
Πανεπιστήμιο Θεσσαλίας

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

# ΟΜΑΔΟΠΟΙΗΣΗ ΣΕ ΥΠΕΡΓΡΑΦΗΜΑΤΑ

*Διπλωματική Εργασία*

*Απόστολος Νασίου*

Επιβλέποντες καθηγητές

Κατσαρός Δημήτριος, Επίκουρος Καθηγητής

Ποταμιάνος Γεράσιμος, Αναπληρωτής Καθηγητής

2018,Βόλος


Πανεπιστήμιο Θεσσαλίας

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

# ΟΜΑΔΟΠΟΙΗΣΗ ΣΕ ΥΠΕΡΓΡΑΦΗΜΑΤΑ

*Διπλωματική Εργασία*

*Απόστολος Νασίου*

Επιβλέποντες καθηγητές

Κατσαρός Δημήτριος, Επίκουρος Καθηγητής

Ποταμιάνος Γεράσιμος, Αναπληρωτής Καθηγητής

2018,Βόλος


University of Thessaly

School of engineering

Department of electrical and computer engineering

# CLUSTERING IN HYPERGRAPHS

*Diploma Thesis*

*Apostolos Nasiou*

Supervisor

Katsaros Dimitrios, Assistant Professor

Potamianos Gerasimos, Associate Professor

2018,Volos

# ΕΥΧΑΡΙΣΤΙΕΣ

Καταρχήν, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ.Δημήτρη Κατσαρό για τη βοήθεια του και την καθοδήγηση που μου παρείχε για την εκπόνηση της παρούσας εργασίας. Διότι, με την ολοκλήρωσή της, κλείνει άλλος ένας σημαντικός κύκλος, αυτός των φοιτητικών μου σπουδών έπειτα από την πενταετή φοίτησή μου, με τις επιτυχίες και αποτυχίες, τις χαρές και κακουχίες να τη χαρακτηρίζουν, από την οποία όμως κατάφερα να αποκομίσω πέρα από γνώσεις και δεξιότητες πάνω στον τομέα που ασχολήθηκα και εμπειρίες καθ' όλη τη διάρκεια αυτών των χρόνων.

Εξίσου σημαντική ήταν και η παρουσία και υποστήριξη που μου δόθηκε από την οικογένεια μου και τα κοντινά μου άτομα, γι' αυτό και τους οφείλω ένα μεγάλο ευχαριστώ, οι οποίοι με την κατανόηση τους και την εμπιστοσύνη τους σε μένα με βοήθησαν να συνεχίσω το δύσκολο έργο μου και να το υλοποιήσω με επιτυχία.

# Περιεχόμενα

# Κατάλογος Σχημάτων

# Κατάλογος Πινάκων

# ΠΕΡΙΛΗΨΗ

Η συγκεκριμένη εργασία πραγματεύεται ένα ενδιαφέρον ζήτημα, αυτό της ομαδοποίησης σε υπεργραφήματα, δηλαδή τον τρόπο δημιουργίας κοινοτήτων κορυφών σε υπεργράφημα. Με τον όρο υπεργράφημα αναφερόμαστε στη γενίκευση της έννοιας του γραφήματος. Για την καλύτερη κατανόηση της διαδικασίας αυτής, είναι απαραίτητη η γνώση της θεωρίας γύρω από την ομαδοποίηση και τα υπεργραφήματα και γι'αυτό παρουσιάζονται αναλυτικά. Μέσω, λοιπόν, της μελέτης παρόμοιων θεματικά ερευνών έγινε μια προσπάθεια δημιουργίας ενός αλγορίθμου ομαδοποίησης των υπεργραφημάτων, ενώ παράλληλα επιδιώχτηκε και μια σύγκριση με υπάρχων αλγόριθμο. Μεγάλο μέρος της εργασίας έχει καταλάβει η δημιουργία αλγορίθμων που χρησιμοποιήθηκαν για την υλοποίηση και την επεξεργασία των δεδομένων προς ομαδοποίηση, αλλά και η δημιουργία των μέτρων για την αξιολόγηση των αποτελεσμάτων. Ακόμη, πραγματοποιήθηκε και αναγωγή αλγορίθμου από τη θεωρία των γραφημάτων στη θεωρία των υπεργραφημάτων. Πειράματα, τέλος, διεξήχθησαν για τη λήψη αποτελεσμάτων ακρίβειας αλλά και χρόνου εκτέλεσης των αλγορίθμων.

# ABSTRACT

This specific research is about a very interesting matter, hypergraph clustering , that is the way communities of vertices are created in a hypergraph. By the term hypergraph we are referring to the generalization of graphs. For better understanding of this research it is essential to know the theory of clustering and hypergraphs and for that reason they are presented analytically. So via the study of thematic similar researches there has been an effort of creating an algorithm for hypergraph clustering, while at the same time it was held a compare with an excisting algorithm. Big part of this research is the creation of algorithms that were used for the creation and processing of the data for clustering and also τhe creation of measures for evaluation of the results. Furthermore an algotithn from the graph theory it was converted to fit the hypergraph theory. Experiments have conducted in order to take results for the accuracy and the execution time of the algorithms.

# Κεφάλαιο 1

# ΕΙΣΑΓΩΓΗ

Τα υπεργραφήματα είναι μία γενίκευση των γραφημάτων όπου επεκτείνουν τη δυαδική σχέση των γραφημάτων σε μία πιο πολύπλοκη κ-σχέση. Ένας από τους τρόπους ανάλυσης των υπεργραφημάτων είναι η αναγωγή τους σε γραφήματα όπου η σχέση μεταξύ των δεδομένων είναι δυαδική άρα και πιο απλή προς επεξεργασία. Ενώ από τη μία χάνουμε έτσι πληροφορία από την άλλη έχουμε πολλά εργαλεία για να δουλέψουμε. Στην παρούσα εργασία δε θα γίνει κάποιου είδους αναγωγή, αλλά τα υπεργραφήματα θα αντιμετωπιστούν ως υπεργραφήματα και θα προσπαθήσουμε να χρησιμοποιήσουμε για να το πετύχουμε αυτό σύγχρονους μεθόδους ομαδοποίησης.

## 1.1 Περίγραμμα διπλωματικής εργασίας

Στο **Κεφάλαιο 2: Θεωρητικό Υπόβαθρο** Θα αναφερθώ στις γνώσεις που πρέπει κάποιος να έχει για να κατανοήσει την εργασία σχετικά με τους όρους που θα περιλαμβάνει.

Στο **Κεφάλαιο 3: Βιβλιογραφική Ανασκόπηση** Θα γίνει αναφορά σε προηγούμενες εργασίες και έρευνες που έχουν γίνει στο συγκεκριμένο θέμα και τα ευρήματα τους έχουν αξιοποιηθεί στην παρούσα εργασία.

Στο **Κεφάλαιο 4: Υλοποίηση αλγορίθμων από τη βιβλιογραφία** Θα παρουσιάσω τις

μεθόδους που χρησιμοποίησα και υλοποίησα από αποτελέσματα προηγούμενων ερευνών και έχουν χρησιμότητα στην παρούσα έρευνα.

Στο **Κεφάλαιο 5: Αναφορά στη δουλειά μου** Θα εξηγήσω τις μεθόδους που υλοποίησα για το σκοπό αυτής της εργασίας.

Στο **Κεφάλαιο 6: Αποτελέσματα** Θα παρουσιαστούν και εξηγηθούν τα δεδομένα που μας έδωσαν οι αλγόριθμοι που υλοποιήθηκαν.

