ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη Ομαδοποίηση Κινητών Χρηστών Βάσει Πληροφορίας Θέσης

Γεώργιος Σ. Μπισμπίκης

Επιβλέποντες: **Ευστάθιος Χατζηευθυμιάδης,** Επίκουρος Καθηγητής ΕΚΠΑ
Βασίλειος Παπαταξιάρχης, Υποψήφιος Διδάκτωρ ΕΚΠΑ

ΑΘΗΝΑ

ΑΥΓΟΥΣΤΟΣ 2012

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη Ομαδοποίηση Κινητών Χρηστών Βάσει Πληροφορίας Θέσης

Γεώργιος Σ. Μπισμπίκης
Α.Μ.: Μ1100

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ευστάθιος Χατζηευθυμιάδης,** Επίκουρος Καθηγητής ΕΚΠΑ
Βασίλειος Παπαταξιάρχης, Υποψήφιος Διδάκτωρ ΕΚΠΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Ευστάθιος Χατζηευθυμιάδης,** Επίκουρος Καθηγητής ΕΚΠΑ
Αλέξιος Δελής, Καθηγητής ΕΚΠΑ

Αύγουστος 2012

ΠΕΡΙΛΗΨΗ

Στις υπηρεσίες οι οποίες βασίζονται στη θέση (Location Based Services - LBS) βασικός σκοπός είναι η μεταφορά στοχευμένης πληροφορίας στους χρήστες του δικτύου. Αυτή η πληροφορία είναι άμεσα συσχετισμένη με τη θέση των χρηστών ή των ομάδων του δικτύου. Τα συστήματα αυτά εντοπίζουν τη θέση του εκάστοτε χρήστη μέσω κάποιας φορητής συσκευής που έχει πάνω του (π.χ., κινητό τηλέφωνο, tablet) και εγκαθιδρύουν μια σύνδεση με αυτή. Είναι λογικό πως όσο το πλήθος των χρηστών αυξάνεται και δεδομένης της κινητικότητάς τους, το φόρτο για το δίκτυο γίνεται σημαντικό. Μια ευρέως διαδεδομένη λύση λοιπόν σε αυτό το πρόβλημα είναι η παρακολούθηση των κινούμενων αντικειμένων (moving objects) και πιο συγκεκριμένα η παρακολούθηση κινούμενων ομάδων χρηστών με στόχο τη μείωση της συμφόρησης στο δίκτυο. Αυτό θα μπορούσε να επιτευχθεί ορίζοντας έναν αντιπρόσωπο για τις ομάδες αυτές ο οποίος θα αναλάβει τη μεταφορά της πληροφορίας στα μέλη της ομάδος του. Στην παρούσα διπλωματική εργασία παρουσιάζουμε μια ιεραρχική προσέγγιση όσον αφορά την αυτόματη ομαδοποίηση (hierarchical clustering) των κινούμενων χρηστών καθώς και μια τεχνική κατάδειξης ομάδων πιθανών να διασπαστούν στο κοντινό μέλλον. Στόχος είναι η εντατική παρακολούθηση μόνο των "άτακτων" χρηστών και ομάδων ώστε να μειωθεί ο συνολικός φόρτος του δικτύου διατηρώντας παράλληλα το ποσοστό των μη-ενημερωμένων κόμβων (απώλεια πληροφορίας) σε αποδεκτά επίπεδα. Το σύστημα αξιολογήθηκε πειραματικά με συνθετικά δεδομένα σε περιβάλλον προσομοίωσης κινητικότητας χρηστών και δυναμικής διαμόρφωσης ομάδων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπηρεσίες Βάσει Πληροφορίας Θέσης

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ομαδοποίηση, ιεραρχικοί αλγόριθμοι, κινούμενα αντικείμενα, συγχωνευτικός, ύποπτες ομάδες

ABSTRACT

The main objective of Location Based Services is the delivery of information to mobile users based on their surroundings. These systems detect the location of the user via a mobile terminal that he carries (e.g. mobile phone, table pc) and establish a connection with each terminal. However, as the number of users increases and the users continuously change their position, the system load for location-based content delivery becomes quite significant. A widely used solution to this problem is to monitor the moving objects, and if we want to be more specific, monitoring moving clusters (groups if objects) in order to reduce congestion in the network. This could be achieved by defining a representative user to each of those groups who will take the responsibility of transferring the information to the members of his group (the group he represents). In the context of this thesis we propose a hierarchical approach for automated clustering of mobile users and also an algorithm for denoting groups from clustering result as suspicious, which means that there is great possibility to lose their cohesion(because of a split event) in the near future. The aim is to monitor with high frequency only the suspicious users and groups in order to reduce the overall burden of the network while maintaining the percentage of non-updated users (loss of information) in acceptable levels. The system has been experimentally evaluated by using synthetic data in a simulated users' mobility environment with dynamic group formation support.

SUBJECT AREA: Location Based Services

KEYWORDS: clustering, hierarchical algorithms, moving objects, agglomerative, suspicious clusters

Θα ήθελα να αφιερώσω την εργασία αυτή στην οικογένειά μου που είναι δίπλα μου, με αντέχει και με στηρίζει σε όλα μου τα βήματα

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών (ειδίκευση «Επεξεργασία Σήματος για Επικοινωνίες και Πολυμέσα») του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Αθηνών. Στόχος μας, όταν συζητήσαμε το θέμα για πρώτη φορά, ήταν η εξερεύνηση μιας ερευνητικής περιοχής άμεσα συσχετιζόμενης με τις υπηρεσίες βάσει πληροφορίας θέσης. Πιο συγκεκριμένα σκοπός μας ήταν η ανάπτυξη ενός αποδοτικού σχήματος αυτόματης ομαδοποίησης χρηστών (βάσει πληροφορίας θέσης) σε ομάδες, εντός μιας περιοχής, με σκοπό την παρακολούθηση της κίνησής τους σε επόμενη φάση. Η ενασχόληση με την προσπάθεια επίλυσης ενός προβλήματος το οποίο δεν είναι τόσο θεωρητικό αλλά συναντάται σε καθημερινές εφαρμογές με βοήθησε ιδιαιτέρως στο να ασχοληθώ με ακόμα μεγαλύτερη διάθεση.

Θα ήθελα σε αυτό το σημείο να εκφράσω τις θερμότατες ευχαριστίες μου στον επιβλέποντα της διπλωματικής εργασίας, επίκουρο καθηγητή κ. Ευστάθιο Χατζηευθυμιάδη για την καθοδήγηση και την πολύτιμη συνεισφορά του σε όλη τη διάρκεια εκπόνησής της. Ακόμη, θα ήθελα να ευχαριστήσω θερμά το Βασίλειο Παπαταξιάρχη, υποψήφιο διδάκτορα του Τμήματος Πληροφορικής και Τηλεπικοινωνιών ΕΚΠΑ, για τις χρήσιμες οδηγίες και συμβουλές του καθ' όλη τη διάρκεια της εκπόνησης αυτής της εργασίας. Η συνεχής παρακολούθηση της προόδου της διπλωματικής εργασίας, οι εύστοχες επισημάνσεις τους καθώς και η αμεσότητά τους για την επίλυση οποιουδήποτε προβλήματος προέκυπτε συνετέλεσαν στη διαμόρφωση του τελικού αποτελέσματος.

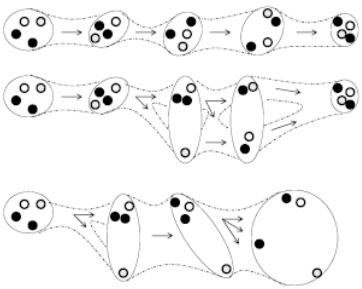
ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή Προβλήματος

Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας επικεντρωθήκαμε στην αντιμετώπιση προβλημάτων που ανακύπτουν σε υπηρεσίες βάσει πληροφορίας θέσης (Location Based Services - LBS). Οι υπηρεσίες αυτές παρέχουν πληροφορίες σε κινούμενους χρήστες βασιζόμενες στην πληροφορία που έχουν για τη τοποθεσία στην οποία βρίσκονται (πληροφορία θέσης, κοντινά αντικείμενα, χωρικά χαρακτηρισμένο περιεχόμενο) [38, 39]. Οι κινούμενοι χρήστες στέλνουν τη τοποθεσία τους σε ένα κεντρικό εξυπηρετητή ο οποίος χρησιμοποιείται για τις υπηρεσίας βασιζόμενες στη θέση (αυτό θα μπορούσε να χαρακτηριστεί και ως το σύστημα). Το σύστημα επεξεργάζεται την πληροφορία για τη θέση του κάθε κινούμενου χρήστη, εγκαθιδρύει μια σύνδεση με κάθε ένα από τα τερματικά των χρηστών και παραδίδει (πιθανότητα εξατομικευμένες) σχετικές πληροφορίες (πχ, πρόγνωση καιρού, νέα για τη κυκλοφοριακή ροή, αθλητικά νέα κα). Επίσης, μπορεί να προωθήσει περιεχόμενο, όπως διαφημιστικά μηνύματα αν ο χρήστης εισέλθει σε μια συγκεκριμένη περιοχή εντός ενός εμπορικού κέντρου ή προειδοποιητικά μηνύματα αν η κατάσταση στην κίνηση του οδικού δικτύου έχει αλλάξει καθώς και τουριστικές πληροφορίες όταν ο χρήστης επισκέπτεται μια νέα περιοχή. Εντούτοις, όσο το πλήθος των χρηστών, που πρέπει να λάβει πληροφορία άμεσα συσχετισμένη με την περιοχή που βρίσκονται, μεγαλώνει, ο φόρτος του συστήματος για την αποστολή πληροφορίας βάσει θέσης γίνεται δύσκολα διαχειρίσιμος.

Σε ένα δίκτυο κινούμενων χρηστών, η κατάτμηση των κινούμενων αντικειμένων σε ομάδες (clusters) είναι ένας ερευνητικός τομέας που έχει τραβήξει την προσοχή της ερευνητικής κοινότητας. Η ομαδοποίηση κινούμενων αντικειμένων (όπως κινούμενοι χρήστες, οχήματα, κ.ά.) ερευνά τη συμπεριφορά της κινητικότητας των αντικειμένων αυτών με στόχο την τοποθέτηση στην ίδια ομάδα, κινούμενων αντικειμένων που δεν είναι μόνο κοντά χωρικά τη δεδομένη χρονική στιγμή, αλλά είναι, επίσης, πιθανό να κινούνται μαζί για μια χρονική περίοδο όπως φαίνεται και στην παρακάτω εικόνα.



Εικόνα 1: Κινητικότητα των κόμβων εντός των ομάδων

Η ομαδοποίηση κινούμενων αντικειμένων έχει εφαρμογή σε πολλές περιοχές όπως είναι οι στοχευμένες πωλήσεις, η εξισορρόπηση του φόρτου του συστήματος, ο έλεγχος της κυκλοφοριακής συμφόρησης καθώς και η πρόβλεψή της, ο εντοπισμός ομάδων και οι εφαρμογές ανταλλαγής μηνυμάτων, ο εντοπισμός θέσης και η πλοήγησης ομάδας, ο εντοπισμός ομάδων για λόγους ασφάλειας και τέλος τα δίκτυα οχημάτων (όπου έχουμε επικοινωνία μεταξύ της υποδομής του συστήματος και των οχημάτων καθώς επίσης και μεταξύ των οχημάτων). Μπορούμε να αξιοποιήσουμε την ομαδοποίηση κινούμενων αντικειμένων για την ελαχιστοποίηση της ποσότητας μηνυμάτων που ανταλλάσσονται μεταξύ ενός υποστηρικτικού συστήματος και των κινούμενων τερματικών. Το σύστημα σε αυτή την περίπτωση, εφόσον εντοπίσει τις ομάδες που έχουν νόημα στο χώρο, στέλνει πληροφορία μόνο στους εκπροσώπους που έχουν επιλεγεί για τις ομάδες αυτές, τον ηγέτη δηλαδή της εκάστοτε ομάδας (Group Leader), από το να έστελνε ξεχωριστά σε κάθε κινούμενο χρήστη την ίδια πληροφορία.

Σε ένα προσανατολισμένο σε ομάδες δίκτυο κινούμενων αντικειμένων η κίνηση του κάθε χρήστη εντός της ομάδας του είναι συσχετισμένη σε μεγάλο βαθμό. Με τον εντοπισμό και το διαχωρισμό των χρηστών σε ομάδες στο χώρο, μπορεί να επιτευχθεί μια πιο λεπτομερής ανάλυση της συμπεριφοράς που παρουσιάζουν οι χρήστες εντός της ομάδας. Ως μια ομάδα θεωρούμε ένα συγκριτικά μικρό αριθμό χρηστών οι οποίοι αλληλεπιδρούν μεταξύ τους απευθείας μέσω επαφής πρόσωπο-με-πρόσωπο ή μέσω μιας επικοινωνίας μικρής ακτίνας με τη βοήθεια Bluetooth. Μια ομάδα ορίζεται στο [16] ως εξής: «Μια ομάδα είναι ένας αριθμός χρηστών που επικοινωνούν μεταξύ τους αρκετά συχνά εντός ενός χρονικού διαστήματος και που είναι αρκετά μικρός έτσι ώστε κάθε ένας από τους χρήστες να μπορεί να επικοινωνήσει με όλους τους υπόλοιπους, όχι έμμεσα, με τη βοήθεια άλλων χρηστών, αλλά πρόσωπο-με-πρόσωπο». Αν και η επικοινωνία πρόσωπο-με-πρόσωπο τονίζεται, ο παραπάνω ορισμός είναι κατά κάποιο τρόπο επεκτάσιμος σε ομάδες όπου η επικοινωνία μεταξύ των μελών τους επιτυγχάνεται μέσω κινητών τερματικών συσκευών (όπως κινητά τηλέφωνα, ταμπλέτες,

φορητούς υπολογιστές κα). Τα μέλη μια ομάδας έχουν αλληλοεξαρτώμενα χαρακτηριστικά όπως κοινούς στόχους, ειδική διάρθρωση της επικοινωνίας καθώς και επίγνωση του σε ποια ομάδα ανήκουν.

Ας σκεφτούμε το εξής παράδειγμα: Οι άνθρωποι γύρω από έναν κινηματογράφο έλαβαν πληροφορίες σχετικά με μια ταινία που παίζεται εκεί αλλά τελικά ανακαλύπτουν πως τα εισιτήρια έχουν εξαντληθεί. Ένας χρήστης (Χ) στέλνει αίτημα προς το σύστημα έτσι ώστε να αναζητήσει άλλους κινηματογράφους που παίζουν την ίδια ταινία. Το σύστημα προτείνει στο χρήστη μια αίθουσα κινηματογράφου, εφόσον βρει αυτό που του ζητήθηκε. Ακολούθως, είτε παρέχει τη πληροφορία αυτή και στους άλλους χρήστες της ομάδας, έτσι ώστε αυτοί να μη χρειάζεται να στείλουν επαναλαμβανόμενα αιτήματα σε αυτό, είτε δε τους παρέχει αυτή την πληροφορία αυτόματα και αφήνει στο χρήστη (Χ) την ευθύνη να την μοιράσει σε άτομα τα οποία βρίσκονται κοντά του (χρήστες δηλαδή με τους οποίους ανήκει στην ίδια ομάδα). Στην πρώτη περίπτωση το σύστημα μειώνει τον αριθμό των αιτήσεων που στέλνονται από τους χρήστες σε αυτό έτσι ώστε να λάβουν την ίδια πληροφορία. Στη δεύτερη περίπτωση, το σύστημα επικοινωνεί μόνο με ένα χρήστη έτσι ώστε να αποσταλεί η πληροφορία και στους υπόλοιπους χρήστες της ομάδας. Στο συγκεκριμένο παράδειγμα και ιδιαίτερα στη $2^{η}$ περίπτωση, ο χρήστης (Χ) παίζει το ρόλο του ηγέτης της ομάδας, όπως αυτός περιγράφηκε παραπάνω. Εντούτοις και στις δύο περιπτώσεις το σύστημα εκμεταλλεύεται τη νομαδική συμπεριφορά των χρηστών έτσι ώστε να μειώσει το φόρτο στο δίκτυο, όσον αφορά την αποστολή αιτήσεων και μηνυμάτων.

Ένα παρόμοιο σενάριο είναι αυτό του εντοπισμού ομάδων χρηστών εντός κάποιου μουσείου. Πιο συγκεκριμένα, έστω πως εντός ενός μουσείου έχουμε χρήστες οι οποίοι μας ενημερώνουν για τη θέση τους. Το σύστημα θα πρέπει να λάβει τη πληροφορία, να δει που βρίσκονται οι ομάδες αυτές εντός του χώρου που παρακολουθεί (εδώ είναι το μουσείο) και να στείλει πληροφορία άμεσα σχετιζόμενη με το έκθεμα που βρίσκεται κοντύτερα στην κάθε ομάδα. Θα μπορούσε και εδώ η πληροφορία να σταλεί σε όλους τους χρήστες των ομάδων, όπου στη συγκεκριμένη περίπτωση το κέρδος είναι πως λαμβάνουν οι χρήστες πληροφορία που μπορεί να κάνει την παρουσία τους στο μουσείο πιο ευχάριστη και εποικοδομητική. Διαφορετικά, μπορεί να σταλεί η πληροφορία που σχετίζεται με τη συγκεκριμένη θέση στον ηγέτη της ομάδας (ο οποίος θεωρητικά θα μπορούσε να είναι και ο ξεναγός της ομάδας αυτής εντός του μουσείου) και αυτός να αναλάβει το διαμοιρασμό του μηνύματος τοπικά, εντός της ομάδας. Σε αυτό το σενάριο υπάρχει και ένα έξτρα όφελος το οποίο αφορά στα επιπλέον μηνύματα

που δεν πρέπει να στείλει τελικά το σύστημα λαμβάνοντας υπόψη του το αποτέλεσμα της ομαδοποίησης. Αυτή η λογική μπορούμε να πούμε πως προσομοιώνει τη λειτουργία των Vehicular Ad-hoc Networks (VANETs).

Η δυσκολία όμως εμφανίζεται όταν η ομάδα που έχει εντοπιστεί κινείται με την πάροδο του χρόνου. Θα μπορούσαμε να χωρίσουμε το πρόβλημα αυτό των υπηρεσιών βάσει πληροφορίας θέσης σε 2 μικρότερα προβλήματα. Στο πρώτο από αυτά, θα πρέπει να βρούμε έναν τρόπο έτσι ώστε όταν οι χρήστες ενημερώνουν για τη θέση τους, το σύστημα να είναι σε θέση να τρέξει αυτόματα έναν αλγόριθμο και να εντοπίσει ποιες ομάδες υπάρχουν τελικά στην περιοχή που παρακολουθεί. Στο δεύτερο πρόβλημα θα πρέπει να βρούμε έναν τρόπο έτσι ώστε να παρακολουθούμε την κίνηση των ομάδων που έχουμε εντοπίσει. Στα πλαίσια της συγκεκριμένης έρευνας εμείς επικεντρωθήκαμε κυρίως στο πως θα μπορούσαμε να αντιμετωπίσουμε την πρώτη περίπτωση δημιουργώντας ταυτόχρονα κάποια πληροφορία την καλύτερη αντιμετώπιση του δεύτερου προβλήματος.

Όσον αφορά την περίπτωση της παρατήρησης των κινούμενων ομάδων, αυτές θα πρέπει να εποπτεύονται ως προς την κίνησή τους, καθώς, όπως φάνηκε και από την Εικόνα 1, τα κινούμενα με διαφορετικές ταχύτητες αντικείμενα εντός μιας ομάδας μπορεί να την καταστήσουν μη συμπαγή μετά από κάποιες χρονικές στιγμές. Σε αυτές τις περιπτώσεις, οι οποίες δε μπορούν να χαρακτηριστούν ως σπάνιες, το σύστημα πρέπει να ελέγχει (τακτικά) αν οι ομάδες που έχουν προκύψει από την αρχική ομαδοποίηση διατηρούν τη δομή τους, η οποία σχετίζεται με το πλήθος των χρηστών που έχει η κάθε ομάδα καθώς και με το ποιοι είναι αυτοί οι χρήστες, έτσι ώστε να μπορεί να στέλνει τη πληροφορία που θέλει μόνο στους ηγέτες των ομάδων αυτών. Επιπροσθέτως, το μεγάλο πλήθος καθώς και η συνεχής αλλαγή της θέσης των χρηστών αυξάνει το φόρτο του συστήματος καθώς πρέπει να υπολογίζονται οι νέες θέσεις αρκετά συχνά, αυξάνοντας έτσι την επιβάρυνση του δικτύου. Αν και η εκμετάλλευση της νομαδικής συμπεριφοράς των χρηστών έχει προταθεί ως τρόπος αντιμετώπισης του προβλήματος που σχετίζεται με το φόρτο του δικτύου [17], η απόφαση του συστήματος να επικοινωνήσει μόνο με τους ηγέτες των ομάδων και όχι με όλα τα μέλη της ομάδας θα πρέπει να επικυρώνεται σε τακτά χρονικά διαστήματα. Τέτοιες αποφάσεις είναι άμεσα συνδεδεμένες με το πόσο συμπαγείς και συνεκτικές είναι οι ομάδες που προκύπτουν από τον αλγόριθμο ομαδοποίησης. Το σύστημα θα πρέπει να είναι αρκετά «ισχυρό» έτσι ώστε να αποφασίζει με μεγάλη σιγουριά για το αν μια ομάδα μένει συμπαγής ή όχι για μια συγκεκριμένη χρονική περίοδο και να παίρνει απόφαση για το αν μπορεί να

βασίσει την επικοινωνία του με την κάθε ομάδα στέλνοντας πληροφορίες μόνο στον ηγέτη της εκάστοτε ομάδας. Αυτό λοιπόν μας έδωσε το κίνητρο για τη δημιουργία ενός μηχανισμού που εκτός από το να εντοπίζει τις ομάδες εντός μιας περιοχής ενδιαφέροντος αυτόματα, θα δίνει και ένα χαρακτηρισμό σε αυτές ο οποίος θα σχετίζεται με το αν η ομάδα που βρέθηκε έχει πιθανότητα αργότερα να χαλάσει τη δομή της έτσι ώστε το σύστημα να παρακολουθεί όλα της τα μέλη.

Η δουλειά μας συνοψίζεται στα εξής: εισάγουμε τη χρησιμοποίηση συγχωνευτικών ιεραρχικών αλγορίθμων για την αυτόματη ομαδοποίηση των χρηστών μέσα σε μια περιοχή παρακολούθησης καθώς και την ενσωμάτωση ενός κριτηρίου μέτρησης της ποιότητας μιας ομαδοποίησης για τον εντοπισμό της καλύτερης (της πιο αντιπροσωπευτικής) από αυτές που δίνει ένας ιεραρχικός αλγόριθμος. Τέλος, παρουσιάζουμε μια μέθοδο κατάδειξης «ύποπτων» ομάδων (όσον αφορά την κινητικότητα των χρηστών που τις απαρτίζουν) από αυτές που προέκυψαν μέσω του εντοπισμού της βέλτιστης ομαδοποίησης. Σκοπός είναι η εκμετάλλευση της συγκεκριμένης κατάδειξης ως επιπλέον πληροφορία στο σύστημά μας για το ποιες ομάδες θα πρέπει να παρακολουθεί μέσω των ηγετών τους μόνο και ποιες πρέπει να τις ελέγχει συνέχεια για τη θέση των μελών τους, κάτι το οποίο αφορά το επίπεδο της παρακολούθησης της κίνησης αυτών των ομάδων στο χώρο.

