθεί με $X = \hat{X}R$ για οποιονδήποτε ορθογώνιο πίνακα $R \in \Re^{C \times C}$. Το γεγονός αυτό μας υποδεικνύει ότι ακόμα και αν η επίλυση του ιδιοσυστήματος μας δώσει ως λύση το στραμμένο σύνολο ιδιοδιανυσμάτων, είναι βέβαιο ότι υπάρχει μια στροφή $\hat{R}$ τέτοια ώστε κάθε σειρά του πίνακα $X\hat{R}$ να έχει μία και μοναδική μη μηδενική καταχώριση. Είναι προφανές ότι, αφού τα ιδιοδιανύσματα του πίνακα L αποτελούν ένωση των ιδιοδιανυσμάτων των ξεχωριστών μπλοκ πινάκων (με κατάλληλη πρόσθεση μηδενικών), αν επιλέξουμε περισσότερα από τα C μεγαλύτερα ιδιοδιανύσματα θα καταλήξουμε σε περισσότερες από μία μη μηδενικές καταχωρήσεις σε μερικές από τις σειρές του πίνακα. Στην αντίθετη περίπτωση που θα επιλέξουμε λιγότερα ιδιοδιανύσματα, δε θα πάρουμε μια βάση που να καλύπτει τον υποχώρο. Επίσης το αν θα υπάρχει μια τέτοια στροφή ή όχι εξαρτάται από τον αρχικό πίνακα X.

Αξίζει να σημειωθεί ότι για κάθε πιθανό πλήθος ομάδων C επιλέγουμε εκείνη τη στροφή που στοιχίζει καλύτερα τις στήλες του πίνακα X με το κανονικό σύστημα συντεταγμένων. Υποθέτουμε ότι $Z \in \Re^{n \times C}$ είναι ο πίνακας που προκύπτει ύστερα από τη στροφή του πίνακα ιδιοδιανυσμάτων X, δηλαδή $Z = XR$, και επίσης ότι $M_i = \max_j Z_{ij}$. Σκοπός μας είναι η εύρεση μιας στροφής R τέτοια ώστε σε κάθε σειρά του Z να υπάρχει το πολύ μια μη μηδενική καταχώρηση. Έτσι καταλήγουμε στον ορισμό μιας συνάρτησης κόστους: $J = \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{Z_{ij}^2}{M_i^2}$ (3).

Συνεπώς αν ελαχιστοποιηθεί η συγκεκριμένη συνάρτηση κόστους σε σχέση με όλες τις πιθανές στροφές θα προκύψει η καλύτερη στοίχιση με το κανονικό σύστημα συντεταγμένων. Αυτό επιτυγχάνεται σύμφωνα με τη μέθοδο *gradient descent* που θα περιγραφεί πιο κάτω. Το πλήθος των ομάδων προκύπτει ως η τιμή εκείνη για την οποία το κόστος που προκύπτει είναι το ελάχιστο. Αν υπάρχουν πολλές τιμές για το πλήθος των ομάδων που οδηγούν στο ίδιο ελάχιστο κόστος τότε επιλέγεται η μεγαλύτερη από αυτές.

Η αναζήτηση για τον αριθμό των ομάδων μπορεί να εκτελεστεί με αυξητικό ρυθμό (incrementally) εξοικονομώντας έτσι, πολύτιμο υπολογιστικό χρόνο. Η διαδικασία ξεκινά στοιχίζοντας τα δύο μεγαλύτερα ιδιοδιανύσματα με τον καλύτερο δυνατό τρόπο και έπειτα σε κάθε βήμα της αναζήτησης (έως ότου φτάσουμε τον μέγιστο αριθμό συστάδων) προσθέτουμε ένα ιδιοδιάνυσμα στα ήδη στραμμένα. Έτσι κάθε βήμα αρχικοποιείται με το αποτέλεσμα της στοίχισης του προηγούμενου αριθμού ομάδων. Η στοίχιση για το νέο αυτό σύνολο ιδιοδιανυσμάτων που προκύπτει, είναι αρκετά γρήγορη αφού η αρχικοποίηση που γίνεται σε κάθε βήμα, θεωρείται αρκετά καλή. Αξίζει να αναφερθεί ότι ο συνολικός χρόνος που απαιτείται για την εκτέλεση αυτής της διαδικασίας είναι ελαφρώς μεγαλύτερος από αυτόν που θα απαιτούνταν για τη στοίχιση όλων των ιδιοδιανυσμάτων με μη-αυξητικό τρόπο.

Στην επόμενη υποενότητα παρουσιάζεται αναλυτικά ο τρόπος που υπολογίζεται η καλύτερη στοίχιση για ένα σύνολο διανυσμάτων σύμφωνα με τη μέθοδο gradient descent.

5.2.2.3 Η μέθοδος gradient descent για τον υπολογισμό της στοίχισης ενός συνόλου διανυσμάτων

Η συνάρτηση κόστους που επιθυμούμε να ελαχιστοποιήσουμε είναι αυτή που περιγράφεται από την εξίσωση (3). Ξεκινάμε υποθέτοντας ότι $m_i = j$ τέτοιο ώστε $Z_{ij} = Z_{im_i} = M_i$. Η μέθοδος gradient descent μας επιτρέπει να αυξήσουμε το κόστος που αντιστοιχεί σε ένα υποσύνολο των σειρών του πίνακα X υπό τον όρο ότι το συνολικό κόστος μειώνεται.

Υποθέτουμε ότι $\widetilde{G}_{i,j,\theta}$ είναι μια στροφή Givens (*Βλέπε Ορισμοί Κεφ. 1*) ακτινίων θ (αντίθετων της φοράς του ρολογιού) στο επίπεδο (i,j) . Θεωρώντας στροφές Givens τέτοιες ώστε $i < j$, μπορούμε να χρησιμοποιήσουμε ένα πιο βολικό σχέδιο απεικόνισης ως εξής: $G_{k,\theta} = \widetilde{G}_{i,j,\theta}$, όπου (i,j) είναι η k-οστή καταχώριση σε μια λεξικογραφική λίστα του $(i,j) \in \{1,2,\ldots,C\}^2$ με $i < j$. Ως εκ τούτου, βρίσκουμε τις στροφές στοίχισης που ελαχιστοποιούν τη συνάρτηση κόστους J πάνω από το $\Theta \in \left[-\frac{\pi}{2},\frac{\pi}{2}\right)^K$. Ο κανόνας για την ενημέρωση του Θ είναι ο ακόλουθος: $\Theta_{k+1} = \Theta_k - \alpha \nabla J\,|_{\Theta=\Theta_k}$, όπου $\alpha \in \Re^+$ είναι το μέγεθος του βήματος.

Στη συνέχεια υπολογίζουμε το *gradient* J και τα όρια για το a κάτω από τα οποία επιτυγχάνεται σταθερότητα. Εισάγουμε το συμβολισμό $U_{(a,b)} = G_{a,\theta_a} G_{a+1,\theta_{a+1}} \ldots G_{b,\theta_b}$ όπου $U_{(a,b)} = I$ αν $b < a, U_k = U_{(k,k)}$, και $V_k = \frac{\partial}{\partial \theta_k} U_k$. Ορίζουμε τον $A^{(k)}$, $1 \le k \le \mathrm{K}$, όπου κάθε στοιχείο του ισούται με $A_{ij}^{(k)} = \frac{\partial Z_{ij}}{\partial \theta_k}$. Εφόσον $Z = XR$ παίρνουμε τη σχέση $A^{(k)} = XU_{(1,k-1)} V_k U_{(k+1,\mathrm{K})}$. Se αυτό το σημείο είμαστε σε θέση να υπολογίσουμε τα στοιχεία ∇J ως εξής:

$$\frac{\partial J}{\partial \theta_k} = \sum_{i=1}^{n}\sum_{j=1}^{C} \frac{\partial}{\partial \theta_k}\frac{Z_{ij}^2}{M_i^2} - 1 = 2\sum_{i=1}^{n}\sum_{j=1}^{C} \frac{Z_{ij}}{M_i^2} A_{ij}^{(k)} - \frac{Z_{ij}^2}{M_i^3}\frac{\partial M_i}{\partial \theta_k}.$$ Αυτό που αξίζει να αναφέρουμε είναι ότι η σύγκλιση επιτυγχάνεται όταν η τιμή της παράστασης $1 - \alpha F_{kl}$ εντοπίζεται εντός του μοναδιαίου κύκλου, όπου $F_{kl} = \left[\frac{\partial^2 J}{\partial \theta_l \partial \theta_k}\right]_{\Theta=0}$.

Για τιμή του Θ=0 έχουμε ότι $Z_{ij} = 0$ για $j \ne m_i$, $Z_{im_i} = M_i$, και $\frac{\partial M_i}{\partial \theta_k} = \frac{\partial Z_{im_i}}{\partial \theta_k} = A_{im_i}^{(k)}$. Αν παραγωγίσουμε αυτή τη σχέση τότε παίρνουμε το ακόλουθο αποτέλεσμα: $\left[\frac{\partial^2 J}{\partial \theta_l \partial \theta_k}\right]_{ij|\Theta=0} = 2\sum_{i=1}^{n}\sum_{j\ne m_i} \frac{1}{M_i^2} A_{ij}^{(k)} A_{ij}^{(l)}$. Με αντικατάσταση

των τιμών για $A_{ij}^{(k)}|_{\Theta=0}$ καταλήγουμε στο ακόλουθο αποτέλεσμα: $F_{kl} = 2 \# i$ s.t. $m_i = i_k$ αν $k = l$ ή $F_{kl} = 0$ σε κάθε άλλη περίπτωση, όπου (i_k, j_k) είναι το ζεύγος (i, j) που αντιστοιχεί στο δείκτη k μέσα στο σχέδιο απεικόνισης που αναφέρθηκε πιο πάνω. Συνεπώς, θεωρώντας ικανοποιητικά μικρό α καταλήγουμε στο γεγονός ότι το αποτέλεσμα της σχέσης $1 - \alpha F_{kl}$ βρίσκεται εντός του μοναδιαίου κύκλου οπότε η σύγκλιση είναι εγγυημένη.

5.3 Περιγραφή του αλγορίθμου ZP

Αλγόριθμος ZP φασματικής συσταδοποίησης

Είσοδος: X // σύνολο N σημείων εισόδου $X = \{\chi_1, \ldots, \chi_n\}$, $\chi_i \in \Re^d$.

Έξοδος: Ομαδοποίηση των N σημείων σε C συστάδες

Βήμα 1: Υπολογισμός της τοπικής παραμέτρου σ_i για κάθε σημείο $\chi_i \in X$ χρησιμοποιώντας την εξίσωση (2).

Βήμα 2: Σχηματισμός του πίνακα γειτνίασης τοπικής κλιμάκωσης $\hat{A} \in \Re^{n \times n}$,όπου κάθε στοιχείο του $\hat{A}_{ij}$ ορίζεται σύμφωνα με την εξίσωση (1) για $i \neq j$ και $\hat{A}_{ii} = 0$.

Βήμα 3: Ορισμός του διαγωνίου πίνακα D με στοιχεία $D_{ii} = \sum_{j=1}^{n} \hat{A}_{ij}$ και κατασκευή του κανονικοποιημένου πίνακα γειτνίασης $L = D^{-\frac{1}{2}} \hat{A} D^{-\frac{1}{2}}$.

Βήμα 4: Υπολογισμός των C μεγαλύτερων διανυσμάτων του πίνακα L, $v_1, \ldots, v_c$, και δημιουργία του πίνακα $V = [v_1, \ldots, v_c] \in \Re^{n \times C}$, όπου C είναι ο μεγαλύτερος πιθανός αριθμός ομάδων.

Βήμα 5: Υπολογισμός της στροφής R που στοιχίζει όσο το δυνατόν καλύτερα τις στήλες του V με το κανονικό σύστημα συντεταγμένων χρησιμοποιώντας το σχήμα gradient descent.

Βήμα 6: Αξιολόγηση του κόστους της στοίχισης για κάθε πιθανή τιμή του πλήθους των ομάδων (μέχρι το μέγιστο αριθμό αυτών που ισούται με C), σύμφωνα με την εξίσωση (3).

Βήμα 7: Ο τελικός αριθμός ομάδων τίθεται ίσος με C_b και θεωρείται ο μέγιστος αριθμός ομάδων με το μικρότερο κόστος στοίχισης.

Βήμα 8: Εξαγωγή του αποτελέσματος στοίχισης Z των C_b μεγαλύτερων ιδιοδιανυσμάτων και ανάθεση του αρχικού σημείου χ_i στη συστάδα c αν και μόνο αν $\max_j \left(Z_{ij}^2\right) = Z_{ic}^2$.

Βήμα 9: Στην περίπτωση ύπαρξης δεδομένων με υψηλό θόρυβο, χρησιμοποιούμε το προηγούμενο αποτέλεσμα για την αρχικοποίηση του αλγορίθμου k-means ώστε να συσταδοποιήσουμε τις σειρές του Z.

5.4 Συμπέρασμα

Η συγκεκριμένη μέθοδος που χρησιμοποιείται από τον αλγόριθμο φασματικής συσταδοποίησης ZP έχει δύο επιθυμητές ιδιότητες. Πρώτον, αφού γίνει η στοίχιση με το κανονικό σύστημα συντεταγμένων, ο αριθμός των φορών που θα επαναληφθεί η διαδικασία συσταδοποίησης με τον αλγόριθμο k-means περιορίζεται. Συνήθως ο αριθμός των επαναλήψεων που απαιτούνται είναι ίσος με 100, ενώ εξαρτάται σημαντικά και από την αρχικοποίησή του. Δεύτερον, τα αποτελέσματα της συσταδοποίησης που προκύπτουν για όλες τις ενδεχόμενες τιμές των ομάδων υπολογίζονται με ελάχιστο επιπρόσθετο υπολογιστικό κόστος.

Η καλή αρχικοποίηση των αλγορίθμων συσταδοποίησης που χρησιμοποιούνται στο Βήμα 9 του αλγορίθμου ZP με τα αποτελέσματα του Βήματος 8.

Κεφάλαιο 6

Φασματική Συσταδοποίηση Κινηματογραφικών Πλάνων

6.1 Γενικά

Σε αυτό το κεφάλαιο προτείνεται ένας αλγόριθμος φασματικής συσταδοποίησης για τη συσταδοποίηση κινηματογραφικών πλάνων με βάση τις οπτικές ομοιότητες και τις χρονικές σχέσεις που ενδεχομένως υπάρχουν, και ο οποίος βασίζεται στον αλγόριθμο φασματικής συσταδοποίησης NJW.

6.2 Θεωρήσεις

6.2.1 Προσδιορισμός του αριθμού των συστάδων

6.2.1.1 Ο αριθμός των συστάδων βάσει του μέσου τετραγωνικού σφάλματος

Στον αλγόριθμο NJW όταν ο αριθμός των συστάδων K, ο οποίος υπολογίζεται πειραματικά, ανταποκρίνεται στον πραγματικό αριθμό των συστάδων K_{ideal} τότε οι γραμμές του πίνακα Y θα συσταδοποιηθούν σε K ορθογώνιες κατευθύνσεις. Συνεπώς τα K αρχικά κεντροειδή: $(Y_i^c)_{i=1,\ldots,K}$ στο Βήμα 5 του αλγορίθμου, μπορούν να επιλεχθούν με την ακόλουθη λογική. Προσδιορίζουμε πρώτα τη γραμμή του πίνακα Y για την οποία ισχύει ότι οι N πρώτοι της γείτονες σχηματίζουν τη πιο συμπαγή συστάδα, και έπειτα με αναδρομικό τρόπο επιλέγουμε τη γραμμή εκείνη της οποίας

το εσωτερικό γινόμενο με τα ήδη υπάρχοντα κεντροειδή είναι το μικρότερο, σύμφωνα με την ακόλουθη σχέση:

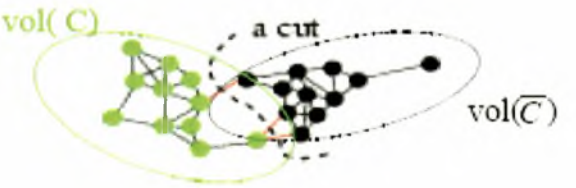
$$Y_{i+1}^{c} = \arg\min_{Y_j} \max_{(Y_l^c)_{l=1:i}} (Y_l^c \bullet Y_j) .$$

Ως γνωστόν αυτό που μας ενδιαφέρει στον αλγόριθμο k-means, που χρησιμοποιείται στο Βήμα 5 του αλγορίθμου NJW, είναι να ελαχιστοποιήσουμε το μέσο τετραγωνικό σφάλμα. Στην περίπτωση όπου ισχύει: $K < K_{ideal}$, η ορθογωνική ιδιότητα ίσως να μην ικανοποιείται γενικά και συνεπώς το μέσο τετραγωνικό σφάλμα μπορεί να είναι ή και να μην είναι το ελάχιστο σε αυτή την περίπτωση. Όταν $K > K_{ideal}$, η ορθογωνική ιδιότητα δεν ισχύει και οδηγούμαστε σε μια υπερσυσταδοποίηση της ιδανικής περίπτωσης. Συνεπώς δεν υπάρχει κάποια ξεκάθαρη υπόδειξη για το ποια θα είναι η συμπεριφορά του μέσου τετραγωνικό σφάλματος.