# Κεφάλαιο 2

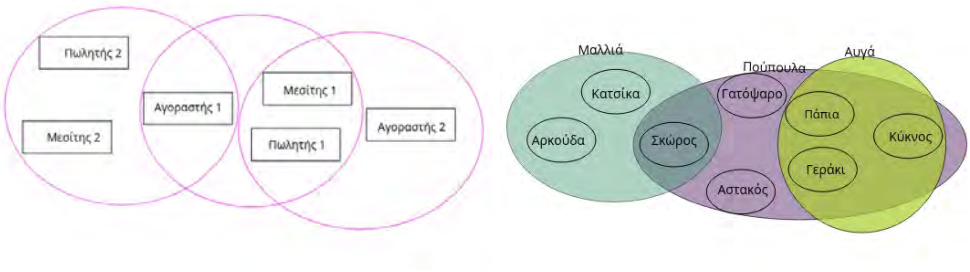
# ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Σκοπός του κεφαλαίου αυτού είναι η ανάλυση και περαιτέρω διασαφήνιση των όρων που θα συναντήσουμε κατά τη διάρκεια μελέτης της εργασίας, εννοιών όπως « Υπεργράφημα » και « Ομαδοποίηση » ώστε να γίνει ευκολότερα κατανοητό το νόημα αυτών.

## 2.1 Υπεργράφημα

Με τον όρο hypergraph που χρησιμοποιείται στην αγγλική βιβλιογραφία και στα ελληνικά θα το μεταφράζαμε ως υπεργράφημα αναφερόμαστε στη γενίκευση ενός γραφήματος όπου μία ακμή μπορεί να ενώσει παραπάνω από δύο κορυφές. Ένα υπεργράφημα συμβολίζεται ως $|V| \times |E|$ όπου $V$ είναι ένα σύνολο στοιχείων που ονομάζονται κορυφές και $E$ είναι ένα σύνολο μη κενών υποσυνόλων του πίνακα $V$ που ονομάζονται ακμές ή στην περίπτωση μας υπερακμές.

Ενώ στα γραφήματα μία ακμή αποτελείται από ένα ζευγάρι κόμβων, στα υπεργραφήματα μία υπερακμή είναι ένα σύνολο κόμβων. Όταν κάθε υπερακμή ενός υπεργραφήματος έχει τον ίδιο βαθμό $k$ δηλαδή το ίδιο πλήθος κορυφών, το υπεργράφημα ονομάζεται $k$-ομοιόμορφο. Οπότε ένα 2-ομοιόμορφο υπεργράφημα είναι ένα γράφημα, ένα 3-ομοιόμορφο υπεργράφημα είναι ένα σύνολο τριπλετών και ούτω καθεξής. Στην παρούσα εργασία δε θα εργαστούμε με υπερακμές που έχουν τον ίδιο βαθμό αλλά τυχαίο.




(α□) 3-ομοιόμορφο (β□) Μη ομοιόμορφο

Σχήμα 2.1: Παράδειγμα υπεργραφημάτων

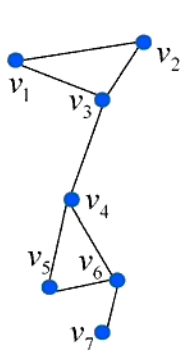
Ένα σταθμισμένο υπεργράφημα είναι ένα υπεργράφημα στο οποίο ένας αριθμός ( βάρος ) έχει ανατεθεί σε κάθε υπερακμή. Όταν δεν έχει ανατεθεί κάποιο βάρος ή το βάρος όλων των υπερακμών είναι ίσο με 1 το υπεργράφημα θεωρείται μη σταθμισμένο.

Ένα κατευθυνόμενο γράφημα είναι ένα γράφημα όπου οι ακμές έχουν κατεύθυνση. Κατευθυνόμενο υπεργράφημα είναι ένα σύνολο $(V, A)$ όπου $V$ είναι ένα πεπερασμένο σύνολο κορυφών και $A$ είναι ένα σύνολο υπερακμών πάνω από το $V$.

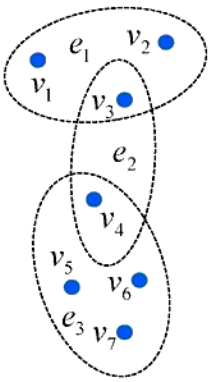
Τα υπεργραφήματα χρησιμοποιούνται για την περιγραφή πολύπλοκων σχέσεων και γι'αυτό χρησιμοποιούνται στη δημιουργία ιατρικών τεστ, στην ομαδοποίηση δεδομένων σε διακομιστές αποθήκευσης δεδομένων, στην τοποθέτηση δικτυακών εφαρμογών σε διακομιστές κτλ. Υπεργράφημα έχουμε όταν θέλουμε να περιγράψουμε τη σχέση μεταξύ αγοραστή-πωλητή-μεσίτη [3] όπως φαίνεται στο σχήμα 2.1α□. Έχουμε ένα 3-ομοιόμορφο υπεργράφημα. Με αυτό τον τρόπο απεικόνισης μπορούμε να δούμε όλα τα άτομα που σχετίζονται με μια αγορά. Στο σχήμα 2.1β□ βλέπουμε ένα μη-ομοιόμορφο υπεργράφημα. Εκεί μπορούμε να δούμε τη σχέση των ζώων με βάση κάποιο χαρακτηριστικό τους.

Στο Σχήμα 2.2 βλέπουμε ένα απλό παράδειγμα για κάποιες από τις έννοιες που αναφέρθηκαν παραπάνω. Άρα συνοψίζοντας, στην εργασία αυτή θα ασχοληθούμε με μη κατευθυνόμενα, μη σταθμισμένα και μη ομοιόμορφα υπεργραφήματα.

Σχήμα 2.2: Παράδειγμα (α) γραφήματος , (β) υπεργραφήματος , (γ) πίνακα ακμών υπεργραφήματος




(a) Simple graph




(b) Hypergraph **G**

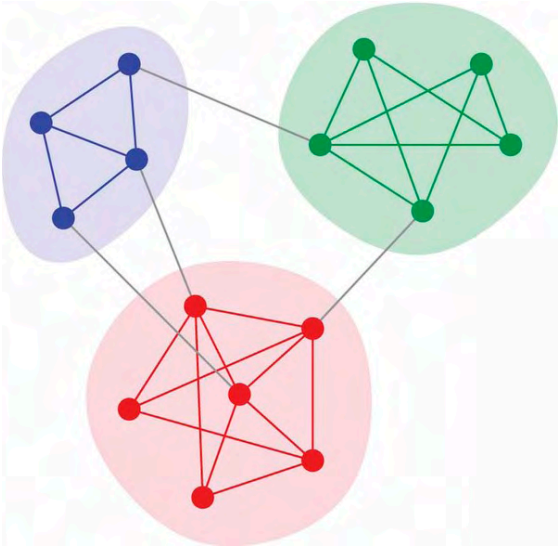
| | $e_1$ | $e_2$ | $e_3$ |
|---|---|---|---|
| $v_1$ | 1 | 0 | 0 |
| $v_2$ | 1 | 0 | 0 |
| $v_3$ | 1 | 1 | 0 |
| $v_4$ | 0 | 1 | 1 |
| $v_5$ | 0 | 0 | 1 |
| $v_6$ | 0 | 0 | 1 |
| $v_7$ | 0 | 0 | 1 |