1.2 Σχετική Έρευνα

1.2.1 Ομαδοποίηση Στατικών Χωρικών Δεδομένων

Η ομαδοποίηση στατικών χωρικών δεδομένων (πχ., στατικών σημείων) είναι ένα καλά μελετημένο θέμα στη βιβλιογραφία. Μεγάλο πλήθος παραδειγμάτων ομαδοποίησης τέτοιων τύπων δεδομένων έχει προταθεί με διαφορετικούς ορισμούς και κριτήρια αξιολόγησης, βάσει του αντικειμένου που αφορά η ομαδοποίηση. Μέθοδοι κατάτμησης, όπως η $k-medoids$ μέθοδος [7, 8], χωρίζουν τα αντικείμενα σε k ομάδες και επαναληπτικά ανταλλάσουν αντικείμενα μεταξύ τους μέχρι η ποιότητα των ομάδων να μη βελτιώνεται πλέον. Αρχικά, επιλέγονται k medoids τυχαία από το σετ των δεδομένων που έχουμε. Κάθε αντικείμενο τοποθετείται στην ομάδα που έχει το κοντινότερο medoid και η ποιότητα των ομάδων ορίζεται αθροίζοντας τις αποστάσεις όλων των σημείων από το κοντινότερο σε αυτά medoid. Έπειτα, το medoid αντικαθίσταται από ένα τυχαίο σημείο και η αλλαγή οριστικοποιείται μόνο αν οδηγεί σε ομάδες με καλύτερη ποιότητα.

Ένα τοπικό βέλτιστο επιτυγχάνεται έπειτα από μια μεγάλη ακολουθία ανεπιτυχών αντικαταστάσεων. Η παραπάνω διαδικασία επαναλαμβάνεται για ένα πλήθος σετ από τυχαία αρχικά medoids και οι ομάδες οριστικοποιούνται βάσει εκείνου του σετ που έδωσε το μεγαλύτερο τοπικό βέλτιστο.

Μια άλλη κατηγορία τεχνικών ομαδοποίησης είναι οι ιεραρχικές μέθοδοι και πιο συγκεκριμένα οι συγχωνευτικές (agglomerative) οι οποίες ορίζουν τις ομάδες με έναν «από κάτω προς τα πάνω» τρόπο, υποθέτοντας αρχικά πως όλα τα αντικείμενα ορίζουν μια ομάδα το καθένα και σταδιακά συγχωνεύουν το κοντινότερο ζεύγος ομάδων μέχρι να μείνει ένας επιθυμητό πλήθος ομάδων βάσει κάποιου κριτηρίου διακοπής. Αλγόριθμοι όπως οι BIRCH [9], που θα αναλυθεί εκτενέστερα παρακάτω, και CURE [10] προτάθηκαν έτσι ώστε να βελτιώσουν την επεκτασιμότητα της συγχωνευτικής ιεραρχικής ομαδοποίησης καθώς και την ποιότητα των κατατμήσεων των δεδομένων που προκύπτουν από τέτοιες μεθόδους. Ο C2P [11] είναι ένας άλλος αλγόριθμος, όμοιος με τον CURE, ο οποίος ενσωματώνει αλγορίθμους κοντινότερου ζεύγους και χρησιμοποιεί ένα χωρικό ευρετήριο για τη βελτίωση της επεκτασιμότητάς του.

Μέθοδοι βασιζόμενοι στην πυκνότητα εντοπίζουν πυκνές περιοχές στο χώρο, όπου τα αντικείμενα είναι πολύ κοντά μεταξύ τους και τα διαχωρίζει από περιοχές χαμηλής πυκνότητας. Ο DBSCAN [25] είναι ο πιο αντιπροσωπευτικός αλγόριθμος αυτής της κατηγορίας. Αρχικά, επιλέγει ένα σημείο p από το σύνολο των δεδομένων. Μια χωρική επερώτηση εύρους, για περιοχή κέντρου p και ακτίνας ε εφαρμόζεται για να εξακριβωθεί αν στη γειτονιά του p περιλαμβάνεται ένα πλήθος σημείων τουλάχιστον ίσο με $MinPts$ (δηλαδή αν πρόκειται για μια πυκνή περιοχή). Αν ισχύει κάτι τέτοιο τότε τα σημεία αυτά τοποθετούνται στην ίδια ομάδα με το σημείο p και αυτή η διαδικασία επαναλαμβάνεται για τα νέα σημεία της ομάδας αυτής. Ο DBSCAN συνεχίζει μέχρι να μην μπορεί να επεκταθεί περαιτέρω η συγκεκριμένη ομάδα και την καταδεικνύει ως μια συμπαγή περιοχή. Η διαδικασία επαναλαμβάνεται για όλα τα σημεία που δεν έχουν ελεγχθεί μέχρι να εντοπιστούν όλες οι ομάδες καθώς και οι ακραίες τιμές. Άλλη μια μέθοδος της κατηγορίας αυτής είναι το σύστημα OPTICS [12]. Λειτουργεί παρόμοια με τον DBSCAN αλλά δεν εντοπίζει το διαχωρισμό του συνόλου των δεδομένων σε ομάδες. Αντιθέτως, επιστρέφει μια διάταξη των σημείων του συνόλου δεδομένων η οποία χρησιμοποιείται σε ένα δεύτερο βήμα για τον εντοπισμό ομάδων για διάφορες τιμές του ε, δηλαδή της ακτίνας ελέγχου για την πυκνότητα μιας περιοχής.

Αν και οι παραπάνω μέθοδοι μπορούν να χρησιμοποιηθούν για να βρεθεί η στιγμιαία απεικόνιση των ομάδων για μια δοθείσα χρονική στιγμή, δε μπορούν να

χρησιμοποιηθούν απευθείας στην αναγνώριση κινούμενων αντικειμένων αλλά ούτε και στην παρακολούθησή τους. Μια απλή τεχνική που συναντάται συχνά είναι η ομαδοποίηση των δεδομένων περιοδικά. Εντούτοις, αν η περίοδος είναι μικρή τότε η προσέγγιση αυτή είναι αρκετά ακριβή. Αυτό συμβαίνει διότι η προσπάθεια που καταναλώθηκε για τον εντοπισμό της προηγούμενης ομαδοποίησης δεν αξιοποιήθηκε αναλόγως. Αντιθέτως, αν η περίοδος είναι μεγάλη τότε για αρκετό χρονικό διάστημα δεν θα υπάρχει ενημερωμένη πληροφορία για την ομαδοποίηση των δεδομένων. Τέλος, ένα άλλο στοιχείο των παραπάνω προσεγγίσεων είναι πως, για την περίπτωση που εξετάζουμε, απαιτούν από το σύστημα να τους ορίσει το πόσες ομάδες πρέπει να εντοπίσουν και σε μερικούς από τους αλγορίθμους αυτούς χρειάζεται να δοθούν και κάποιοι αρχικοί εκπρόσωποι των ομάδων. Στη δική μας προσέγγιση, αντιθέτως, επιτρέπουμε στα συστήματα που λειτουργούν βάσει πληροφορίας θέσης να εντοπίζουν αυτόματα τις ομάδες εντός του χώρου που παρακολουθούν οποιαδήποτε στιγμή κρίνεται απαραίτητο πως κάτι τέτοιο χρειάζεται.

1.2.2 Ομαδοποίηση Χωροχρονικών Δεδομένων

Οι προηγούμενες μέθοδοι για την ομαδοποίηση χωροχρονικών δεδομένων εστιάζουν κυρίως στην ομαδοποίηση τροχιών όμοιου σχήματος. Η μονοδιάστατη έκδοση αυτού του προβλήματος αντιστοιχεί στην ομαδοποίηση χρονοσειρών που παρουσιάζουν παρόμοιες κινήσεις. Όπως είναι κατανοητό όμως το πρόβλημα της ομαδοποίησης τροχιών είναι διαφορετικό από τον εντοπισμό και την παρακολούθηση ομάδων που κινούνται. Η κύρια διαφορά εντοπίζεται στο γεγονός πως μια ομάδα από τροχιές έχει ένα σταθερό πλήθος αντικειμένων κατά τη διάρκεια της ζωής της σε αντίθεση με τα μέλη μιας κινούμενης ομάδας τα οποία και μπορεί να μεταβληθούν κατά τη διάρκεια ζωής της. Για το λόγο αυτό η έρευνα έχει επικεντρωθεί πλέον στη συνεχή ομαδοποίηση κινούμενων αντικειμένων. Η συνεχής ομαδοποίηση κινούμενων αντικειμένων λαμβάνει υπόψη της την τρέχουσα θέση του κάθε αντικειμένου σε μια ομάδα και συνδυάζει μια πρόβλεψη για τη μελλοντική του θέση σε σχέση με την ομάδα αυτή. Επιπροσθέτως, σε ένα σχήμα ομαδοποίησης κινούμενων αντικειμένων θα πρέπει να είναι δυνατός ο εντοπισμός οποιασδήποτε αλλαγής υπάρξει στη δομή των ομάδων κατά τη διάρκεια της συνεχούς παρακολούθησης της κίνησης του πληθυσμού. Αυτό είναι υποχρεωτικό έτσι ώστε να υπάρχει καλύτερη αντίληψη για την κίνηση των αντικειμένων.

Το μοντέλο που προτείνεται στο [14] υιοθετεί τις μικρο-ομάδες [9] για τα κινούμενα αντικείμενα και διατηρεί δυναμικά οριοθετημένα παράθυρα από ομάδες έτσι ώστε να

εντοπίσει αλλαγές στις ήδη σχηματισμένες ομάδες. Αυτό επιτυγχάνεται με τη συνεχή αναφορά της θέσης του κάθε αντικειμένου. Όμως, το πλήθος των ελέγχων για τη διατήρηση της δομής των ομάδων (έλεγχος για το αν ένα κινούμενο αντικείμενο βρίσκεται εκτός του οριοθετημένου παραθύρου) είναι απαγορευτικό όσον αφορά την υπολογιστική πολυπλοκότητα. Δοθείσης μιας κινούμενης μικρο-ομάδας με n αντικείμενα, οι έλεγχοι για τη διατήρηση της δομής της, οι οποίοι αφορούν τα αντικείμενα που βρίσκονται στα άκρα του οριοθετημένου παραθύρου, μπορεί να γίνουν και $O(n)$ φορές κατά τη διάρκεια της κίνησης όπως αναφέρεται στο [15]. Το προτεινόμενο σχήμα στο [15] διευρύνει την έννοια του χαρακτηριστικού ομαδοποίησης (clustering feature - CF) και του δέντρου χαρακτηριστικών (CF tree) του αλγορίθμου BIRCH για την εφαρμογή σε συνεχή ομαδοποίηση κινούμενων αντικειμένων. Και αυτό το σχήμα όμως απαιτεί συνεχή αναφορά της θέσης των κινούμενων αντικειμένων έτσι ώστε να τροφοδοτείται η βαθμίδα πρόβλεψης για κάθε κινούμενο αντικείμενο καθώς και για την παρακολούθηση των ήδη σχηματισμένων ομάδων έτσι ώστε να ληφθούν αποφάσεις συγχώνευσης ή διάσπασης.

Οι παραπάνω μέθοδοι έχουν όλες ως κοινό αρνητικό την απαίτηση το σύστημα να προκαθορίζει το πόσες ομάδες θα πρέπει να εντοπιστούν, έστω και έμμεσα με την εισαγωγή κάποιον παραμέτρων. Θα ήταν προτιμότερο να υπάρχει ένα σύστημα το οποίο θα εντοπίζει τις ομάδες αυτόματα και θα είναι σε θέση να παρακολουθήσει την κίνησή τους χωρίς την οποιαδήποτε παρέμβαση. Αυτό λοιπόν είναι που προσπαθούμε να επιτύχουμε με το σύστημά μας. Με τη χρήση του συγχωνευτικού ιεραρχικού αλγορίθμου και ενός κριτηρίου εύρεσης της βέλτιστης ομαδοποίησης είναι σε θέση να εντοπίζει χωρίς μεγάλα σφάλματα τις ομάδες που υπάρχουν σε μια περιοχή και προσφέρει επιπλέον πληροφορία όσον αφορά τις πιο «κινητικές» ομάδες έτσι ώστε να μπορούν να παρακολουθούνται με μεγαλύτερη ακρίβεια.

Στην προσέγγισή μας το σύστημα αρχικά εντοπίζει αυτόματα τις ομάδες που βρίσκονται στο χώρο εποπτείας του κάποια χρονική στιγμή και στη συνέχεια θα θέλαμε να τις παρακολουθεί κάνοντας περιοδικά εκ νέου ομαδοποίηση των δεδομένων έτσι ώστε να επαναορίζει τις ομάδες. Δηλαδή, παρατηρούμε πως υπάρχει μια αντιμετώπιση του προβλήματος σε δύο επίπεδα. Όπως γίνεται κατανοητό, παίζει πολύ μεγάλο ρόλο το κάθε πότε θα γίνεται αυτή η επαναομαδοποίηση καθώς πρόκειται για μια αρκετά κοστοβόρα διαδικασία σε υπολογιστική πολυπλοκότητα. Η λειτουργία της επαναομαδοποίησης θα μπορούσε να ενεργοποιείται με δύο τρόπους: (α) να έχει περάσει ένα προκαθορισμένο χρονικό διάστημα (περίοδος T) από την προηγούμενη

ομαδοποίηση και (β) στις ομάδες που έχουν δημιουργηθεί να παρατηρηθεί κάποια συγχώνευση ή διάσπαση. Το δεύτερο επίπεδο δε θα ερευνηθεί στα πλαίσια αυτής της διπλωματικής εργασίας εκτενώς, αλλά θα παρουσιαστεί το πως σκεφτόμαστε να χρησιμοποιηθεί η λύση που προτείνουμε για το πρώτο επίπεδο, έτσι ώστε να ενισχυθεί η σωστή λειτουργία του δεύτερου επιπέδου. Εν κατακλείδι, θέλουμε μέσω του προτεινόμενου συστήματος να ισορροπήσουμε την ποιότητα της ομαδοποίησης και τη διατήρηση της δομής της με τη μείωση της κίνησης των μηνυμάτων στο δίκτυο.

1.3 Στόχοι Εργασίας

Το αντικείμενο μελέτης της παρούσας εργασίας σχετίζεται με τις υπηρεσίες βάσει πληροφορίας θέσης και αφορά την αυτόματη εύρεση ομάδων χρηστών σε μια περιοχή εποπτείας και την παρακολούθησή τους. Οι στόχοι της εργασίας συνοψίζονται στην (α) ενσωμάτωση ενός ιεραρχικού σχήματος ομαδοποίησης για τον αυτόματο εντοπισμό ομάδων χρηστών στο χώρο ενδιαφέροντος μιας τέτοιας υπηρεσίας καθώς και (β) την εύρεση μιας μεθόδου κατάδειξης ύποπτων ομάδων στο χώρο, με σκοπό την εκμετάλλευση αυτής της πληροφορίας για παρακολούθηση των δημιουργηθέντων ομάδων στη συνέχεια.

1.4 Οργάνωση Εργασίας

Η οργάνωση της εργασίας έχει ως εξής: στο 2^{ο} Κεφάλαιο γίνεται μια γενική αναφορά στον όρο «ομαδοποίηση». Περιγράφεται τη σημαίνει ο όρος ομαδοποίηση, ποιες κατηγορίες αλγορίθμων ομαδοποίησης υπάρχουν, πώς μπορεί να επιλεγεί ο κατάλληλος αλγόριθμος από αυτούς κάθε φορά, ενώ τέλος δίνεται μια εισαγωγή σχετικά με την ομαδοποίηση επάνω σε ροές δεδομένων. Στο 3^{ο} Κεφάλαιο, που αποτελεί και τον πυρήνα της παρούσας εργασίας, παρουσιάζονται οι ιεραρχικοί αλγόριθμοι και εκτενέστερα οι συγχωνευτικοί ιεραρχικοί αλγόριθμοι. Στη συνέχεια, δίνεται το μέτρο βάσει του οποίου οι αλγόριθμοι αυτοί επιλέγουν αυτόματα την καλύτερη ομαδοποίηση και τέλος παρουσιάζεται η μέθοδος κατάδειξης ύποπτων ομάδων που αναπτύξαμε. Το 4^{ο} Κεφάλαιο περιλαμβάνει την πειραματική αξιολόγηση του προτεινόμενου συστήματος. Στο πρώτο μέρος του παρουσιάζονται οι μετρικές που χρησιμοποιήθηκαν για την πειραματική αποτίμηση του συστήματος, το εργαλείο που χρησιμοποιήθηκε για την παραγωγή των συνόλων δεδομένων που χρειαζόμασταν για να ελέγξουμε το σύστημά μας καθώς και η μεθοδολογία πραγματοποίησης των πειραμάτων. Στο δεύτερο και

τελευταίο μέρος του 4ου Κεφαλαίου παρουσιάζονται και σχολιάζονται τα αποτελέσματα των πειραμάτων που έγιναν για τον έλεγχο του συστήματος. Η εργασία ολοκληρώνεται με το 5ο Κεφάλαιο στο οποίο παρουσιάζονται συγκεντρωτικά τα συμπεράσματα της μελέτης και αναφέρονται κάποια ανοιχτά θέματα που αφορούν το συγκεκριμένο σύστημα καθώς και την περιοχή έρευνας, για την επίλυση μερικών από τα προβλήματα της οποίας αρχικά σχεδιάστηκε.

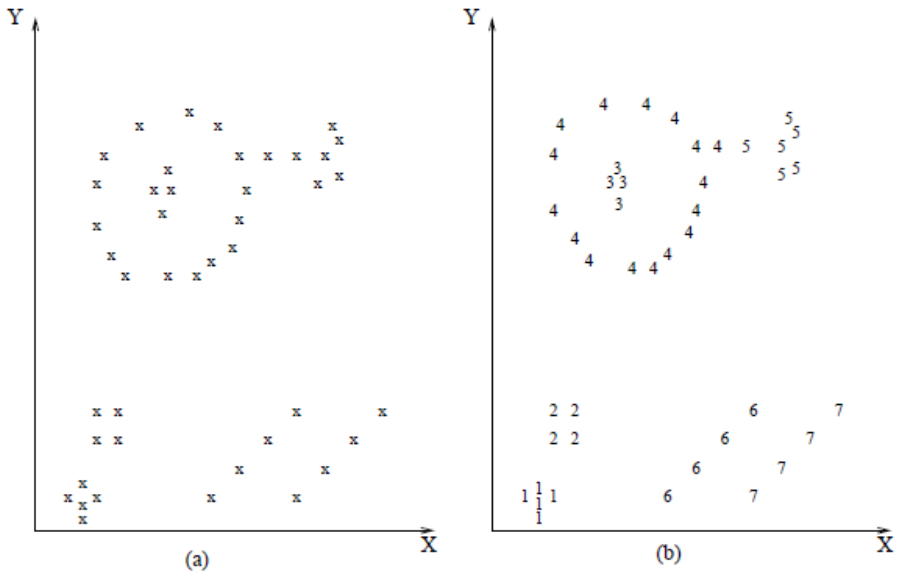
ΚΕΦΑΛΑΙΟ 2

ΟΜΑΔΟΠΟΙΗΣΗ

2.1 Εισαγωγή

Η ανάλυση δεδομένων (data analysis) βρίσκεται πίσω από πολλές εφαρμογές της πληροφορικής, είτε στο κομμάτι του σχεδιασμού είτε στο κομμάτι των online λειτουργιών τους. Οι διαδικασίες της ανάλυσης δεδομένων μπορούν να διαχωριστούν σε διερευνητικές (exploratory procedures) και επιβεβαιωτικές (confirmatory procedures) με βάσει τη διαθεσιμότητα των κατάλληλων μοντέλων για τις πηγές δεδομένων. Ένα βασικό στοιχείο για τους δύο τύπους διεργασιών στην ανάλυση δεδομένων (είτε πρόκειται για σχηματισμό υπόθεσης είτε για λήψη απόφασης) είναι η ομαδοποίηση ή ταξινόμηση των μετρήσεων η οποία βασίζεται είτε (i) στο πόσο καλά ταιριάζει σε ένα υποτιθέμενο μοντέλο, είτε (ii) στις φυσικές ομαδοποιήσεις (clustering) που εντοπίζονται μέσω της ανάλυσης. Η ανάλυση κατά συστάδες (cluster analysis) είναι η οργάνωση μιας συλλογής προτύπων (τα οποία συνήθως αναπαρίστανται ως ένα διάνυσμα μετρήσεων ή ως ένα σημείο σε πολυδιάστατο χώρο) σε ομάδες βάσει της ομοιότητας που έχουν μεταξύ τους. Διαισθητικά, τα πρότυπα (patterns) που ανήκουν σε μια έγκυρη ομάδα (valid group) είναι περισσότερο όμοια μεταξύ τους συγκριτικά με εκείνα τα πρότυπα που ανήκουν σε διαφορετικές ομάδες. Ένα παράδειγμα ομαδοποίησης φαίνεται και στην ακόλουθη Εικόνα 2. Τα πρότυπα που αποτελούν την είσοδο φαίνονται στην Εικόνα 2(a) και οι επιθυμητές ομάδες είναι ορατές στην Εικόνα 2(b). Εδώ, τα σημεία που ανήκουν στις ίδιες ομάδες έχουν τις ίδιες ετικέτες για να διαχωρίζονται από εκείνα που ανήκουν σε διαφορετικές ομάδες. Η ποικιλία τεχνικών για την αναπαράσταση δεδομένων, τη μέτρηση της εγγύτητας (ομοιότητα ή ανομοιότητα) καθώς και την ομαδοποίηση των στοιχείων των δεδομένων έχει δημιουργήσει μια πλούσια λίστα μεθόδων ομαδοποίησης, κάτι όμως που προκαλεί πολλές φορές σύγχυση. Είναι πολύ σημαντικό να καταστεί σαφής η διαφορά μεταξύ της ομαδοποίησης (κατηγοριοποίηση χωρίς επίβλεψη – unsupervised classification) και της διακρίνουσας ανάλυσης (κατηγοριοποίηση υπό επίβλεψη – supervised classification). Στην κατηγοριοποίηση (classification) υπό επίβλεψη, παρέχεται μια συλλογή προτύπων που φέρουν ετικέτες (labels), οι οποίες υποδηλώνουν την κατηγορία στην οποία ανήκουν, και το πρόβλημα έχει να κάνει με το να ορίσουμε την ετικέτα ενός νέου προτύπου το οποίο δεν φέρει ετικέτα, δηλαδή δε δίνεται πληροφορία σχετικά με την

κατηγορία στην οποία ανήκει. Τυπικά, τα δοσμένα πρότυπα με ετικέτες χρησιμοποιούνται για να μάθουμε την περιγραφή των κατηγοριών οι οποίες στη συνέχεια χρησιμοποιούνται για να βάλουμε ετικέτες στα νέα πρότυπα που εμφανίζονται. Από την άλλη μεριά, στην περίπτωση της ομαδοποίησης (clustering), ο σκοπός είναι να χωριστούν σε ομάδες που έχουν νόημα, πρότυπα τα οποία δεν φέρουν ετικέτες. Κατά μία έννοια, οι ετικέτες και εδώ έχουν σχέση και με τις ομάδες, απλά αυτή η κατηγορία ετικετών προέρχεται αποκλειστικά από τα δεδομένα που έχουμε ως είσοδο κάθε φορά.