6.2.1.2 Ο αριθμός των συστάδων βάσει του φασματικού ιδιοκενού

Όπως έχει ήδη αναφερθεί, το κριτήριο του φασματικού ιδιοκενού πρόκειται για ένα σημαντικό μέτρο στις φασματικές μεθόδους. Ένας εναλλακτικός ορισμός από αυτόν που ήδη δόθηκε σε προηγούμενο κεφάλαιο είναι ο ακόλουθος: $\delta(A) = 1 - \frac{\lambda_2}{\lambda_1}$ όπου λ_1 και λ_2 είναι οι δύο μεγαλύτερες ιδιοτιμές του πίνακα A. Όπως έχει ήδη αναφερθεί το φασματικό ιδιοκενό χρησιμοποιείται για να αξιολογηθεί η σταθερότητα του μεγαλύτερου ιδιοδιανύσματος ή των k μεγαλύτερων ιδιοδιανυσμάτων ενός πίνακα, στην περίπτωση που η μεγαλύτερη ιδιοτιμή είναι πολλαπλότητας k. Το φασματικό ιδιοκενό σχετίζεται με μία σταθερά που δείχνει το πόσο συμπαγείς είναι οι συστάδες και ονομάζεται *σταθερά Cheeger*. Για τον προσδιορισμό αυτής της συσχέτισης ορίζουμε αρχικά την τιμή αποκοπής της διαμέρισης $(C, \overline{C})$ ενός γράφου με πίνακα γειτνίασης A ως εξής: $Cut_A(C, \overline{C}) = \sum_{i \in C} \sum_{j \notin C} A_{ij}$. Επίσης ορίζουμε το μέγεθος του υποσυνόλου C ως εξής: $Vol_A(C) = \sum_{i \in C} \sum_{j \in C} A_{ij}$. Επιπλέον ορίζουμε το μέγεθος ϕ για τη διαμέριση

$(C,\overline{C})$ ως εξής: $\phi_A(C)=\frac{Cut_A(C,\overline{C})}{\min(Vol_A(C),Vol_A(\overline{C}))}$. Από εδώ η *σταθερά Cheeger* h_G ορίζεται ως εξής: $h_G(A)=\min_C \phi_A(C)$ και ισχύει ότι $h_G(A)\geq\frac{1}{2}\delta(A)$. Αν θέλαμε να δώσουμε μια ποιοτική εξήγηση για τον ορισμό της σταθεράς Cheeger, θα μπορούσαμε να πούμε ότι το μέγεθος ϕ εκφράζει το πόσο καλά χωρίζει το σύνολο των κόμβων του γράφου η διαμέριση $(C,\overline{C})$, και η εύρεση του ελάχιστου πάνω στο σύνολο *C* αντιστοιχεί στην καλύτερη διαμέριση. Επομένως, αν υπάρχει μια διαμέριση για την οποία: α) τα βάρη A_{ij} των ακμών του γράφου κατά πλάτος της διαμέρισης, είναι μικρά και β) κάθε μία από τις περιοχές της διαμέρισης έχει ικανοποιητικό μέγεθος, τότε η σταθερά Cheeger θα είναι μικρή. Έτσι λοιπόν ξεκινώντας με τιμή K=1 θα θέλαμε να επιλέξουμε το πιο απλό μοντέλο συσταδοποίησης για το οποίο οι εξαγόμενες συστάδες είναι αρκετά συμπαγείς.



Εικόνα 7.2.2.1

Εδώ παρουσιάζεται η έννοια της σταθεράς Cheeger. Συγκεκριμένα αν η τιμή της σταθεράς είναι μικρή τότε υπάρχει μια καλή αποκοπή μέσα στο γράφο αλλιώς αν είναι μεγάλη τότε ο γράφος είναι αρκετά συμπαγής και είναι δύσκολο να διασπαστεί σε δύο υποσύνολα.

Κάτι τέτοιο ισοδυναμεί με ένα από τα ακόλουθα: α) η σταθερά Cheeger είναι αρκετά μεγάλη (*Εικόνα 7.2.2.1*) για κάθε μία συστάδα ή β) το φασματικό ιδιοκενό είναι μεγάλο για όλες τις συστάδες. Έτσι βάσει αυτών των θεωρήσεων προκύπτει το ακόλουθο κριτήριο: $\delta_K=\min_{i\in 1\ldots K}\delta\left(L\left(A_K^{(ii)}\right)\right)$ (1) , όπου $A_K^{(ii)}$ είναι οι υποπίνακες που εξάγονται από τον πίνακα *A* σύμφωνα με το μοντέλο που υποδεικνύεται από τον φασματικό αλγόριθμο και L είναι ο γνωστός κανονικοποιημένος πίνακας γειτνίασης

του αλγορίθμου NJW. Η τιμή του K στην εξίσωση (1) επιλέγεται έτσι ώστε το δ_K να ξεπερνά μια τιμή κατωφλίου.

6.2.1.3 Ο αριθμός των συστάδων βάσει της σχετικής αποκοπής

Το μέτρο που ορίστηκε σύμφωνα με την εξίσωση (1) λόγω του ότι λαμβάνει υπόψη μόνο την πληροφορία που εντοπίζεται εντός των συστάδων, παρουσιάζει ένα μειονέκτημα. Συγκεκριμένα, όταν σε ένα τμήμα των δεδομένων δεν είναι σαφής η διαχώριση των συστάδων τότε μπορεί να υπάρξει υπερεκτίμηση του αριθμού των συστάδων έτσι ώστε όλες οι συστάδες να θεωρούνται συμπαγείς. Από εδώ προκύπτει ένα νέο κριτήριο που χαρακτηρίζει τη συνολική ποιότητα μια συσταδοποίησης και ορίζεται ως εξής:

$$rcut_K = \frac{\sum_{k=1}^{K}\sum_{l=1,l\neq k}^{K}\sum_{i\in S_k}\sum_{j\in S_l} A_{ij}}{\sum_i\sum_j A_{ij}}$$

. Επιλέγεται το μεγαλύτερο K για το οποίο η τιμή του rcut βρίσκεται κάτω από ένα κατώφλι.

6.2.2 Τρόπος αναπαράστασης των κινηματογραφικών πλάνων και εξαγωγή χαρακτηριστικών

Τα κινηματογραφικά πλάνα συνήθως περιλαμβάνουν περισσότερες από μια απεικονίσεις, και αυτό συμβαίνει εξαιτίας της κίνησης που υπεισέρχεται λόγω της κάμερας. Συνεπώς μπορεί να χρειάζονται περισσότερα του ενός πλαίσια-κλειδιά (key-frames) για να αναπαρασταθεί η εσωτερική ποικιλία απεικόνισης που υπάρχει. Ο αριθμός και η ποιότητα των πλαισίων-κλειδιών σαφώς επηρεάζουν την απόδοση της συσταδοποίησης αλλά, στην παρούσα φάση θα θεωρήσουμε ότι ο αριθμός τους ισούται με 5 ($N_{kf} = 5$). Το i- οστό πλαίσιο-κλειδί f_i ενός βίντεο χαρακτηρίζεται από ένα ιστόγραμμα $h_i = \{h_{ij}\}$, όπου το j τρέχει πάνω από ένα σύνολο (r,g,b) από ομοιόμορφα κβαντισμένα χρώματα. Επίσης μπορεί να υπάρχει ανάγκη εξαγωγής

μόνο ενός τμήματος του πλαισίου οπότε τότε για το ιστόγραμμα h_i το j τρέχει πάνω από το σύνολο (r,g,b,p) , όπου p είναι μια συγκεκριμένη περιοχή του πλαισίου.

6.2.3 Υπολογισμός της ομοιότητας

Ο πίνακας γειτνίασης A κατασκευάζεται απευθείας από το σύνολο όλων των πλαισίων-κλειδιών που εντοπίζονται σε ένα βίντεο. Το μέτρο ομοιότητας μεταξύ των πλαισίων-κλειδιών θα πρέπει να αντανακλά τη γνώση που σχετίζεται με το συγκεκριμένο χώρο εφαρμογής που μελετάται. Για παράδειγμα στην περίπτωση των οικιακών βίντεο το περιεχόμενο δεν έχει κάποιους περιορισμούς και έτσι μέτρα ομοιότητας που βασίζονται σε απεικονίσεις καθολικών σκηνών είναι μια εύλογη και αρκετά καλή επιλογή. Στην περίπτωση όμως που έχουμε για παράδειγμα κάποιο βίντεο ποδοσφαίρου μπορούν να οριστούν πιο συγκεκριμένα μέτρα ομοιότητας.

6.2.3.1 Παράδειγμα υπολογισμού του πίνακα ομοιότητας για οικιακά βίντεο

Τα οικιακά βίντεο περιέχουν ακολουθίες από συνεχόμενα και χρονικά διατεταγμένα πλάνα που μπορούν να οργανωθούν σε ομάδες που συνήθως σχετίζονται με διακεκριμένες σκηνές. Η οπτική ομοιότητα και η χρονική διάταξη είναι δύο από τα βασικά κριτήρια που μας επιτρέπουν να προσδιορίσουμε τις συστάδες σε συλλογές από βίντεο, όταν δεν είναι γνωστή κάποια πληροφορία σχετικά με το περιεχόμενο τους. Έτσι ορίζουμε τον πίνακα γειτνίασης A ως εξής:

$A_{ij} = A_{ij}^{\upsilon} A_{ij}^{t}$ με $A_{ij}^{\upsilon} = e^{-\frac{d_{\upsilon}^2(f_i,f_j)}{2\sigma_{\upsilon}^2}}$ και $A_{ij}^{t} = e^{-\frac{d_t^2(f_i,f_j)}{2\sigma_t^2}}$, όπου A_{ij} είναι ο πίνακας γειτνίασης μεταξύ των πλαισίων-κλειδιών f_i και f_j, και d_{υ}, d_t είναι τα μέτρα οπτικής και χρονικής ομοιότητας, ενώ $\sigma_{\upsilon}^2, \sigma_t^2$ είναι η οπτική και η χρονική παράμετρος κλιμάκωσης αντίστοιχα.

Η οπτική ομοιότητα υπολογίζεται σύμφωνα με τη μετρική που βασίζεται στον *συντελεστή Bhattacharyya*, και πιο συγκεκριμένα $d_{\upsilon}(f_i, f_j) = (1 - \rho_{BT}(h_i, h_j))^{\frac{1}{2}}$, όπου με ρ_{BT} συμβολίζουμε τον συντελεστή Bhattacharyya. Ο συντελεστής Bhattacharyya ορίζεται ως εξής: $\rho_{BT} = \sum_k (h_{ik} h_{jk})^{\frac{1}{2}}$ με το άθροισμα να τρέχει πάνω από όλα τα χρώματα. Αξίζει να σημειωθεί ότι η συγκεκριμένη μετρική έχει αποδειχθεί ότι είναι αρκετά αξιόπιστη για τη σύγκριση χρωματικών κατανομών.

Η χρονική ομοιότητα εκμεταλλεύεται το γεγονός ότι απομακρυσμένα πλάνα κατά μήκος του χρονικού άξονα είναι λιγότερο πιθανό να ανήκουν στην ίδια σκηνή. Αν συμβολίσουμε με $d_t(f_i, f_j)$ μεταξύ των δύο πλαισίων-κλειδιών f_i και f_j τη χρονική ομοιότητα τότε αυτή ορίζεται ως εξής: $d_t(f_i, f_j) = \frac{\|f_j| - |f_i\|}{|\upsilon c|}$, όπου το $|f_i|$ υποδηλώνει τον απόλυτο αριθμό για το πλαίσιο f_i μέσα στο βίντεο, και $|\upsilon c|$ υποδηλώνει τη διάρκεια ολόκληρου του βίντεο σε αριθμό πλαισίων. Να σημειωθεί ότι για τις τιμές τόσο του d_{υ} όσο και του d_t το διάστημα από το οποίο λαμβάνουν τιμές είναι το [0,1].

6.2.3.2 Παράδειγμα υπολογισμού του πίνακα ομοιότητας για βίντεο ποδοσφαίρου

Στην περίπτωση των βίντεο ποδοσφαίρου, οι πληροφορίες χωρικής γνώσης μπορούν να χρησιμοποιηθούν για την ανάλυση του περιεχομένου τους και της χρονικής δομής. Στα πειράματα που παρουσιάζονται στο κεφάλαιο των πειραμάτων, χρησιμοποιούνται μόνο πληροφορίες καθολικής εμφάνισης, αγνοώντας άλλες χρήσιμες πληροφορίες όπως κίνηση κάμερας, δραστηριότητα κίνησης, εντοπισμός συγκεκριμένων περιοχών όπως για παράδειγμα το γρασίδι. Σε αυτή την περίπτωση εφαρμογής , επειδή τα απομακρυσμένα πλάνα μπορεί να ανήκουν στην ίδια σκηνή, ορίζουμε την ομοιότητα A_{ij} δύο πλαισίων-κλειδιών f_i και f_j μόνο ως εξάρτηση της οπτικής ομοιότητας, δηλαδή $A_{ij} = A_{ij}^{\upsilon}$.

6.3 Ο αλγόριθμος φασματικής συσταδοποίησης κινηματογραφικών πλάνων

Έχοντας κάνει όλες τις προηγούμενες θεωρήσεις είμαστε πλέον σε θέση να παρουσιάσουμε τον συγκεκριμένο αλγόριθμο φασματικής συσταδοποίησης κινηματογραφικών πλάνων.

Αλγόριθμος φασματικής συσταδοποίησης κινηματογραφικών πλάνων

Είσοδος: Τα πλαίσια-κλειδιά των κινηματογραφικών πλάνων

Έξοδος: Ομαδοποίηση των κινηματογραφικών πλάνων σε K συστάδες

Βήμα 1: Αρχικοποίηση του αριθμού των συστάδων K σε τιμή ίση με 1 και αναδρομική επανάληψη για αριθμό φορών K των Βημάτων 2 – 4.

Βήμα 2: Υπολογισμός των συστάδων $(S_1,\ldots,S_K)$ με βάση τον αλγόριθμο NJW.

Βήμα 3: Υπολογισμός της τιμής του δ_K ως το ελάχιστο από τα φασματικά ιδιοκενά των συστάδων S^i.

Βήμα 4: Τερματισμός όταν ισχύει δ_K = T (όπου T είναι το κατώφλι) και απόδοση του αποτελέσματος αλλιώς συνέχισε.

6.4 Συμπέρασμα

Ο συγκεκριμένος αλγόριθμος υποστηρίζει την αυτόματη επιλογή του αριθμού των συστάδων κάτι που όπως έχει ήδη αναφερθεί αποτελεί ένα ανοιχτό ερευνητικό θέμα. Επίσης τα αποτελέσματα που εξάγονται κατά τον τερματισμό του μπορούν να θεωρηθούν συγκρίσιμα με αυτά της ανθρώπινης εκτέλεσης.