(c) Incidence matrix **H**

## 2.2 Ομαδοποίηση

Συναντάμε αρκετούς ορισμούς για την ομαδοποίηση, αλλά ο πιο αντιπροσωπευτικός κατά την άποψη μου είναι ότι αποτελεί μία διαδικασία ομαδοποίησης ενός συνόλου δεδομένων με τέτοιο τρόπο ώστε τα αντικείμενα στην ίδια ομάδα/συστάδα να είναι πιο σχετικά/όμοια μεταξύ τους σε σχέση με στοιχεία άλλων συστάδων. Στη θεωρία των γραφημάτων οι συστάδες αναφέρονται ως κοινότητες και έχουν ακριβώς τις ίδιες ιδιότητες. Στην ανάλυση μας, μας ενδιαφέρουν οι μη επικαλυμμένες συστάδες. Επικάλυψη έχουμε όταν μία κορυφή ανήκει σε παραπάνω από μία συστάδες. Και ως επέκταση θα εξετάσουμε μόνο μη επικαλυμμένους αλγορίθμους. Στο Σχήμα 2.2 βλέπουμε χωρισμένες χρωματικά τις κοινότητες/συστάδες ενός γραφήματος.

Σχήμα 2.3: Παράδειγμα κοινοτήτων/συστάδων σε γράφημα



# Κεφάλαιο 3

# ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Για τη διεκπεραίωση της παρούσας εργασίας έχουν μελετηθεί ποικίλα άρθρα και έρευνες προηγούμενων ετών το εννοιολογικό περιεχόμενο των οποίων κινείται γύρω από τις έννοιες υπεργραφήματα, ομαδοποίηση και έλεγχος σφαλμάτων. Από αυτά έγινε επιλογή συγκεκριμένων, των οποίων η θεματική σχετίζεται περισσότερο με την εργασία. Παρακάτω στο κεφάλαιο αυτό θα γίνει μια συνοπτική αναφορά σε καθεμία εργασία που μελετήθηκε και έχει χρησιμοποιηθεί ως υλικό.

Με τον όρο κεντρικότητα, σημαντική έννοια που θα μας απασχολήσει στη συνέχεια, αναφερόμαστε στη σημαντικότητα μίας θέσης σε ένα δίκτυο. Ο συγκεκριμένος τύπος σημαντικότητας διαφέρει από μέτρο σε μέτρο. Μία θέση μπορεί να είναι σημαντική γιατί η πληροφορία περνάει από αυτή (ενδιάμεση κεντρικότητα (Freeman ,1979) ) ή επειδή μπορεί εύκολα να επικοινωνήσει με άλλα μέλη του δικτύου (εγγύτητας κεντρικότητα (Freeman ,1979) ) ή επειδή από μόνη της συνδέεται σε άλλα κεντρικά σημεία. Στην εργασία [3] ασχολούνται με το τελευταίο και το επεκτείνουν στη θεωρία των υπεργραφημάτων. Αυτό το μέτρο θα χρησιμοποιηθεί στην παρούσα εργασία για την αναγωγή ενός αλγορίθμου από τη θεωρία των υπεργραφημάτων στη θεωρία των γραφημάτων.

Η ενδιάμεση κεντρικότητα ενός κόμβου $v$ δίνεται από τη σχέση:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

όπου $\sigma_{st}$ είναι ο συνολικός αριθμός των σύντομων μονοπατιών από τον κόμβο $s$ στον κόμβο $t$ και $\sigma_{st}(v)$ ο αριθμός από αυτά τα μονοπάτια που περνάνε από το $v$.

Η εγγύτητας κεντρικότητα ενός κόμβου $v$ δίνεται από τη σχέση:

$$C(v) = \frac{1}{\sum_u d(v,u)}$$

ο $d(v,u)$ αριθμός των ακμών στο σύντομο μονοπάτι που συνδέει τις κορυφές $v$ και $u$.

Στο [6] γίνεται εκτενής ανάλυση πολλών αλγορίθμων ομαδοποίησης. Το συγκεκριμένο βιβλίο επικεντρώνεται στην εξήγηση των αλγορίθμων αυτών, ενώ ταυτόχρονα αναφέρει και τους τύπους δεδομένων εισαγωγής που έχουν νόημα να αναλυθούν από τον κάθε αλγόριθμο και θα αποδώσουν αξιοπρεπή αποτελέσματα. Βάσει αυτού του βιβλίου έγινε η επιλογή των μεθόδων που χρησιμοποιήθηκαν στην παρούσα εργασία.

Το [8] είναι ένα άρθρο που ερευνά την απόδοση των αλγορίθμων υπό συγκεκριμένες συνθήκες σφάλματος στα δεδομένα εισαγωγής. Η μέτρηση της απόδοσης των αλγορίθμων στις συνθήκες των εκάστοτε δεδομένων γίνεται με τη χρήση διάφορων μέτρων, εκ των οποίων κάποια από αυτά εφαρμόζονται και στην παρούσα εργασία.

Η [9] είναι μια έρευνα που ασχολείται με την ανάλυση των υπεργραφημάτων ως μία οντότητα πολύπλοκων σχέσεων μεταξύ των δεδομένων σε αντίθεση με άλλες έρευνες που κάνουν αναγωγή σε γραφήματα τα οποία έχουν δυαδικές σχέσεις μεταξύ των δεδομένων τους. Κατά την διεξαγωγή της υλοποιεί τη γενίκευση του αλγορίθμου που χρησιμοποιείται στην προαναφερθείσα περίπτωση. Αυτός ο αλγόριθμος θα αξιοποιηθεί ως μέτρο σύγκρισης για τον αλγόριθμο που δημιουργήθηκε στην παρούσα εργασία.

Γενικότερα, η ομαδοποίηση γραφημάτων έχει μεγάλη ιστορία στο σχεδιασμό VLSI [2]. Εκεί, οι κορυφές αντιστοιχούν σε στοιχεία κυκλώματος και οι υπερακμές αντιστοιχούν στην καλωδίωση η οποία μπορεί να συνδέει παραπάνω από δύο στοιχεία. Η εύρεση του ελάχιστου κόστους κοψίματος επιτρέπει το διαχωρισμό των στοιχείων σε ενότητες με ελάχιστες ενδοσυνδέσεις (interconnections). Η εισαγωγή της ομαδοποίησης των υπεργραφημάτων στην υπολογιστική όραση (computer vision) και στη μηχανική μάθηση (machine learning) είναι σχετικά πρόσφατη [5][1] . Η ανάλυση διάφορων υπαρχόντων μεθόδων ομαδοποίησης υπεργραφημάτων και παρουσίασης των υπεργραφημάτων με

τη μορφή κλίκας σε κανονικό γράφημα έχει γίνει από διάφορες έρευνες [4][7].

# Κεφάλαιο 4

# ΥΛΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΑΠΟ ΤΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Σε αυτό το κεφάλαιο θα εξηγήσω τους αλγορίθμους που έχουν αξιοποιηθεί στα πλαίσια της εργασίας αλλά έχουν δημιουργηθεί από προηγούμενες έρευνες.