Εικόνα 2: Ομαδοποίηση Δεδομένων

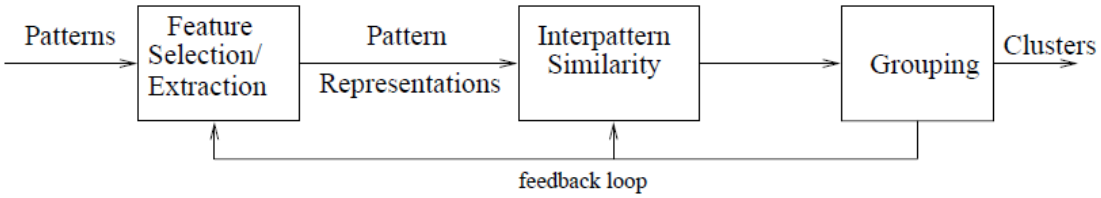
Η ανάλυση κατά συστάδες, είναι χρήσιμη σε πολλές περιπτώσεις διερευνητικής ανάλυσης προτύπων όπως είναι η ομαδοποίηση, η λήψη απόφασης και η εκμάθηση μηχανής. Τέτοιες καταστάσεις είναι η εξόρυξη δεδομένων, η κατάτμηση εικόνας, η ανάκτηση εγγράφων και η κατηγοριοποίηση προτύπων. Ωστόσο, σε πολλά από αυτά τα προβλήματα, η ύπαρξη πρότερης γνώσης (π.χ., εμπειρικά μοντέλα) σχετικά με τα δεδομένα είναι αρκετά περιορισμένη και επίσης πρέπει να γίνονται όσο το δυνατόν λιγότερες υποθέσεις σχετικά με τα δεδομένα. Υπό αυτούς τους περιορισμούς, η μεθοδολογία διαχωρισμού σε συστάδες είναι κάτι παραπάνω από κατάλληλη για την εξερεύνηση των σχέσεων μεταξύ των δεδομένων έτσι ώστε να πραγματοποιηθεί (προκαταρκτικά έστω) μια αξιολόγηση της δομής τους. Ο όρος «ομαδοποίηση» χρησιμοποιείται σε αρκετά ερευνητικά πεδία για να περιγράψει τις μεθόδους για τη δημιουργία ομάδων σε δεδομένα χωρίς ετικέτες.

2.2 Τα Βήματα μιας Διαδικασίας Ομαδοποίησης

Τυπικά, η διαδικασία ομαδοποίησης περιλαμβάνει τα ακόλουθα βήματα:

(1) Αναπαράσταση των προτύπων (προαιρετικά περιλαμβάνει και εξαγωγή ή/και επιλογή χαρακτηριστικών)
(2) Προσδιορισμός ενός μέτρου εγγύτητας κατάλληλου για τη μορφή των δεδομένων
(3) Ομαδοποίηση
(4) Αφαίρεση δεδομένων (αν χρειάζεται) και
(5) Αξιολόγηση της εξόδου (αν χρειάζεται)

Στην Εικόνα 3 απεικονίζεται μια τυπική ακολουθία των τριών πρώτων βημάτων από αυτά που δόθηκαν παραπάνω η οποία συμπεριλαμβάνει και ένα μονοπάτι ανατροφοδότησης για το αποτέλεσμα της διαδικασίας ομαδοποίησης, το οποίο ακολούθως μπορεί να επηρεάσει την εξαγωγή χαρακτηριστικών καθώς και τους υπολογισμούς για το μέτρο ομοιότητας που έχει επιλεγεί.



Εικόνα 3: Βήματα Ομαδοποίησης

Η αναπαράσταση προτύπων (pattern representation) σχετίζεται με το πλήθος των κατηγοριών, τον αριθμό των διαθέσιμων προτύπων καθώς και τον αριθμό, τον τύπο και την κλίμακα των χαρακτηριστικών που υπάρχουν διαθέσιμα για τον αλγόριθμο ομαδοποίησης. Βέβαια, κάποια από αυτή την πληροφορία μπορεί να μην είναι διαχειρίσιμη από το χρήστη που υλοποιεί ένα τέτοιο αλγόριθμο. Η επιλογή χαρακτηριστικού (feature selection) είναι η διαδικασία εύρεσης του πιο αποτελεσματικού υποσυνόλου από τα αρχικά χαρακτηριστικά για να χρησιμοποιηθεί στην ομαδοποίηση. Η εξαγωγή χαρακτηριστικού (feature extraction) είναι η χρήση μίας ή περισσότερων παραλλαγών των χαρακτηριστικών εισόδου για την παραγωγή ενός νέου και σημαντικότερου (αντιπροσωπευτικότερου) χαρακτηριστικού. Μία (εννίοτε και οι δύο) από αυτές τις μεθόδους μπορεί να χρησιμοποιηθεί για την εύρεση ενός αντιπροσωπευτικού συνόλου χαρακτηριστικών που θα χρησιμοποιηθεί στη διαδικασία της ομαδοποίησης.

Η εγγύτητα μεταξύ των προτύπων (pattern proximity) συνήθως μετριέται με τη βοήθεια μιας συνάρτησης απόστασης η οποία ορίζεται για ζευγάρια προτύπων. Υπάρχει μια πληθώρα από μετρικές απόστασης οι οποίες χρησιμοποιούνται από τις διάφορες ερευνητικές κοινότητες. Ένα απλό μέτρο υπολογισμού της απόστασης, όπως είναι η Ευκλείδεια απόσταση, συχνά χρησιμοποιείται για την αποτύπωση της ανομοιότητας μεταξύ δύο προτύπων, ενώ άλλα μέτρα ομοιότητας μπορούν να χρησιμοποιηθούν για το χαρακτηρισμό της εννοιολογικής ομοιότητας μεταξύ των προτύπων. Οι διάφορες μετρικές απόστασης, που σχετίζονται με την περιοχή έρευνας των υπηρεσιών βάσει πληροφορίας θέσης, θα αναλυθούν παρακάτω και πιο συγκεκριμένα στο κομμάτι της περιγραφής του συγχωνευτικών ιεραρχικών αλγορίθμων.

Το κομμάτι της ομαδοποίησης μπορεί να επιτευχθεί με διαφόρους τρόπους. Το αποτέλεσμα αυτής της διαδικασίας μπορεί να είναι αυστηρό (κατάτμηση των δεδομένων σε ομάδες) ή ασαφές (το κάθε πρότυπο έχει ένα βαθμό συμμετοχής σε κάθε μία από τις ομάδες που προέκυψαν). Οι ιεραρχικοί αλγόριθμοι (hierarchical algorithms) παράγουν ένα εμφωλευμένο σύνολο από διαφορετικές κατατμήσεις των δεδομένων με τη βοήθεια ενός κριτηρίου για τη συγχώνευση (συγχωνευτικός) ή το διαχωρισμό (διαιρετικός) ομάδων βασισμένο στις ομοιότητες (ή ανομοιότητες) μεταξύ τους. Οι αλγόριθμοι κατάτμησης (partitioning algorithms) εντοπίζουν την κατάτμηση του συνόλου των δεδομένων με τρόπο τέτοιο ώστε να βελτιστοποιείται κάποιο κριτήριο ομαδοποίησης, κάτι το οποίο συνήθως έχει τοπική εμβέλεια. Άλλες τεχνικές για τη διαδικασία της ομαδοποίησης περιλαμβάνουν πιθανοτικές (probabilistic) και βασισμένες σε θεωρία γράφων (graph-theoretic) μεθόδους.

Η αφαίρεση δεδομένων (data abstraction) είναι η διαδικασία εύρεσης μιας συμπαγούς και αντιπροσωπευτικής αναπαράστασης ενός συνόλου δεδομένων. Εδώ, η απλότητα σχετίζεται είτε με το κομμάτι της αυτόματης ανάλυσης, έτσι ώστε μια μηχανή να μπορεί να εκτελέσει αποδοτικά περαιτέρω επεξεργασία, είτε είναι ανθρωποκεντρική, έτσι ώστε η αναπαράσταση που παράγεται να είναι κατανοητή και διαισθητικά ελκυστική. Στο πλαίσιο της ομαδοποίησης, η τυπική αφαίρεση δεδομένων αναφέρεται σαν μια συμπαγής αναπαράσταση κάθε ομάδας που τελικά προκύπτει, συνήθως με τη μορφή προτύπων που παίζουν το ρόλο του εκπροσώπου, όπως είναι το κεντροειδές (centroid) μιας ομάδας.

Πώς όμως αξιολογείται η έξοδος ενός αλγορίθμου ομαδοποίησης; Τι χαρακτηρίζει ένα «καλό» αποτέλεσμα ομαδοποίησης και τι ένα «κακό»; Όλοι οι αλγόριθμοι ομαδοποίησης, εφόσον τους δοθούν δεδομένα εισόδου θα παράξουν κάποιο

αποτέλεσμα που θα περιέχει τις ομάδες στις οποίες διαχώρισαν τα δεδομένα ανεξάρτητα από το αν τελικά τα δεδομένα που τους δόθηκαν περιέχουν ή όχι ομάδες. Αν τα αρχικά δεδομένα όντως περιέχουν ομάδες, τότε κάποιοι αλγόριθμοι ομαδοποίησης θα δώσουν «καλύτερες» ομάδες σε σχέση με άλλους. Η αξιολόγηση λοιπόν του αποτελέσματος μιας διαδικασίας ομαδοποίησης έχει πολλές πτυχές. Η μία από αυτές στην πραγματικότητα είναι η εκτίμηση του τομέα των δεδομένων παρά του αλγορίθμου ομαδοποίησης αυτού καθεαυτού. Αυτό σημαίνει πως δεδομένα τα οποία δεν περιέχουν ομάδες δεν θα πρέπει να επεξεργάζονται από τέτοιους αλγορίθμους. Εντούτοις, η μελέτη της τάσης των ομάδων (cluster tendency), όπου τα δεδομένα εισόδου εξετάζονται για να διαπιστωθεί αν υπάρχει κάποια αξία από την ανάλυσή τους σε ομάδες, πριν αυτή πραγματοποιηθεί, είναι μια σχετικά μη ενεργή ερευνητικά περιοχή, γι' αυτό και δεν θα αναλυθεί περαιτέρω.

Η εγκυρότητα μιας ομαδοποίησης (clustering validity), είναι ουσιαστικά η διαδικασία αξιολόγησης του αποτελέσματος που προκύπτει από έναν αλγόριθμο ομαδοποίησης. Συχνά, αυτή η ανάλυση χρησιμοποιεί ένα συγκεκριμένο κριτήριο βελτιστοποίησης, εντούτοις τα περισσότερα από αυτά τα κριτήρια συνήθως βασίζονται στην υποκειμενικότητα. Επομένως, λίγες είναι οι περιπτώσεις που μπορούν να χρησιμοποιηθούν ως «χρυσός κανόνας» και αυτές είναι σε πολύ καλά ορισμένες ερευνητικές υποπεριοχές. Οι αξιολογήσεις εγκυρότητας είναι αντικειμενικές και πραγματοποιούνται ώστε να προσδιορίσουν αν ένα αποτέλεσμα ομαδοποίησης έχει νόημα. Γενικά, μια δομή ομαδοποίησης είναι έγκυρη αν δε μπορεί να έχει προκύψει κατά τύχη ή ως τεχνούργημα ενός αλγορίθμου ομαδοποίησης. Όταν χρησιμοποιούνται στατιστικές προσεγγίσεις στην ομαδοποίηση, η επικύρωση επιτυγχάνεται με τη προσεκτική εφαρμογή στατιστικών μεθόδων και τον έλεγχο υποθέσεων. Υπάρχουν τρεις κύριες μέθοδοι επικύρωσης:

- Η *εξωτερική αξιολόγηση* της εγκυρότητας (external assessment of validity) συγκρίνει την προκύπτουσα δομή με μια η οποία έχει προκύψει νωρίτερα.
- Η *εσωτερική αξιολόγηση* της εγκυρότητας (internal assessment of validity) προσπαθεί να προσδιορίσει αν η προκύπτουσα δομή είναι εγγενώς κατάλληλη για τα δεδομένα εισόδου.
- Τέλος, ο *σχετικός έλεγχος*(relative test) περιλαμβάνει τη σύγκριση δύο δομών και υπολογίζει τη σχετική τους αξία.

2.3 Το Δίλημμα των Χρηστών

Η διαθεσιμότητα μιας πληθώρας αλγορίθμων ομαδοποίησης στη βιβλιογραφία μπορεί πολύ εύκολα να συγχύσει ένα χρήστη στην προσπάθειά του να επιλέξει τον κατάλληλο αλγόριθμο για το πρόβλημα που έχει να αντιμετωπίσει. Κάποια κοινώς παραδεκτά κριτήρια για τη σύγκριση μεταξύ αλγορίθμων τα οποία εισήχθησαν από τους Fisher και Van Ness στο [28] βασίζονται στα ακόλουθα:

- Στον τρόπο με τον οποίο δημιουργούνται οι ομάδες
- Στη δομή των δεδομένων και
- Στην ευαισθησία της τεχνικής του αλγορίθμου σε αλλαγές οι οποίες δεν επηρεάζουν τη δομή των δεδομένων

Ωστόσο, δεν υπάρχει κάποια σύγκριση ή μέθοδος ανάλυσης των αλγορίθμων ομαδοποίησης η οποία να μπορεί να απαντήσει σε καίρια ερωτήματα όπως:

- Πώς πρέπει να κανονικοποιηθούν τα δεδομένα εισόδου;
- Ποιό μέτρο εγγύτητας είναι το καταλληλότερο να χρησιμοποιηθεί έχοντας μια συγκεκριμένη κατάσταση να αντιμετωπίσουμε;
- Πώς μπορούμε να εκμεταλλευτούμε τη γνώση πάνω σε ένα συγκεκριμένο πεδίο για ένα πρόβλημα ομαδοποίησης;
- Πώς μπορεί ένα πολύ μεγάλο πλήθος δεδομένων εισόδου να ομαδοποιηθεί αποδοτικά;

Υπό αυτό το πρίσμα των δυσκολιών, ένας αρκετά ενημερωμένος χρήστης στο κομμάτι των αλγορίθμων ομαδοποίησης θα είναι σε θέση να αναγνωρίσει τα θετικά και τα αρνητικά της εκάστοτε τεχνικής και τελικά να κάνει την επιλογή της κατάλληλης τεχνικής ή ακολουθίας από τεχνικές για την όσο το δυνατόν σωστότερη αντιμετώπιση ενός συγκεκριμένου προβλήματος.

Δεν υπάρχει προς το παρόν κάποια τεχνική ομαδοποίησης η οποία να είναι καθολικά εφαρμόσιμη για την αποκάλυψη της ποικιλίας των δομών που μπορεί να υπάρχουν σε ένα σύνολο δεδομένων. Ας αναλογιστούμε το παράδειγμα με τα δισδιάστατα δεδομένα που παρουσιάζεται στην Εικόνα 2(a). Δε μπορούν όλες οι τεχνικές ομαδοποίησης να εντοπίσουν τις ομάδες που παρουσιάζονται εκεί με την ίδια ευκολία και αυτό διότι οι αλγόριθμοι περιέχουν υπονοούμενες υποθέσεις που σχετίζονται με το σχήμα των ομάδων ή με διαμορφώσεις πολλαπλών ομάδων βασιζόμενοι στις μετρικές εγγύτητας και στα κριτήρια ομαδοποίησης που χρησιμοποιούν.

Οι άνθρωποι μπορούν να ανταγωνιστούν σε απόδοση αυτόματες διαδικασίες ομαδοποίησης σε δεδομένα με δύο διαστάσεις, αλλά τα περισσότερα πραγματικά προβλήματα περιλαμβάνουν ομαδοποίηση σε μεγαλύτερες διαστάσεις. Είναι αρκετά δύσκολο για τους ανθρώπους να αποτυπώσουν διαισθητικά τα δεδομένα σε περισσότερες των τριών διαστάσεων. Επιπροσθέτως, τα δεδομένα σπανίως ακολουθούν ιδανική δόμηση (π.χ., σφαιρική, γραμμική, κτλ.) όπως αυτή που παρουσιάζεται στην Εικόνα 2. Αυτό εξηγεί το μεγάλο αριθμό αλγορίθμων οι οποίοι συνεχίζουν να εμφανίζονται στη βιβλιογραφία. Κάθε ένας από αυτούς τους αλγορίθμους ομαδοποίησης αποδίδει καλύτερα συγκριτικά με τους υπόλοιπους για μια συγκεκριμένη κατανομή προτύπων. Η κατηγοριοποίηση των αλγορίθμων ομαδοποίησης δεν είναι κάτι απλό και εύκολο. Στην πραγματικότητα, πολλές από τις ομάδες αλγορίθμων αρκετά συχνά επικαλύπτονται. Αν λοιπόν θέλαμε να έχουμε μια γενική κατηγοριοποίηση αυτών των αλγορίθμων αυτή θα ήταν η εξής:

Clustering Algorithms

- Hierarchical Methods
  - Agglomerative Algorithms
  - Divisive Algorithms
- Partitioning Methods
  - Relocation Algorithms
  - Probabilistic Clustering
  - K-medoids Methods
  - K-means Methods
  - Density-Based Algorithms
    - Density-Based Connectivity Clustering
    - Density Functions Clustering
- Grid-Based Methods
- Methods Based on Co-Occurrence Categorical Data
- Constraint Based Clustering
- Clustering Algorithms Used in Machine Learning
  - Gradient Descent and Artificial Neural Networks
  - Evolutionary Methods
- Scalable Clustering Algorithms
- Algorithms for High Dimensional Data

- Subspace Clustering
- Projection Techniques
- Co-Clustering Techniques

Είναι πολύ σημαντικό για το χρήστη ενός αλγορίθμου ομαδοποίησης όχι μόνο να καταλαβαίνει σε βάθος τη συγκεκριμένη τεχνική που χρησιμοποιείται, αλλά να γνωρίζει και λεπτομέρειες που σχετίζονται με τη συλλογή των δεδομένων που χρησιμοποιεί. Όσο καλύτερη γνώση έχει ο χρήστης για την προέλευση των δεδομένων τόσο καλύτερα θα μπορέσει να αναπαραστήσει τις ομάδες που εμφανίζονται σε αυτά. Μάλιστα, η πληροφορία σε σχέση με αυτό το κομμάτι μπορεί να χρησιμοποιηθεί για τη βελτίωση της ποιότητας της εξαγωγής χαρακτηριστικών, του υπολογισμού της μετρικής εγγύτητας που χρησιμοποιείται, της ομαδοποίησης καθώς και της αναπαράστασης των ομάδων.

Έχοντας αναλογιστεί τα παραπάνω λοιπόν, και μετά από εκτενή μελέτη των υπαρχόντων αλγορίθμων στη συγκεκριμένη ερευνητική περιοχή, οδηγηθήκαμε στο συμπέρασμα πως για το πρόβλημα που έχουμε να αντιμετωπίσουμε (το οποίο σχετίζεται με τον εντοπισμό και την παρακολούθηση κινούμενων ομάδων σε έναν εποπτευόμενο χώρο) θα ήταν ιδανική η χρήση ιεραρχικών αλγορίθμων ταξινόμησης με κάποιο κριτήριο για το βέλτιστο εντοπισμό των ομάδων.

2.4 Ομαδοποίηση Ροών Δεδομένων

2.4.1 Εισαγωγή

Όπως αναφέρθηκε και σε προηγούμενες ενότητες, ομαδοποίηση είναι η διαδικασία διαχωρισμού σε ξεχωριστές ομάδες των αντικειμένων που παρατηρούνται, έτσι ώστε οι κοινές ιδιότητες των αντικειμένων εντός των ομάδων να είναι πολλές και μεταξύ των ομάδων λίγες. Οι μέθοδοι ομαδοποίησης χρησιμοποιούνται ευρέως στην εξόρυξη δεδομένων. Χρησιμοποιούνται είτε για μια προσεκτικότερη ματιά στα δεδομένα όσον αφορά την κατανομή τους είτε ως ένα βήμα προεπεξεργασίας πριν την εφαρμογή άλλων αλγορίθμων. Οι πιο κοινότυπες προσεγγίσεις των αλγορίθμων αυτών χρησιμοποιούν μετρικές απόστασης μεταξύ των παραδειγμάτων ως κριτήριο της ομοιότητας που παρουσιάζουν. Συνήθως μάλιστα, απαιτούν χώρο αποθήκευσης ο οποίος είναι τετραγωνικός σε σχέση με τον αριθμό των παρατηρήσεων, το οποίο είναι απαγορευτικό στην περίπτωση που έχουμε ροές δεδομένων (data streams) οι οποίες χαρακτηρίζονται από το τεράστιο πλήθος πληροφορίας και προτύπων που φέρουν.

Το πρόβλημα ομαδοποίησης ροών δεδομένων ορίζεται ως η δυσκολία του να βρεθεί και να διατηρηθεί μια συνεχώς συμπαγής και αντιπροσωπευτική ομαδοποίηση των δεδομένων που έχουν παρατηρηθεί μέχρι στιγμής, σε μικρό χρονικό διάστημα και περιορισμένο χώρο. Τα εμπόδια προκύπτουν από τη συνεχή εμφάνιση-ανανέωση των δεδομένων καθώς και από την ανάγκη που υπάρχει για την επεξεργασία τους σε πραγματικό χρόνο. Αυτοί οι περιορισμοί οδηγούν στην ανάγκη για συνεχή ομαδοποίηση έτσι ώστε να διατηρείται η δομή των ομάδων που εξελίσσονται όσο περνάει ο χρόνος. Επιπροσθέτως, οι ροές δεδομένων μπορεί να διαφοροποιούνται συνεχώς με αποτέλεσμα να εμφανίζονται νέες ομάδες και άλλες να εξαφανίζονται αντικατοπτρίζοντας την δυναμική φύση που έχουν αυτές οι ροές.

Στις κύριες και πιο διαδεδομένες προσεγγίσεις ομαδοποίησης στην ανάλυση των δεδομένων σε συστάδες, βάσει και του διαχωρισμού που παρουσιάστηκε παραπάνω, έχουμε αλγορίθμους όπως:

- *Αλγόριθμοι κατάτμησης* (partitioning algorithms): Αυτοί οι αλγόριθμοι κατασκευάζουν μια κατάτμηση ενός συνόλου αντικειμένων σε k ομάδες, έτσι ώστε να ελαχιστοποιείται μια αντικειμενική συνάρτηση. Τέτοια παραδείγματα είναι αλγόριθμοι όπως k-means [30], k-medoids [7, 8] κ.ά.

- *Αλγόριθμοι μικρο-ομαδοποίησης* (micro-clustering algorithms): Χωρίζουν τη διαδικασία της ομαδοποίησης σε δύο φάσεις, όπου η πρώτη φάση είναι online και συνοψίζει τη ροή δεδομένων σε τοπικά μοντέλα(micro-clusters) ενώ η δεύτερη φάση παράγει ένα καθολικό μοντέλο ομάδων το οποίο προκύπτει από τα τοπικά μοντέλα της πρώτης φάσης. Παραδείγματα τέτοιων αλγορίθμων είναι οι BIRCH [9] και CluStream [31].

- *Αλγόριθμοι βασισμένοι στην πυκνότητα* (density-based algorithms): Αυτοί οι αλγόριθμοι, όπως προδίδεται και από το όνομά τους, βασίζονται στη συνδεσιμότητα μεταξύ των περιοχών και των συναρτήσεων πυκνότητας. Τέτοιοι τύποι μεθόδων, όπως είναι οι DBSCAN [25] και OPTICS [12], εντοπίζουν ομάδες αυθαίρετου σχήματος.

- *Αλγόριθμοι βασισμένοι στο πλέγμα* (grid-based algorithms): Βασίζονται σε μια πολύ-επίπεδη δομή διακριτότητας. Αντιμετωπίζουν το χώρο των στιγμιότυπων

σαν δομημένα πλέγματα. Τέτοιοι είναι αλγόριθμοι όπως οι Fractal Clustering [32] και STING [33].