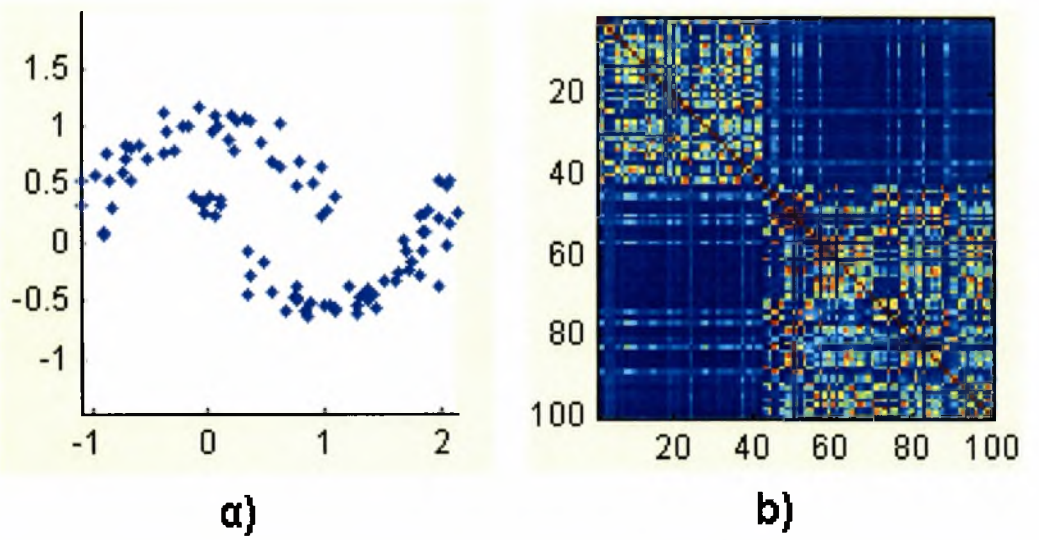
Η συγκεκριμένη μέθοδος μπορεί να βελτιωθεί ως προς τα αποτελέσματα που δίνει, αν χρησιμοποιηθούν καλύτερες αποστάσεις ομοιότητας μεταξύ των διαφόρων πλάνων ή εικόνων. Πιο συγκεκριμένα αυτό μπορεί να επιτευχθεί χρησιμοποιώντας στοιχεία όπως η κίνηση ή η υφή. Παρόλα αυτά, η εύρεση αποδοτικών τρόπων για το συνδυασμό αυτών των στοιχείων μέσα σε έναν πίνακα ομοιότητας καθώς και η επίδραση που θα έχει πάνω στον αλγόριθμο συσταδοποίησης ένα τέτοιο εγχείρημα, αποτελούν ακόμη ένα ανοιχτό ερευνητικό πεδίο. Ωστόσο σε μια συγκεκριμένη εφαρμογή μπορούν να επιλεγούν αποστάσεις ομοιότητας που ανταποκρίνονται στη φύση της εφαρμογής και μπορούν να οδηγήσουν τον αλγόριθμο σε καλύτερα και πιο ακριβή αποτελέσματα.

Κεφάλαιο 7

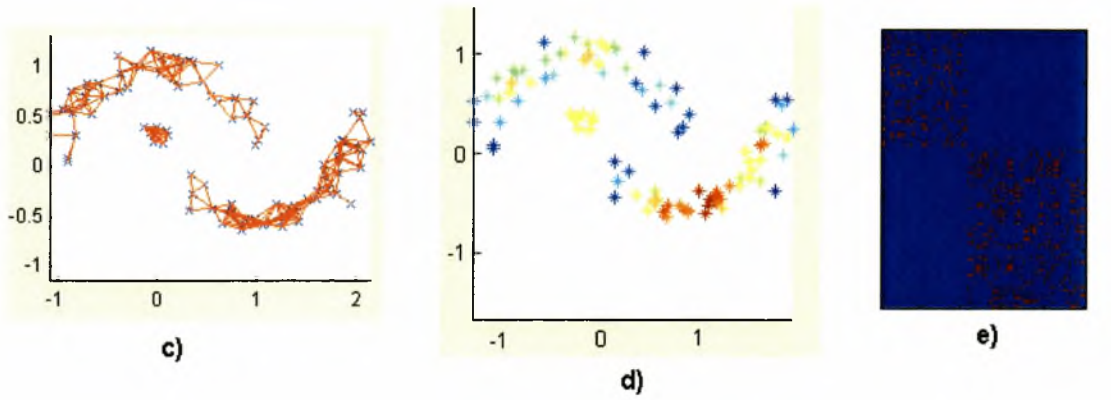
Πειραματικά Αποτελέσματα

7.1 Γράφοι ομοιότητας, πίνακες γειτνίασης και πίνακας ομοιότητας ενός συνόλου δεδομένων

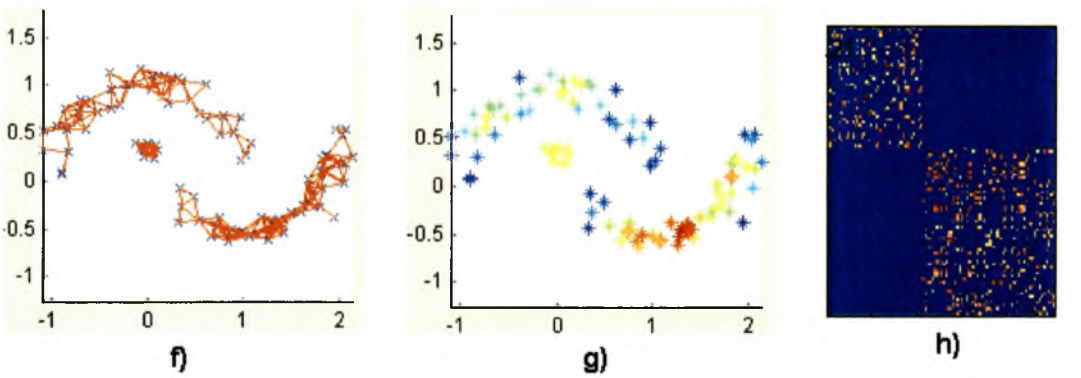
Στην παρακάτω αλληλουχία εικόνων φαίνονται οι διαφορετικοί γράφοι ομοιότητας, οι πίνακες γειτνίασης για κάθε γράφο και ο πίνακας ομοιότητας για ένα σύνολο δεδομένων που αποτελείται από 100 σημεία παρουσίας Γκαουσιανό θορύβου με διακύμανση 0.01 και με συνάρτηση ομοιότητας αυτή του Γκαουσιανού πυρήνα. Η τιμή της παραμέτρου σ της Γκαουσιανής συνάρτησης τέθηκε σε τιμή ίση με 0.5.



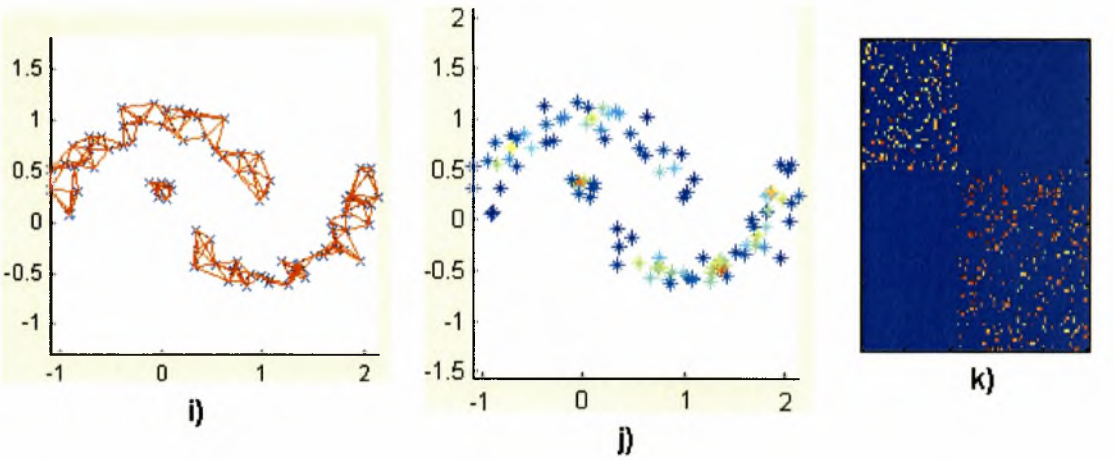
Εικόνα a) Τα αρχικά δεδομένα δύο διαστάσεων, Εικόνα b) Γραφική αναπαράσταση του πίνακα ομοιότητας των σημείων. Τα pixel που έχουν χρώμα ερυθρό υποδηλώνουν υψηλό βαθμό ομοιότητας ενώ αυτά που έχουν χρώμα μπλε υποδηλώνουν χαμηλό βαθμό ομοιότητας.



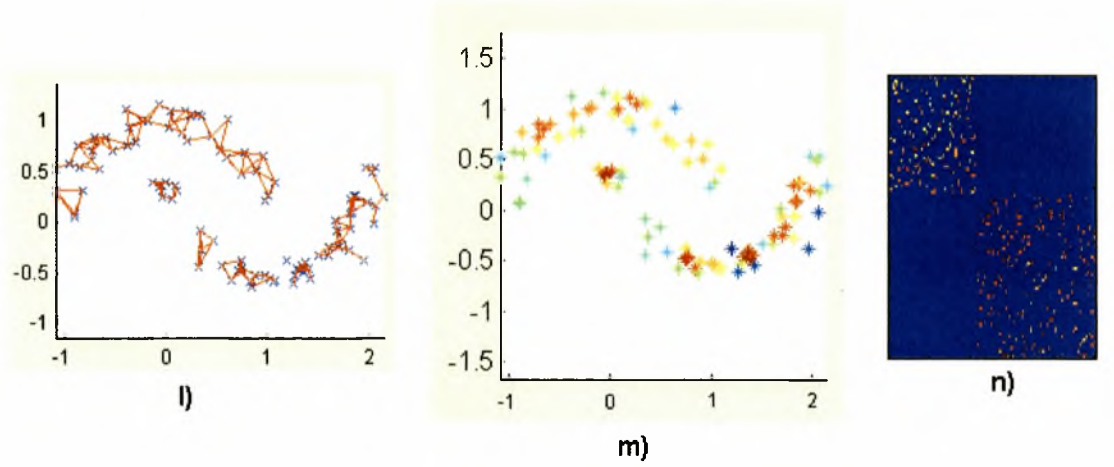
Εικόνα c) Ο μη βεβαρημένος γράφος της e-γειτονιάς για ε=0.27271, Εικόνα d) Απεικόνιση του βαθμού των κορυφών του μη βεβαρημένου γράφου e-γειτονιάς. Το ερυθρό χρώμα ισοδυναμεί με υψηλό βαθμό για την κορυφή ενώ το μπλε χρώμα ισοδυναμεί με χαμηλό βαθμό για την κορυφή, Εικόνα e) Ο πίνακας γειτνίασης για το μη βεβαρημένο γράφο της ε-γειτονιάς.



Εικόνα f) Ο βεβαρημένος γράφος της ε-γειτονιάς για ε=0.27271, Εικόνα g) Απεικόνιση του βαθμού κορυφών του βεβαρημένου γράφου της ε-γειτονιάς, Εικόνα h) Ο πίνακας γειτνίασης για το βεβαρημένο γράφο της e-γειτονιάς.



Εικόνα i) Ο συμμετρικός γράφος των k κοντινότερων γειτόνων για k=5, Εικόνα j) Απεικόνιση του βαθμού των κορυφών του συμμετρικού γράφου των k κοντινότερων γειτόνων, Εικόνα k) Ο πίνακας γειτνίασης για το συμμετρικό γράφο των k κοντινότερων γειτόνων.



Εικόνα l) Ο αμοιβαίος γράφος των k κοντινότερων γειτόνων για k=5. Παρατηρούμε, όπως είναι και φυσικό άλλωστε, ότι ο αμοιβαίος γράφος των k κοντινότερων γειτόνων είναι υποσύνολο του αντίστοιχου συμμετρικού, Εικόνα m) Απεικόνιση του βαθμού των κορυφών του αμοιβαίου γράφου των k κοντινότερων γειτόνων, Εικόνα n) Ο πίνακας γειτονιάς για τον αμοιβαίο γράφο των k κοντινότερων γειτόνων.

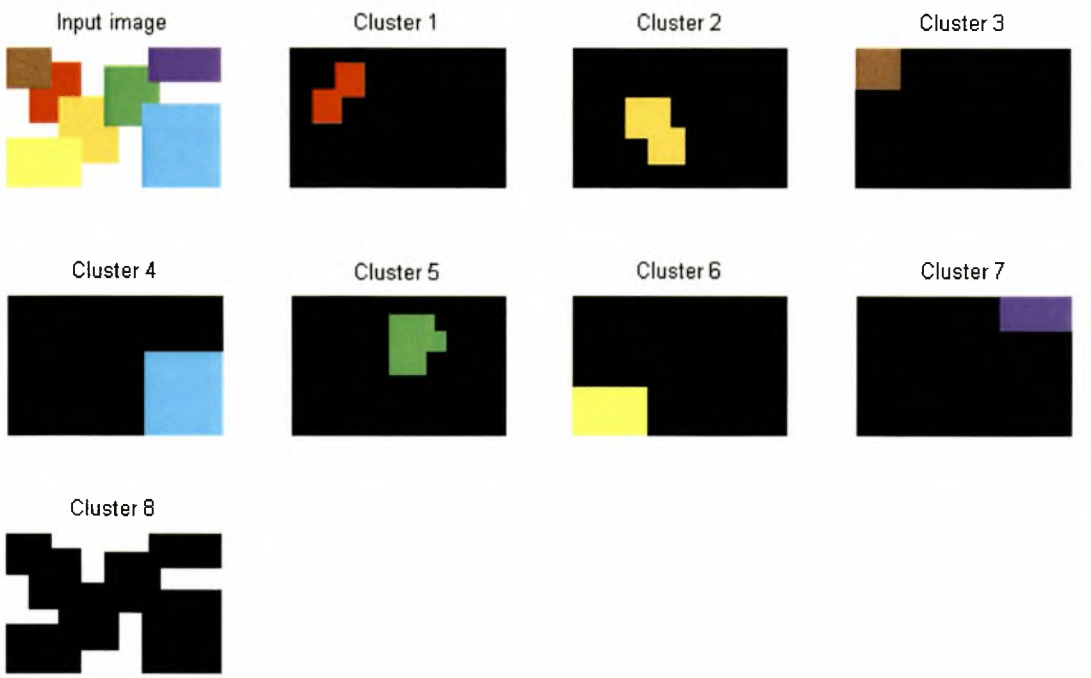
7.2 Υπολογισμός του αριθμού των συστάδων στον αλγόριθμο NJW

7.2.1 Θεωρήσεις

Στα ακόλουθα πειράματα που σχετίζονται με τον αλγόριθμο NJW ο k-means αλγόριθμος της εφαρμογής Matlab εκτελέστηκε με βάση τις ακόλουθες τιμές για τις αντίστοιχες παραμέτρους:

1. Η τιμή της παραμέτρου *EmptyAction* τέθηκε ίση με *singleton*. Η έννοια της συγκεκριμένης αρχικοποίησης είναι ότι αν υπάρχει κάποιο σημείο μέσα στο σύνολο δεδομένων που εντοπίζεται σε αρκετή απόσταση μακριά από το κεντροειδές τότε δημιουργείται μια ξεχωριστή συστάδα για αυτό το σημείο.
2. Η τιμή της παραμέτρου *Replicates* τέθηκε ίση με *30*. Η συγκεκριμένη τιμή υποδηλώνει των αριθμό των φορών που θα επαναληφθεί η συσταδοποίηση με τη χρήση διαφορετικού συνόλου κεντροειδών για την κάθε νέα αρχικοποίηση του αλγορίθμου.
3. Η τιμή της παραμέτρου *Maxiter* τέθηκε ίση με *400*. Με αυτό τον τρόπο υποδηλώνεται ο αριθμός των επαναλήψεων του αλγορίθμου k-means ώστε να επιτευχθεί η επιθυμητή σύγκλιση.
4. Η τιμή της παραμέτρου *Start* τέθηκε ίση με *cluster*. Η τιμή *cluster* υποδηλώνει ότι επιτελείται μια αρχική συσταδοποίηση σε ένα δείγμα του συνόλου δεδομένων, της τάξης του 10%, ούτως ώστε να υπολογιστούν οι θέσεις των αρχικών κεντροειδών.
5. Η τιμή της παραμέτρου *Distance* τέθηκε ίση με *sqEuclidean*. Πρόκειται για το μέτρο απόστασης που θα χρησιμοποιήσει ο αλγόριθμος k-means στο χώρο διάστασης p, ώστε να αποφανθεί κατά πόσο δύο σημεία του χώρου είναι όμοια μεταξύ τους ή όχι. Στη συγκεκριμένη περίπτωση το μέτρο απόστασης, που χρησιμοποιείται, είναι η τετραγωνική Ευκλείδεια απόσταση.