## 4.1 Φασματική ομαδοποίηση

Από τον ορισμό που δώσαμε για τα υπεργραφήματα και από [9] έχουμε ένα πίνακα $H$ με εγγραφές $h(v, e) = 1$ αν $v \in e$ και 0 αλλιώς, ο οποίος αποτελεί τον πίνακα υπερακμών του υπεργραφήματος $G$. Συμπερασματικά έχουμε:

$d(v) = \sum_{e \in E} w(e)h(v, e)$

$\delta(e) = \sum_{v \in V} h(v, e)$

Όπου $D_v$ και $D_e$ διαγώνιοι πίνακες που περιέχουν το βαθμό των κορυφών και των υπερακμών αντίστοιχα. Έστω $W$ ο διαγώνιος πίνακας που περιέχει τα βάρη. Έστω $A$ ο πίνακας γειτνίασης που ορίζεται ως $A = HWH^T - D_v$

Χρησιμοποιώντας τα παραπάνω καταλήγουμε στον πίνακα

$$\Delta = I - D_v^{-1/2} H W H^T D_v^{-1/2}$$

που χαρακτηρίζεται ως κανονικοποιημένος Laplacian πίνακας για ένα μη κατευθυνόμενο υπεργράφημα.

Η εργασία [9] χρησιμοποιεί μια γενίκευση της φασματικής ομαδοποίησης (spectral clustering) για υπεργραφήματα για τη δημιουργία των συστάδων του υπεργραφήματος. Η βάση ενός αλγορίθμου φασματικής ομαδοποίησης είναι ο Laplacian πίνακας που ορίσαμε παραπάνω.

Για την εκτέλεση του αλγορίθμου φασματικής ομαδοποίησης ορίζουμε αρχικά τις $k$ συστάδες που θέλουμε να δημιουργήσει. Ο αλγόριθμος έχει τα εξής βήματα:

**Βήμα 1**: Εύρεση του Laplacian πίνακα $\Delta$ με τον παραπάνω τύπο
**Βήμα 2**: Υπολογισμός των πρώτων $k$ ιδιοδιανυσμάτων $u_1, \ldots, u_k$ του πίνακα $\Delta$
**Βήμα 3**: Έστω $U \in \mathbb{R}^{n \times k}$ ο πίνακας που περιέχει τα διανύσματα ως $u_1, u_2, \ldots, u_k$ στήλες.
**Βήμα 4**: Για $i = 1, \ldots, n$, έστω $y_i \in \mathbb{R}^k$ το διάνυσμα που ανταποκρίνεται στην $i$-οστή σειρά του $U$
**Βήμα 5**: Ομαδοποιούμε τα σημεία $(y_i)_{i=1,\ldots,n}$ $in$ $\mathbb{R}^k$ με τον αλγόριθμο k-means σε συστάδες $C_1, \ldots, C_k$

## 4.2 Δημιουργία τυχαίων δεδομένων

Στο [6] γίνεται αναφορά και χρήση ενός αλγορίθμου για τη δημιουργία τυχαίων πολυδιάστατων δεδομένων. Σε αυτή την εργασία θα χρησιμοποιηθεί αυτός ο αλγόριθμος για τη δημιουργία των τυχαίων δεδομένων προς εισαγωγή στους αλγορίθμους. Με αυτό τον αλγόριθμο δημιουργούνται σφαιρικά δομημένες Γκαουσιανές (Gaussian) συστάδες. Η συγκεκριμένη υλοποίηση του αλγορίθμου μας εγγυάται ότι οι συστάδες δεν θα επικαλύπτουν η μία την άλλη πέραν από ένα συγκεκριμένο ποσοστό, αλλά η κάθε συστάδα θα περιέχει τουλάχιστον ένα συγκεκριμένο αριθμό στοιχείων και επίσης καθορίζεται και ο ακριβής αριθμός συστάδων που θέλουμε να έχουν τα δεδομένα μας. Εφόσον δημιουρ-

γούμε εμείς τα δεδομένα έχουμε τη γνώση της πραγματικής δομής των συστάδων. Άρα οι παράμετροι του αλγορίθμου είναι:

$$
\begin{aligned}
n &: \textit{ο συνολικός αριθμός των σημείων που θέλουμε να δημιουργηθούν} \\
d &: \textit{οι διαστάσεις που θέλουμε να έχουν τα δεδομένα} \\
\sigma &: \textit{η έκταση που θέλουμε να έχουν οι συστάδες} \\
n_{min} &: \textit{τα ελάχιστα στοιχεία που θέλουμε να έχει η κάθε συστάδα} \\
I_o &: \textit{ο δείκτης επικάλυψης των συστάδων} \\
c &: \textit{ο αριθμός των συστάδων, πρέπει } n_{min}c \leq n
\end{aligned}
$$

Ο σκοπός είναι να δημιουργήσουμε ένα σύνολο $n$ στοιχείων σε έναν $d$- διάστατο χώρο διατεταγμένα σε $c$ συστάδες με τουλάχιστον $n_{min}$ στοιχεία ανά συστάδα. Τα σημεία κάθε συστάδας ξεχωριστά είναι ανεξάρτητα δείγματα $d$- διάστατων Γκαουσιανών κατανομών συγκεντρωμένα σε τυχαία επιλεγμένα κέντρα και έχοντας ως πίνακα συνδιακύμανσης:

$$\sigma^2 I, \ \textit{όπου } I \textit{ ένας } d \times d \textit{ μοναδιαίος πίνακας}$$

Η επικάλυψη μεταξύ δύο συστάδων $i$ και $j$, $O(i,j)$ καθορίζεται ως το σφάλμα $Bayes$ της ακόλουθης δοκιμής υποθέσεων (hypothesis-testing) του προβλήματος που ταξινομεί ένα $d$-διάνυσμα ,$X$, σύμφωνα με :

$$
\begin{aligned}
H_i &: X \textit{ κατανεμημένο ως } N(\mu_i, \sigma^2 I) \\
H_j &: X \textit{ κατανεμημένο ως } N(\mu_j, \sigma^2 I)
\end{aligned}
$$

όπου $\mu_i$ και $\mu_j$ είναι τα κέντρα των δύο συστάδων. Ο κανόνας $Bayes$ είναι καλά ορισμένος για τέτοιου είδους προβλήματα στη βιβλιογραφία. Έστω $n_i$, $n_j$ ο αριθμός των δεδομένων που βρίσκονται στις συστάδες $i$ και $j$ αντίστοιχα και έστω $N_i$ και $N_j$ να δηλώνουν τον αριθμό των δεδομένων των συστάδων $i$ και $j$ που ταξινομήθηκαν λανθασμένα από τον κανόνα του $Bayes$ , άρα η επικάλυψη των δύο συστάδων καθορίζεται

ως:

$$O(i,j) = \frac{N_i + N_j}{n_i + n_j}$$

Η επικάλυψη κυμαίνεται από $0$ έως $1$ όπου $0$ σημαίνει καλά χωρισμένες συστάδες και $1$ σημαίνει ότι οι συστάδες συμπίπτουν. Η παράμετρος επικάλυψης $_o$ καθορίζει τη μέγιστη επικάλυψη μεταξύ κάθε ζεύγους συστάδων και δεν επιτρέπεται να υπερβεί την τιμή $_o$. Ο αλγόριθμος έχει τα εξής βήματα:

**Βήμα 1**: Καθορίζεται το μέγεθος των συστάδων $\{n_i, n_2, \ldots, n_c\}$ όπου

$$\sum_{k=1}^{c} n_k = n \textit{ και } n_k \geq n_{min}\ \forall k$$

Αυτό επιτυγχάνεται θέτοντας το $n_k$ ίσο με $n_{min}\ \forall k$, έπειτα επιλέγοντας τυχαία συστάδες και αυξάνοντας το μέγεθός τους κατά $1$ έως ότου το άθροισμα των στοιχείων όλων των συστάδων να είναι ίσο με $n$.