- *Αλγόριθμοι βασισμένοι σε μοντέλα* (model-based algorithms): Προσπαθούν να εντοπίσουν το μοντέλο που ταιριάζει καλύτερα σε όλες τις ομάδες κάτι το οποίο τους καθιστά ιδιαιτέρως αποδοτικούς στην εννοιολογική ομαδοποίηση. Τέτοιοι είναι αλγόριθμοι όπως ο COBWEB [34] και ο SOM [35].

2.4.2 Απαιτήσεις για Ομαδοποίηση Ροών Δεδομένων

Για την ομαδοποίηση ροών δεδομένων στο [5] έχουν καταγραφεί τέσσερις βασικές απαιτήσεις βάσει της έρευνας που έχει διεξαχθεί:

i. Συμπαγής αναπαράσταση (compactness of representation)
ii. Γρήγορη και συνεχής επεξεργασία νέων δεδομένων
iii. Παρακολούθηση των αλλαγών στα μοντέλα ομαδοποίησης
iv. Σαφής και γρήγορος εντοπισμός των ακραίων τιμών (outliers)

Οι παραπάνω κατηγορίες θα αναλυθούν εκτενώς στις υποενότητες που ακολουθούν.

2.4.2.1 Συμπαγής Αναπαράσταση (Compactness of Representation)

Από τη στιγμή που τα δεδομένα καταφθάνουν ασταμάτητα σε μια τοποθεσία, οποιοσδήποτε αλγόριθμος ομαδοποίησης στοχεύει στο να τα επεξεργαστεί δεν έχει τη πολυτέλεια μιας μακροσκελούς περιγραφής των ομάδων που έχουν βρεθεί μέχρι εκείνη τη στιγμή που τα επεξεργάζεται. Γενικότερα, το να βασίζουν την απόφαση στο που θα τοποθετήσουν το επόμενο σημείο στη λίστα των ομάδων που έχουν βρεθεί μέχρι στιγμής δεν αποτελεί επιλογή. Αυτή η λίστα αυξάνει χωρίς κάποιο περιορισμό όσο νέα σημεία εμφανίζονται και θα κατανάλωνε όλη τη διαθέσιμη μνήμη του μηχανήματος. Επιπροσθέτως, από τη στιγμή που υπάρχει επιμονή για την επεξεργασία νέων στιγμιότυπων online, ο έλεγχος του που ανήκει ένα νέο πρότυπο στις ήδη δημιουργημένες ομάδες, συγκρίνοντάς το με την αναπαράστασή τους στη δευτερεύουσα μνήμη, δεν είναι εφικτός. Για τον λόγο αυτό, ένας αλγόριθμος ο οποίος αναλαμβάνει να εκτελέσει ομαδοποίηση σε ροές δεδομένων πρέπει να παρέχει μια αναπαράσταση των ομάδων που τελικά εντοπίζει η οποία όχι μόνο να είναι συμπαγής

αλλά να μην αυξάνει και αισθητά όσο επεξεργάζονται νέα δεδομένα. Αξίζει να σημειωθεί πως ακόμα και μια γραμμική αύξηση της αναπαράστασης αυτής θα ήταν αφόρητη.

2.4.2.2 Γρήγορη και Συνεχής Επεξεργασία Νέων Δεδομένων

Η ανάγκη για ταχύτητα και επεξεργασία κατά τη διάρκεια της εμφάνισης νέων δεδομένων είναι πέρα από απαραίτητη αν αναλογιστούμε την online φύση του ζητήματος. Η τοποθέτηση των νέων σημείων θα πρέπει να βασίζεται στην αξιολόγηση μιας συνάρτησης. Αυτή η συνάρτηση πρέπει να ακολουθεί τους εξής δύο περιορισμούς:

- Η τοποθέτηση ενός νέου σημείου δεν πρέπει να αποφασίζεται μέσω μιας συνάρτησης η οποία απαιτεί σύγκριση με όλα τα δεδομένα τα οποία έχουν επεξεργαστεί κατά το παρελθόν.
- Η συνάρτηση η οποία αποφασίζει για την τοποθέτηση των νέων σημείων πρέπει να επιδεικνύει καλή απόδοση.

Βάσει του πρώτου περιορισμού γίνεται για μια ακόμη φορά ορατή η ανάγκη για συμπαγή αναπαράσταση των ομάδων που έχουν δημιουργηθεί μέχρι στιγμής. Μιλάει όμως και για μια συνάρτηση η οποία μπορεί να εκτιμηθεί με τη χρήση αυτής της αναπαράστασης. Ο δεύτερος περιορισμός απλά δίνει έμφαση στην ανάγκη χρησιμοποίησης μιας συνάρτησης με καλή πολυπλοκότητα, δηλαδή για παράδειγμα, γραμμικής σε σχέση με το μέγεθος της αναπαράστασης που έχει επιλεγεί.

2.4.2.3 Σαφής και Γρήγορος Εντοπισμός «Ακραίων Τιμών» (Outliers)

Αυτή η απαίτηση είναι ίσως και η λιγότερη κατανοητή διαισθητικά από τις τέσσερις. Η λέξη «ακραίες τιμές» έχει μπει σε εισαγωγικά καθώς χρησιμοποιείται συνήθως για να περιγράψει σημεία τα οποία δεν φαίνεται να ταιριάζουν αρκετά καλά σε κάποια από τις ομάδες που ο αλγόριθμος έχει εντοπίσει μέχρι στιγμής.

Ο λόγος για τον οποίο προστέθηκε η απαίτηση αυτή μπορεί να εξηγηθεί από τη δυναμική φύση των ροών δεδομένων. Είναι αρκετά πιθανό η ροή δεδομένων να παρουσιάζει διαφορετικές τάσεις κατά τη διάρκεια του κύκλου ζωής της, συνεπώς τα σημεία τα οποία ελήφθησαν μια χρονική στιγμή μπορεί να μην ταιριάζουν καλά στο μοντέλο ομαδοποίησης το οποίο έχει σχηματιστεί μέχρι εκείνη την ώρα.

Για το λόγο αυτό, ο αλγόριθμος που θα χρησιμοποιηθεί, θα πρέπει να είναι σε θέση να το προσδιορίσει με ένα σαφή και γρήγορο τρόπο. Για την ακρίβεια, η συνάρτηση η

οποία εκτιμάει τη θέση στην οποία θα τοποθετηθεί κάθε νέο σημείο θα πρέπει να έχει ορισμένο ένα εύρος για την περίπτωση των «ακραίων τιμών».

Άμεσα συνυφασμένη με αυτή την απαίτηση είναι η ανάγκη να αποφασίσουμε τι θα γίνεται με αυτά τα σημεία. Είναι μια απόφαση η οποία συσχετίζεται άμεσα με την εφαρμογή που έχουμε. Σε μερικές περιπτώσεις, αν βρεθούν αρκετά τέτοια σημεία (το πλήθος ορίζεται και πάλι αναλόγως) μπορεί να χρειάζεται να εγκαταλείψουμε την τρέχουσα ομαδοποίηση και να προβούμε στη δημιουργία μιας νέας. Ένα παράδειγμα του παραπάνω θα μπορούσε να είναι μια ροή με μετεωρολογικά δεδομένα στην οποία επαρκής αριθμός από ακραίες τιμές υποδηλώνει μια νέα τάση η οποία πρέπει να αναπαρασταθεί από μία νέα ομάδα. Αντίστοιχα, σε μια ροή με θέσεις ανθρώπων που κινούνται σε ένα χώρο, αυτό θα μπορούσε να σημαίνει πως αρκετοί από τους χρήστες έχουν αλλάξει αρκετά τη θέση τους, οπότε θα ήταν καλύτερο να χτιστεί μια νέα δομή ομαδοποίησης η οποία και θα βασίζεται στις τελευταίες μετρήσεις που λάβαμε για τις θέσεις τους. Σε άλλες περιπτώσεις, μπορεί απλά να χρειαστεί να επανασχεδιάσουμε τα όρια των ήδη εντοπισμένων ομάδων. Ένα παράδειγμα αυτής της περίπτωσης θα μπορούσε να είναι μια ροή από δεδομένα η οποία περιλαμβάνει μετρήσεις για εντοπισμένα περιστατικά αρρώστιας. Ακραίες τιμές μπορεί να υποδεικνύουν την εξάπλωση μιας επιδημίας σε μια ευρύτερη γεωγραφική περιοχή και όχι τη δημιουργία μιας νέας τάσης, άρα την ανάγκη δημιουργίας μια νέας ομάδας.

2.4.2.4 Παρακολούθηση των Αλλαγών στα Μοντέλα Ομαδοποίησης

Καθώς νέες μετρήσεις καταφθάνουν, πρέπει να είμαστε σε θέση να προσδιορίσουμε αν χρειάζεται να ορισθεί νέο μοντέλο ομαδοποίησης. Υποθέτοντας πως υπάρχει ένας εύκολος τρόπος εντοπισμού των ακραίων τιμών μέσω του αλγορίθμου που χρησιμοποιείται (όπως προσδιορίστηκε σαν απαίτηση στο 2.4.2.3), μπορούμε να κρατάμε πληροφορία για το πόσες τέτοιες τιμές έχουν κάνει την εμφάνισή τους στο πρόσφατο παρελθόν. Για να προσδιοριστεί ποιο είναι το άνω όριο για τέτοιες τιμές που θα πρέπει να επιτευχθεί έτσι ώστε να αλλάξει το μοντέλο ομαδοποίησης, μπορεί να οριστεί μια μετρική η οποία κάνει χρήση των ορίων του Chernoff [6]. Αναλόγως το πρόβλημα στο οποίο χρησιμοποιούμε τέτοιους αλγορίθμους, υπάρχει και η κατάλληλη αντιστοίχηση για το τι θεωρούμε ως ακραίες τιμές όπως έχει αναφερθεί και παραπάνω. Διαισθητικά είναι πιο κατανοητό όταν ως ακραίες τιμές αναφέρονται κόμβοι-πρότυπα οι οποίοι για κάποιο λόγο θεωρούνται μη καλά ορισμένα στην ομάδα που βρίσκονται. Μια άλλη μέθοδος εύρεσης της στιγμής στην οποία θα πρέπει να ξαναχτιστεί η δομή των

ομάδων θα ήταν η παρατήρηση των κεντροειδών των ομάδων που έχουν εντοπιστεί μέχρι στιγμής και τι συσχέτιση έχουν αυτά μεταξύ τους με την πάροδο του χρόνου. Αυτό βρίσκει εφαρμογή κυρίως σε περιπτώσεις όπου έχουμε να παρατηρήσουμε την κίνηση ομάδων μέσα σε ένα χώρο και τις αλληλεπιδράσεις που εμφανίζουν.

Από τι στιγμή που θα εντοπιστεί πως η τρέχουσα δομή των ομάδων δεν είναι πλέον αντιπροσωπευτική μπορεί να ακολουθηθούν δύο τρόποι για τον σχηματισμό των νέων:

A. Επαναομαδοποίηση όλων των παρατηρήσεων που έχουν εμφανιστεί μέχρις στιγμής οι οποίες θα περιλαμβάνουν και εκείνες οι οποίες οδήγησαν στην απόφαση για το ότι η τρέχουσα δομή δεν είναι αντιπροσωπευτική. Προφανώς, μια τέτοια λογική είναι η ποιο ακριβή υπολογιστικά από τις διαθέσιμες. Εντούτοις, μια καλή και συνάμα συμπαγής αναπαράσταση των ομάδων, πριν την εμφάνιση των ακραίων τιμών που οδήγησαν στην απόφαση αυτή, σε συνδυασμό με τη χρήση και των ακραίων τιμών θα μπορούσε να χρησιμοποιηθεί για το σχηματισμό των ομάδων χωρίς να χρειαστεί να επεξεργαστούν ξανά οι προηγούμενες μετρήσεις.
B. Παραμερισμός των παλαιών ομάδων και παραγωγή ενός νέου συνόλου ομάδων με την επεξεργασία μόνο των τελευταίων μετρήσεων που μας ήρθαν.

Είναι κάτι παραπάνω από προφανές πως η δεύτερη μέθοδος είναι λιγότερο πολύπλοκη αλλά το ίδιο αποδοτική με την πρώτη. Επομένως, στην ιδέα που θα παρουσιαστεί στη συνέχεια, αυτή θα ενσωματωθεί για να προσδιορίζονται εκ νέου οι ομάδες όταν η δομή που θα υπάρχει θα θεωρείται μη αντιπροσωπευτική.

2.4.3 Ομαδοποίηση Παραδειγμάτων (Clustering Examples)

Μέχρι τώρα έχουμε δει κάποια από τα σημαντικότερα στοιχεία των αλγορίθμων ομαδοποίησης. Σε τι είναι χρήσιμοι αλλά και ποιες προϋποθέσεις πρέπει να καλύπτουν έτσι ώστε να θεωρηθούν «καλοί». Σε αυτό το κομμάτι θα εξετάσουμε με ποιους τρόπους οι αλγόριθμοι αυτής της κατηγορίας επεξεργάζονται τις ροές των δεδομένων.

Η ομαδοποίηση παραδειγμάτων είναι η πιο συχνά χρησιμοποιούμενη μέθοδος στη μη εποπτευόμενη μάθηση. Οι πιο γνωστές τεχνικές είναι η ομαδοποίηση κατάτμησης, η οποία όμως απαιτεί την γνώση του επιθυμητού αριθμού των ομάδων εξαρχής, καθώς και η ιεραρχική ομαδοποίηση, η οποία δημιουργεί μια ιεραρχία από εμφωλευμένες ομαδοποιήσεις. Διαισθητικά, θα μπορούσαμε να την παρομοιάσουμε με την ομαδοποίηση ανά χρονικές στιγμές των γεγονότων. Δηλαδή, οι ομάδες περιέχουν

δεδομένα με μια ετικέτα συγκεκριμένης χρονικής στιγμής που μοιάζουν με δεδομένα που φέρουν ετικέτες άλλων χρονικών στιγμών. Γίνεται κατανοητό βέβαια, πως θα πρέπει να εξετάζονται τα δεδομένα με χρήση παραθύρου συγκεκριμένου εύρους, όπου μέσα σε αυτό το εύρος θα γίνεται οι ομαδοποίηση και για τις μετρήσεις που έχουμε εκεί και μόνο.

Μια έξυπνη προσέγγιση στην ομαδοποίηση ροών δεδομένων είναι η αναπαράσταση κάθε ομάδας που έχει εντοπιστεί με την χρήση των χαρακτηριστικών ομάδας. Τα χαρακτηριστικά ομάδας, ή αλλιώς μικρο-ομάδα, είναι μια συμπαγής αναπαράσταση ενός συνόλου σημείων που ανήκουν στην ίδια ομάδα. Μια τέτοια δομή αποτελείται από μια τριπλέτα στοιχείων $(\mathrm{N}, \mathrm{LS}, \mathrm{SS})$, τα οποία χρησιμοποιούνται για την αποθήκευση των επαρκών στατιστικών στοιχείων για ένα σύνολο σημείων:

- N είναι το πλήθος των σημείων που ανήκουν σε αυτή την ομάδα
- LS είναι ένα διάνυσμα, ίδιας διάστασης με αυτή των σημείων του συνόλου δεδομένων που κρατάει πληροφορία σχετικά με γραμμικό άθροισμα των N σημείων της ομάδας.
- SS είναι ένα διάνυσμα, ίδιας διάστασης με αυτή των σημείων του συνόλου δεδομένων που κρατάει πληροφορία σχετικά με το τετραγωνικό άθροισμα των N σημείων της ομάδας.

Οι ιδιότητες των δομών των χαρακτηριστικών ομάδων είναι:

- Επαύξηση (incrementality)

  Αν ένα σημείο x προστεθεί στην ομάδα A, τα στατιστικά της δομής ενημερώνονται ως εξής:

$$LS_A \leftarrow LS_A + x$$

$$SS_A \leftarrow SS_A + x^2$$

$$N_A \leftarrow N_A + 1$$

- Προσθετικότητα (additivity)

  Αν οι A και B είναι ασύνδετες ομάδες, η συγχώνευσή τους ισούται με το άθροισμα των στατιστικών τους. Η ιδιότητα της προσθετικότητας μας επιτρέπει να συγχωνεύουμε υποομάδες αυξητικά χωρίς παραπάνω πράξει για τον επαναϋπολογισμό των στατιστικών αυτών για την νέα ομάδα που δημιουργείται:

$$LS_C \leftarrow LS_A + LS_B$$

$$SS_C \leftarrow SS_A + SS_B$$

$$N_C \leftarrow N_A + N_B$$

Μια καταχώρηση δομής χαρακτηριστικών ομάδας έχει επαρκή πληροφορία για τον υπολογισμό των μετρικών (νόρμες):

$$L_1 = \sum_{i=1}^{n} \left| x_{a_i} - x_{b_i} \right|$$

$$L_2 = \sqrt{\sum_{i=1}^{n} \left(x_{a_i} - x_{b_i} \right)^2}$$

καθώς και βασικών μέτρων που χαρακτηρίζουν μια ομάδα όπως:

- Κεντροειδές (centroid), το οποίο ορίζεται ως το κέντρο βάρους της ομάδας:

$$\vec{XO} = \frac{LS}{N}$$

- Ακτίνα (radius), η οποία ορίζεται ως η μέση απόσταση των μελών της ομάδας από το κεντροειδές:

$$R = \sqrt{\frac{\sum_{i=1}^{N} \left(\vec{x}_i - \vec{XO} \right)^2}{N}}$$

2.4.3.1 Αλγόριθμοι Κατάτμησης

2.4.3.1.1 Ο Αλγόριθμος του Ηγέτη (The Leader Algorithm)

Ο απλούστερος αλγόριθμος κατάτμησης ενός περάσματος είναι γνωστός ως ο Αλγόριθμος του Ηγέτη [36]. Χρησιμοποιεί ένα κατώφλι το οποίο ορίζεται από το χρήστη και προσδιορίζει τη μέγιστη επιτρεπτή απόσταση μεταξύ ενός παραδείγματος και του κεντροειδούς. Σε κάθε βήμα, ο αλγόριθμος αναθέτει το τρέχον παράδειγμα στην πιο όμοια ομάδα (τον ηγέτη) αν η απόσταση μεταξύ τους είναι κάτω από το κατώφλι που έχει οριστεί. Αλλιώς, το ίδιο το παράδειγμα προστίθεται ως ηγέτης. Πρόκειται για έναν αλγόριθμο ενός περάσματος, γρήγορο και δεν απαιτεί εκ των προτέρων γνώση για τον αριθμό των ομάδων. Εντούτοις, είναι ένας ασταθής αλγόριθμος καθώς η απόδοσή του εξαρτάται πολύ από τη σειρά με την οποία εμφανίζονται τα παραδείγματα και στη σωστή εικασία σχετικά με την τιμή του κατωφλίου, κάτι το οποίο απαιτεί πρότερη γνώση του συνόλου των δεδομένων.

2.4.3.1.2 Αλγόριθμος k-means Ενός Περάσματος (Single Pass k-means)

Ο αλγόριθμος *k-means* [30] είναι ο πιο ευρέως χρησιμοποιούμενος αλγόριθμος ομαδοποίησης. Κατασκευάζει μια κατάτμηση του συνόλου των δεδομένων σε k ομάδες, η οποία ελαχιστοποιεί μια αντικειμενική συνάρτηση, συνήθως συνάρτηση τετραγωνικού σφάλματος, κάτι το οποίο υποδηλώνει ομάδες σφαιρικού σχήματος. Η παράμετρος εισόδου k είναι σταθερή και πρέπει να δίνεται στην αρχή του τρεξίματος του αλγορίθμου, πράγμα που περιορίζει την ικανότητά εφαρμογής του σε περιπτώσεις ροών δεδομένων που συνεχώς εξελίσσονται.

Έτσι λοιπόν προτείνεται στη βιβλιογραφία ο αλγόριθμος *k-means* ενός περάσματος [30]. Η κύρια ιδέα περιλαμβάνει την ύπαρξη ενός αποθηκευτικού χώρου στον οποίο τα στοιχεία του συνόλου δεδομένων θα κρατούνται σε μια πιο συνεπτυγμένη μορφή. Η ροή δεδομένων επεξεργάζεται κομματιαστά, με τη χρήση κάποιου προκαθορισμένου παραθύρου. Όλος ο διαθέσιμος αποθηκευτικός χώρος που έχει οριστεί γεμίζεται με δεδομένα από την ροή. Χρησιμοποιώντας λοιπόν τα δεδομένα που βρίσκονται στον αποθηκευτικό χώρο, βρίσκουμε k κέντρα έτσι ώστε το άθροισμα των αποστάσεων των σημείων του χώρου από το κοντινότερο κέντρο τους να ελαχιστοποιείται. Μόνο τα k κεντροειδή (που αντιπροσωπεύουν το αποτέλεσμα της ομαδοποίησης) διατηρούνται μαζί με τις αντίστοιχες δομές χαρακτηριστικών για τις ομάδες που αναπαριστούν. Στις επόμενες επαναλήψεις, ο αποθηκευτικός χώρος αρχικοποιείται με τα k κεντροειδή που

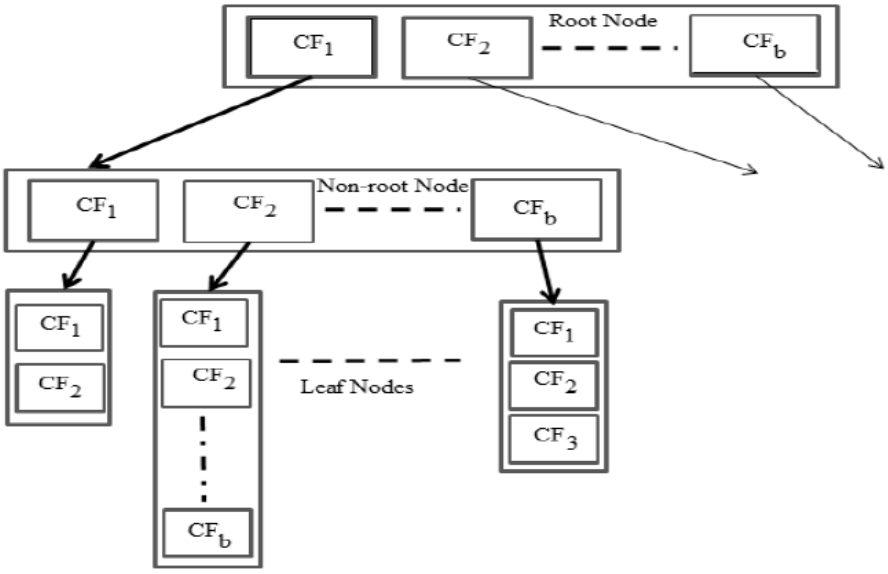
βρέθηκαν στο ακριβώς προηγούμενο βήμα, σταθμισμένα με τις αντίστοιχες δομές χαρακτηριστικών ομάδας καθώς και με τα νέα δεδομένα από τη ροή. Ο συγκεκριμένος αλγόριθμος βελτιώνει την απόδοσή του όσο νέα δεδομένα καταφθάνουν και ο αποθηκευτικός χώρος που χρησιμοποιεί είναι σταθερού μεγέθους. Όπως και προηγουμένως, έτσι και εδώ υπάρχει το ζήτημα της τιμής που πρέπει να εισαχθεί από το χρήστη, οπότε η εφαρμογή του αλγορίθμου σε προβλήματα ομαδοποίησης ροών δεδομένων δεν είναι ιδανική.