7.2.2 Σχέση της πολλαπλότητας των ιδιοτιμών του κανονικοποιημένου πίνακα γειτνίασης και του αριθμού των συστάδων

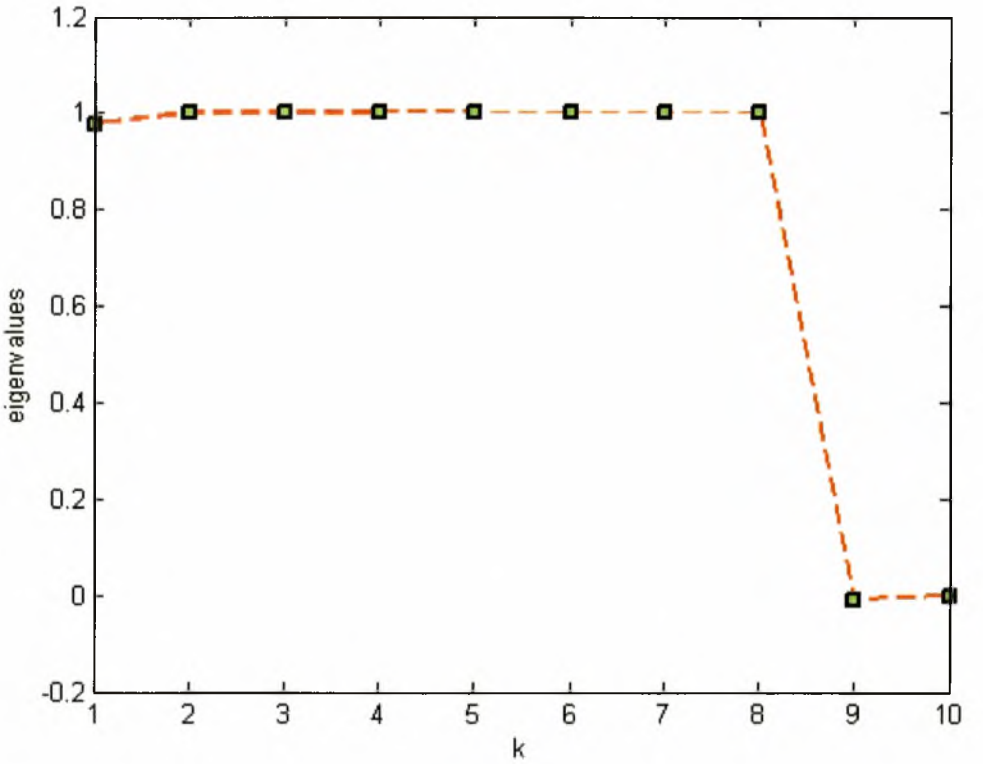


Εικόνα 7.2.1

Η αρχική εικόνα εισόδου και οι 8 συστάδες που δημιουργήθηκαν μετά την εκτέλεση του αλγορίθμου φασματικής συσταδοποίησης NJW για τιμή της καθολικής παραμέτρου σ ίση με 0.06. Η τιμή της παραμέτρου υπολογίστηκε πειραματικά.

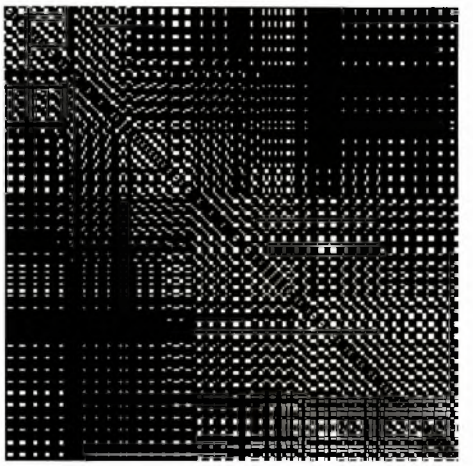
Αν και ο αριθμός των συστάδων στο συγκεκριμένο παράδειγμα είναι εύκολο να προσδιοριστεί, εντούτοις, μπορεί να αποδειχθεί πειραματικά, ότι ο αριθμός των συστάδων είναι ίσος με 8 με βάση τη θεωρία του βαθμού πολλαπλότητας της μεγαλύτερης ιδιοτιμής του κανονικοποιημένου πίνακα γειτνίασης L. Στην *Εικόνα 7.2.2* απεικονίζονται οι 10 μεγαλύτερες ιδιοτιμές του κανονικοποιημένου πίνακα γειτνίασης L που αντιστοιχούν στην πιο πάνω είσοδο του πειράματος. Αυτό που παρατηρούμε είναι ότι η μεγαλύτερη ιδιοτιμή του πίνακα L, που ισούται πάντα με 1, έχει πολλαπλότητα 8 (αν και η πρώτη ιδιοτιμή δε φαίνεται να είναι ακριβώς ίση με 1,

αυτό έχει να κάνει με το σφάλμα υπολογισμού της εφαρμογής Matlab), οπότε ο αριθμός των τελικών συστάδων θα ισούται με 8.



Εικόνα 7.2.2

Στην *Εικόνα 7.2.3* ο πίνακας γειτνίασης που παρουσιάζεται, αντιστοιχεί στην προηγούμενη εικόνα εισόδου και δεν προσεγγίζει κατά πολύ τη μορφή ενός διαγωνίου μπλοκ πίνακα. Η μορφή μπλοκ πίνακα γειτνίασης παρουσιάζεται όταν οι συστάδες στα δεδομένα μας έχουν τη τάση να είναι καλά σχηματισμένες και διαχωρισμένες πριν ακόμα εκτελεστεί η πράξη της συσταδοποίησης. Τα δεδομένα εισόδου δεν συμμορφώνονται με αυτή την υπόθεση καθώς οι 7 συστάδες τις εικόνας βρίσκονται εντός μιας μεγαλύτερης συστάδας που έχει χρώμα λευκό. Συνεπώς η μορφή που έχει ο συγκεκριμένος πίνακας γειτνίασης δεν υποδηλώνει κάτι για το πώς είναι διατεταγμένα τα δεδομένα εισόδου.



Εικόνα 7.2.3

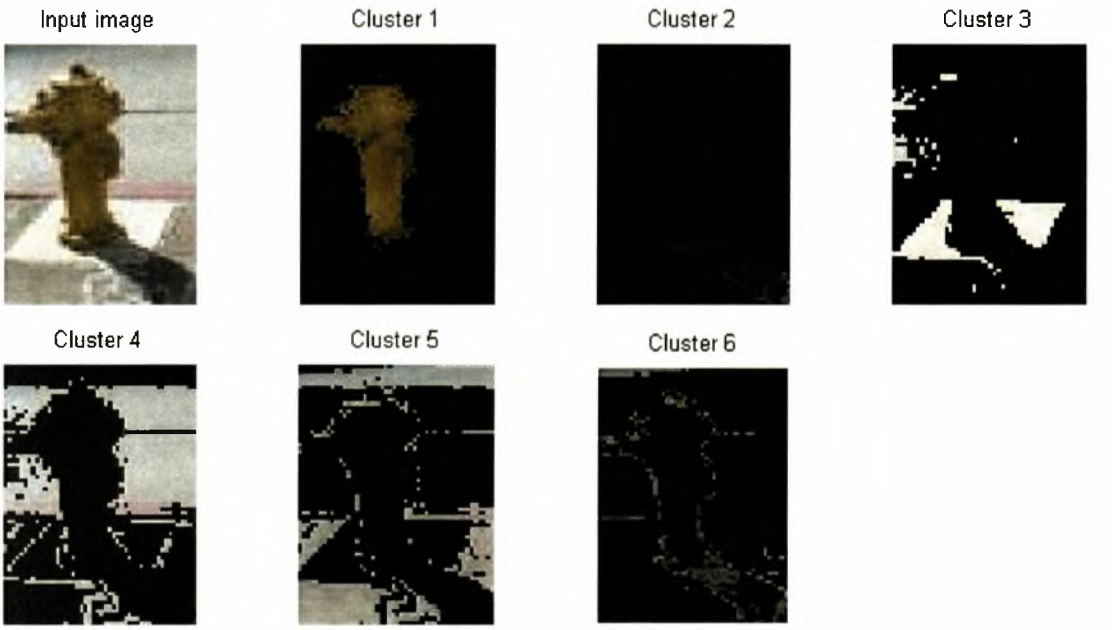
Η αποτελεσματικότητα του φασματικού ιδιοκενού εξαρτάται κατά πολύ από την μορφή που έχει ο πίνακας γειτνίασης. Αν η μορφή του δεν είναι κοντά στη μορφή ενός διαγωνίου μπλοκ πίνακα τότε το ευρετήριο του φασματικού ιδιοκενού ενδεχομένως να αποτύχει κατά τον υπολογισμό του ακριβή αριθμού των συστάδων.

7.2.3 Ο αριθμός των συστάδων βάσει του φασματικού ιδιοκενού

7.2.3.1 Παράδειγμα εξαγωγής αξιόπιστων αποτελεσμάτων βάσει του φασματικού ιδιοκενού

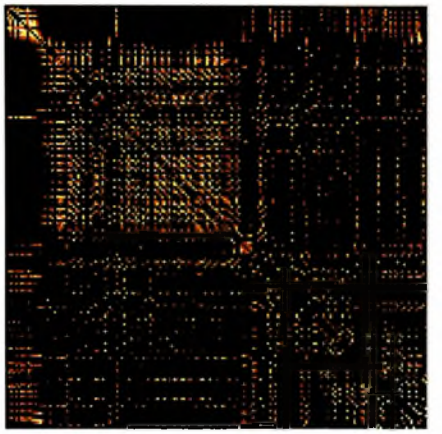
Στη συνέχεια παρουσιάζουμε μια πιο ρεαλιστική είσοδο στον αλγόριθμο NJW που επιβεβαιώνει την αποτελεσματικότητα της θεωρίας του φασματικού ιδιοκενού, ενώ αμέσως μετά παρουσιάζεται μια άλλη είσοδος στον ίδιο αλγόριθμο, όπου το ευρετήριο του φασματικού ιδιοκενού αποτυγχάνει. Και στις δύο εισόδους αυτό που μπορεί εύκολα να παρατηρηθεί είναι ότι και οι δύο εικόνες δεν χαρακτηρίζονται από κάποια σχετική ομαλότητα στα χρώματα, αλλά, σε ορισμένες περιοχές τους εμφανίζεται κάποιο χρώμα ως «θόρυβος» μέσα σε κάποιο άλλο που επικρατεί στην εκάστοτε περιοχή της εικόνας. Εκτός όμως από το συγκεκριμένο είδος θορύβου, στο εσωτερικό της κάθε συστάδας που έχει δημιουργηθεί, διαφαίνεται και ένα άλλο είδος θορύβου. Συγκεκριμένα υπάρχει και ένα σχετικά μικρό ποσοστό από στοιχεία της εικόνας που θα έπρεπε να έχουν τοποθετηθεί σε κάποια άλλη συστάδα, αλλά αυτό

συμβαίνει εξαιτίας δύο σημαντικών παραγόντων που επηρεάζουν άμεσα τα αποτελέσματα του πειράματος: 1) ο συγκεκριμένος αλγόριθμος προτείνει τη χρήση καθολικής παραμέτρου s και όχι αντίστοιχων τοπικών 2) ο ίδιος ο αλγόριθμος συσταδοποίησης k-means, που χρησιμοποιείται στο τελευταίο στάδιο του αλγορίθμου, επιβάλλει από μόνος του κάποιο σφάλμα υπολογισμού καθώς εξαρτάται κατά πολύ από τα δεδομένα αρχικοποίησης με τα οποία θα τροφοδοτηθεί.



Εικόνα 7.2.4

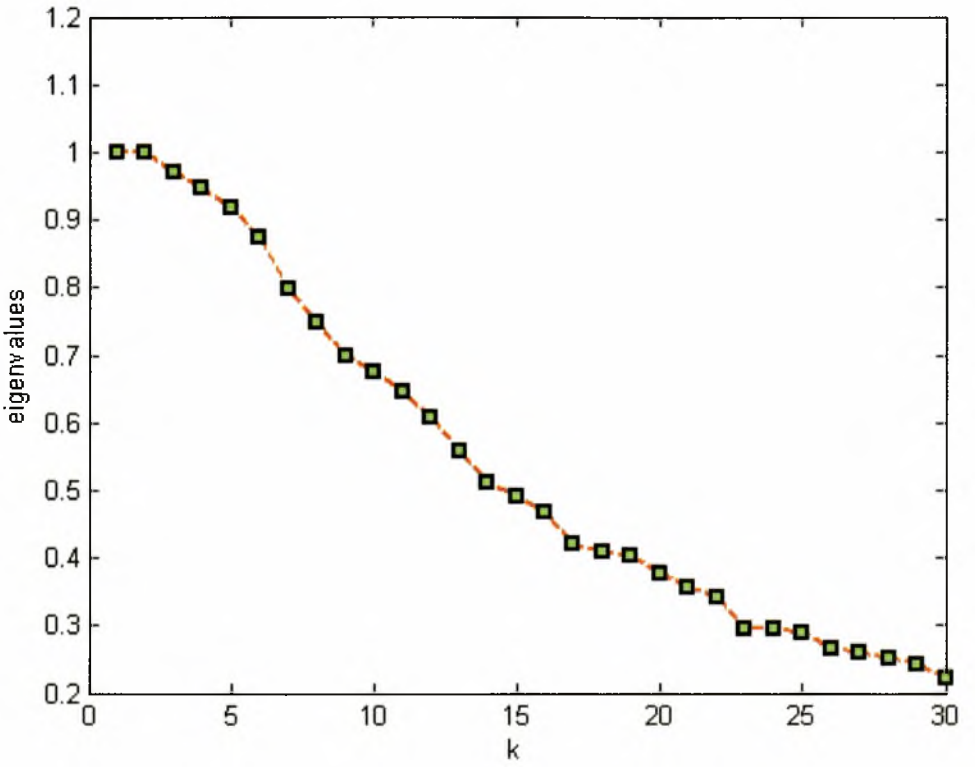
Η αρχική εικόνα εισόδου και οι 6 συστάδες που δημιουργήθηκαν μετά την εκτέλεση του αλγορίθμου φασματικής συσταδοποίησης NJW για τιμή της καθολικής παραμέτρου σ ίση με 0.08. Η τιμή της παραμέτρου υπολογίστηκε πειραματικά.



Εικόνα 7.2.5

Ο πίνακας γειτνίασης για την προηγούμενη εικόνα εισόδου. Όπως φαίνεται η μορφή του συγκεκριμένου πίνακα εισόδου πλησιάζει πολύ τη μορφή ενός πίνακα τύπου διαγωνίου μπλοκ. Συνεπώς το κριτήριο του φασματικού ιδιοκενού θα δώσει καλά αποτελέσματα σε αυτή την περίπτωση.