Επαναλαμβάνουμε τα βήματα 2 έως 5 για $i$ από $1$ έως $c$.

**Βήμα 2**: Δημιουργούμε τυχαίο κέντρο $\mu_i$
**Βήμα 3**: Δημιουργούμε $n_i$ δεδομένα γύρω από το κέντρο $\mu_i$ σύμφωνα με την $N(\mu_i, \sigma^2 I)$ κατανομή
**Βήμα 4**: Αν κάποια επικάλυψη $O(i, i-1), O(i, 1-2), \ldots, O(i, 1)$ ξεπερνάει την τιμή $I_o$ επαναλαμβάνουμε τα βήματα $2$ και $3$
**Βήμα 5**: Αν υλοποιήσουμε $50$ επαναλήψεις και δεν έχουμε επιτύχει στη δημιουργία νέου κέντρου συστάδας αυξάνουμε την τιμή $I_o$ στη μέγιστη επικάλυψη που συναντήσαμε σε αυτές τις $50$ επαναλήψεις και επαναλαμβάνουμε τα βήματα $2$ έως $4$

Η επιλογή του συγκεκριμένου αλγορίθμου έγινε γιατί μας δίνει τη δυνατότητα να δημιουργήσουμε δεδομένα με όποια μορφή θέλουμε και είναι αναγκαία για να μελετήσουμε τις ιδιότητες και αδυναμίες του κάθε αλγορίθμου.

## 4.3 Μέτρο κεντρικότητας

Για τον υπολογισμό του μέτρου κεντρικότητας που αναφέρεται στο [3] χρειαζόμαστε τον πίνακα υπερακμών ή κορυφών του υπεργραφήματος που θέλουμε να αναλύσουμε. Εμείς εργαζόμαστε με τον πίνακα υπερακμών. Αν και υπάρχουν διαφορές μεταξύ των δυο υπολογισμών, η διαφορά δεν είναι σημαντική. Έστω $H$ ο πίνακας υπερακμών ενός υπεργραφήματος και έστω $x$ και $y$ οι τιμές κεντρικότητας για τα στοιχεία σειράς και γραμμής , ή αλλιώς των κορυφών και των υπερακμών και $\lambda$ η μεγαλύτερη ιδιοτιμή του πίνακα $H$. Άρα μπορούμε να πούμε ότι:

$$E^T x = \lambda y \quad Ey = \lambda x$$

Τα διανύσματα $x$ και $y$ είναι ιδιοδιανύσματα από διαφορετικούς πίνακες, όμως και τα δυο σχετίζονται με το ίδιο ιδιοτιμή $\lambda^2$:

$$EE^T x = \lambda^2 x \quad E^T Ey = \lambda^2 y$$

## 4.4 Μέτρο ακρίβειας

Στο [8] γίνεται αναφορά σε ένα μέτρο για τη μέτρηση της απόδοσης του κάθε αλγορίθμου. Αυτό το μέτρο χρειάζεται γνώση της πραγματικής μορφής των συστάδων και μπορεί να εφαρμοστεί μετά την εκτέλεση του αλγορίθμου. Αυτό το μέτρο θα χρησιμοποιηθεί και σε αυτή την εργασία για τον υπολογισμό της απόδοσης των αλγορίθμων.

Το μέτρο μπορεί να πάρει τιμές από $0.0$ έως $1.0$ όπου η τιμή $1.0$ αντιστοιχεί στην τέλεια δημιουργία των συστάδων. Το στατιστικό βασίζεται σε ένα τετραγωνικό πίνακα $n \times n$ όπου $n$ ο αριθμός των δεδομένων προς ομαδοποίηση. Οι τιμές του πίνακα είναι είτε $0$ είτε $1$.

Έστω $i, j \in V$, η τιμή $\delta(i, j)$ είναι ίση με $1$ όταν οι δύο κορυφές ανήκουν στην συστάδα και στη λύση του αλγορίθμου και στην πραγματική λύση αλλά και ακόμη όταν δυο κορυφές δεν ανήκουν στην ίδια συστάδα ούτε στη λύση του αλγορίθμου ούτε στην

πραγματική λύση. Και διαφορετικά, δηλαδή όταν οι κορυφές είναι ομαδοποιημένες μαζί στη μία λύση και σε διαφορετικές συστάδες στην άλλη, η τιμή είναι $0$. Οπότε το $0$ σε μία καταχώρηση του πίνακα συμβολίζει την αποτυχία του αλγορίθμου να ομαδοποιήσει σωστά αυτό το ζευγάρι κορυφών. Το στατιστικό υπολογίζεται ως:

$$\frac{\sum_{j=1}^{n}\sum_{i>j}^{n}\delta(i,j)}{n(n-1)/2}$$

Το θετικό αυτού του αλγορίθμου , και ο λόγος που επιλέχθηκε , είναι ότι δε κοιτάει αν το στοιχείο βρίσκεται στην ίδια συστάδα βάσει της ετικέτας που έχει δώσει ο αλγόριθμος αλλά κοιτάει αν η δομή της συστάδας έχει αποτυπωθεί στη λύση του αλγορίθμου.

# Κεφάλαιο 5

# ΑΝΑΦΟΡΑ ΣΤΗ ΔΟΥΛΕΙΑ ΜΟΥ

## 5.1 Μορφοποίηση δεδομένων

Βάσει του ορισμού για τα υπεργραφήματα μπορούμε να θεωρήσουμε κάθε πίνακα $n \times m$ όπου οι τιμές του είναι $0$ ή $1$ ένα πίνακα ακμών ενός γραφήματος όπου έχει $n$ κορυφές και $m$ ακμές. Αν όμως κάθε ακμή από $i = 1, \ldots, m$ και $|i| \geq 3$ μπορούμε να θεωρήσουμε κάθε ακμή ότι είναι μία υπερακμή και άρα έχουμε ένα υπεργράφημα. Τα σύνολα δεδομένων όμως δεν είναι απαραίτητο και δεν είναι σύνηθες να έχουν μόνο χαρακτηριστικά που παίρνουν τιμές $0$ και $1$ αλλά οποιοδήποτε ακέραιο αριθμό, δεκαδικούς αριθμούς αλλά και χαρακτήρες.

Οπότε αν μπορέσουμε να επεξεργαστούμε τα χαρακτηριστικά οποιουδήποτε συνόλου δεδομένων ώστε να περιέχουν μόνο τιμές $0$ και $1$ μπορούμε να δημιουργήσουμε υπεργράφημα για να εξετάσουμε τους αλγορίθμους των υπεργραφημάτων που έχουμε υλοποιήσει. Η δομή των συστάδων του συνόλου που θα μετασχηματιστεί δεν αλλάζει καθώς αλλάζουμε μόνο τη μορφή των χαρακτηριστικών για να ταιριάζει στην περίπτωση μας.

Άρα με την παραπάνω σκέψη δημιουργήθηκε ένας αλγόριθμος που μπορεί να μετασχηματίσει ένα οποιοδήποτε σύνολο δεδομένων σε ένα πίνακα υπερακμών. Ο αλγόριθμος μορφοποίησης συνόλου δεδομένων έχει ως ακολούθως:

**Βήμα 1**: Αναγνώριση των διαφορετικών τύπων δεδομένων που υπάρχουν στο σύνολο δεδομένων προς μορφοποίηση

Επαναλαμβάνουμε τα βήματα 2 και 3 για κάθε τύπο που βρήκαμε στο βήμα 1.