2.4.3.2 Ιεραρχικοί Αλγόριθμοι

Ένα πολύ σημαντικό επίτευγμα σε αυτή την ερευνητική περιοχή είναι το σύστημα BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) [9]. Αυτό το σύστημα, συμπιέζει τα δεδομένα, κατασκευάζοντας μια ιεραρχική δομή γνωστή ως δέντρο χαρακτηριστικών ομάδας όπου κάθε κόμβος του είναι μια πλειάδα η οποία περιλαμβάνει επαρκή στατιστικά στοιχεία που περιγράφουν ένα σύνολο δεδομένων και συγκεντρώνει όλη τη πληροφορία των υποομάδων σε ένα δέντρο χαρακτηριστικών ομάδας. Το BIRCH δουλεύει μόνο με συνεχή δεδομένα και έχει σχεδιαστεί για μεγάλα σύνολα δεδομένων λαμβάνοντας σοβαρά υπόψη του περιορισμούς σε χρονικούς και χωρικούς πόρους. Για παράδειγμα, δεν γίνεται χρήση όλων των σημείων του συνόλου δεδομένων για την ομαδοποίηση καθώς συμπαγείς περιοχές δεδομένων αντιμετωπίζονται ως υποομάδες. Συνήθως κάνει διπλό πέρασμα των δεδομένων για την κατασκευή του δέντρου χαρακτηριστικών ομάδων αλλά μπορεί να λειτουργήσει και με μονό πέρασμα. Το σύστημα αυτό περιλαμβάνει δύο φάσεις κατά την εφαρμογή του. Αρχικά, διατρέχει τη βάση των δεδομένων έτσι ώστε να κατασκευάσει μια αρχική δομή του δέντρου στη μνήμη, μια πολυεπίπεδη σύμπτυξη των δεδομένων η οποία προσπαθεί να διατηρήσει μια κληρονομική ιεραρχική δόμηση των δεδομένων. Στη δεύτερη φάση χρησιμοποιεί έναν αυθαίρετο αλγόριθμο ομαδοποίησης για την ομαδοποίηση των φύλων του δέντρου χαρακτηριστικών ομάδας που έχει δημιουργηθεί στην πρώτη φάση τρεξίματος του συστήματος.

Και αυτός ο αλγόριθμος απαιτεί δύο παραμέτρους τις οποίες εισάγει ο χρήστης και αυτές είναι (α) ο παράγοντας διακλάδωσης B ή αλλιώς το μέγιστο αριθμό καταχωρήσεων που μπορεί να υπάρχουν σε ένα κόμβο του δέντρου που δεν είναι φύλλο και (β) τη μέγιστη διάμετρο (ή ακτίνα) T οποιασδήποτε δομής χαρακτηριστικών ομάδας που ανήκει σε κόμβο-φύλλο του δέντρου. Η μέγιστη διάμετρος T προσδιορίζει τα παραδείγματα που μπορούν να απορροφηθούν από μια δομή χαρακτηριστικών

ομάδας. Όσο αυξάνεται αυτή η παράμετρος τόσο περισσότερα παραδείγματα μπορούν να υπάρχουν σε μια μικρο-ομάδα κάτι το οποίο έχει ως αποτέλεσμα να παράγονται μικρότερα δέντρα χαρακτηριστικών ομάδας (CF-trees). Σε κάθε εσωτερικό κόμβο του δέντρου, το παράδειγμα ακολουθεί το κοντινότερο μονοπάτι χαρακτηριστικών ομάδας (closest-CF path) λαμβάνοντας υπόψη είτε τη νόρμα L_1 είτε τη νόρμα L_2 ο ορισμός των οποίων δόθηκε παραπάνω. Αν η κοντινότερη δομή χαρακτηριστικών ομάδας σε κόμβο-φύλλο δεν μπορέσει να απορροφήσει το παράδειγμα (λόγω της μέγιστης διαμέτρου T που έχει οριστεί), δημιουργείται μια νέα καταχώρηση για τέτοια δομή. Αν δεν υπάρχει χώρος για νέο φύλλο τότε διασπάται ο γονικός κόμβος. Ένας κόμβος-φύλλο μπορεί να επεκταθεί βάσει των περιορισμών που τίθενται από τις παραμέτρους B και T. Η διαδικασία αποτελείται από την εύρεση των δύο μακρινότερων δομών χαρακτηριστικών και τη δημιουργία δύο νέων κόμβων-φύλλων. Όταν πηγαίνουν προς τα πάνω, οι δομές χαρακτηριστικών ομάδων ενημερώνονται.



Εικόνα 4: Το δέντρο χαρακτηριστικών ομάδων του συστήματος BIRCH

Το σύστημα BIRCH προσπαθεί να εντοπίσει τις καλύτερες δυνατές ομάδες βάσει της διαθέσιμης μνήμης ενώ παράλληλα ελαχιστοποιεί τις εισόδους και εξόδους. Το δέντρο των δομών χαρακτηριστικών ομάδας αυξάνει βάσει της συγκέντρωσης και με ένα πέρασμα των δεδομένων φτάνει μια πολυπλοκότητα της τάξης του $O(N)$. Εντούτοις, έχει αποδειχτεί [37] πως δεν αποδίδει καλά σε περίπτωση που έχουμε ομάδες οι οποίες δεν έχουν σφαιρικό σχήμα. Τέλος, όπως και στους αλγορίθμους κατάτμησης, έτσι και

εδώ πρέπει να εισάγει ο χρήστης τιμές σε παραμέτρους που επηρεάζουν την απόδοση και τα αποτελέσματα του αλγορίθμου.

2.4.4 Ομαδοποίηση Μεταβλητών (Clustering Variables)

Η περισσότερη έρευνα σχετικά με τη συνεχή ομαδοποίηση πάνω από ροές δεδομένων έχει γίνει στην ομαδοποίηση των παραδειγμάτων και όχι στην ομαδοποίηση των μεταβλητών. Η ομαδοποίηση των μεταβλητών/πεδίων είναι ένα αρκετά χρήσιμο εργαλείο για κάποιες εφαρμογές όπως είναι τα δίκτυα αισθητήρων, τα δίκτυα κοινωνικής δικτύωσης, η ζήτηση ηλεκτρικού ρεύματος, η παρακολούθηση κινητών χρηστών κα. Η διαφορά μεταξύ της ομαδοποίησης παραδειγμάτων και της ομαδοποίησης μεταβλητών δεν είναι ουσιώδης στην τμηματική ομαδοποίηση (με χρήση χρονικών παραθύρων δηλαδή) καθώς τα παραδείγματα και οι μεταβλητές μπορούν να αλλάξουν θέση με μια απλή μετατόπιση (transpose). Στο πλαίσιο των ροών δεδομένων μεγάλων ταχυτήτων η μετατόπιση των γραμμών σε στήλες του πίνακα που αναπαριστά τα δεδομένα δεν είναι κάτι απλό και εύκολο υπολογιστικά. Η μετατόπιση ή μετατροπή είναι ένας τελεστής χειρισμού μπλοκ. Η ομαδοποίηση πεδίων στις ροές δεδομένων απαιτεί άλλες τεχνικές και αλγορίθμους.

Η βασική ιδέα πίσω από την ομαδοποίηση πεδίων σε ροές χρονοσειρών είναι η εύρεση μεταβλητών οι οποίες παρουσιάζουν όμοια συμπεριφορά με την πάροδο του χρόνου. Αυτή η ομοιότητα συνήθως μετριέται με όρους απόστασης μεταξύ χρονοσειρών, με χρήση μετρικών όπως είναι η Ευκλείδεια απόσταση ή η συσχέτιση. Εντούτοις, όταν εφαρμόζεται ομαδοποίηση πεδίων σε ροές δεδομένων, το σύστημα δε μπορεί ποτέ να έχει πλήρη γνώση των δεδομένων, από τη στιγμή που αυτά συνεχώς εξελίσσονται και επιπροσθέτως δεν είναι αποδοτικά τα πολλαπλά περάσματα πάνω από τα δεδομένα. Επομένως, οι αποστάσεις αυτές θα πρέπει να υπολογίζονται σταδιακά. Έστω $X = \langle x_1, x_2, \dots \dots, x_n \rangle$ πως είναι το πλήρες σετ από τις n ροές δεδομένων και πως $X^t = \langle x_1^t, x_2^t, \dots \dots, x_n^t \rangle$ είναι το παράδειγμα το οποίο περιλαμβάνει τις παρατηρήσεις για όλες τις ροές x_i τη δεδομένη χρονική στιγμή t. Ο στόχος ενός συστήματος σταδιακής ομαδοποίησης για πολλαπλές χρονοσειρές είναι να εντοπίσει (και να καθιστά διαθέσιμη οποιαδήποτε χρονική στιγμή t) μια κατάτμηση P αυτών των ροών, όπου οι ροές στην ίδια ομάδα έχουν την τάση να είναι πιο όμοιες μεταξύ τους σε σχέση με τις ροές διαφορετικών ομάδων. Σε μια ιεραρχική προσέγγιση του προβλήματος, με το προνόμιο του να μην χρειάζεται να προσδιοριστεί εκ των προτέρων ο αριθμός των ομάδων, ο

στόχος είναι να διατηρείται συνεχώς μια δομημένη ιεραρχία των ομάδων. Ένα παράδειγμα κατάτμηση μπορεί να οριστεί ως $P^t = \{\{\{x_1\},\{x_3,x_5\}\},\{x_2,x_4\}\}$, το οποίο σημαίνει πως οι ροές δεδομένων x_1, x_3 και x_5 έχουν μια ομοιότητα μεταξύ τους (συγκεκριμένα οι x_3 και x_5 έχουν κάπως μεγαλύτερη) ενώ ταυτόχρονα είναι λιγότερο όμοιες με τις x_2 και x_4. Οι περισσότερες δημοσιεύσεις ερευνών αναφέρονται στην ομαδοποίηση των ροών δεδομένων κατά παραδείγματα ενώ υπάρχουν πολύ λίγες που σχετίζονται με την ομαδοποίηση κατά πεδία. Μία από τις πρώτες δουλειές που παρουσιάστηκε σε αυτό το ερευνητικό πεδίο ήταν το σύστημα ODAC [24].

2.4.4.1 Μια Ιεραρχική Προσέγγιση – Το σύστημα ODAC

Σε αυτή τη παράγραφο θα παρουσιαστεί το σύστημα ODAC (Online Divisive-Agglomerative Clustering) [24]. Πρόκειται για έναν αλγόριθμο σταδιακής/συνεχούς ομαδοποίησης ροών από χρονοσειρές ο οποίος κατασκευάζει μια ιεραρχική δενδροειδή δομή από ομάδες με μια στρατηγική από πάνω προς τα κάτω (top-down). Τα φύλλα είναι οι ομάδες που προέκυψαν και κάθε ένα από αυτά ομαδοποιεί ένα σύνολο από πεδία/μεταβλητές. Η ένωση των φύλλων είναι το πλήρες σύνολο μεταβλητών που υπάρχουν. Η τομή των φύλλων είναι το κενό σύνολο.

Το σύστημα κάνει χρήση δύο σημαντικών τελεστών για επέκταση (expansion) και συγχώνευση της δενδροειδούς δομής, οι οποίοι βασίζονται στη δυναμική της διεργασίας που παράγει δεδομένα. Η επέκταση μιας ομάδας συνήθως συναντάται σε στατικές περιόδους της ροής. Όταν μια ομάδα λαμβάνει περισσότερη πληροφορία μπορούμε να προσδιορίσουμε ομάδες με μεγαλύτερη λεπτομέρεια. Σε μη-στάσιμες περιόδους, όταν η δομή της συσχέτισης της ροής αλλάζει, το σύστημα μπορεί να ανιχνεύσει ότι η δομή της συσχέτισης των πιο πρόσφατων δεδομένων διαφέρει από εκείνη που παρατηρήθηκε στο παρελθόν. Σε αυτή τη περίπτωση ο τελεστής συγχώνευσης ενεργοποιείται και δύο αμφιθαλείς ομάδες συγχωνεύονται σε μία νέα.

Το σύστημα ODAC συνεχώς παρακολουθεί τη διάμετρο των ομάδων που έχουν ανιχνευθεί μέχρι στιγμής. Η διάμετρος μιας ομάδας ορίζεται ως η μέγιστη απόσταση μεταξύ των μεταβλητών της ομάδας. Για κάθε ομάδα που έχει εντοπιστεί, το σύστημα βρίσκει τις δύο μεταβλητές οι οποίες προσδιορίζουν τη διάμετρο αυτής της ομάδας. Αν η διάμετρος ξεπερνά ένα ευρετικό άνω όριο, το σύστημα διασπά την ομάδα και αναθέτει κάθε μία από τις μεταβλητές σε μία από τις νεοδημιουργηθέντες ομάδες και γίνεται μεταβλητή-άξονας για εκείνη την ομάδα. Στη συνέχεια, όλες οι εναπομείναντες

μεταβλητές στη παλιά ομάδα ανατίθενται στη νέα ομάδα που έχει τον κοντινότερο άξονα στη μεταβλητή. Στα νέα φύλλα, ξεκινάνε νέες στατιστικές πληροφορίες, υποθέτοντας πως μόνο η ερχόμενη πληροφορία θα είναι χρήσιμη έτσι ώστε να αποφασιστεί αν αυτή η ομάδα πρέπει να διασπαστεί. Κάθε κόμβος c_k θα αναπαριστά σχέσεις μεταξύ ροών χρησιμοποιώντας παραδείγματα της μορφή $X^{i_k \dots s_k}$, όπου i_k είναι η χρονική στιγμή που δημιουργήθηκε ο κόμβος και s_k η χρονική στιγμή που διασπάστηκε (ή η τρέχουσα χρονική στιγμή για κόμβους φύλλα). Αυτή η δυνατότητα αυξάνει την ικανότητα του συστήματος να ελέγχει αν η διάσπαση που αποφασίστηκε προηγουμένως αντιπροσωπεύει ακόμα τη δομή των μεταβλητών. Αν οι διάμετροι των παιδιών-φύλλων είναι μεγαλύτερες από τη διάμετρο του πατέρα-κόμβου, τότε η απόφαση που πάρθηκε προηγουμένως μπορεί να μην αντικατοπτρίζει πλέον τη δομή των δεδομένων. Τότε, το σύστημα συγχωνεύει τα δύο φύλλα στον πατέρα-κόμβο και αρχίζει ξανά τα στατιστικά.

Αυτή είναι μια γενική παρουσίαση του συστήματος ODAC και δεν θα μπούμε σε περαιτέρω λεπτομέρειες που σχετίζονται με πιο σύνθετα κομμάτια του. Ο λόγος που αναφέρθηκε, είναι πως από τη στιγμή που μπορούμε να αντιμετωπίσουμε την ομαδοποίηση κινητών χρηστών (και ομάδων) ως ομαδοποίηση μεταβλητών σε ροές δεδομένων, έχει νόημα να εξετάσουμε ένα συναφές σύστημα. Στη δική μας προσέγγιση έχουμε πάρει ιδέες από το συγκεκριμένο σύστημα, αλλά αντιμετωπίζουμε τη ροή των δεδομένων (πληροφορία θέσης) που έχουμε με διαφορετικό τρόπο. Προσπαθούμε να βρούμε μια περίοδο όπου δεν θα χρειάζεται να παρατηρούμε όλους τους κόμβους του δικτύου μας αλλά μόνο ένα μέρος από αυτούς. Σε κάθε βήμα, παίρνουμε ένα παράδειγμα με μετρήσεις και το μετατοπίζουμε έτσι ώστε να αντιμετωπίσουμε τις τιμές των κόμβων ως παραδείγματα της ίδιας χρονικής στιγμής.

ΚΕΦΑΛΑΙΟ 3

ΣΥΣΤΗΜΑ ΑΥΤΟΜΑΤΗΣ ΟΜΑΔΟΠΟΙΗΣΗ ΚΑΙ ΕΝΤΟΠΙΣΜΟΥ ΥΠΟΠΤΩΝ ΟΜΑΔΩΝ

3.1 Εισαγωγή

Στα συστήματα υπηρεσιών βάσει πληροφορίας θέσης υπάρχει αποστολή πληροφορίας σε χρήστες, η οποία είναι άμεσα συσχετισμένη με τη τοποθεσία τους και με το τι τους περιβάλλει. Είναι εμφανές πως αν μπορούσαμε να εντοπίσουμε ομάδες οι οποίες κινούνται κοντά στις ίδιες περιοχές, θα μειώναμε αρκετά την αποστολή πληροφορίας εντός του δικτύου, καθώς μπορούμε να δώσουμε εντολή σε ένα κόμβο της ομάδας αυτής (ηγέτης) να διανείμει την πληροφορία στους γείτονές του. Χρειαζόμαστε λοιπόν έναν αλγόριθμο ο οποίος θα λαμβάνει ως είσοδο τις συντεταγμένες των χρηστών στη περιοχή που λειτουργεί αυτή η υπηρεσία και θα παράγει ως έξοδο τη κατανομή τους σε ομάδες. Βέβαια, αυτή η διαδικασία πρέπει να γίνεται εντελώς αυτόματα. Δε μπορεί το σύστημα να γνωρίζει εκ των προτέρων πόσες ομάδες έχει να εντοπίσει εντός μιας περιοχής, επομένως θέλουμε έναν αλγόριθμο ομαδοποίησης ο οποίος θα τρέχει χωρίς εποπτεία και θα εντοπίζει τις ομάδες.

Οι αλγόριθμοι κατάτμησης (k-means, k-medoids, KNN, κτλ) είναι ευρέως διαδεδομένοι για την πολύ καλή απόδοση που παρουσιάζουν ως προς την ποιότητα των ομάδων που εντοπίζουν καθώς δίνουν ομάδες που είναι πολύ συμπαγείς, καλά διαχωρίσιμες στο χώρο και σφαιρικού σχήματος συνήθως. Εντούτοις, τέτοια είδη αλγορίθμων δεν θα μπορούσαν να χρησιμοποιηθούν στο σύστημά μας καθώς απαιτούν ως είσοδο το πλήθος των ομάδων που πρέπει να εντοπίσουν καθώς και έναν εκπρόσωπο για τις ομάδες αυτές έτσι ώστε να αρχίσουν τη διαδικασία της ομαδοποίησης. Έτσι, καταλήξαμε στην ενσωμάτωση ενός ιεραρχικού αλγορίθμου ομαδοποίησης με τη χρησιμοποίηση κάποιου κριτηρίου για την επιλογή της βέλτιστης ομαδοποίησης, καθώς όπως θα δούμε στη συνέχεια, οι αλγόριθμοι αυτοί παράγουν μια ακολουθία από ομαδοποιήσεις καθώς τρέχουν.

Οι ιεραρχικοί αλγόριθμοι ομαδοποίησης χωρίζονται σε δύο κύριες κατηγορίες βάσει της λογικής που χτίζουν τις ομάδες τους:

- Συγχωνευτικούς Ιεραρχικούς Αλγορίθμους
- Διαιρετικούς Ιεραρχικούς Αλγορίθμους

Οι συγχωνευτικοί ιεραρχικοί αλγόριθμοι λειτουργούν με μια λογική από πάνω προς τα κάτω (bottom-up) για την ομαδοποίηση των δεδομένων. Θεωρούν κάθε σημείο του συνόλου δεδομένων πως ανήκει αρχικά σε μια ομάδα και επαναληπτικά ενώνουν τις ομάδες ανά δύο, βάσει τους ποιες είναι πιο κοντά κάθε φορά, μέχρι στο τέλος να προκύψει μια ομάδα που περιέχει όλα τα σημεία. Με την αντίστροφη ακριβώς λογική, οι διαιρετικοί αλγόριθμοι ομαδοποιούν τα δεδομένα από πάνω προς τα κάτω (top-down). Θεωρούν αρχικά πως όλα τα σημεία του συνόλου δεδομένων ανήκουν σε μία ομάδα μόνο και επαναληπτικά διαχωρίζουν βέλτιστα την ομάδα που είναι περισσότερο ανομοιογενής σε σχέση με τις υπόλοιπες. Τα πλεονεκτήματα των ιεραρχικών αλγορίθμων περιλαμβάνουν:

- Ενσωματωμένη ευελιξία όσον αφορά το επίπεδο διακριτότητας
- Ευκολία στη διαχείριση οποιασδήποτε μορφής μετρικής που βασίζεται στην ομοιότητα ή την απόσταση
- Δυνατότητα εφαρμογής σε οποιονδήποτε τύπο χαρακτηριστικού

Ενώ, τα μειονεκτήματα μιας ιεραρχικής ομαδοποίησης συνοψίζονται στα ακόλουθα:

- Ασάφεια όσον αφορά τα κριτήρια τερματισμού της διαδικασίας
- Οι περισσότεροι από αυτούς τους αλγορίθμους δεν επανεξετάζουν τις ενδιάμεσες ομάδες που έχουν παραχθεί κατά τη διάρκεια εφαρμογής τους για περαιτέρω βελτίωση

Θα έλεγε κανείς πως εφόσον εμείς θέλουμε να εντοπίσουμε ομάδες ατόμων εντός μιας περιοχής και θεωρώντας πως το πλήθος των ομάδων αυτών συνήθως είναι αρκετά μικρό συγκριτικά με το μέγεθος του συνόλου των δεδομένων, οι διαιρετικοί συγχωνευτικοί αλγόριθμοι θα ήταν καταλληλότεροι έναντι των συγχωνευτικών. Λόγω της μεγάλης πολυπλοκότητας αυτής της κατηγορίας διαιρετικών αλγορίθμων, τελικά κάτι τέτοιο δεν είναι και πολύ ρεαλιστικό. Αξίζει να αναφερθεί στο σημείο αυτό πως η πολυπλοκότητα ενός συγχωνευτικού ιεραρχικού αλγορίθμου είναι της τάξης του $O(N^3)$ ενώ ενός διαιρετικού της τάξης του $O(2^n)$, κάτι το οποίο μας δείχνει πως όταν το πλήθος του συνόλου των δεδομένων (*N*) που έχουμε να ομαδοποιήσουμε ξεπεράσει την τιμή του 10, οι συγχωνευτικοί αλγόριθμοι γίνονται πιο αποδοτικοί. Εκτός αυτού, ένα άλλο χαρακτηριστικό των συγχωνευτικών αλγορίθμων, καθώς αυτοί προχωρούν βήμα-βήμα, είναι πως δημιουργούν μια κατάτμηση των δεδομένων όπου για κάθε ομάδα αυτής της κατάτμησης έχουμε πληροφορία για το ποιες ομάδες την απαρτίζουν, από τα προηγούμενα βήματα του αλγορίθμου. Αυτή η πληροφορία, όπως θα παρουσιαστεί στη συνέχεια, είναι πολύ χρήσιμη για τον αλγόριθμό μας έτσι ώστε να καταδείξει αν κάποιες

από τις ομάδες που προέκυψαν στη βέλτιστη ομαδοποίηση είναι «ύποπτες» όσον αφορά την κινητικότητα των μελών τους. Για τους παραπάνω λόγους, επομένως, επιλέξαμε τη χρησιμοποίηση ενός συγχωνευτικού ιεραρχικού σχήματος ώστε να εντοπίζονται αυτόματα οι ομάδες των ατόμων που βρίσκονται μέσα σε μια περιοχή εποπτείας.

3.2 Συγχωνευτικός Ιεραρχικός Αλγόριθμος

3.2.1 Εισαγωγή

Όπως έχει προαναφερθεί, οι ιεραρχικοί αλγόριθμοι είναι διαφορετικής φιλοσοφίας σε σχέση με τους υπόλοιπους αλγορίθμους ομαδοποίησης καθώς αντί να δίνουν ως αποτέλεσμα μια ομαδοποίηση για τα δεδομένα που εξετάζουν, παράγουν μια ιεραρχία εμφωλευμένων ομαδοποιήσεων [4]. Πριν περιγράψουμε τη βασική ιδέα λειτουργία του συγχωνευτικού ιεραρχικού αλγορίθμου ας δούμε πρώτα κάποιους ορισμούς γι' αυτά που θα χρησιμοποιηθούν κατά τη διάρκεια αυτής της παραγράφου.