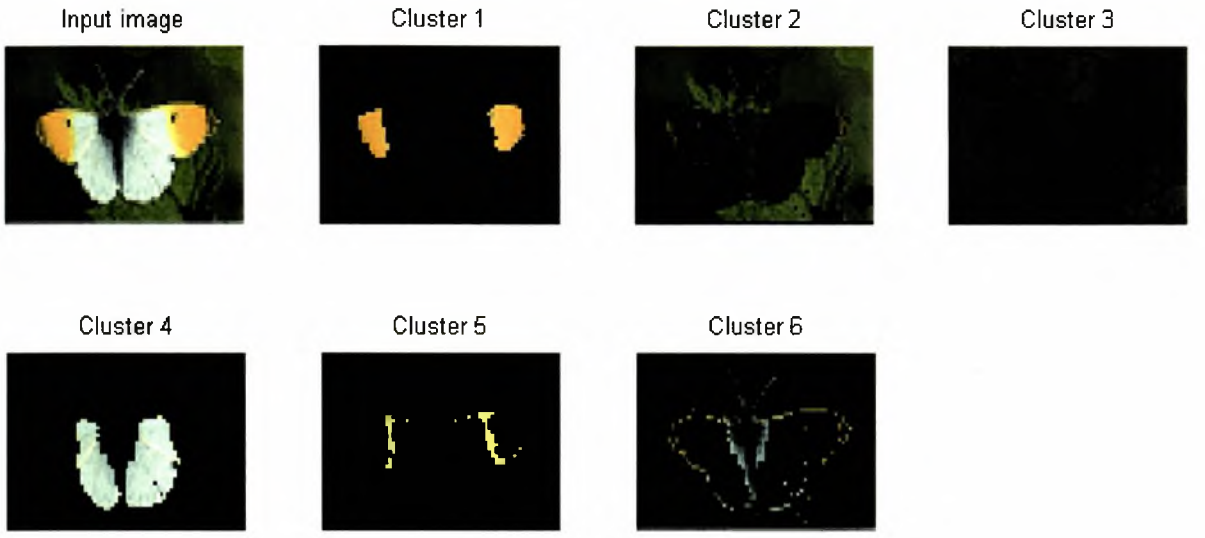
Στην *Εικόνα 7.2.6* απεικονίζονται οι 30 μεγαλύτερες ιδιοτιμές του πίνακα L. Αυτό που παρατηρούμε είναι ότι μεταξύ των ιδιοτιμών λ_6 και λ_7 η ποσότητα $\Delta_\kappa = |\lambda_\kappa - \lambda_{\kappa+1}|$ μεγιστοποιείται. Συνεπώς ο αριθμός των τελικών συστάδων της εικόνας θα είναι 6.



Εικόνα 7.2.6

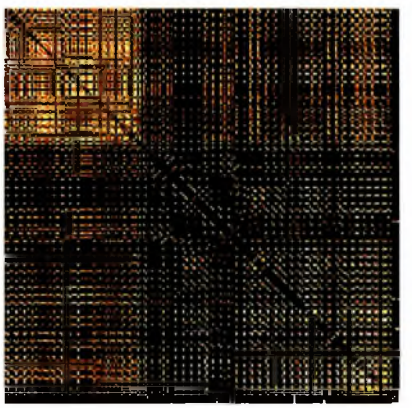
7.2.3.2 Παράδειγμα εξαγωγής μη αξιόπιστων αποτελεσμάτων βάσει του φασματικού ιδιοκενού

Στη συνέχεια παρουσιάζουμε μια εικόνα εισόδου για την οποία η χρήση της θεωρίας του φασματικού ιδιοκενού αποτυγχάνει να προσδιορίσει μια καλή προσέγγιση για τον αριθμό των τελικών συστάδων.

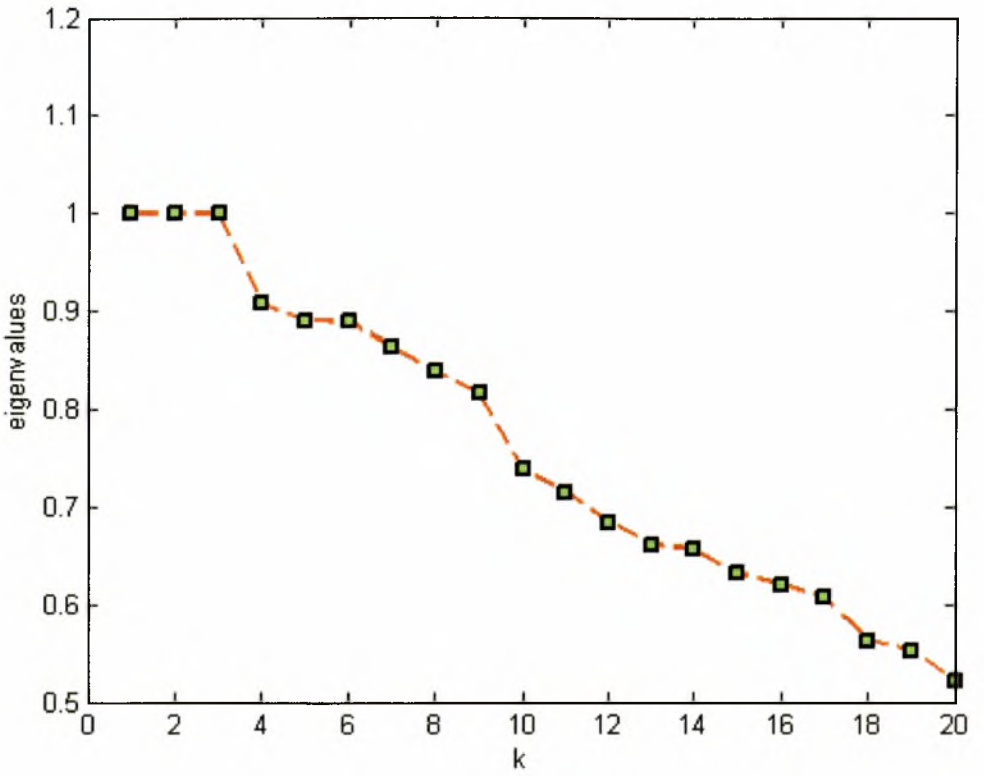


Εικόνα 7.2.7

Η αρχική εικόνα εισόδου και οι 6 συστάδες που δημιουργήθηκαν μετά την εκτέλεση του αλγορίθμου NJW για τιμή καθολικής παραμέτρου ίση με 0.06. Η τιμή της παραμέτρου υπολογίστηκε πειραματικά.



Ο πίνακας γειτνίασης για την παραπάνω εικόνα εισόδου. Όπως φαίνεται η μορφή του πίνακα γειτνίασης δε παραπέμπει στη μορφή ενός πίνακα τύπου διαγωνίου μπλοκ, οπότε είναι σίγουρο ότι το κριτήριο του φασματικού ιδιοκενού σε αυτή την περίπτωση αποτυγχάνει κατά τον υπολογισμό του ακριβή αριθμού των συστάδων.

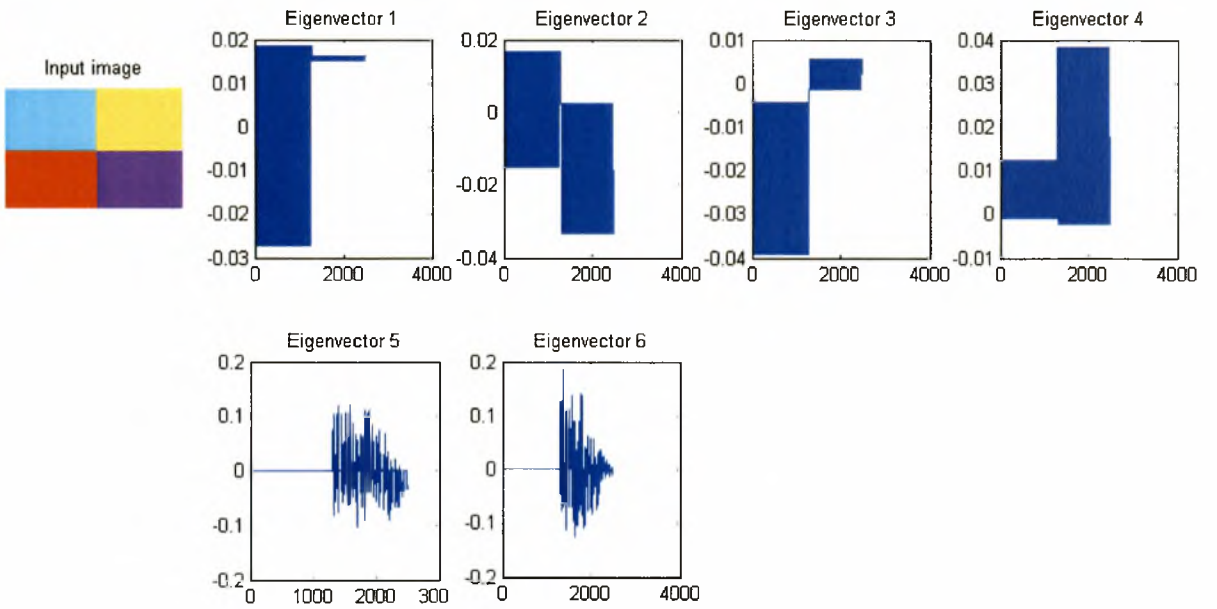


Εικόνα 7.2.8

Η Εικόνα 7.2.8 απεικονίζει της 20 μεγαλύτερες ιδιοτιμές του πίνακα L για την παραπάνω εικόνα εισόδου. Αυτό που παρατηρούμε είναι ότι μεταξύ των ιδιοτιμών λ_3 *και* λ_4 *η ποσότητα* $\Delta_\kappa = \left|\lambda_\kappa - \lambda_{\kappa+1}\right|$ *μεγιστοποιείται. Συνεπώς ο αριθμός των τελικών συστάδων σύμφωνα με τη θεωρία του φασματικού ιδιοκενού ισούται με 3, κάτι που προφανώς δεν ισχύει.*

7.3 Πληροφορία που δίνουν τα c μεγαλύτερα ιδιοδιανύσματα στον αλγόριθμο NJW σχετικά με την κατανομή των δεδομένων

Πολλές φορές ανάλογα με την εικόνα εισόδου, μπορούμε να εξάγουμε κάποιο συμπέρασμα για την διάταξη των στοιχείων μιας εικόνας μέσω των c μεγαλύτερων ιδιοδιανυσμάτων του πίνακα L.

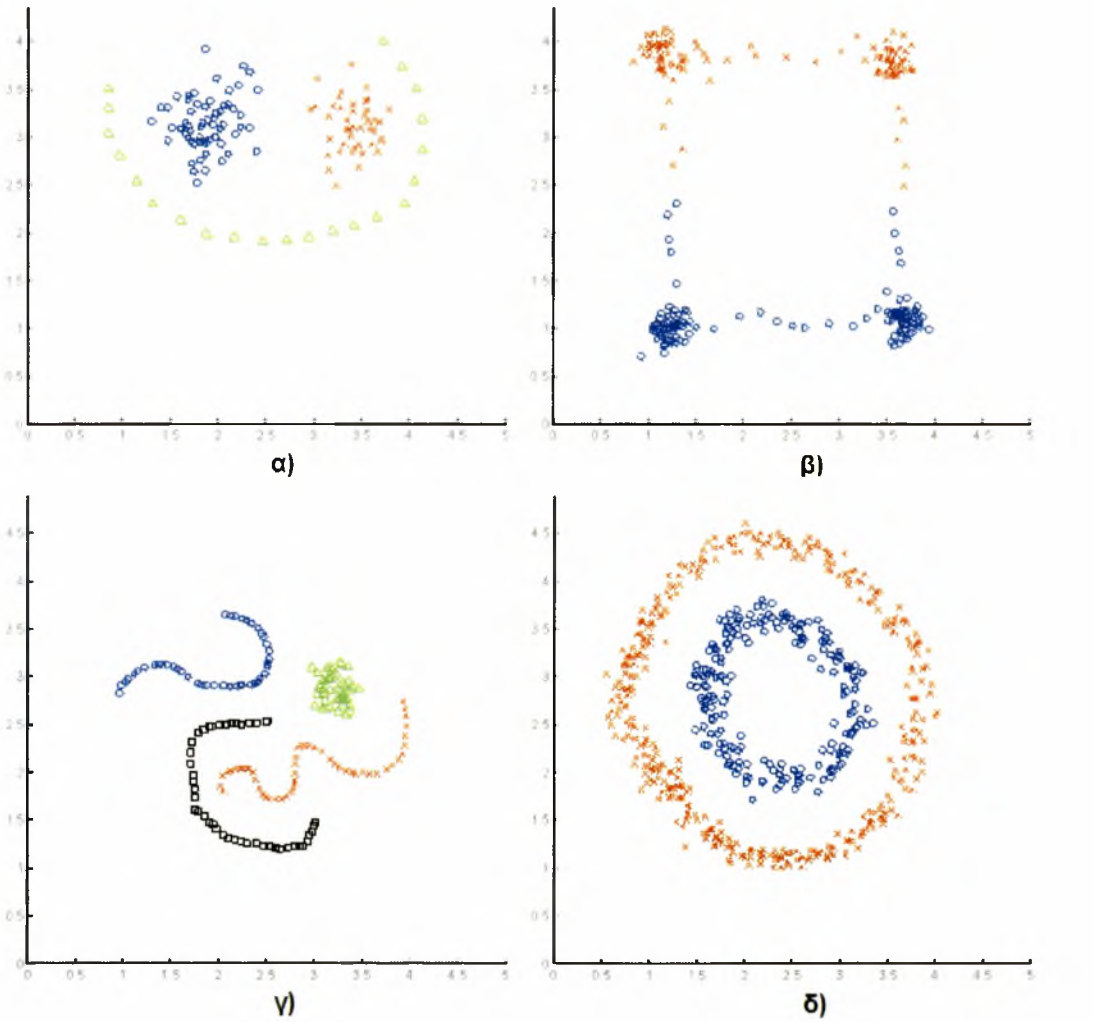


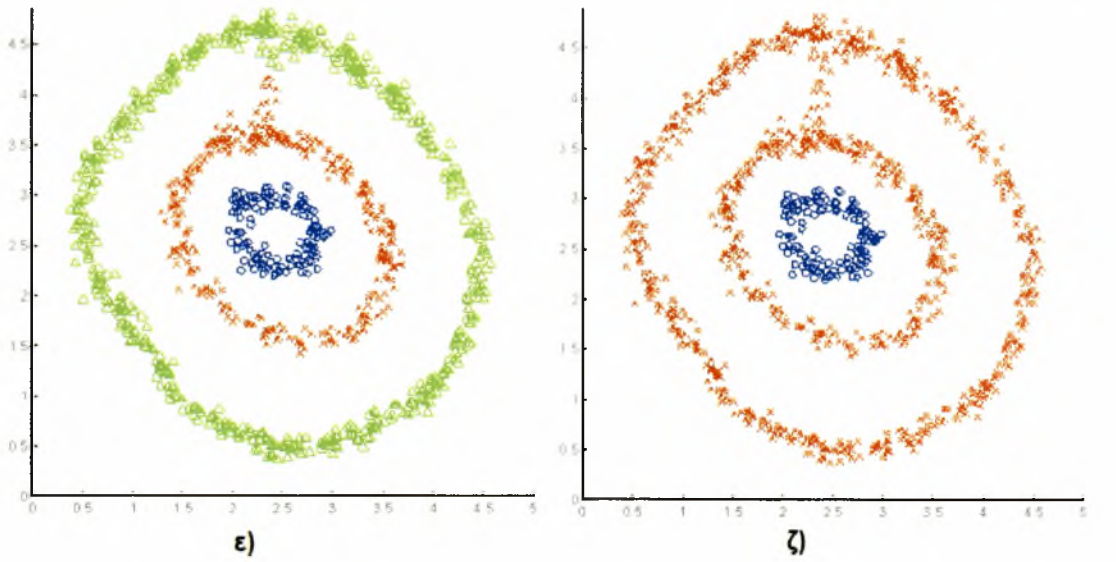
Εικόνα 7.2.9

Η αρχική εικόνα εισόδου για τον αλγόριθμο NJW και τα 6 μεγαλύτερα ιδιοδιανύσματα του πίνακα L που υπολογίζονται από τον αλγόριθμο. Όπως φαίνεται τα 4 μεγαλύτερα ιδιοδιανύσματα είναι αυτά που μας παρέχουν τη περισσότερη πληροφορία για το πώς είναι ο σχετικός διαχωρισμός των δεδομένων εισόδου. Επίσης το 5^o και το 6^o ιδιοδιάνυσμα παρουσιάζονται για να υποδείξουν ότι δεν παρέχουν κάποια αξιόλογη πληροφορία σχετικά με τα δεδομένα και πως ο κατάλληλος αριθμός μεγαλύτερων ιδιοδιανυσμάτων είναι αυτός που δίνει τη πιο σημαντική πληροφορία για τα στοιχεία (pixel) μιας εικόνας.