**Βήμα 2**: Έλεγχος του τύπου δεδομένων. Για κάθε διαφορετικό τύπο δεδομένων ακολουθείται διαφορετική επεξεργασία.

**Περίπτωση 1: Χαρακτήρας**. Αν ο τύπος δεδομένων είναι χαρακτήρας δημιουργούνται εικονικές μεταβλητές ( dummy variables) για κάθε τιμή ξεχωριστά.

**Περίπτωση 2: Δεκαδικός Αριθμός**. Σε αυτή την περίπτωση το σύνολο τιμών χωρίζεται σε διαστήματα με βάση τα εκατοστημόρια $[0.01, 0.25, 0.5, 0.75, 0.95, 0.99]$. Δημιουργούνται οι ανάλογες εικονικές μεταβλητές ανά διάστημα. Το κάθε στοιχείο έχει τιμή $1$ για το διάστημα που βρίσκεται.

**Περίπτωση 3: Ακέραιος**. Αν είναι ακέραιος και οι διαφορετικές τιμές που παίρνει το χαρακτηριστικό είναι λιγότερες από εννιά δημιουργούνται εικονικές μεταβλητές όπως στην Περίπτωση 1. Αλλιώς δημιουργούνται μεταβλητές όπως στην Περίπτωση 2

**Βήμα 3**: Μετά από την επεξεργασία του κάθε χαρακτηριστικού διαγράφεται η στήλη αφού δεν έχει πλέον σημασία γιατί έχει μετασχηματιστεί ισοδύναμα.

## 5.2 Ομαδοποίηση

Για την ομαδοποίηση των υπεργραφημάτων υπάρχουν δύο τρόποι αντιμετώπισης: σύμφωνα με τον πρώτο τρόπο ως ένα κλασικό πρόβλημα ομαδοποίησης που συνεπάγεται τη χρήση σύγχρονων μεθόδων ομαδοποίησης ή σύμφωνα με το δεύτερο με αναγωγή αλγορίθμων ομαδοποίησης, δημιουργίας κοινοτήτων, από τη θεωρία των γραφημάτων στη θεωρία των υπεργραφημάτων. Στην ενότητα 5.2.1 θα αναλυθεί η πρώτη περίπτωση και στην ενότητα 5.2.2 η δεύτερη περίπτωση.

### 5.2.1 Σύγχρονοι μέθοδοι ομαδοποίησης

Η επιλογή κατάλληλου αλγορίθμου για την περίπτωση που εξετάζουμε δεν είναι εύκολη καθώς το σύνολο δεδομένων προς ομαδοποίηση αποτελείται από τον πίνακα υπερακμών $H$ ενός υπεργραφήματος $G$ , όπου οι τιμές του είναι $0$ και $1$. Βάσει του [6] όπου γίνεται εκτενής αναφορά στους αλγορίθμους ομαδοποίησης επιλέχθηκε η χρήση ιεραρχικής μεθόδου ομαδοποίησης(hierachical clustering).

Η μέθοδος ιεραρχικής ομαδοποίησης είναι μία διαδικασία που μετασχηματίζει ένα πίνακα αποστάσεων (proximity matrix) σε μία αλληλουχία από εμφολευμένα χωρίσματα(nested partitions). Η ιεραρχική ομαδοποίηση είναι η μέθοδος και όχι ο αλγόριθμος. Ο αλγόριθμος καθορίζεται όταν οριστούν το μέτρο υπολογισμού των κέντρων και το μέτρο υπολογισμού των αποστάσεων των σημείων μεταξύ τους.

Ο πίνακας αποστάσεων υπολογίζεται από το μέτρο απόστασης. Για την εύρεση όμως της απόστασης των δύο σημείων δε μπορούμε να χρησιμοποιήσουμε μέτρα όπως την ευκλείδεια απόσταση ή την απόσταση μανχάταν. Ως δεδομένα εισόδου θα έχουμε τον πίνακα υπερακμών και οι τιμές του δεν επιτρέπουν την επιλογή τέτοιων μέτρων. Οπότε καταλήξαμε στην απόσταση jaccard και στην απόσταση dice. Η απόσταση jaccard ορίζεται ως εξής:

Έστω $v_1, v_2$ διανύσματα όπου οι τιμές του είναι οι γραμμές από τον πίνακά υπερακμών $H$ ενός υπεργραφήματος $G$.

$M_{11}$: αντιπροσωπεύει τον συνολικό αριθμό χαρακτηριστικών που $v_1$ και $v_2$ έχουν τιμή $1$
$M_{01}$: αντιπροσωπεύει τον συνολικό αριθμό χαρακτηριστικών όπου το χαρακτηριστικό του $v_1$ έχει τιμή $0$ και το χαρακτηριστικό $v_2$ έχει τιμή $1$
$M_{10}$: αντιπροσωπεύει τον συνολικό αριθμό χαρακτηριστικών όπου το χαρακτηριστικό του $v_1$ έχει τιμή $1$ και το χαρακτηριστικό $v_2$ έχει τιμή $0$

Άρα ο συντελεστής ομοιότητας jaccard είναι:

$$J = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}$$

και εν τέλει η απόσταση jaccard είναι:

$$d_J = 1 - J$$

Η απόσταση dice ορίζεται ως εξής:

$$DSC = \frac{2|v_1 \cap v_2|}{|v_1| + |v_2|}$$

όταν εφαρμόζεται σε δυαδικά δεδομένα, όπως στην περίπτωση μας, χρησιμοποιούμε τον ορισμό θετικά αληθές(true positive TP) , ψευδές θετικό (false positive FP), ψευδές αρνητικό (false negative FN) έχουμε:

$$DSC = \frac{2TP}{2TP + FP + FN}$$

και η απόσταση ορίζεται ως:

$$d_{DSC} = 1 - DSC$$

Ως μέθοδος υπολογισμού τής απόστασης των κέντρων των συστάδων χρησιμοποείται η μέθοδος μέσων (average).

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{|u| * |v|}$$

για κάθε σημείο $i$ και $j$ όπου $|u|$ και $|v|$ είναι το πλήθος των στοιχείων των συστάδων $u$ και $v$ αντίστοιχα.

Άρα έχουμε δύο αλγορίθμους ιεραρχικής ομαδοποίησης όπου ο ένας έχει μέτρο απόστασης των σημείων την απόσταση jaccard και ο άλλος την απόσταση dice και οι δύο έχουν μέτρο απόστασης κέντρων τη μέθοδο μέσων.

### 5.2.2 Girvan-Newman σε υπεργραφήματα

Ο Girvan-Newman είναι ένας αλγόριθμος που ανιχνεύει κοινότητες αφαιρώντας σταδιακά ακμές από το αρχικό δίκτυο. Τα συνδεδεμένα στοιχεία του παραμένων δικτύου είναι οι κοινότητες. Αντί να προσπαθήσει να δημιουργήσει ένα μέτρο για να αποφασίσει ποια από τις ακμές είναι η πιο κεντρική στις κοινότητες, ο αλγόριθμος εστιάζει στις ακμές που είναι πιο πιθανόν ανάμεσα σε κοινότητες. Αυτός ο αλγόριθμος εφαρμόζεται για την εύρεση κοινοτήτων σε γραφήματα.