Αν ένα σύνολο από N l-διάστατα δεδομένα που πρέπει να ομαδοποιηθούν, συμβολιστεί ως

$$X = \{x_i, i = 1, \dots, N\}$$

τότε ο ορισμό μιας ομαδοποίησης ορίζεται ως

$$\Re = \{C_j, j = 1, \dots, m\}$$

όπου $C_j \subseteq X$.

Μια ομαδοποίηση $\Re_1$ η οποία περιέχει k ομάδες ορίζεται ως εμφωλευμένη στην ομαδοποίηση $\Re_2$ η οποία περιέχει $r\ (< k)$ ομάδες, αν κάθε ομάδα της $\Re_1$ είναι ένα υποσύνολο κάποιου συνόλου στην $\Re_2$. Πρέπει επίσης να αναφερθεί πως τουλάχιστον μία ομάδα της $\Re_1$ είναι καθαρό υποσύνολο της $\Re_2$. Σε αυτή την περίπτωση γράφουμε πως $\Re_1 \sqsubset \Re_2$. Για παράδειγμα, η ομαδοποίηση $\Re_1 = \{\{x_1, x_3\}, \{x_4\}, \{x_2, x_5\}\}$ θεωρείται εμφωλευμένη στην $\Re_2 = \{\{x_1, x_3, x_4\}, \{x_2, x_5\}\}$. Αντιθέτως, η ομαδοποίηση $\Re_1$ δεν είναι εμφωλευμένη στην $\Re_3 = \{\{x_1, x_4\}, \{x_3\}, \{x_2, x_5\}\}$ αλλά ούτε και στην $\Re_4 = \{\{x_1, x_2, x_4\}, \{x_3, x_5\}\}$. Τέλος, είναι προφανές πως μια ομαδοποίηση δεν μπορεί να είναι εμφωλευμένη στον εαυτό της.

Οι ιεραρχικοί αλγόριθμοι ομαδοποίησης παράγουν μια ιεραρχία εμφωλευμένων ομαδοποιήσεων. Πιο συγκεκριμένα, οι αλγόριθμοι αυτοί περιλαμβάνουν $N-1$ βήματα, όσο δηλαδή και το πλήθος των διανυσμάτων δεδομένων που έχουν να ομαδοποιήσουν, μειωμένο κατά 1. Σε κάθε βήμα t, μια νέα ομαδοποίηση προκύπτει η οποία και βασίζεται στην ομαδοποίηση που προέκυψε στο βήμα $t-1$. Όπως έχει προαναφερθεί, υπάρχουν δύο κύριες κατηγορίες σε αυτούς του αλγορίθμους, οι συγχωνευτικοί και οι διαιρετικοί αλγόριθμοι.

Η αρχική ομαδοποίηση $\Re_0$ ενός συγχωνευτικού ιεραρχικού αλγορίθμου αποτελείται από N ομάδες, όπου κάθε μία περιέχει ένα μόνο στοιχείο του συνόλου X. Στο πρώτο βήμα, παράγεται η ομαδοποίηση $\Re_1$ η οποία περιέχει $N-1$ ομάδες έτσι ώστε $\Re_0 \sqsubset \Re_1$. Αυτή η διαδικασία συνεχίζεται μέχρι να προκύψει και η τελευταία δυνατή ομαδοποίηση $\Re_{N-1}$ όπου περιλαμβάνει μία και μόνο ομάδα που έχει στο εσωτερικό της όλα τα στοιχεία του συνόλου X. Για την ιεραρχία των ομαδοποιήσεων που προέκυψαν από τα βήματα του συγχωνευτικού ιεραρχικού αλγορίθμου ισχύει ότι:

$$\Re_0 \sqsubset \Re_1 \sqsubset \cdots \sqsubset \Re_{N-1}$$

Οι διαιρετικοί ιεραρχικοί αλγόριθμοι ακολουθούν ακριβώς το αντίθετο μονοπάτι. Σε αυτή τη κατηγορία ιεραρχικών αλγορίθμων, η αρχική ομαδοποίηση $\Re_0$ περιλαμβάνει μία ομάδα που περιέχει όλα τα στοιχεία του συνόλου X. Στο πρώτο βήμα, παράγεται η ομαδοποίηση $\Re_1$ η οποία αποτελείται από δύο ομάδες για τις οποίες ισχύει $\Re_1 \sqsubset \Re_0$. Αυτή η διαδικασία συνεχίζεται μέχρι να προκύψει και η τελευταία δυνατή ομαδοποίηση $\Re_{N-1}$ όπου περιλαμβάνει N ομάδες που η κάθε μία έχει και ένα στοιχείο από το σύνολο X. Σε αυτή την περίπτωση, για την ιεραρχία των ομαδοποιήσεων που προέκυψαν από τα βήματα του διαιρετικού ιεραρχικού αλγορίθμου ισχύει ότι:

$$\Re_{N-1} \sqsubset \Re_{N-2} \sqsubset \cdots \sqsubset \Re_0$$

Στην επόμενη υποενότητα, θα γίνει εκτενής ανάλυση των συγχωνευτικών ιεραρχικών αλγορίθμων.

3.2.2 Συγχωνευτικοί Αλγόριθμοι

Έστω $g(C_i, C_j)$ μια συνάρτηση η οποία ορίζεται για όλους τους πιθανούς συνδυασμούς ζευγαριών στο σύνολο δεδομένων X. Η συγκεκριμένη συνάρτηση ποσοτικοποιεί την εγγύτητα μεταξύ των ομάδων C_i και C_j. Επίσης, έστω πως t είναι το τρέχον επίπεδο της

ιεραρχίας. Το γενικό συγχωνευτικό σχήμα επομένως μπορεί να περιγραφεί ως ακολούθως:

Γενικευμένο Συγχωνευτικό Σχήμα (ΓΣΣ)

- Αρχικοποίηση:
  - Επιλογή $\Re_0 = \{C_i = \{x_i\}, i = 1, \dots, N\}$ ως αρχικής ομαδοποίησης
  - $t = 0$
- Επανέλαβε:
  - $t = t + 1$
  - Μεταξύ όλων των πιθανών ζευγαριών ομάδων (C_r, C_s) στην ομαδοποίηση $\Re_{t-1}$ βρες αυτό, έστω το (C_i, C_j), για το οποίο ισχύει:

$$g(C_i, C_j) \begin{cases} \min\limits_{r,s} g(C_r, C_s), & \text{αν η } g \text{ είναι μια συνάρτηση ανομοιότητας} \\ \max\limits_{r,s} g(C_r, C_s), & \text{αν η } g \text{ είναι μια συνάρτηση ομοιότητας} \end{cases}$$

  - Δημιούργησε τη νέα ομάδα $C_q = C_i \cup C_j$ και παρήγαγε τη νέα ομαδοποίηση $\Re_t = (\Re_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$
- Μέχρι όλα τα αρχικά στοιχεία του συνόλου να ανήκουν στην ίδια ομάδα

Είναι ξεκάθαρο πως αυτό το σχήμα δημιουργεί μια ιεραρχία από N ομαδοποιήσεις, έτσι ώστε κάθε μία να είναι εμφωλευμένη σε όλες τις επιτυχημένες ομαδοποιήσεις, δηλαδή,

$$\Re_{t_1} \sqsubset \Re_{t_2}, \text{ για } t_1 < t_2, t_2 = 1, \dots, N-1$$

Εναλλακτικά, μπορούμε να ισχυριστούμε πως αν δύο στοιχεία συγχωνευθούν σε μια ομάδα στο επίπεδο t της ιεραρχίας των ομαδοποιήσεων, θα παραμείνουν στην ίδια ομάδα για όλες τις ομαδοποιήσεις που θα ακολουθήσουν στα επόμενα βήματα της ιεραρχίας. Αυτή είναι μια άλλη οπτική γωνία της ιδιότητας της εμφώλευσης.

Ένα μειονέκτημα της ιδιότητας αυτής είναι πως δεν υπάρχει τρόπος να ανακάμψουμε από μια «μέτρια» ομαδοποίηση η οποία μπορεί να συνέβη σε ένα χαμηλότερο επίπεδο της ιεραρχίας [26]. Είναι όμως κάτι το οποίο, αν και αρνητικό γι' αυτή την κατηγορία αλγορίθμων, μπορεί να μας προσφέρει μια αρκετά χρήσιμη πληροφορία για τις περιπτώσεις των υπηρεσιών βασισμένων σε πληροφορία θέσης, όπως θα δούμε στην ενότητα 3.4.

Σε κάθε επίπεδο t της ιεραρχίας υπάρχουν $N-t$ ομάδες. Επομένως, για να προσδιοριστεί το ζεύγος των ομάδων που θα συγχωνευθεί στο επίπεδο $t+1$, πρέπει να ελεγχθούν $\binom{N-t}{2} = \frac{(N-t)(N-t-1)}{2}$ ζεύγη ομάδων. Προκύπτει λοιπόν πως το συνολικό πλήθος των ζευγών που πρέπει να ελεγχθούν σε όλη τη διάρκεια της διαδικασίας ομαδοποίησης είναι

$$\sum_{t=0}^{N-1} \binom{N-t}{2} = \sum_{k=1}^{N} \binom{k}{2} = \frac{(N-1)N(N+1)}{6}$$

κάτι το οποίο μας δείχνει επίσης πως το σύνολο των πράξεων που χρειάζονται από ένα συγχωνευτικό ιεραρχικό σχήμα είναι ανάλογο του N^3. Εντούτοις, η ακριβής πολυπλοκότητα του αλγορίθμου εξαρτάται άμεσα από τον ορισμό της συνάρτησης g.

3.2.2.1 Ορισμός Χρήσιμων Ποσοτήτων

- Μήτρα Εγγύτητας (Proximity Matrix)

Υπάρχει μια περαιτέρω κατηγοριοποίηση των συγχωνευτικών αλγορίθμων σε αυτούς που βασίζονται στη θεωρία μητρών (πινάκων) και σε εκείνους που βασίζονται στη θεωρία συνόλων. Η δική μας προσέγγιση έγινε βάσει της λειτουργία των συγχωνευτικών ιεραρχικών αλγορίθμων επάνω στη θεωρία μητρών. Για το λόγο αυτό θα δοθεί έμφαση μόνο στη συγκεκριμένη κατηγορία, αφού πρώτα οριστούν κάποιες χρήσιμες ποσότητες που θα συναντήσουμε αργότερα. Η *μήτρα προτύπων* $D(X)$ είναι μια μήτρα $N \times l$, της οποίας η $i-$*οστή* είναι το (μεταφερμένο) $i-$*οστό* διάνυσμα από το σύνολο δεδομένων X. Η *μήτρα ομοιότητας* (ή *ανομοιότητας*) $P(X)$ είναι μια μήτρα $N \times N$ της οποίας το (i,j) στοιχείο αντιπροσωπεύει την τιμή της ομοιότητας $s(x_i, x_j)$ (ή της ανομοιότητας $d(x_i, x_j)$) μεταξύ των διανυσμάτων x_i και x_j. Πολλές φορές αναφέρεται και ως μήτρα εγγύτητας έτσι ώστε να καλύψει και τις δύο περιπτώσεις του μέτρου εγγύτητας (ομοιότητα ή ανομοιότητα δηλαδή). Γενικά, η μήτρα P είναι συμμετρική. Όταν η μήτρα P είναι συμμετρική, τα διαγώνια στοιχεία της είναι ίσα με τη μέγιστη τιμή της συνάρτησης ομοιότητας s ή με τη μικρότερη τιμή της συνάρτησης ανομοιότητας d αντιστοίχως. Για μια μήτρα προτύπων $D(X)$ υπάρχουν περισσότερες από μια μήτρες εγγύτητας $P(X)$ ανάλογα με την επιλογή του μέτρου εγγύτητας $p(x_i, x_j)$. Εντούτοις, αποφασίζοντας για το πως θα ορίζεται το $p(x_i, x_j)$, μπορεί εύκολα να παρατηρηθεί πως δοθείσης μιας μήτρας προτύπων υπάρχει μια και μόνο άμεσα συσχετιζόμενη μήτρα

εγγύτητας. Αντιθέτως, μια μήτρα εγγύτητας μπορεί να αντιστοιχεί σε πολλές μήτρες προτύπων. Ας δούμε σε αυτό το σημείο ένα παράδειγμα έτσι ώστε να είναι λίγο πιο κατανοητά τα παραπάνω.

Παράδειγμα 3.1

Έστω $X = \{x_i, i = 1, \dots, 5\}$ όπου $x_1 = [1,1]^T, x_2 = [2,1]^T, x_3 = [5,4]^T, x_4 = [6,5]^T$ και $x_5 = [6.5,6]^T$. Η μήτρα προτύπων του X είναι

$$D(X) = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 5 & 4 \\ 6 & 5 \\ 6.5 & 6 \end{bmatrix}$$

και η αντίστοιχη μήτρα ανομοιότητας, όταν ως μέτρο ανομοιότητας χρησιμοποιείται η Ευκλείδεια απόσταση, είναι

$$P(X) = \begin{bmatrix} 0 & 1 & 5 & 6.4 & 7.4 \\ 1 & 0 & 4.2 & 5.7 & 6.7 \\ 5 & 4.2 & 0 & 1.4 & 2.5 \\ 6.4 & 5.7 & 1.4 & 0 & 1.1 \\ 7.4 & 6.7 & 2.5 & 1.1 & 0 \end{bmatrix}$$

Όταν ως μέτρο εγγύτητας χρησιμοποιείται η μετρική ομοιότητας Tanimoto, η μήτρα ομοιότητας του X γίνεται

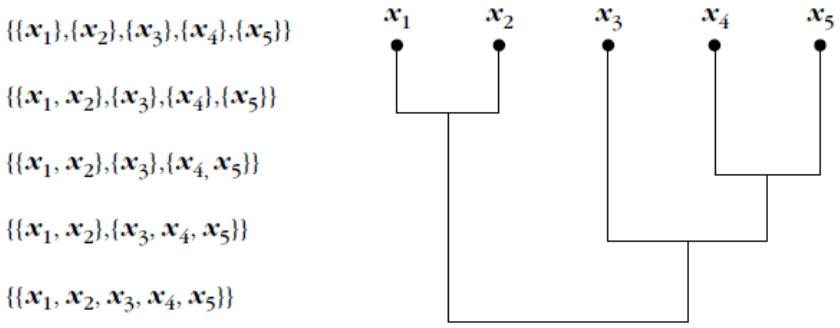
$$P'(X) = \begin{bmatrix} 1 & 0.75 & 0.26 & 0.21 & 0.18 \\ 0.75 & 1 & 0.44 & 0.35 & 0.20 \\ 0.26 & 0.44 & 1 & 0.96 & 0.90 \\ 0.21 & 0.35 & 0.96 & 1 & 0.98 \\ 0.18 & 0.20 & 0.90 & 0.98 & 1 \end{bmatrix}$$

Βλέπουμε πως στη μήτρα ανομοιότητας $P(X)$ όλα τα διαγώνια στοιχεία είναι ίσα με 0, εφόσον $d_2(x, x) = 0$, ενώ στην μήτρα ομοιότητας $P'(X)$ όλα τα διαγώνια στοιχεία είναι ίσα με 1, εφόσον $s_T(x, x) = 1$.

- Δενδρόγραμμα

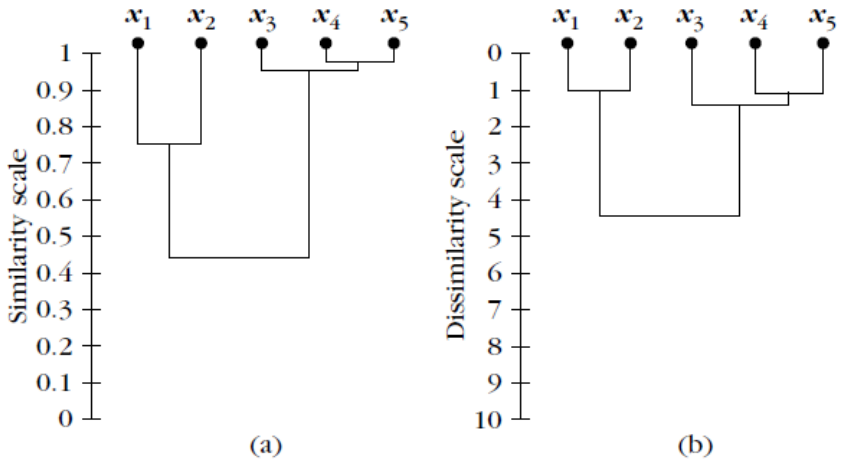
Ένα δενδρόγραμμα κατωφλίων ή πιο απλά ένα **δενδρόγραμμα** (dendrogram), είναι ένα αποδοτικό μέσο για την αναπαράσταση της ακολουθίας των ομαδοποιήσεων που παράγονται από έναν συγχωνευτικό (ή διαιρετικό) ιεραρχικό αλγόριθμο. Για να αποσαφηνιστεί σαν έννοια, ας αναλογιστούμε το σύνολο δεδομένων από το

Παράδειγμα 3.1. Ορίζοντας τη συνάρτηση $g(C_i, C_j)$ ως την ελάχιστη από τις αποστάσεις όλων των ζευγών (C_i, C_j), προκύπτει πολύ εύκολα πως αν το μέτρο ανομοιότητας μεταξύ δύο ομάδων είναι η Ευκλείδεια απόσταση, τότε η ακολουθία των ομαδοποιήσεων που παίρνουμε ως αποτέλεσμα από το γενικευμένο συγχωνευτικό σχήμα είναι αυτή που φαίνεται στην Εικόνα 5. Στο πρώτο βήμα του αλγορίθμου, τα πρότυπα x_1 και x_2 συγχωνεύονται και δημιουργούν μια νέα ομάδα. Στο δεύτερο βήμα, συγχωνεύονται τα πρότυπα x_4 και x_5 δημιουργώντας και αυτά μια νέα ομάδα. Στο τρίτο βήμα, το πρότυπο x_3 συγχωνεύεται με την ομάδα $\{x_4, x_5\}$ και τέλος, στο τέταρτο βήμα οι ομάδες $\{x_1, x_2\}$ και $\{x_3, x_4, x_5\}$ συγχωνεύονται σε μία ομάδα που περιέχει όλα τα στοιχεία του συνόλου X. Κάθε βήμα του γενικευμένου συγχωνευτικού σχήματος (ΓΣΣ) αντιστοιχεί και σε ένα διαφορετικό επίπεδο του δενδρογράμματος. Κόβοντας το δενδρόγραμμα σε ένα συγκεκριμένο επίπεδο οδηγούμαστε και σε ένα διαφορετικό αποτέλεσμα ομαδοποίησης.



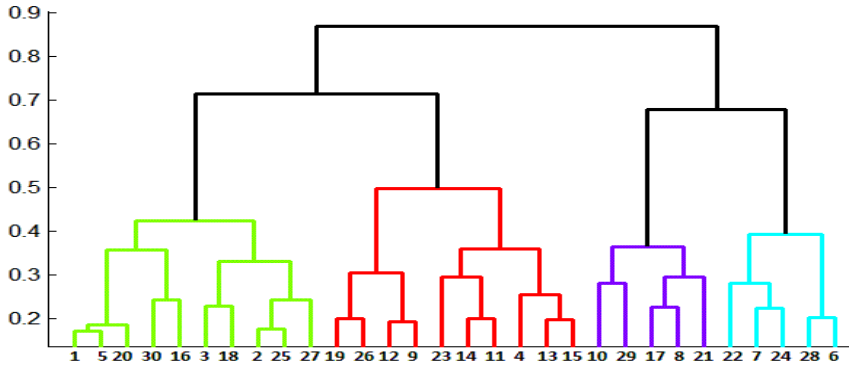
Εικόνα 5: Η ιεραρχία των ομαδοποιήσεων για το σύνολο δεδομένων X του Παραδείγματος 3.1 και το αντίστοιχο δενδρόγραμμα

Το *δενδρόγραμμα εγγύτητας* είναι ένα δενδρόγραμμα που λαμβάνει υπόψη του το επίπεδο εγγύτητας στο οποίο δύο ομάδες συγχωνεύτηκαν σε μία για πρώτη φορά. Όταν χρησιμοποιείται ένα μέτρο ανομοιότητας (ή ομοιότητας), το δενδρόγραμμα εγγύτητας ονομάζεται δενδρόγραμμα ανομοιότητας (ή ομοιότητας). Αυτό το εργαλείο μπορεί να χρησιμοποιηθεί ως μια αναπαράσταση της φυσικής ή της τεχνητής μορφής των ομάδων σε κάθε επίπεδο. Δηλαδή, μπορεί να δώσει μια ιδέα για την ομαδοποίηση που ταιριάζει περισσότερο στο σύνολο δεδομένων που έχουμε. Στην Εικόνα 6 μπορούμε να δούμε τα δενδρογράμματα ομοιότητας και ανομοιότητας για το σύνολο δεδομένων του Παραδείγματος 3.1 όταν χρησιμοποιούνται οι μήτρες εγγύτητας $P'(X)$ και $P(X)$ αντίστοιχα.



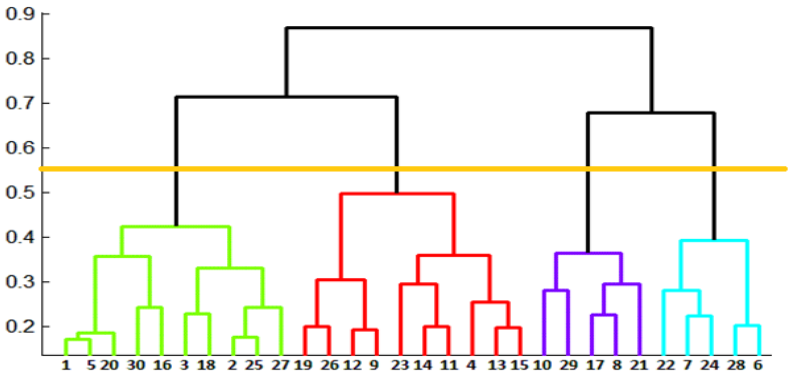
Εικόνα 6: (a)Το δενδρόγραμμα εγγύτητας (ομοιότητας) για το σύνολο X χρησιμοποιώντας τη μήτρα εγγύτητας P'(X) του Παραδείγματος 3.1, (b)Το δενδρόγραμμα εγγύτητας (ανομοιότητας) για το σύνολο X χρησιμοποιώντας τη μήτρα εγγύτητας P(X) του Παραδείγματος 3.1

Πριν προχωρήσουμε σε πιο ενδελεχή ανάλυση των συγχωνευτικών ιεραρχικών αλγορίθμων, πρέπει να τονιστεί κάτι. Όπως εξηγήθηκε προηγουμένως, οι ιεραρχικοί αλγόριθμοι, προσδιορίζουν μια ακολουθία ομαδοποιήσεων αντί για μία και μόνο ομαδοποίηση. Ο προσδιορισμός επομένως ολόκληρου του δενδρογράμματος μπορεί να είναι πολύ χρήσιμος σε μερικές εφαρμογές όπως η βιολογική ταξινόμηση. Εντούτοις, σε άλλες εφαρμογές ενδιαφερόμαστε μόνο για την ομαδοποίηση που ταιριάζει καλύτερα στα δεδομένα που έχουμε. Οπότε, αν κάποιος σκοπεύει να κάνει χρήση των ιεραρχικών αλγορίθμων για τέτοιες περιπτώσεις, όπως έχει αναφερθεί πολλές φορές μέχρι τώρα, θα πρέπει να προσδιορίσει ποια ομαδοποίηση από την παραγόμενη ιεραρχία είναι η πιο κατάλληλη για τα δεδομένα του. Ισοδύναμα, πρέπει να είναι σε θέση να προσδιορίσει το κατάλληλο επίπεδο στο οποίο πρέπει να κοπεί το δενδρόγραμμα που αντιστοιχεί στην προκύπτουσα ιεραρχία ομαδοποιήσεων. Έστω λοιπόν πως έχουμε μια περίπτωση όπου ομαδοποιούμε με τη βοήθεια ενός συγχωνευτικού ιεραρχικού αλγορίθμου χρήστες που βρίσκονται σε μια περιοχή ενδιαφέροντος της εφαρμογής μας και ως μέτρο εγγύτητας χρησιμοποιείται η Ευκλείδεια απόσταση που είναι μια μετρική ανομοιότητας. Το δενδρόγραμμα ανομοιότητας που θα παίρναμε, με την ιεραρχία των ομαδοποιήσεων, φαίνεται στην Εικόνα 7. Στο συγκεκριμένο δενδρόγραμμα στον άξονα των x έχουμε τις ταυτότητες των χρηστών (user ids) ενώ στον άξονα των y τις τιμές της μετρικής ανομοιότητας. Παρατηρούμε πως όσο τα βήματα του συγχωνευτικού ιεραρχικού αλγορίθμου προχωρούν συγχωνεύονται ομάδες που έχουν μεγαλύτερη ανομοιότητα. Παρατηρώντας το συγκεκριμένο δενδρόγραμμα θα μπορούσε ο χρήστης



Εικόνα 7: Δενδρόγραμμα ανομοιότητας από ένα παράδειγμα συνόλου δεδομένων που προέρχεται από την παρακολούθηση χρηστών σε ένα χώρο

να ζητήσει ο αλγόριθμος να διακόψει τη λειτουργία του όταν θα έχουν παραχθεί 4 ομάδες ή όταν πάνε να συγχωνευθούν ομάδες με ανομοιότητα μεγαλύτερη του 0.55, καθώς έχει παρατηρήσει από τα δεδομένα του πως μια τέτοια ομαδοποίηση είναι αυτή που ταιριάζει καλύτερα. Δηλαδή, θα θέλαμε να κόψουμε το δενδρόγραμμα ανομοιοτήτων όπως φαίνεται στην Εικόνα 8.