7.4 Περιπτώσεις συνόλων δεδομένων όπου ο αλγόριθμος k-means αποτυγχάνει στην συσταδοποίηση

Σε αυτή την ενότητα παρουσιάζονται δεδομένα εισόδου, όπου δεν αποτελούν κυρτές περιοχές από τη φύση τους, και δεν μπορούν να συσταδοποιηθούν επιτυχώς με την άμεση εφαρμογή του αλγορίθμου k-means, γι' αυτό και χρησιμοποιούνται τεχνικές φασματικής συσταδοποίησης. Στην ακόλουθη εικόνα παρουσιάζονται ορισμένα παραδείγματα μη κυρτών περιοχών.



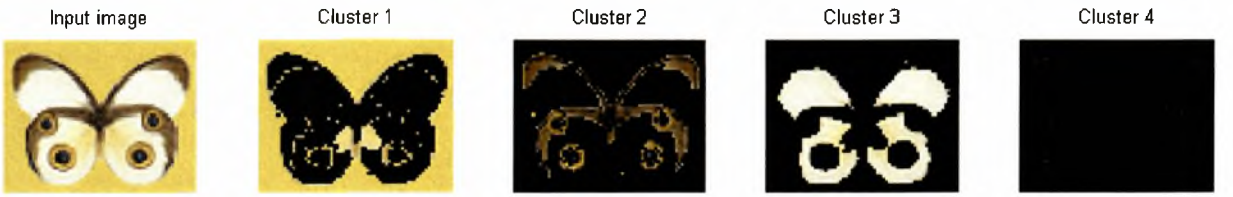


Για τα διάφορα δεδομένα εισόδου των εικόνων (α) – (ζ) έχει εφαρμοστεί ο αλγόριθμος NJW και έχουν προκύψει τα συσταδοποιημένα σύνολα δεδομένων που αποτελούνται αντίστοιχα από: α) 3 συστάδες β) 2 συστάδες γ) 4 συστάδες δ) 2 συστάδες ε) 3 συστάδες ζ) 2 συστάδες. Κάθε φορά ο αριθμός των επιθυμητών συστάδων δινόταν ως είσοδος από τον ίδιο το χρήστη.

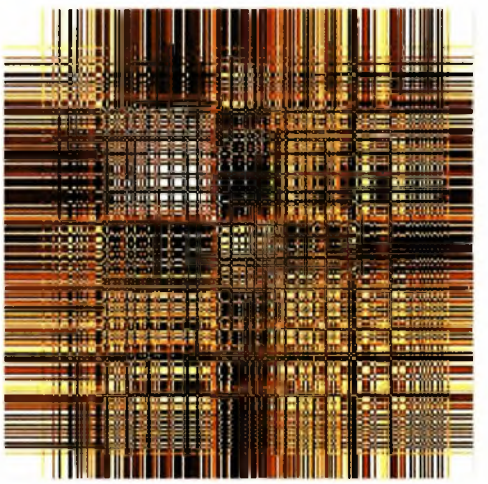
7.5 Φασματική συσταδοποίηση με χρήση των μεθόδων των Fischer και Poland

7.5.1 Συσταδοποίηση με χρήση του πίνακα αγωγιμότητας C

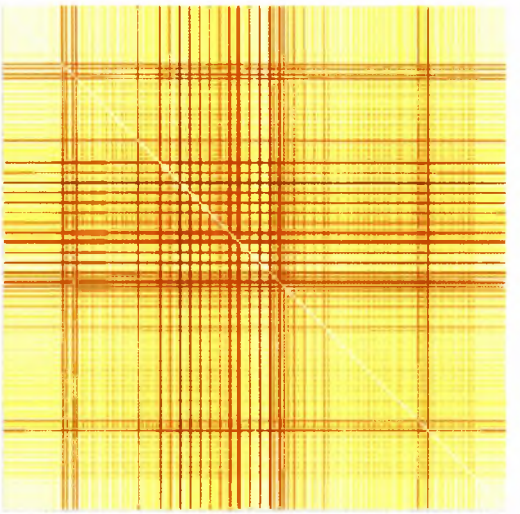
Σε αυτή την ενότητα παρουσιάζεται το αποτέλεσμα της εκτέλεσης του αλγορίθμου φασματικής συσταδοποίησης αλλά με τη χρήση του πίνακα αγωγιμότητας αντί του κανονικοποιημένου πίνακα γειτνίασης L. Επίσης γίνεται απεικόνιση του πίνακα γειτνίασης A και του πίνακα αγωγιμότητας C. Τέλος γίνεται η απεικόνιση των 4 μεγαλύτερων ιδιοδιανυσμάτων του πίνακα αγωγιμότητας C,



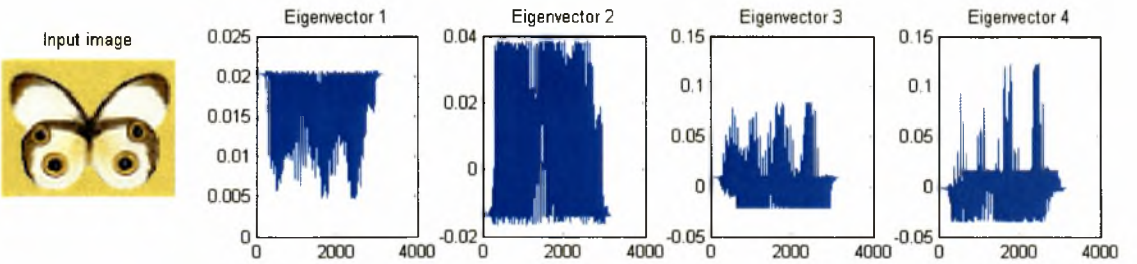
Η αρχική εικόνα εισόδου και οι 4 συστάδες που δημιουργήθηκαν ύστερα από την εφαρμογή του αλγορίθμου φασματικής συσταδοποίησης ο οποίος χρησιμοποιεί τον πίνακα αγωγιμότητας C και όχι τον πίνακα Laplace, ενώ η τιμή της καθολικής παραμέτρου σ για τη συνάρτηση του Γκαουσιανού πυρήνα τέθηκε ίση με 79. Αυτό που παρατηρούμε είναι ότι για να δώσει αυτό το αποτέλεσμα συσταδοποίησης ο συγκεκριμένος αλγόριθμος χρειάστηκε να θέσουμε την τιμή του σ σε μια αρκετά μεγάλη τιμή, ώστε να επιτευχθεί σύγκλιση του αλγορίθμου.



Ο πίνακας γειτνίασης A που προκύπτει για την παραπάνω είσοδο-εικόνα. Ο συγκεκριμένος πίνακας δεν έχει κάποια ευδιάκριτη δομή διαγωνίου μπλοκ πίνακα.



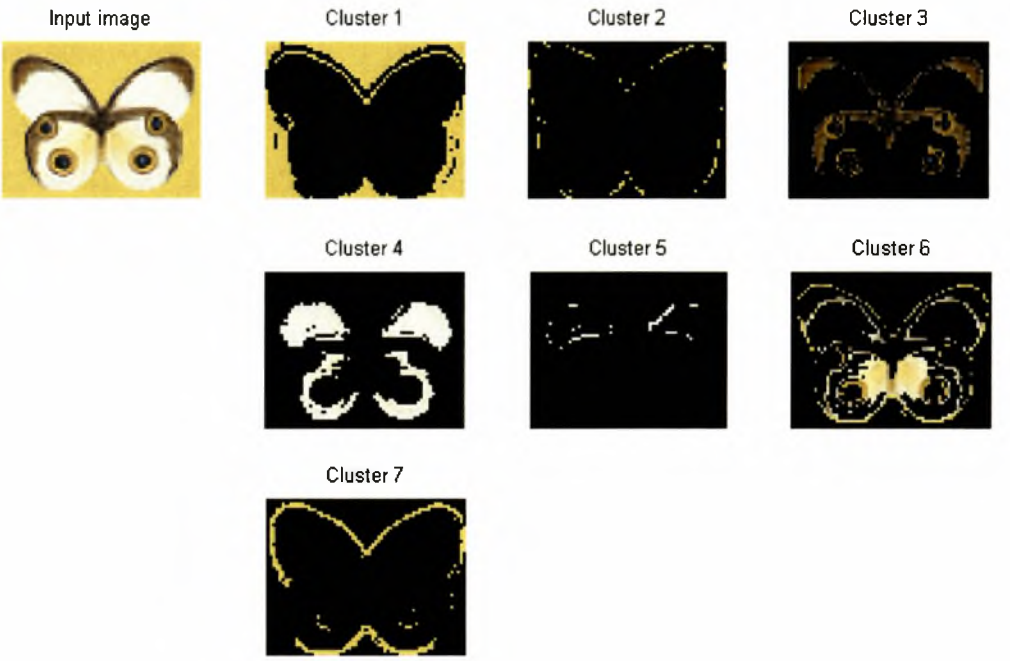
Ο πίνακας αγωγιμότητας C που αντιστοιχεί στην παραπάνω είσοδο-εικόνα και κατασκευάστηκε με βάση τον πίνακα γειτνίασης A.



Απεικόνιση των τεσσάρων μεγαλύτερων ιδιοδιανυσμάτων του πίνακα αγωγιμότητας C, από τα οποία το 1^o,το 3^o και το 4^o είναι αυτά που μας δίνουν την περισσότερη πληροφορία για το πως είναι διαχωρισμένα τα δεδομένα μεταξύ τους.

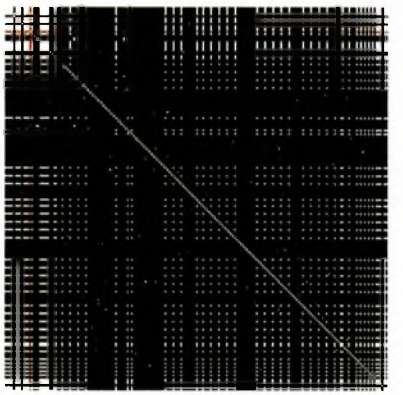
7.5.2 Ασυμμετρική φασματική συσταδοποίηση

Σε αυτή την ενότητα παρουσιάζεται ένα παράδειγμα εκτέλεσης του αλγορίθμου ασυμμετρικής φασματικής συσταδοποίησης που πρότειναν οι Fischer και Poland. Συγκεκριμένα παρουσιάζονται οι συστάδες που προέκυψαν για μια εικόνα εισόδου, ο πίνακας $\widetilde{A}$, ο συμμετρικός πίνακας γειτνίασης A και ο πίνακας αγωγιμότητας C.

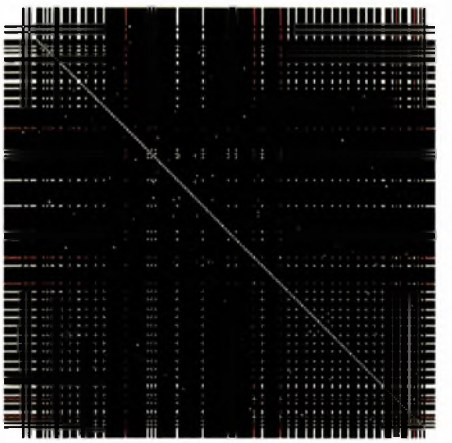


Η αρχική εικόνα εισόδου και οι 7 συστάδες που δημιουργήθηκαν ύστερα από την εκτέλεση του αλγορίθμου ασυμμετρικής συσταδοποίησης των Fischer και Poland με τιμή παραμέτρου τ ίση με 7, καθώς η διάσταση D των δεδομένων είναι 3 και η εξ' ορισμού τιμή της παραμέτρου τ υπολογίζεται ως γνωστόν από τον τύπο $\tau = 2 \times D + 1$. *Ο αλγόριθμος έδωσε ένα αρκετά καλό αποτέλεσμα συσταδοποίησης αφού στο κυρίως χρώμα του περιγράμματος της εικόνας τοποθετήθηκαν και κάποια pixels διαφορετικής απόχρωσης τα οποία εντοπίστηκαν από τον αλγόριθμο και τοποθετήθηκαν στις συστάδες 2 και 7. Η ίδια λογική εφαρμόστηκε και για το κυρίως χρώμα της συστάδας 4 γι' αυτό και προέκυψε η συστάδα 5 που αφορούσε τον εισαχθέντα θόρυβο στο κυρίως χρώμα. Στις συστάδες 3 και 6 αν και φαίνεται ότι κάποια pixels ενδεχομένως να μην ταιριάζουν στη συγκεκριμένη συστάδα εντούτοις τα χρώματα έχουν αναμιχθεί και*

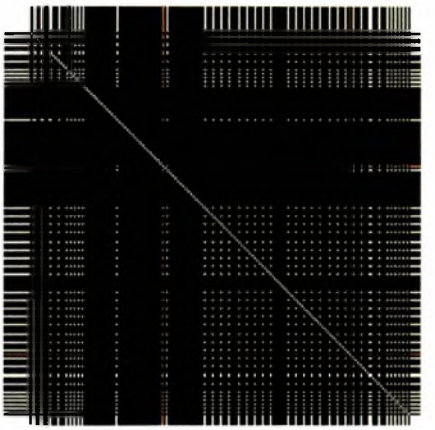
ομοιογενοποιηθεί για να δώσουν μια καινούργια εκδοχή ενός χρώματος. Άλλωστε αυτό που μας ενδιαφέρει είναι η μέση τιμή του χρώματος μιας συστάδας από την οποία δεν πρέπει να αποκλίνουν πολύ τα pixels της εικόνας.



Ο πίνακας $\widetilde{A}$ *για τη συγκεκριμένη εικόνα εισόδου βάσει του οποίου κατασκευάζεται ο συμμετρικός πίνακας γειτνίασης A.*



Ο συμμετρικός πίνακας γειτνίασης A όπως προκύπτει από τη σχέση $A_{ij} = \min\left\{\widetilde{A}_{ij}, \widetilde{A}_{ji}\right\}$.