Για να εφαρμόσουμε τον παρόντα αλγόριθμο στα υπεργραφήματα θα πρέπει να μετασχηματίσουμε λίγο τα μέτρα που χρησιμοποιεί. Θα πρέπει να έχουμε μέτρα που έχουν σημασία στη θεωρία των υπεργραφημάτων. Άρα θα πρέπει να μετασχηματίσουμε το μέτρο με το οποίο αφαιρούνται οι ακμές. Στην περίπτωση μας ο αλγόριθμος θα αφαιρεί την ακμή με τη μεγαλύτερη κεντρικότητα βάσει του μέτρου κεντρικότητας που αναφέρθηκε στην ενότητα 4.3

# Κεφάλαιο 6

# ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρκετά πειράματα έχουν διεξαχθεί χρησιμοποιώντας το συνδυασμό των παραπάνω μεθόδων. Αρχικά δημιουργείται ένα σύνολο δεδομένων όπως αναφέρθηκε στην ενότητα 4.2 και αυτό μορφοποιείται όπως αναφέρθηκε στην ενότητα 5.1. Τα δεδομένα αυτά εισάγονται στους αλγορίθμους που αναφέρθηκαν στην ενότητα 4.1 και 5.2.1.

Οι αλγόριθμοι ελέγχθηκαν για διάφορα σύνολα δημιουργημένα τυχαία, με διάφορους παραμέτρους. Δημιουργήθηκαν όλα τα πιθανά σενάρια για αριθμό διαστάσεων $[4, 6, 9]$ και πλήθος συστάδων $[3, 4, 5, 6]$ με σταθερές τις τιμές συνολικών σημείων $n = 500$ και διασποράς των συστάδων $s = 0.2$. Με το μετασχηματισμό του συνόλου δεδομένων που εξηγήθηκε παραπάνω έχω 9 καινούριες στήλες/υπερακμές για κάθε διάσταση, άρα θα έχω τελικά $[36, 54, 81]$ υπερακμές για κάθε αριθμό διαστάσεων αντίστοιχα.

Στον Πίνακα 6.1 βλέπουμε τα αποτελέσματα από τις εκτελέσεις. Στις 12 περιπτώσεις που εξετάστηκαν στις μισές οι αλγόριθμοι που υλοποιήθηκαν ξεπέρασαν τον αλγόριθμο που είχαμε ως συγκριτικό. Το θετικό αυτών των αλγορίθμων σε σχέση με τον αλγόριθμο φασματικής ομαδοποίησης είναι ο χρόνος εκτέλεσης που θα δούμε παρακάτω.

| methods | d=4,c=3 | d=6,c=3 | d=9,c=3 | d=4,c=4 | d=6,c=4 | d=9,c=4 | d=4,c=5 | d=6,c=5 | d=9,c=5 | d=4,c=6 | d=6,c=6 | d=9,c=6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| average jaccard | 0.68 | 0.78 | 1 | 0.73 | 0.97 | 1 | 0.74 | 0.83 | 1 | 0.78 | 0.92 | 1 |
| average dice | 0.67 | 0.77 | 1 | 0.74 | 0.97 | 1 | 0.77 | 0.84 | 1 | 0.76 | 0.9 | 1 |
| kmeans | 0.8 | 0.81 | 0.99 | 0.72 | 0.97 | 1 | 0.79 | 0.84 | 1 | 0.78 | 0.9 | 1 |

Πίνακας 6.1: Αποτελέσματα επιτυχίας

| methods | d=20,c=3 | d=22,c=3 | d=25,c=3 | d=20,c=4 | d=22,c=4 | d=25,c=4 | d=20,c=5 | d=22,c=5 | d=25,c=5 | d=20,c=6 | d=22,c=6 | d=25,c=6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| average jaccard | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| average dice | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| kmeans | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Πίνακας 6.2: Πολλές υπερακμές ακρίβεια

| methods | d=20,c=3 | d=22,c=3 | d=25,c=3 | d=20,c=4 | d=22,c=4 | d=25,c=4 | d=20,c=5 | d=22,c=5 | d=25,c=5 | d=20,c=6 | d=22,c=6 | d=25,c=6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| average jaccard | $5.75 \cdot 10^{-4}$ | $7.87 \cdot 10^{-5}$ | $3.75 \cdot 10^{-5}$ | $3.08 \cdot 10^{-5}$ | $5.35 \cdot 10^{-5}$ | $5.04 \cdot 10^{-3}$ | $4.53 \cdot 10^{-5}$ | $2.67 \cdot 10^{-5}$ | $2.68 \cdot 10^{-3}$ | $3.37 \cdot 10^{-5}$ | $1.5 \cdot 10^{-5}$ | $5.62 \cdot 10^{-4}$ |
| average dice | $6.93 \cdot 10^{-6}$ | $1.23 \cdot 10^{-6}$ | $9.83 \cdot 10^{-5}$ | $4.12 \cdot 10^{-5}$ | $9.54 \cdot 10^{-6}$ | $2.54 \cdot 10^{-3}$ | $3.59 \cdot 10^{-5}$ | $1.61 \cdot 10^{-5}$ | $1.88 \cdot 10^{-4}$ | $2.47 \cdot 10^{-3}$ | $2.33 \cdot 10^{-5}$ | $3.57 \cdot 10^{-5}$ |
| kmeans | $6.35 \cdot 10^{-4}$ | $5.99 \cdot 10^{-5}$ | $7.61 \cdot 10^{-6}$ | $7.52 \cdot 10^{-3}$ | $1.24 \cdot 10^{-2}$ | $7.55 \cdot 10^{-3}$ | $1.09 \cdot 10^{-2}$ | $1.51 \cdot 10^{-4}$ | $4.35 \cdot 10^{-5}$ | $9.77 \cdot 10^{-3}$ | $6.57 \cdot 10^{-4}$ | $1.08 \cdot 10^{-5}$ |

Πίνακας 6.3: Πολλές υπερακμές χρόνος εκτέλεσης

Βλέποντας τον πίνακα των αποτελεσμάτων μπορούμε να παρατηρήσουμε ότι το ποσοστό επιτυχίας για λίγες διαστάσεις, δηλαδή για λίγες υπερακμές, είναι όχι απλά μικρότερο από το ποσοστό του αλγορίθμου προς σύγκριση αλλά μικρό. Αντίθετα, για πολλές διαστάσεις, δηλαδή για πολλές υπερακμές, η ακρίβεια μεγιστοποιείται. Για να γίνει καλύτερη κατανόηση αυτών των φαινομένων πραγματοποιήθηκαν περισσότερα πειράματα στις δυο αυτές περιπτώσεις.

Αρχικά η περίπτωση των πολλών υπερακμών. Στη νέα σειρά πειραμάτων χρησιμοποιήθηκαν δεδομένα με αριθμό διαστάσεων $[20, 22, 25]$ και πλήθος συστάδων $[3, 4, 5, 6]$ με σταθερές τις τιμές συνολικών σημείων $n = 1000$ και διασποράς των συστάδων $s = 0.2$. Όπως βλέπουμε στον πίνακα 6.2 έχουμε πλήρη επιτυχία. Το γεγονός αυτό οφείλεται στο πλήθος των υπερακμών που δημιουργούνται και στους δυνατούς συνδυασμούς $0$ και $1$ που μπορούμε να έχουμε εφόσον τώρα έχουμε εννέα στοιχεία σε κάθε μια από τις είκοσι ακμές, λαμβάνοντας υπόψιν τον μικρότερο αριθμό διαστάσεων. Άρα έχουμε $9^20$ συνδυασμούς και επειδή η πιθανότητα δυο ίδιων ακολουθιών είναι πολύ μικρή έχουμε καλύτερο υπολογισμό διαφοράς μεταξύ των σημείων προς ομαδοποίηση.