Εικόνα 8: Τομή του δενδρογράμματος σε επίπεδο τέτοιο ώστε να πάρουμε ως τελική την ομαδοποίηση που περιέχει 4 ομάδες

Βέβαια, το να γνωρίζει το σύστημα πόσες ομάδες θέλει να έχει είναι κάτι που δεν συναντάται συχνά στις πραγματικές εφαρμογές παρακολούθησης ομάδων για υπηρεσίες που βασίζονται στη θέση, αν όμως ισχύει τότε θα ήταν προτιμότερο να χρησιμοποιηθεί κάποιος αλγόριθμος κατάτμησης αντί κάποιος ιεραρχικός καθώς έχει πολύ μεγαλύτερη πολυπλοκότητα. Χρησιμοποιώντας όμως κάποιο καθολικό κριτήριο, το οποίο ελέγχοντας την ιεραρχία των ομαδοποιήσεων θα μπορούσε να αποφαίνεται για το ποια είναι η ομαδοποίηση με το «μεγαλύτερο» νόημα για το σύνολο των

δεδομένων, οι ιεραρχικοί αλγόριθμοι μετατρέπονται πραγματικά στους καταλληλότερους για ενσωμάτωση στο σύστημά μας έτσι ώστε να έχουμε αυτόματο εντοπισμό των ομάδων σε τέτοιου είδους υπηρεσίες.

3.2.2.2 Συγχωνευτικοί Αλγόριθμοι Βασιζόμενοι στη Θεωρία Μητρών

Όπως αναφέρθηκε και προηγουμένως, από τις δύο κατηγορίες συγχωνευτικών ιεραρχικών αλγορίθμων, η προσέγγισή μας ενσωματώνει αυτή που κάνει χρήση της θεωρίας μητρών στα βήματά της. Αυτοί οι αλγόριθμοι μπορούν να θεωρηθούν ως ειδική περίπτωση του γενικευμένου συγχωνευτικού σχήματος (ΓΣΣ). Η είσοδος σε τέτοια σχήματα είναι μια $N \times N$ μήτρα ανομοιότητας, έστω $P_0 = P(X)$, η οποία προκύπτει από το σύνολο X. Σε κάθε επίπεδο, t, όταν δύο ομάδες συγχωνεύονται σε μία, το μέγεθος της μήτρας ανομοιότητας P_t γίνεται $(N-t) \times (N-t)$. Η P_t προκύπτει από την P_{t-1} με (α) τη διαγραφή των δύο γραμμών και των δύο στηλών που αντιστοιχούν στις ομάδες που συγχωνεύονται και έπειτα με (β) την εισαγωγή μιας νέας σειράς και μιας νέας στήλης που περιέχει τις αποστάσεις μεταξύ της νεοσύστατης ομάδας και των παλαιότερων (που δεν επηρεάζονται σε αυτό το επίπεδο) ομάδων. Έτσι λοιπόν, δε χρειάζεται σε κάθε βήμα συγχώνευσης του αλγορίθμου να υπολογίζονται όλες οι αποστάσεις ξανά, αλλά με τη χρήση της μήτρας ανομοιότητας αρκεί ο υπολογισμός της απόστασης της νέας ομάδας από τις υπόλοιπες αυτού του επιπέδου και μόνο. Γενικά, η απόσταση μεταξύ της νεοσύστατης ομάδας C_q (το αποτέλεσμα της συγχώνευσης των ομάδων C_i και C_j) και μιας ομάδας η οποία σε αυτό το επίπεδο των ομαδοποιήσεων δεν επηρεάζεται C_s, είναι μια συνάρτηση της μορφής

$$d(C_q, C_s) = f(d(C_i, C_s), d(C_j, C_s), d(C_i, C_j))$$

Η διαδικασία ενημέρωσης της μήτρας ανομοιότητας επομένως δικαιολογεί το όνομά της ως *Αλγόριθμος Ενημέρωσης Μήτρας (Matrix Updating Algorithm - MUA)*, που συναντάται συχνά στη βιβλιογραφία. Ας δούμε λοιπόν το αλγοριθμικό σχήμα που υλοποιήθηκε στα πλαίσια αυτής τη διπλωματικής εργασίας, όπου και πάλι το t υποδηλώνει το τρέχον επίπεδο της ιεραρχίας των ομαδοποιήσεων.

Αλγόριθμος Ενημέρωσης Μήτρας (ΑΕΜ)

- Αρχικοποίηση:
  - $\Re_0 = \{C_i = \{x_i\}, i = 1, \dots, N\}$

- $P_0 = P(X)$
- $t = 0$

- Επανέλαβε:
  - $t = t + 1$
  - Εντόπισε ομάδες C_i, C_j τέτοιες ώστε
$$d(C_i, C_j) = \min_{r,s=1\dots N, r\neq s} d(C_r, C_s)$$
  - Συγχώνευσε τις ομάδες C_i, C_j σε νέα ομάδα C_q και σχημάτισε τη νέα ομαδοποίηση $\Re_t = (\Re_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$
  - Προσδιόρισε τη μήτρα εγγύτητας (ανομοιότητας) P_t από την P_{t-1} όπως περιγράφηκε παραπάνω
- Μέχρι το σχηματισμό της ομαδοποίησης $\Re_{N-1}$, δηλαδή μέχρι όλα τα στοιχεία του συνόλου X να ανήκουν στην ίδια ομάδα

Είναι εμφανές πως το παραπάνω σχήμα είναι στο πνεύμα του γενικευμένου συγχωνευτικού σχήματος, με τη διαφορά πως τώρα έχουμε τη χρήση της μήτρας ανομοιότητας όπου ορίζει τη διαδικασία με μεγαλύτερη σαφήνεια.

3.2.3 Μετρικές Εγγύτητας (Proximity Measures)

Όπως έχει φανεί και στις προηγούμενες παραγράφους, πρόκειται για ένα μέτρο το οποίο ποσοτικοποιεί το πόσο όμοια ή διαφορετικά είναι δύο διανύσματα χαρακτηριστικών. Θα πρέπει φυσικά όλα τα χαρακτηριστικά να συμμετέχουν με την ίδια βαρύτητα στη διαμόρφωση της τιμής αυτής και δεν υπάρχει κάποιο που να επικρατεί έναντι των άλλων και να συμβάλει με μεγαλύτερο βάρος. Αυτό λαμβάνεται υπόψη στη φάση της προεπεξεργασίας. Τα μέτρα αυτά χωρίζονται σε 2 κύριες κατηγορίες οι οποίες είναι τα μέτρα ομοιότητας (similarity measures) και τα μέτρα ανομοιότητας (dissimilarity measures). Αν και υπάρχουν αρκετά διαφορετικά μέτρα, εντούτοις δεν έχει βρεθεί κάποιο μέχρι τώρα το οποίο να κυριαρχεί έναντι των υπολοίπων. Ουσιαστικά, ανάλογα με το πρόβλημα που έχουμε να αντιμετωπίσουμε, επιλέγουμε και το κατάλληλο μέτρο.

3.2.3.1 Μετρικές Ανομοιότητας (Dissimilarity Measures)

Και οι ιεραρχικές μέθοδοι αλλά και οι μέθοδοι κατάτμησης χρησιμοποιούν διαφορετικά μέτρα ομοιότητας και ανομοιότητας (συνήθως αποστάσεις). Η συνηθισμένη L_p απόσταση

$$d(x,y) = \|x-y\|_p, \|z\|_p = \left(\sum\nolimits_{j=1:d} |z_j|^p\right)^{1/p}, \|z\| = \|z\|_2 \qquad (\mathbf{3.1})$$

χρησιμοποιείται για αριθμητικά δεδομένα, με $1 \leq p < \infty$, όπου η χαμηλότερη τιμή για το p αντιστοιχεί σε μια πιο εύρωστη εκτίμηση (συνεπώς επηρεάζεται λιγότερο από ακραίες τιμές). Αν θεωρήσουμε αυτά τα μέτρα ως παραμέτρους, τότε είναι μια παράμετρος η οποία καθορίζει τον τρόπο με τον οποίο μετριέται η απόσταση μεταξύ δύο σημείων ή ομάδων της εισόδου. Οι πιο διαδομένες επιλογές είναι οι ακόλουθες:

- <u>Ευκλείδεια Απόσταση:</u>

Η μέθοδος της Ευκλείδειας απόστασης (Euclidean Distance) μετρά την απόσταση μεταξύ δύο σημείων μέσω της μικρότερης ευθείας που υπάρχει μεταξύ τους. Αν τα σημεία που έχουμε είναι τα $X(x_1, x_2, \dots, x_n)$ και $Y(y_1, y_2, \dots, y_n)$ τότε ο τύπος που δίνει την Ευκλείδεια απόσταση μεταξύ τους είναι ο ακόλουθος:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Δηλαδή, η Ευκλείδεια απόσταση μεταξύ δύο σημείων είναι η ρίζα του αθροίσματος των τετραγώνων των διαφορών των γνωρισμάτων των σημείων αυτών. Είναι με διαφορά η πιο δημοφιλής επιλογή ως μέτρο ανομοιότητας, όπου χρησιμοποιείται και στην αντικειμενική συνάρτηση του αλγορίθμου k-means (το άθροισμα των τετραγώνων των αποστάσεων μεταξύ σημείων και κεντροειδών) η οποία έχει μια ξεκάθαρη στατιστική σημασία για τη συνολική διακύμανση μεταξύ των ομάδων.

- <u>Ευκλείδεια Τετραγωνική Απόσταση:</u>

Η μέθοδος αυτή υπολογίζεται παρόμοια με την Ευκλείδεια απόσταση με την διαφορά του ότι δεν υπάρχει η ρίζα. Όπως είναι λογικό ο υπολογισμός της απόστασης μεταξύ δύο σημείων με αυτή τη μέθοδο είναι γρηγορότερος απ' ότι με την Ευκλείδεια

απόσταση. Εντούτοις, αν και αλγόριθμοι όπως ο k-means δεν επηρεάζονται αν η Ευκλείδεια απόσταση αντικατασταθεί με την Ευκλείδεια τετραγωνική απόσταση υπάρχει περίπτωση οι ιεραρχικοί αλγόριθμοι να δώσουν διαφορετικά αποτελέσματα με την χρήση αυτών των δύο μέτρων.

$$d(X,Y) = \sum_{i=1}^{n} (x_i - y_i)^2$$

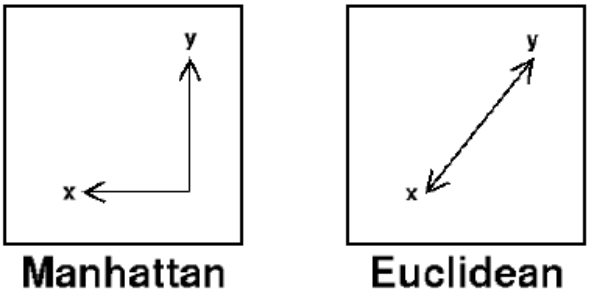
- Απόσταση Manhattan:

Η συνάρτηση της μεθόδου αυτής υπολογίζει την απόσταση που θα διανύαμε για τη μετάβασή μας από το ένα σημείο στο άλλο αν ακολουθούσαμε ένα μονοπάτι με μορφή πλέγματος. Αν τα σημεία που έχουμε είναι τα $X(x_1, x_2, \dots, x_n)$ και $Y(y_1, y_2, \dots, y_n)$ τότε ο τύπος που δίνει την Manhattan απόσταση μεταξύ τους είναι ο ακόλουθος:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|$$

όπου n είναι το πλήθος των χαρακτηριστικών που έχουν τα σημεία αυτού του συνόλου δεδομένων και x_i, y_i είναι οι τιμές της $i-οστής$ μεταβλητής στα σημεία X και Y.

Η παρακάτω εικόνα παρουσιάζει τη διαφορά που έχουν η Ευκλείδεια Απόσταση και η Απόσταση Manhattan ως προς τον τρόπο που προσδιορίζουν την απόσταση μεταξύ δύο σημείων.



Εικόνα 9: Ευκλείδεια και Manhattan απόσταση

- Απόσταση Mahalanobis:

Είναι μια μετρική για τη μέτρηση της απόστασης μεταξύ των σημείων η οποία βασίζεται στη συσχέτιση μεταξύ των μεταβλητών και έτσι μπορεί να εντοπίζει και να αναλύει διαφορετικά πρότυπα. Δίνει μια εικόνα για την ομοιότητα ενός άγνωστου συνόλου

δειγμάτων σε σχέση με ένα γνωστό (όσον αφορά την κατηγορία που αντιπροσωπεύουν). Ως μέτρο, διαφέρει από την Ευκλείδεια απόσταση στο γεγονός πως λαμβάνει υπόψη του τις συσχετίσεις εντός του συνόλου δεδομένων χωρίς να διαφοροποιεί τις κλίμακες.

Η απόσταση Mahalanobis ενός προτύπου πολλών μεταβλητών $x = (x_1, x_2, \dots, x_n)^T$ από μια ομάδα προτύπων με μέση τιμή $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$ και μήτρα συνδυασποράς S ορίζεται ως:

$$d_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}$$

Επίσης, μπορεί να οριστεί ως μέτρο ανομοιότητας μεταξύ δύο τυχαίων διανυσμάτων $\vec{x}$ και $\vec{y}$ της ίδια κατανομής και με μήτρα συνδυασποράς S ως εξής:

$$d_M(\vec{x}, \vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}$$

Μάλιστα, αν η μήτρα συνδυασποράς μεταξύ των διανυσμάτων αυτών ταυτίζεται με τη μήτρα ταυτότητας (identity matrix) η απόσταση Mahalanobis εμπίπτει στην Ευκλείδεια απόσταση. Τέλος, αν η μήτρα συνδυασποράς είναι διαγώνια, τότε η προκύπτουσα μετρική απόστασης ονομάζεται *κανονικοποιημένη (normalized) Ευκλείδεια απόσταση*:

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

όπου s_i είναι η τυπική απόκλιση των x_i και y_i πάνω στο σετ των δειγμάτων.

Όλες οι προαναφερθείσες μετρικές απόστασης υποθέτουν ανεξαρτησία μεταξύ των μεταβλητών (διαγώνια μήτρα συνδιασποράς S). Η απόσταση Mahalanobis από την άλλη χρησιμοποιείται σε αλγορίθμους, όπως ο ORCLUS, οι οποίοι δεν κάνουν αυτή την υπόθεση.

Η απόσταση η οποία επιστρέφει το μέγιστο των απολύτων διαφορών μεταξύ των συντεταγμένων των σημείων επίσης συναντάται συχνά και αντιστοιχεί για $p = \infty$ στη Σχέση (3.1). Σε αρκετές εφαρμογές (π.χ., ανάλυση προφίλ) τα σημεία μετατρέπονται στην ίδια κλίμακα έτσι ώστε να έχουν μια κοινή νόρμα, οπότε το μέτρο εγγύτητας είναι ουσιαστικά η γωνία μεταξύ των σημείων:

$$d(x, y) = \arccos\left(x^T y / \|x\| \cdot \|y\|\right)$$

Αυτό το μέτρο χρησιμοποιείται σε συγκεκριμένα εργαλεία, όπως το DIGNET, καθώς και σε εφαρμογές όπως η εξόρυξη κειμένου.

3.2.3.1.1 Haversine Formula

Πρέπει να τονιστεί πως αυτές οι μετρικές θα πρέπει να είναι σε θέση να υπολογίσουν απόσταση μεταξύ σημείων ή ομάδων η οποία να έχει νόημα σε σχέση με τις μονάδες μέτρησης που χρησιμοποιούνται για την αναπαράσταση των πεδίων κάθε σημείου από αυτές τις ομάδες. Πιο συγκεκριμένα, σε εφαρμογές όπου τα δεδομένα μας είναι στίγματα θέσης κινούμενων αντικειμένων σε ένα χώρο παρακολούθησης είναι πολύ πιθανό οι τιμές που παίρνουμε να είναι σε γεωγραφικές(GPS) συντεταγμένες, κάτι το οποίο σημαίνει πως αν θέλουμε να υπολογίσουμε σωστά τις αποστάσεις μεταξύ αυτών των σημείων θα πρέπει να λάβουμε υπόψη μας αυτή τη παράμετρο. Για το λόγο αυτό χρησιμοποιείται η Haversine Formula [29]. Πρόκειται για μια εξίσωση πολύ σημαντική στην πλοήγηση η οποία δίνει την απόσταση μεταξύ δύο σημείων επάνω στην επιφάνεια της γης η οποία μετράται κατά μήκος μιας διαδρομής επάνω στην επιφάνεια της σφαίρας (και όχι περνώντας μέσα από τη σφαίρα). Η θέση των σημείων σε τέτοιες επιφάνειες ορίζεται από το γεωγραφικό μήκος και πλάτος τους επομένως γίνεται κατανοητό πως από τη στιγμή που δεν βρισκόμαστε στο καρτεσιανό επίπεδο συντεταγμένων, δεν θα μπορούσαμε να πάρουμε την απόστασή τους με τη χρήση της Ευκλείδειας απόστασης ως μέτρο ανομοιότητας. Η συγκεκριμένη φόρμουλα είναι μια ειδική περίπτωση μιας πιο γενικευμένης φόρμουλας στο χώρο της σφαιρικής τριγωνομετρίας, η οποία είναι γνωστή ως ο νόμος των Haversines (Law of Haversines) και συνδέει τις πλευρές και τις γωνίες σφαιρικών τριγώνων.

$$haversin(c) = haversin(a - b) + \sin(a)\sin(b)\, haversin(C)$$

Για οποιαδήποτε δύο σημεία πάνω σε μια σφαιρική επιφάνεια η τιμή για το Haversine της κεντρικής γωνίας που σχηματίζεται μεταξύ τους δίνεται από τη σχέση:

$$haversin\left(\frac{d}{r}\right) = haversin(\varphi_2 - \varphi_1) + cos(\varphi_1)\cos(\varphi_2) haversin(\psi_2 - \psi_1) \qquad (\mathbf{3.2})$$

όπου,

- $haversin$ είναι η συνάρτηση Haversine:

$$haversin(\theta) = sin(\theta/2)^2 = \frac{1 - cos\theta}{2}$$

- d είναι η απόσταση μεταξύ των δύο σημείων πάνω από μια διαδρομή στην επιφάνεια της σφαίρας όμως
- r *είναι η ακτίνα της σφαίρας*
- φ_1, φ_2: το γεωγραφικό πλάτος των δύο σημείων
- ψ_1, ψ_2: το γεωγραφικό μήκος των δύο σημείων

Επιπροσθέτως, ο λόγος $^{d}/_{r}$ στο αριστερό μέλος της εξίσωσης είναι η εσωτερική γωνία των δύο σημείων με μονάδα μέτρησης $radians$. Η μετατροπή από $degrees$ σε $radians$ γίνεται, όπως είναι γνωστό, με τον πολλαπλασιασμό της τιμής σε $degrees$ με τον λόγο $^{\pi}/_{180}$. Λύνοντας λοιπόν τη σχέση 3.2 ως προς την απόσταση d που είναι και το ζητούμενο, έχουμε:

$$d = r \cdot haversin^{-1}(h) = 2r \cdot \arcsin\left(\sqrt{h}\right), h = haversin\left(\frac{d}{r}\right)$$

Ή πιο λεπτομερώς:

$$d = 2r\ arcsin\left(\sqrt{haversin(\varphi_2 - \varphi_1) + cos(\varphi_1)\cos(\varphi_2)haversin(\psi_2 - \psi_1)}\right)$$

$$\Rightarrow d = 2r\ arcsin\left(\sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + cos(\varphi_1)\cos(\varphi_2)sin^2\left(\frac{\psi_2 - \psi_1}{2}\right)}\right)$$

3.2.3.2 Μέτρικες Ομοιότητας (Similarity Measures)

Αναφορικά με τα μέτρα ομοιότητας, ο τύπος

$$s(x, y) = \frac{1}{1 + d(x, y)}$$

προσδιορίζει την ομοιότητα μεταξύ αριθμητικών πεδίων. Άλλες επιλογές είναι:

- Cosine:

$$s_{cos}(x, y) = \frac{x^T y}{\|x\|\|y\|}$$

- Dice coefficients:

$$s_{Dice}(x, y) = \frac{2x^T y}{\|x\|^2 + \|y\|^2}$$

- Εκθετική απόσταση (distance exponent):

$$s_{exp}(x, y) = \exp\left(-\|x - y\|^a\right)$$

Στρέφοντας τώρα τη προσοχή μας στα κατηγοριοποιημένα δεδομένα, υπάρχει ένα πλήθος μετρικών για την ομοιότητα μεταξύ των κατηγοριοποιημένων μεταβλητών/πεδίων. Υποθέτοντας δυαδικές μεταβλητές με τιμές $a, \beta = \pm$, έστω $d_{a\beta}$ ένα πλήθος πεδίων που έχουν ως αποτέλεσμα a στο x και β στο y. Τότε οι δείκτες R και J από τους *Rand* και *Jaccard* (οι οποίοι είναι επίσης γνωστοί και ως *Tanimoto*) ισούνται με:

$$R(x,y) = \frac{d_{++} + d_{--}}{d_{++} + d_{+-} + d_{-+} + d_{--}}$$

και

$$J(x,y) = \frac{d_{++}}{d_{++} + d_{+-} + d_{-+}}$$

Πρέπει να δοθεί έμφαση στο γεγονός πως ο δείκτης Jaccard αντιμετωπίζει τις θετικές και τις αρνητικές τιμές ασύμμετρα, κάτι το οποίο τον καθιστά ως το καταλληλότερο μέτρο για δεδομένα συναλλαγών όπου το «+» σημαίνει πως ένα αντικείμενο είναι παρόν. Ορίζεται ως το κλάσμα των κοινών στοιχείων μεταξύ δύο συναλλαγών προς το συνολικό αριθμό των στοιχείων των συναλλαγών. Χρησιμοποιείται ακόμη στο συνεργατικό φιλτράρισμα, στην ακολουθιακή ανάλυση, στην εξόρυξη κειμένου και στην αναγνώριση προτύπων.