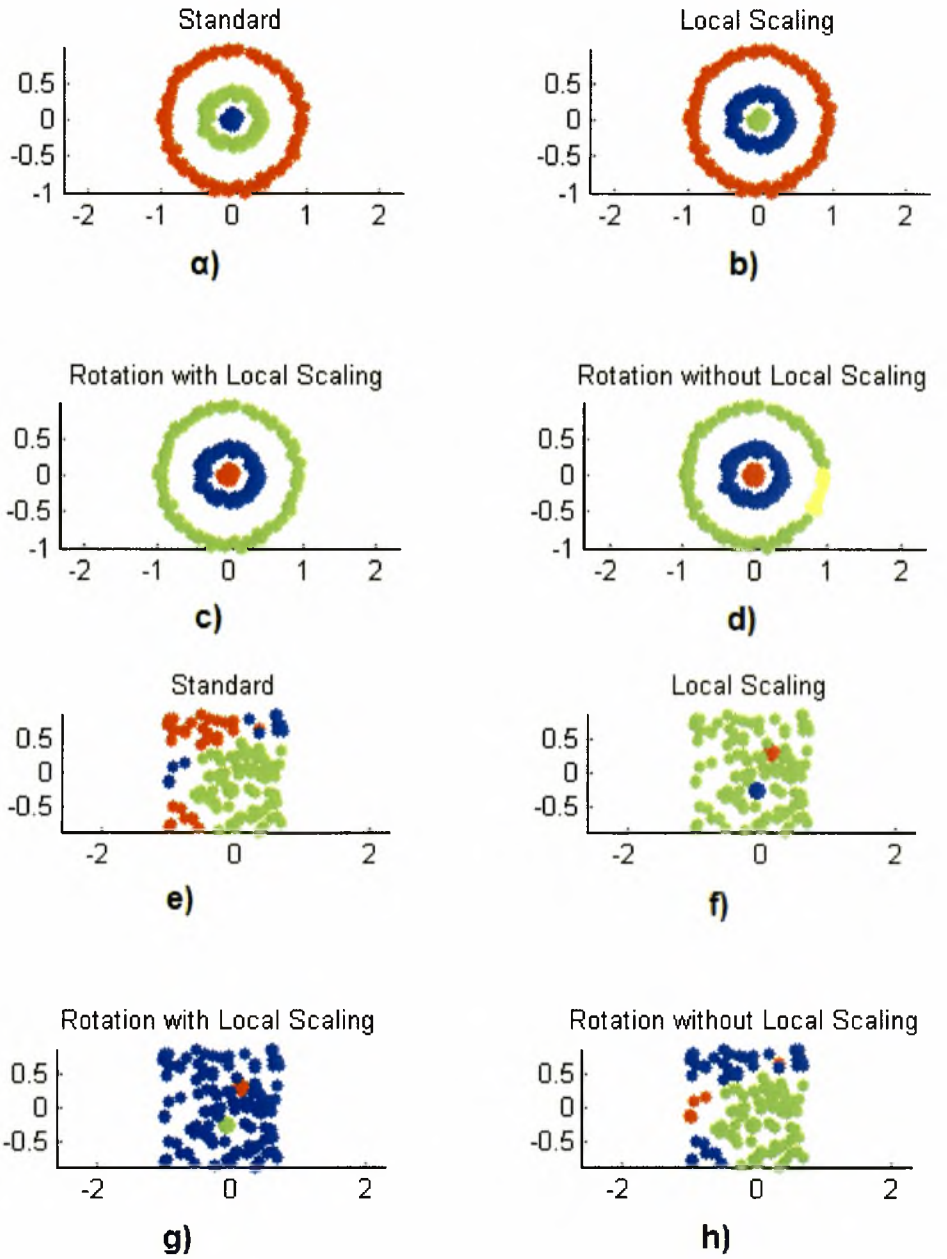


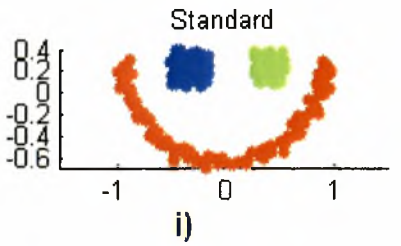
Ο πίνακας αγωγιμότητας C όπως προέκυψε από το συμμετρικό πίνακα γειτνίασης A.

7.6 Φασματική συσταδοποίηση με τον αλγόριθμο ZP

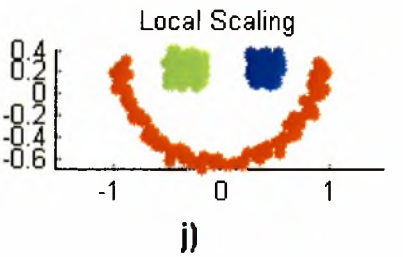
7.6.1 Σύγκριση του αλγορίθμου NJW με τον αλγόριθμο ZP

Σε αυτή την ενότητα παρουσιάζονται κάποια παραδείγματα εκτέλεσης για συγκεκριμένες εισόδους στον αλγόριθμο NJW και στον αλγόριθμο ZP. Πιο συγκεκριμένα όμως δίνεται μια οπτική σύγκριση των αποτελεσμάτων που δίνει ο αλγόριθμος NJW και ο αλγόριθμος ZP σε τρεις εκδοχές του: 1) με την ύπαρξη μόνο τοπικής κλιμάκωσης, που ουσιαστικά πρόκειται για επέκταση του αλγορίθμου NJW, αφού στη συγκεκριμένη εκδοχή χρησιμοποιείται ο πίνακας γειτνίασης τοπικής κλιμάκωσης των Zelnik και Perona 2) με τοπική κλιμάκωση και χρήση της στροφής των ιδιοδιανυσμάτων 3) χωρίς τοπική κλιμάκωση και με χρήση της στροφής των ιδιοδιανυσμάτων. Στις δύο τελευταίες εκδοχές χρησιμοποιείται ένας μέγιστος αριθμός συστάδων προς εξερεύνηση και στη συνέχεια επιλέγεται αυτός που επιφέρει το μικρότερο κόστος στοίχισης.

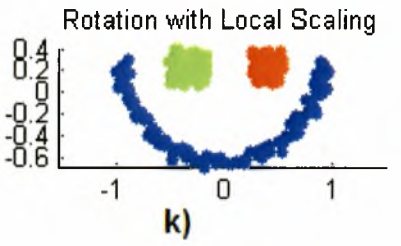

Standard
Local Scaling
Rotation with Local Scaling
Rotation without Local Scaling
Standard
Local Scaling
Rotation with Local Scaling
Rotation without Local Scaling
a)
b)
c)
d)
e)
f)
g)
h)



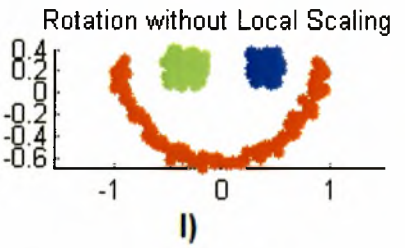
i)



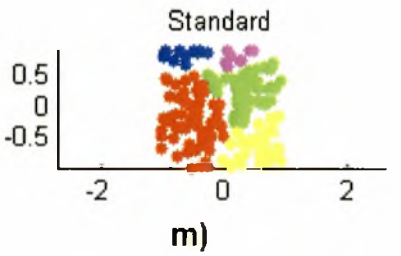
j)



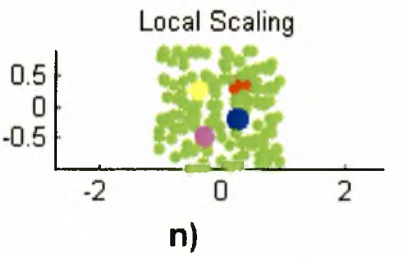
k)



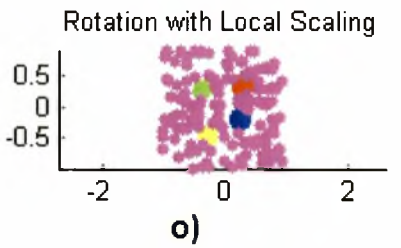
l)



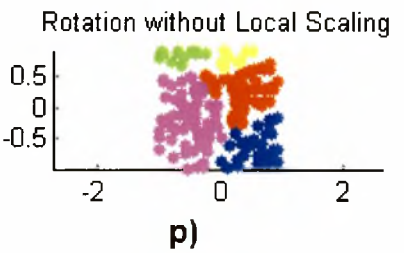
m)



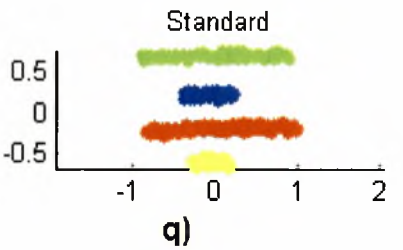
n)



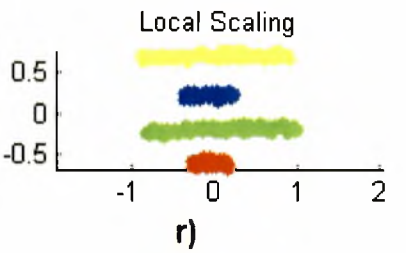
o)



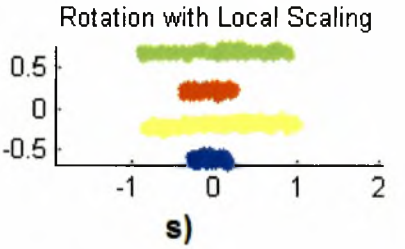
p)



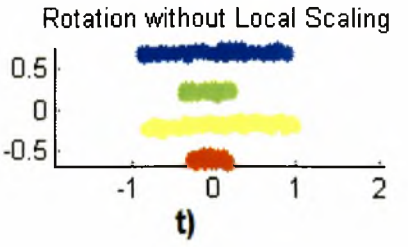
q)



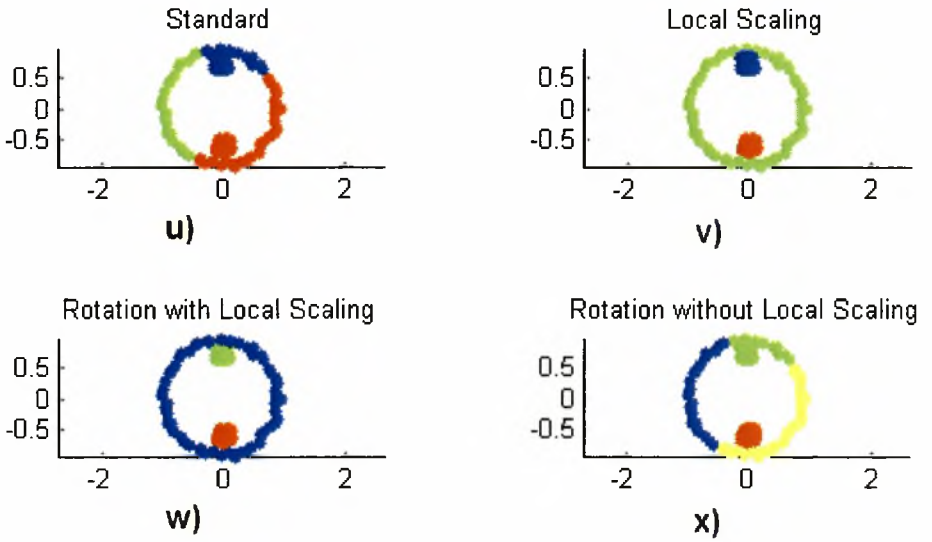
r)



s)



t)



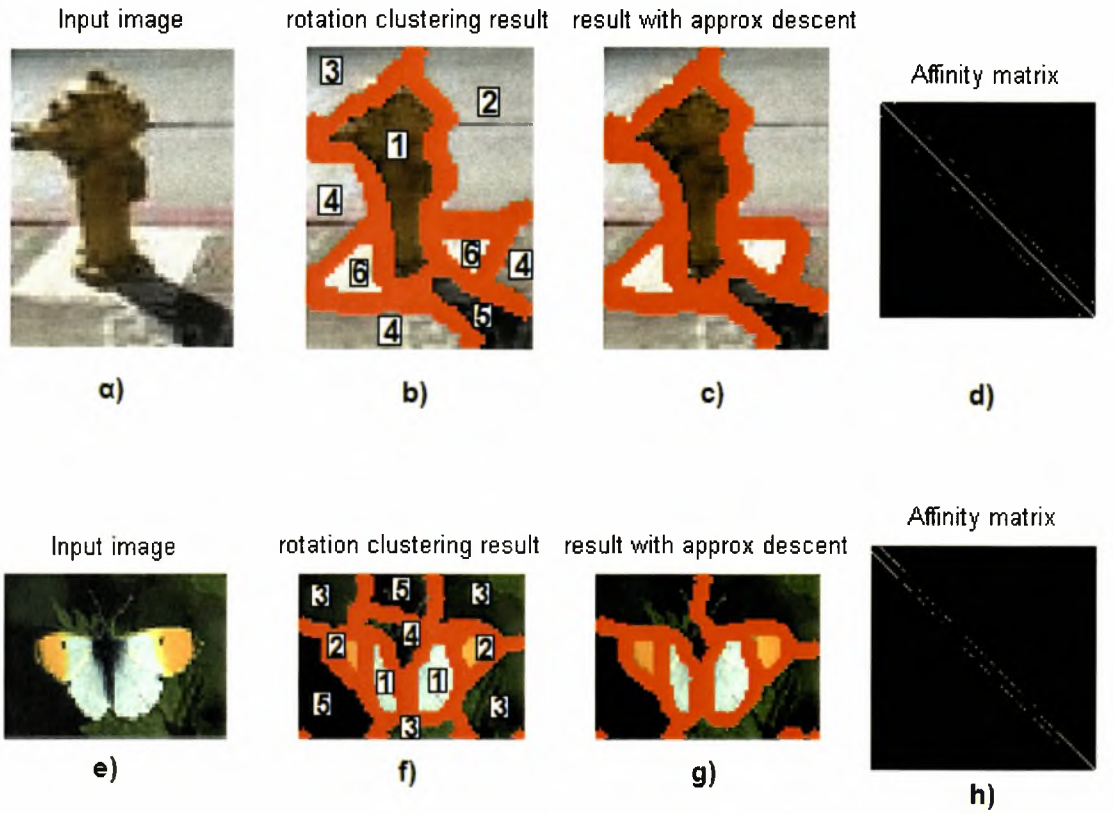
Στις εικόνες (a), (e), (i), (m), (q) και (u) η κλασσική συσταδοποίηση με τον αλγόριθμο NJW εκτελείται με παράμετρο καθολικής κλιμάκωσης 0.04. Ο μέγιστος αριθμός συστάδων για τον οποίο γίνεται έλεγχος από τον αλγόριθμο ZP, ώστε να μπορέσει αποφανθεί για τη βέλτιστη λύση, είναι 5. Τα διαφορετικά χρώματα υποδεικνύουν και διαφορετικές συστάδες για τα δεδομένα εισόδου, ενώ ταυτόχρονα υποδηλώνουν και τον αριθμό των συστάδων των δεδομένων.

Αυτό που μπορούμε να παρατηρήσουμε στα παραπάνω αποτελέσματα είναι ότι όταν χρησιμοποιούμε τον αλγόριθμο ZP με στροφή ιδιοδιανυσμάτων χωρίς τοπική κλιμάκωση ενδεχομένως να πάρουμε διαφορετικά αποτελέσματα από την αντίστοιχη εκδοχή του αλγορίθμου με τοπική κλιμάκωση. Επίσης σε κάποιες περιπτώσεις η χρήση μόνο του πίνακα τοπικής κλιμάκωσης στον αλγόριθμο φασματικής συσταδοποίησης των NJW, ενδεχομένως να δίνει χειρότερα αποτελέσματα από αυτά που δίνει ο ίδιος αλγόριθμος με τη χρήση κάποιας συγκεκριμένης τιμής για την καθολική παράμετρο κλιμάκωσης.

7.6.2 Εφαρμογή του αλγορίθμου ZP πάνω σε ψηφιακές εικόνες

Στη συνέχεια παρουσιάζουμε τα τελικά αποτελέσματα συσταδοποίησης του αλγορίθμου ZP που προέκυψαν για διαφορετικές ψηφιακές εικόνες εισόδου. Εκτός

των άλλων παρουσιάζεται τόσο το αποτέλεσμα που προέκυψε από την αυξητική μέθοδο gradient descent, που χρησιμοποιείται στο Βήμα 5 του αλγορίθμου, όσο και ο πίνακας γειτνίασης τοπικής κλιμάκωσης για κάθε είσοδο.