Στον πίνακα 6.2 βλέπουμε τους χρόνους που αντιστοιχούν στην παραπάνω εκτέλεση.

Στην άλλη περίπτωση όπου έχουμε λίγες διαστάσεις έχουμε το αντίθετο φαινόμενο από αυτό που παρατηρήσαμε παραπάνω. Έχουμε λίγες διαστάσεις και λίγους πιθανούς συνδυασμούς, $9^3$ στην περίπτωση των τριών διαστάσεων και χίλια συνολικά σημεία άρα ο πίνακας αποστάσεων δεν έχει μεγάλο εύρος τιμών που συνεπάγεται στη μικρή ακρίβεια. Για τέσσερις διαστάσεις η ακρίβεια του αλγορίθμου αυξάνεται, επιβεβαιώνοντας την εξήγησή μας.

| methods | d=3,c=3 | d=4,c=3 | d=3,c=4 | d=4,c=4 | d=3,c=5 | d=4,c=5 | d=3,c=6 | d=4,c=6 |
|---|---|---|---|---|---|---|---|---|
| average jaccard | 0.69 | 0.8 | 0.7 | 0.73 | 0.67 | 0.79 | 0.73 | 0.77 |
| average dice | 0.61 | 0.76 | 0.7 | 0.73 | 0.69 | 0.79 | 0.76 | 0.76 |
| kmeans | 0.76 | 0.78 | 0.71 | 0.76 | 0.71 | 0.8 | 0.75 | 0.77 |

Πίνακας 6.4: Λίγες υπερακμές ακρίβεια

| methods | d=3,c=3 | d=4,c=3 | d=3,c=4 | d=4,c=4 | d=3,c=5 | d=4,c=5 | d=3,c=6 | d=4,c=6 |
|---|---|---|---|---|---|---|---|---|
| average jaccard | $3.95 \cdot 10^{-5}$ | $5.26 \cdot 10^{-5}$ | $1.42 \cdot 10^{-5}$ | $7.36 \cdot 10^{-5}$ | $3.77 \cdot 10^{-7}$ | $6.68 \cdot 10^{-6}$ | $2.53 \cdot 10^{-3}$ | $5.79 \cdot 10^{-5}$ |
| average dice | $1.83 \cdot 10^{-5}$ | $1.29 \cdot 10^{-5}$ | $2.52 \cdot 10^{-5}$ | $5.86 \cdot 10^{-4}$ | $7.96 \cdot 10^{-5}$ | $6.78 \cdot 10^{-6}$ | $2.57 \cdot 10^{-3}$ | $4.98 \cdot 10^{-4}$ |
| kmeans | $8.32 \cdot 10^{-6}$ | $4.72 \cdot 10^{-5}$ | $1.48 \cdot 10^{-4}$ | $7.09 \cdot 10^{-3}$ | $5.83 \cdot 10^{-5}$ | $1.78 \cdot 10^{-5}$ | $1.91 \cdot 10^{-3}$ | $9.69 \cdot 10^{-5}$ |

Πίνακας 6.5: Λίγες υπερακμές χρόνος εκτέλεσης

Στον πίνακα 6.2 παρατηρούμε τους χρόνους που αντιστοιχούν στην εκτέλεση με τις λίγες διαστάσεις.

Για τη λήψη του χρόνου εκτέλεσης κάθε αλγορίθμου λήφθηκε ο μέσος χρόνος από 10 εκτελέσεις του κάθε αλγορίθμου. Το σύνολο των δεδομένων που χρησιμοποιήθηκε γι' αυτή τη μέτρηση είχε 5000 εγγραφές και 12 διαστάσεις άρα 108 υπερακμές. Ο χρόνος εκτέλεσης βγήκε σε σχέση με το χρόνο εκτέλεσης του αλγορίθμου σύγκρισης φασματικής ομαδοποίησης.

Στον πίνακα 6.6 παρατηρούμε τον χρόνο εκτέλεσης των αλγορίθμων ιεραρχικής ομαδοποίησης σε σχέση με τον χρόνο εκτέλεσης του αλγορίθμου φασματικής ομαδοποίησης. Στη στήλη time ratio βλέπουμε πόσες φορές είναι πιο γρήγορος ο κάθε αλγόριθμος από τον αλγόριθμο σύγκρισης.

Ο αλγόριθμος που αναφέρθηκε στο 5.2.2 δεν επέφερε αποτελέσματα στα οποία να έχουμε μη μοναδιαίες συστάδες. Οπότε τα αποτελέσματα του δεν ενδιαφέρουν την παρούσα έρευνα και γι' αυτό δε συμμετείχε στην παραπάνω σύγκριση.

Συμπερασματικά, προκύπτει ότι ο αλγόριθμος πετυχαίνει μεγάλη ακρίβεια για δεδομένα με πολλές υπερακμές, ενώ αποτυγχάνει στην αντίθετη περίπτωση και έχει πολύ μικρό-

| methods | time ratio |
|---|---|
| average jaccard | 39.0645 |
| average dice | 38.1383 |

Πίνακας 6.6: Χρόνοι εκτέλεσης

τερο χρόνο εκτέλεσης από τον αλγόριθμο φασματικής ομαδοποίησης.

# ΒΙΒΛΙΟΓΡΑΦΙΑ


[1] S. Agarwal et al. “Beyond pairwise clustering”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 2. June 2005, 838–845 vol. 2. DOI: 10.1109/CVPR.2005.89.

[2] Charles J. Alpert and Andrew B. Kahng. “Recent Directions in Netlist Partitioning: A Survey”. In: *Integr. VLSI J.* 19.1-2 (Aug. 1995), pp. 1–81. ISSN: 0167-9260. DOI: 10.1016/0167-9260(95)00008-4. URL: http://dx.doi.org/10.1016/0167-9260(95)00008-4.

[3] Phillip Bonacich, Annie Cody Holdren, and Michael Johnston. “Hyper-edges and multidimensional centrality”. In: 26 (July 2004), pp. 189–203.

[4] Guangliang Chen and Gilad Lerman. “Spectral Curvature Clustering (SCC)”. In: *International Journal of Computer Vision* 81.3 (Mar. 2009), pp. 317–330. ISSN: 1573-1405. DOI: 10.1007/s11263-008-0178-9. URL: https://doi.org/10.1007/s11263-008-0178-9.

[5] V. M. Govindu. “A tensor decomposition for geometric grouping and segmentation”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. June 2005, 1150–1157 vol. 1. DOI: 10.1109/CVPR.2005.50.

[6] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN: 0-13-022278-X.

[7] Hairong Liu, Longin Jan Latecki, and Shuicheng Yan. “Robust Clustering as Ensembles of Affinity Relations”. In: *NIPS*. Curran Associates, Inc., 2010, pp. 1414–1422.

[8] Glenn Milligan. “An Examination of the Effect of Six Types of Error Perturbation of Fifteen Clustering Algorithms”. In: 45 (Feb. 1980), pp. 325–342.

[9] D. Zhou, J. Huang, and B. Schölkopf. *Beyond Pairwise Classification and Clustering Using Hypergraphs*. Tech. rep. 143. Max Planck Institute for Biological Cybernetics, Aug. 2005.