Εικόνες (a) και (e): οι εικόνες εισόδου για τον αλγόριθμο ZP. Εικόνες (b) και (f): το αποτέλεσμα της τελικής συσταδοποίησης του αλγορίθμου για καθεμία από τις δύο εικόνες. Εικόνες (c) και (g): το αποτέλεσμα που προκύπτει για κάθε εικόνα στο Βήμα 5 της φασματικής συσταδοποίησης ZP, όπου εφαρμόζεται η μέθοδος gradient descent. Εικόνες (d) και (h): οι πίνακες γειτνίασης τοπικής κλιμάκωσης, οι οποίοι έχουν τη μορφή διαγωνίου μπλοκ πίνακα.

7.7 Συσταδοποίηση κινηματογραφικών σκηνών

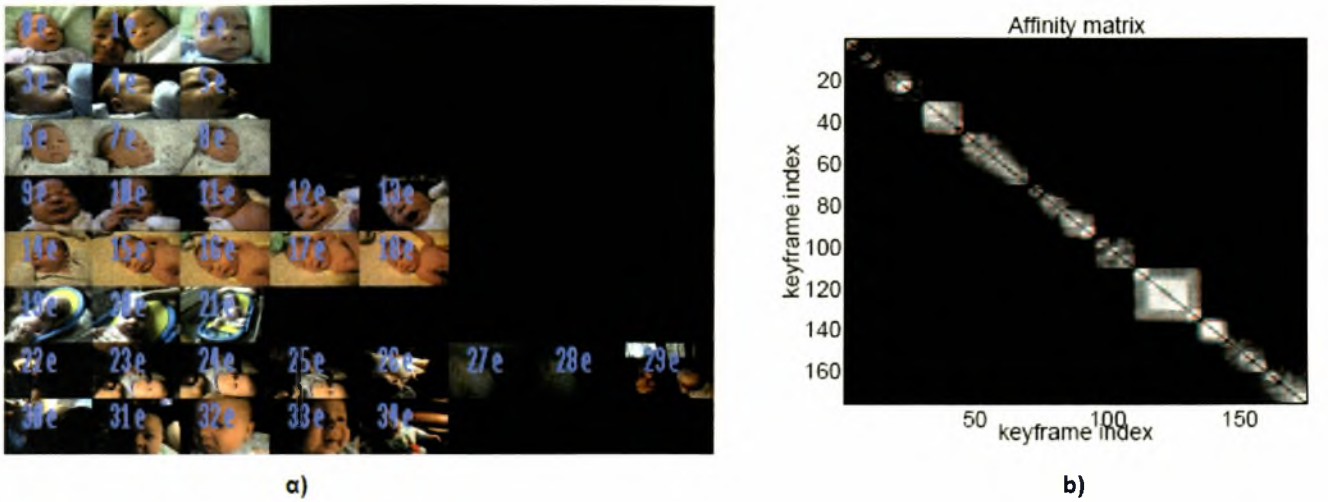
7.7.1 Γενικά

Σε αυτή την ενότητα παρουσιάζουμε τα αποτελέσματα που προέκυψαν ύστερα από την εκτέλεση του αλγορίθμου φασματικής συσταδοποίησης κινηματογραφικών σκηνών για δύο βίντεο. Το πρώτο βίντεο έχει οικιακή θεματολογία και πρόκειται για ένα βίντεο 20 λεπτών, μορφής MPEG-1, που περιέχει 430 πλάνα, ενώ το κατώφλι που καλείται να ικανοποιεί το φασματικό ιδιοκενό έχει τιμή $\delta_K = 0.15$. Το δεύτερο βίντεο πρόκειται για ένα βίντεο ποδοσφαίρου διάρκειας 10 λεπτών που περιέχει 86 πλάνα. Το κατώφλι που καλείται να ικανοποιεί το φασματικό ιδιοκενό σε αυτή την περίπτωση έχει τιμή $\delta_K = 0.15$.

7.7.2 Παράδειγμα οικιακού βίντεο

7.7.2.1 Τιμές παραμέτρων, πίνακας γειτνίασης και τελική συσταδοποίηση

Για το βίντεο της *Εικόνας 7.7.2.1 (a)* οι παράμετροι κλιμάκωσης σ_v και σ_t υπολογίστηκαν με τον ακόλουθο τρόπο. Η τιμή της παραμέτρου σ_v τέθηκε ίση με την τιμή 0.25, που αντιπροσωπεύει ένα καλό κατώφλι για το διαχωρισμό ενδοσυσταδικών και διασυσταδικών κατανομών ομοιότητας σε οικιακά βίντεο. Επίσης πειράματα έχουν δείξει ότι κατά μέσο όρο το 70% των σκηνών από οικιακά βίντεο συντίθενται από 4 ή λιγότερα πλάνα. Συνεπώς η τιμή της παραμέτρου σ_t τέθηκε ίση με το μέσο χρονικό διαχωρισμό μεταξύ τεσσάρων πλάνων ενός δοθέντος βίντεο. Στην *Εικόνα 7.7.2.1 (b)* φαίνεται ο πίνακας γειτνίασης που αντιστοιχεί στο βίντεο της *Εικόνας 7.7.2.1 (a)*.



Εικόνα 7.7.2.1

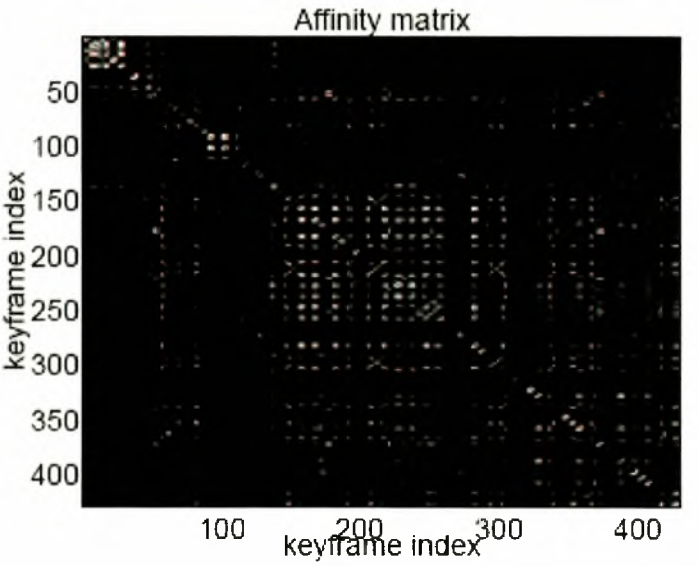
Στην Εικόνα 7.7.2.1 (α) παρουσιάζεται το αποτέλεσμα της συσταδοποίησης που προέκυψε ύστερα από την εκτέλεση του αλγορίθμου φασματικής συσταδοποίησης κινηματογραφικών πλάνων. Όπως φαίνεται ο αριθμός των τελικών συστάδων είναι ίσος με 8 ενώ για κάθε πλάνο απεικονίζεται μόνο ένα πλαίσιο-κλειδί. Στην Εικόνα 7.7.2.1 (b) παρουσιάζεται ο πίνακας γειτνίασης που αντιστοιχεί σε αυτή την είσοδο. Τα πιο έντονα σημεία αντιστοιχούν σε μεγάλη κατά ζεύγη ομοιότητα. Ο πίνακας παρουσιάζει μια αρκετά ευδιάκριτη μορφή διαγωνίου μπλοκ πίνακα. Αυτό συμβαίνει κυρίως επειδή παρόμοια πλάνα συνήθως αντιστοιχούν σε γειτονικά πλάνα και επειδή ο όρος ομοιότητας που εξαρτάται από το χρόνο, περιορίζει την έκταση των εκτός διαγωνίου όρων.

7.7.3 Παράδειγμα βίντεο ποδοσφαίρου

7.7.3.1 Τιμές παραμέτρων, πίνακας γειτνίασης και σύγκριση της φασματικής με τη k-means συσταδοποίηση

Στο συγκεκριμένο παράδειγμα η παράμετρος κλιμάκωσης σ_v για τον αλγόριθμο φασματικής συσταδοποίησης τέθηκε σε μια τιμή ίση με 0.25. Εκτός από τον πίνακα γειτνίασης και τις τελικές συστάδες που προέκυψαν από τον αλγόριθμο φασματικής συσταδοποίησης, στην *Εικόνα 7.7.3.1 (c)* παρουσιάζεται και το

αποτέλεσμα της συσταδοποίησης που προέκυψε από τον k-means αλγόριθμο, ώστε να γίνει μια οπτική σύγκριση.



Εικόνα 7.7.3.1 α)

Ο πίνακας γειτνίασης που προέκυψε για τα δεδομένα ποδοσφαίρου. Το μπλοκ αποτέλεσμα του πίνακα προέκυψε λόγω της εναλλαγής μεταξύ μακρινών (αυτά που περιέχουν σκηνές με χορτοτάπητα) και κοντινών πλάνων. Η δεύτερη κατηγορία πλάνων οδηγεί σε λιγότερες ενδοσυσταδικές ομοιότητες και επομένως γι' αυτό το λόγο δημιουργούνται λιγότερο έντονα μπλοκ μέσα στον πίνακα γειτνίασης (τόσο στα διαγώνια όσα και στα μη διαγώνια στοιχεία του).

Εικόνα 7.7.3.1 b)

Οι 10 συστάδες που προέκυψαν ύστερα από την εφαρμογή του αλγορίθμου φασματικής συσταδοποίησης. Να σημειωθεί ότι παρουσιάζεται μόνο ένα πλαίσιο-κλειδί για κάθε πλάνο.

Εικόνα 7.7.3.1 c)

Οι 10 συστάδες που δημιουργήθηκαν ύστερα από την εκτέλεση του αλγορίθμου k-means. Αυτό που μπορούμε εύκολα να παρατηρήσουμε είναι ότι μερικές από τις συστάδες που δημιουργήθηκαν, διαφέρουν ως προς το περιεχόμενο αυτών που δημιουργήθηκαν με τον αλγόριθμο φασματικής συσταδοποίησης. Να σημειωθεί επίσης ότι για κάθε πλάνο παρουσιάζεται μόνο ένα πλαίσιο-κλειδί.

7.8 Συμπεράσματα – Μελλοντική Έρευνα

Σε ορισμένες εφαρμογές, ενδεχομένως να γνωρίζουμε τον αριθμό των τελικών συστάδων που θα πρέπει να δημιουργηθούν, αλλά αυτό που πάντα δεν είναι εκ των προτέρων γνωστό είναι η τιμή των παραμέτρων κλιμάκωσης. Συνεπώς αυτές οι παράμετροι είτε θα πρέπει να υπολογιστούν πειραματικά από τον ίδιο το χρήστη είτε αυτόματα από κάποιον αλγόριθμο φασματικής συσταδοποίησης που υποστηρίζει έναν τέτοιο υπολογισμό. Βέβαια κάθε φορά που ανατίθεται σε έναν αλγόριθμο ο αυτόματος υπολογισμός μιας παραμέτρου, αυτός ο υπολογισμός θα πρέπει να γίνεται αποδοτικά και με μικρό υπολογιστικό κόστος.

Η φασματική συσταδοποίηση έχει προτείνει αρκετές λύσεις για τον αυτόματο εντοπισμό του αριθμού των συστάδων και τη δημιουργία ακριβών και ποιοτικών συστάδων. Η εύρεση όμως αποδοτικών αποτελεσμάτων σε περιπτώσεις όπου τα δεδομένα εμπεριέχουν θόρυβο δεν είναι πάντα εύκολη. Ιδανικά θα θέλαμε ο θόρυβος να τοποθετείται μέσα σε μια διαφορετική συστάδα αλλά κάτι τέτοιο εξαρτάται από την εκάστοτε εφαρμογή που μελετάται κάθε φορά. Πάντως γίνονται αρκετές έρευνες για την επίτευξη αυτού του στόχου και θεωρητικά δεν είμαστε αρκετά μακρυά από την επίτευξή του.

Αλγόριθμοι φασματικής συσταδοποίησης θα μπορούσαν να χρησιμοποιηθούν για τη μελέτη του optical flow μιας ακολουθίας από frames που προκύπτει ύστερα από τον τεμαχισμό ενός βίντεο. Αν θέλαμε να δώσουμε έναν περιεκτικό ορισμό για τον όρο optical flow, θα μπορούσαμε να πούμε ότι είναι η έννοια που προσεγγίζει την κίνηση των αντικειμένων μέσα σε μία οπτική αναπαράσταση. Επίσης η χρήση του optical flow σε συνδυασμό με τις φασματικές μεθόδους θα βοηθούσε στον εντοπισμό και τη συσταδοποίηση των διαφορετικών τροχιών (trajectories) που εμφανίζονται μέσα σε ένα βίντεο που περιέχει μια πληθώρα κινούμενων αντικειμένων.

Γενικά θα μπορούσαμε να πούμε ότι οι αλγόριθμοι φασματικής συσταδοποίησης είναι αρκετά υποσχόμενες μέθοδοι για την εξαγωγή αξιόπιστων αποτελεσμάτων σε αρκετές εφαρμογές ακόμα και στην ανίχνευση κάποιου καρκινώματος στο δέρμα.

Βιβλιογραφία-Αναφορές

[1] A. Ng, M. Jordan and Y.Weiss "On spectral clustering: Analysis and an algorithm" *In Advances in Neural Information Processing Systems 14*, 2001

[2] von Luxburg U., "A tutorial on spectral clustering", Technical report, No TR-149, Max-Planck-Institut fór biologische Kybernetik, 2007

[3] Xiang T. S. Gong, "Spectral clustering with eigenvector selection", Pattern Recognition, vol. 41, no 3, pp. 1012-1029, 2008.

[4] Zelnik-Manor L., P. Perona, "Self-Tuning spectral clustering", Advances in Neural Information Processing Systems, vol. 17, 2005.

[5] Bach F. R. and M. I. Jordan, "Learning spectral clustering", In Thrun S. and Soul L. editors. NIPS'16, Cambridge, MA, MIT Press, 2004.

[6] Amplifying the Block Matrix Structure for Spectral Clustering In M. van Otterlo et al. (Editors), Proceedings of the 14th Annual Machine Learning Conference of Belgium and the Netherlands, pp. 21-28, 2005.

[7] J.M. Odobez, D. Gatica-Perez, M. Guillemot, "video shot clustering using spectral methods," in *3th Workshop on Content-Based Multimedia Indexing(CBMI)*, sept 2003.

[8] *Spectral clustering.*
http:// www.cs.biu.ac.il/~louzouy/courses/seminar/talk9.ppt

[9] *Cluster analysis.*
http://en.wikipedia.org/wiki/Data_clustering

[10] *Demo on Graph-based algorithms in Machine Learning.*
http://agbs.kyb.tuebingen.mpg.de/wikis/mlss07/MatthiasHeinUlrikeVonLuxburg

[11] *ZP Clustering.*
http://www.vision.caltech.edu/lihi/Demos/SelfTuningClustering.html

[12] *Derek greene, "Graph partitioning and spectral clustering".*
https://www.cs.tcd.ie/research_groups/mlg/kdp/presentations/Greene_MLG04.ppt

[13] *Givens Rotation.*
http://en.wikipedia.org/wiki/Givens_rotation

[14] *Image Clustering.*
http://www.cs.bilkent.edu.tr/~canf/CS533/CS533Spr06stuPresent/imageClustering.ppt

[15] *Spectral Clustering.*
http://www.stat.washington.edu/wxs/Stat593-s03/Studentpresentations/SpectralClustering2.ppt

[16] *Introduction to spectral clustering.*
http://lagis-vi.univ-ille1.fr/~lm/classpec/reunion_28_02_08/Introduction_to_spectral_clustering.pdf

[17] *Linkage in Hierarchical clustering.*
http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/Hier_Linkage.htm

[18] *Hierarchical clustering.*
http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

[19] *Biclustering*
http://en.wikipedia.org/wiki/Biclustering

[20] *Fuzzy c-means presentation*

http://ce.sharif.edu/~m_amiri/download/yfcmc/y_fcmc_**presentation**_v0.8.ppt

[21] *Video Shot Clustering Using Spectral Methods*

http://www.idiap.ch/~guillemo/CBMI/poster.pdf

[22] *Πίνακας συνδιακύμανσης*

http://en.wikipedia.org/wiki/Covariance_matrix