Τα μέτρα εγγύτητας μεταξύ δύο ομάδων που μπορούν να προκύψουν από τις γειτνιάσεις των σημείων των ομάδων αυτών, είναι στην ουσία οι μετρικές σύνδεσης που θα αναφερθούν αναλυτικότερα στη συνέχεια. Επειδή το πρόβλημα που θέλουμε να αντιμετωπίσουμε περιγράφεται καλύτερα με τον όρο της «απόστασης» μεταξύ χρηστών και ομάδων, από τα μέτρα εγγύτητας θα χρησιμοποιήσουμε μόνο μέτρα ανομοιότητας.

3.2.4 Μετρικές Σύνδεσης (Linkage Metrics)

Η ιεραρχική ομαδοποίηση αρχικοποιεί ένα σύστημα ομάδων είτε σαν ένα σύνολο από ομάδες που περιέχουν μόνο ένα μέλος (συγχωνευτικός αλγόριθμος) , είτε ως μία ομάδα η οποία περιέχει όλο το σύνολο δεδομένων (διαιρετικός αλγόριθμος) και προχωράει επαναληπτικά με τη συνένωση ή το διαχωρισμό των πιο κατάλληλων ομάδων μέχρι να ικανοποιηθεί το κριτήριο διακοπής όπως αναφέρθηκε και προηγουμένως. Η καταλληλότητα μιας ομάδας για να διασπαστεί εξαρτάται από την ομοιότητα (ή την ανομοιότητα) που παρουσιάζουν τα μέλη της, ενώ η συμμετοχή της σε μια συνένωση εξαρτάται από την ομοιότητα (ή την ανομοιότητα) των μελών της σε σχέση με τα μέλη

των άλλων ομάδων που έχουν σχηματιστεί. Αυτό αντικατοπτρίζει τη γενική υπόθεση πως οι ομάδες αποτελούνται από μέλη (σημεία του συνόλου δεδομένων) τα οποία είναι όμοια μεταξύ τους. Ένα γνωστό μέτρο ανομοιότητας μεταξύ δύο σημείων είναι η απόσταση που έχουν όπως είδαμε προηγουμένως. Διαισθητικά, για τη μετρική σύνδεσης θα μπορούσαμε να πούμε ότι καθορίζει το τρόπο με τον οποίο θα μετρηθεί η ομοιότητα (ή ανομοιότητα) μεταξύ των ομάδων, εφόσον έχει γίνει ήδη η επιλογή του μέτρου εγγύτητας.

Για τη συνένωση ή το διαχωρισμό υποσυνόλων από σημεία αντί μεμονωμένων σημείων, η απόσταση μεταξύ μεμονωμένων σημείων πρέπει να γενικευτεί σε απόσταση μεταξύ υποσυνόλων από σημεία. Ένα μέτρο εγγύτητας το οποίο προέρχεται από αυτή την παρατήρηση είναι η ***μετρική σύνδεσης*** (linkage metric). Ο τύπος της μετρικής που επιλέγεται να χρησιμοποιηθεί επηρεάζει σε μεγάλο βαθμό το συγχωνευτικό ιεραρχικό αλγόριθμο, δεδομένου του ότι αντικατοπτρίζει το κομμάτι της συνδεσιμότητας και της απόστασης. Στις βασικότερες μετρικές σύνδεσης μεταξύ των ομάδων περιλαμβάνονται η μονή σύνδεση (single link), η σύνδεση μέσου όρου (average link) και η πλήρης σύνδεση (complete link). Μία ακόμα μέθοδος σύνδεσης είναι αυτή των κεντροειδών (centroid link) η οποία αν και δεν συγκαταλέγεται στις θεμελιώδεις βρίσκει πολλούς υποστηρικτές στις εφαρμογές ομαδοποίησης με ιεραρχικούς αλγορίθμους, λόγου του χαμηλού κόστους υπολογισμού της. Το υποκείμενο μέτρο ανομοιότητας (συνήθως η απόσταση) υπολογίζεται για κάθε ζευγάρι σημείων όπου το ένα σημείο ανήκει σε μία ομάδα και το άλλο σε διαφορετική. Μια ειδική λειτουργία όπως το ελάχιστο (μονή σύνδεση), ο μέσος όρος (μέση σύνδεση) ή το μέγιστο (πλήρης σύνδεση) εφαρμόζεται στα κατά ζεύγη μέτρα ανομοιότητας:

$$d(C_1, C_2) = operation\{d(x, y) | x \in C_1, y \in C_2\}$$

Όλες οι προαναφερθέντες περιπτώσεις μετρικών σύνδεσης μπορούν να προκύψουν ως περιπτώσεις της Lance-Williams φόρμουλας ενημέρωσης [27]:

$$d(C_q, C_s) = a_i d(C_i, C_s) + a_j d(C_j, C_s) + b d(C_i, C_j)$$
$$+ c\left|d(C_i, C_s) - d(C_j, C_s)\right| \quad \textbf{(3.3)}$$

Τα $a, b \, και \, c$ είναι συντελεστές που αντιστοιχούν σε διαφορετικά ήδη μετρικών σύνδεσης. Αυτή η φόρμουλα εκφράζει μια μετρική μεταξύ της ένωσης δύο ομάδων (νέα ομάδα) και μιας τρίτης ομάδας από την άποψη των βασικών συστατικών στοιχείων. Η Lance-Williams φόρμουλα έχει βαρύνουσα σημασία καθώς καθιστά τον χειρισμό των μέτρων ομοιότητας (ή ανομοιότητας) υπολογιστικά πραγματοποιήσιμο. Όταν η μετρική

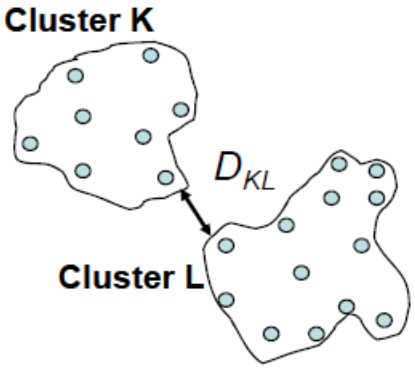
εγγύτητας είναι η απόσταση αυτές οι μέθοδοι κρατάνε το πόσο κοντά βρίσκονται μεταξύ τους οι ομάδες. Επομένως, μια οπτική η οποία βασίζεται στην ομοιότητα και η οποία εξετάζει την συνδεσιμότητα μεταξύ των μελών εντός μια ομάδα είναι εφικτή. Μέσω λοιπόν της Σχέσης (3.3) θα παρουσιάσουμε τις σημαντικότερες από τις μετρικές σύνδεσης (έχοντας πάντα στο μυαλό μας πως χρησιμοποιήθηκαν μέτρα ανομοιότητας για τον ορισμό της απόστασης μεταξύ δύο σημείων), καθώς και πως αυτές προκύπτουν από τη Σχέση (3.3) για διαφορετικές τιμές των παραμέτρων a_i, a_j, b και c.

- <u>Μέθοδος Μονής Σύνδεσης:</u>

Προκύπτει από τη σχέση (3.3) θέτοντας $a_i = {}^1/_2, a_j = {}^1/_2, b = 0, c = -{}^1/_2$. Σε αυτή την περίπτωση,

$$d(C_q, C_s) = \min\{d(C_i, C_s), d(C_j, C_s)\}$$

Δηλαδή, η απόσταση μεταξύ δύο ομάδων ορίζεται βάσει της απόστασης μεταξύ των δύο κοντινότερων στοιχείων των ομάδων αυτών



Εικόνα 10: Αναπαράσταση της μεθόδου Μονής Σύνδεσης

Άρα, $$D_{KL} = \min\big(d(x_i, y_i)\big), \forall\, i\epsilon C_K, j\epsilon C_L$$

Η ομαδοποίηση μονής σύνδεσης είναι γνωστή και ως ομαδοποίηση κοντινότερων γειτόνων. Έχει την τάση να κόβει τα άκρα των κατανομών πριν καταλήξει στο διαχωρισμό των κυρίως ομάδων ενώ λόγω των μη αυστηρών περιορισμών σχετικά με το σχήμα των ομάδων που εντοπίζει σαν μέθοδος, θυσιάζει τη δυνατότητα εντοπισμού συμπαγών ομάδων και ως αντάλλαγμα προσφέρει ανίχνευση επιμηκυμένων και ακανόνιστων ομάδων. Το κύριο αρνητικό όμως της μεθόδου αυτής είναι το φαινόμενο της αλυσίδας. Πιο συγκεκριμένα, υπάρχει περίπτωση κάποιες ομάδες να

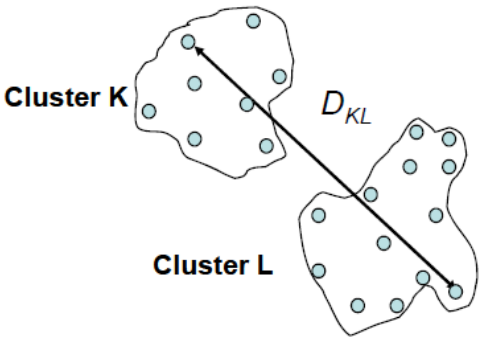
συγχωνεύονται λόγω του ότι έχουν ένα μικρό πλήθος των μελών τους πολύ κοντά ενώ οι κύριες μάζες των μελών τους είναι απομακρυσμένες. Θα λέγαμε επομένως πως το κριτήριο μονής σύνδεσης για συγχώνευση χαρακτηρίζεται ως τοπικό καθώς δίνεται προσοχή μόνο στην περιοχή όπου οι δύο ομάδες έρχονται πιο κοντά μεταξύ τους ενώ τα πιο απομακρυσμένα κομμάτια αλλά και η γενικότερη δομή των ομάδων δε λαμβάνονται υπόψη.

- <u>Μέθοδος Πλήρους Σύνδεσης:</u>

Προκύπτει από τη σχέση (3.3) θέτοντας $a_i = {}^1\!/_2 , a_j = {}^1\!/_2 , b = 0, c = {}^1\!/_2$. Σε αυτή την περίπτωση,

$$d(C_q, C_s) = \max\{d(C_i, C_s), d(C_j, C_s)\}$$

Δηλαδή, στη μέθοδο πλήρους σύνδεσης η ομοιότητα μεταξύ δύο ομάδων εξαρτάται από την ομοιότητα των πιο ανόμοιων μελών τους. Αυτό σημαίνει πως η απόσταση μεταξύ δύο ομάδων ορίζεται ως η απόσταση μεταξύ των δύο μακρινότερων μελών τους.



Εικόνα 11: Αναπαράσταση της μεθόδου Πλήρους Σύνδεσης

Άρα, $$D_{KL} = \max\big(d(x_i, y_i)\big), \forall\, i \epsilon C_K, j \epsilon C_L$$

Αυτό ισοδυναμεί με την επιλογή του ζεύγους ομάδων όπου δίνει ομάδα με τη μικρότερη διάμετρο έναντι των υπολοίπων συνδυασμών. Εν αντιθέσει με το κριτήριο μονής σύνδεσης, το κριτήριο αυτό δε χαρακτηρίζεται ως τοπικής εμβέλειας καθώς εδώ η συνολική δομή της ομαδοποίησης επηρεάζει τις αποφάσεις για συγχώνευση. Το αποτέλεσμα είναι η δημιουργία ομάδων αρκετά συμπαγών με μικρές διαμέτρους αντί για επιμηκυμένες και άτακτες ομάδες, κάτι όμως που προκαλεί μεγάλη ευαισθησία στην εμφάνιση ακραίων τιμών. Ένα μέλος το οποίο βρίσκεται αρκετά μακριά από το κέντρο

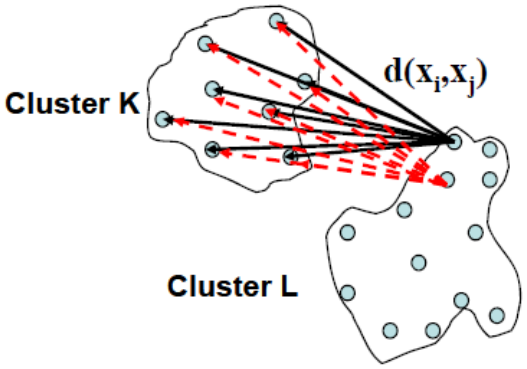
της ομάδας μπορεί να αυξήσει τη διάμετρο των υποψήφιων για συγχώνευση ομάδων δραματικά και να αλλάξει εντελώς τη τελική ομαδοποίηση. Τέλος, αξίζει να αναφερθεί πως σαν τρόπος σύνδεσης έχει την τάση να εντοπίζει ομάδες οι οποίες τελικά έχουν σχεδόν ίδιες διαμέτρους και μεταξύ τους απέχουν αρκετά.

- <u>Μέθοδος Μέσης Σύνδεσης:</u>

Προκύπτει από τη σχέση (3.3) θέτοντας $a_i = {}^1/_2, a_j = {}^1/_2, b = 0, c = 0$. Σε αυτή την περίπτωση,

$$d(C_q, C_s) = \frac{1}{2}(d(C_i, C_s) + d(C_j, C_s)\}$$

Η μέθοδος μέσης σύνδεσης αποτελεί ένα συμβιβασμό μεταξύ του προβλήματος της μεθόδου πλήρους σύνδεσης που αφορά τις ακραίες τιμές καθώς και του προβλήματος της μεθόδου της μονής σύνδεσης που έχει σχέση με την τοπικότητα του κριτηρίου όπως είδαμε παραπάνω. Η απόσταση μεταξύ δύο ομάδων εδώ, σε αντίθεση με τις δύο προηγούμενες μετρικές που παρουσιάστηκαν, δε λαμβάνει υπόψη της μόνο ένα σημείο από κάθε ομάδα. Αντιθέτως, η απόσταση μεταξύ των ομάδων ορίζεται ως ο μέσος όρων των αποστάσεων μεταξύ όλων των συνδυασμών των μελών των ομάδων ανά δύο, χωρίς να υπάρχουν διπλοί υπολογισμοί.



Εικόνα 12: Αναπαράσταση της μεθόδου Σύνδεσης Μέσου Όρου

Άρα, $$D_{KL} = \frac{1}{n_K n_L} \sum_{i \epsilon C_K} \sum_{j \epsilon C_L} d(x_i, x_j)$$

Ως μέθοδος τείνει να ενώνει ομάδες με μικρές διακυμάνσεις και θα μπορούσε να χαρακτηριστεί ως μεροληπτική ως προς τη δημιουργία ομάδων με παρόμοιες διακυμάνσεις. Λόγω του ότι για τον υπολογισμό της απόστασης των δύο ομάδων

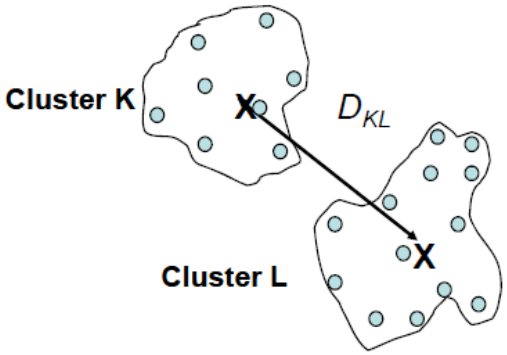
λαμβάνει υπόψη της όλα τα μέλη τους, αντί να χρησιμοποιεί μόνο ένα από κάθε ομάδα, δεν επηρεάζεται τόσο πολύ από τις ακραίες τιμές οι οποίες έχουν άμεσο αντίκτυπο στις μεθόδους μονής και πλήρους σύνδεσης. Εντούτοις, είναι μια υπολογιστικά ακριβή μέθοδος λόγω αυτού του χαρακτηριστικού της.

- <u>Μέθοδος Σύνδεσης Κεντροειδών:</u>

Θέτοντας στη σχέση (3.3) τις τιμές $a_i = \frac{n_i}{n_i+n_j}, a_j = \frac{n_j}{n_i+n_j}, b = -\frac{n_i n_j}{(n_i+n_j)^2}, c = 0$,όπου n_i και n_j είναι το πλήθος των μελών των ομάδων i και j αντίστοιχα, η απόσταση μεταξύ δύο ομάδων ορίζεται ως

$$d(C_q, C_s) = \frac{n_i}{n_i + n_j} d(C_i, C_s) + \frac{n_j}{n_i + n_j} d(C_j, C_s) - \frac{n_i n_j}{(n_i + n_j)^2} d(C_i, C_j)$$

Δηλαδή, η μέθοδος σύνδεσης κεντροειδών για να υπολογίσει την απόσταση μεταξύ δύο ομάδων, βρίσκει τη θέση των ιδεατών κεντροειδών των ομάδων αυτών και ως απόσταση μεταξύ των ομάδων ορίζει την απόσταση των κεντροειδών βάσει του μέτρου εγγύτητας που χρησιμοποιείται.



Εικόνα 13: Αναπαράσταση της μεθόδου Σύνδεσης Κεντροειδών

Άρα, $$D_{KL} = d(\overline{x_K}, \overline{x_L})$$

όπου $\overline{x_K}$ και $\overline{x_L}$ είναι τα κεντροειδή των ομάδων K και L τα οποία υπολογίζονται ως εξής:

$$\overline{x_K} = \frac{1}{n_K} \sum_{i=1}^{n_K} x_{K_i} \quad και \quad \overline{x_L} = \frac{1}{n_L} \sum_{j=1}^{n_L} x_{L_j}$$

Λόγω του ότι η μέθοδος αυτή συγκρίνει τα μέσα των ομάδων επηρεάζεται λιγότερο από κάθε άλλη μέθοδο σύνδεσης ιεραρχικής ομαδοποίησης από τις ακραίες τιμές.

Υπάρχουν όμως και περιπτώσεις όπου μπορεί να μην αποδώσει και τόσο καλά όσο η μέθοδος σύνδεσης Ward (η οποία δεν παρουσιάζεται) ή η μέθοδος μέσης σύνδεσης. Αξίζει να σημειωθεί πως αν υπάρχει συγχώνευση δύο ομάδων με διαφορετικά μεγέθη μέσω της μεθόδου σύνδεσης κεντροειδών στη νέα ομάδα που δημιουργείται υπάρχει η τάση να επικρατεί η ομάδα με το μεγαλύτερο μέγεθος που σημαίνει πως στο κεντροειδές της νέας ομάδας θα συνεισφέρουν περισσότερο τα μέλη της μεγαλύτερης ομάδας και έτσι αυτό θα είναι μετατοπισμένο προς το μέρος της.

Εν κατακλείδι, από τις μεθόδους σύνδεσης ομάδων που παρουσιάστηκαν, αυτή που τελικά ενσωματώθηκε στη προσέγγισή μας είναι η μέθοδος των κεντροειδών λόγω της απλότητας υπολογισμού της, των καλών αποτελεσμάτων που παράγει και επίσης για το λόγω του ότι, όπως θα δείξουμε στην επόμενη παράγραφο, μπορούμε να χρησιμοποιήσουμε τα κεντροειδή των ομάδων σε κάθε αποτέλεσμα ομαδοποίησης έτσι ώστε να βρεθεί ο καλύτερος διαχωρισμός των στοιχείων του συνόλου N σε ομάδες.

3.3 Κριτήριο Βέλτιστης Ομαδοποίησης

Το θετικό των ιεραρχικών αλγορίθμων είναι πως μπορούν να δώσουν μια εικόνα της σχέσης που έχουν οι ομάδες που δημιουργούνται μεταξύ τους. Δηλαδή, μπορεί κανείς κοιτώντας το δέντρο με τις συγχωνεύσεις ή τους διαχωρισμούς που έχουν γίνει (αναλόγως αν έχουμε συγχωνευτικό ή διαιρετικό ιεραρχικό αλγόριθμο ομαδοποίησης) να εξάγει πληροφορία σχετικά με το ποιες ομάδες απαρτίζουν μια μεγαλύτερη ομάδα (για τους συγχωνευτικούς) ή ανάλογα από ποια μεγαλύτερη ομάδα προήλθαν δύο νέες (για τους διαιρετικούς). Οι πιο διαδεδομένοι τρόποι για να διακοπεί ένας ιεραρχικός αλγόριθμος είναι οι εξής:

- Να προσδιοριστεί εξαρχής ο αριθμός των ομάδων που θέλουμε να έχουμε
- Να δοθεί ένα κατώφλι βάσει της μετρικής που χρησιμοποιείται για τον προσδιορισμό των αποστάσεων μεταξύ των ομάδων . Για παράδειγμα, να πούμε πως θέλουμε να σταματήσουμε τον συγχωνευτικό ιεραρχικό αλγόριθμο, όταν πάει να συγχωνεύσει σε μία ομάδα, ομάδες με απόσταση μεγαλύτερη ενός κατωφλίου d και να μας δώσει ως αποτέλεσμα την ομαδοποίηση που προέκυψε στο προηγούμενο βήμα του αλγορίθμου.
- Να δοθεί το πλήθος των βημάτων που θέλουμε να εκτελέσει ο αλγόριθμος. Με το που φτάσουμε στο βήμα που έχουμε ορίσει ως τελευταίο, ο αλγόριθμος

επιστρέφει ως ομαδοποίηση την κατάτμηση των δεδομένων που έχει εντοπίσει μέχρι εκείνη τη χρονική στιγμή.

Γίνεται κατανοητό όμως πως και για τα τρία παραπάνω κριτήρια θα πρέπει να υπάρξει παρέμβαση από το χρήστη για να δώσει αυτές τις τιμές. Αυτό προαπαιτεί ο χρήστης να γνωρίζει είτε το πλήθος των ομάδων που τελικά θέλει να εντοπίσει, είτε να ορίζει ένα κατώφλι απόστασης το οποίο θα δίνει ένα αποτέλεσμα που έχει νόημα βάσει του τι θέλει να εξάγει ως πληροφορία. Επειδή όμως θέλουμε να εντοπίσουμε χρήστες αυτόματα μέσα σε μια περιοχή από την οποία παίρνουμε πληροφορία για τη θέση τους μόνο και δεν έχουμε γνώση για το ποια απόσταση θα είχε νόημα μεταξύ των ομάδων σε αυτό τον χώρο καθώς και το πόσες ομάδες έχουμε, δεν μπορούμε να κάνουμε χρήση κανενός από τα παραπάνω κριτήρια. Για το λόγο αυτό, θα παρουσιάσουμε το κριτήριο το οποίο ενσωματώσαμε στον ιεραρχικό αλγόριθμο του συστήματός μας για την εύρεση της βέλτιστης ομαδοποίησης από αυτές που παράγει όσο εκτελείται. Το κριτήριο αυτό λαμβάνει υπόψη του τη συνοχή που έχουν οι ομάδες που δημιουργούνται σε κάθε βήμα σε συνάρτηση με το πόσο διαφέρουν οι ομάδες μεταξύ τους εξωτερικά. Διαισθητικά, για να ορίσουμε μια ομαδοποίηση ως βέλτιστη θα θέλαμε οι ομάδες που θα έχουμε μέσα στη περιοχή που εξετάζουμε να είναι αρκετά συνεκτικές εσωτερικά και να διαφέρουν αρκετά μεταξύ τους εξωτερικά.

Η ομαδοποίηση, όπως έχουμε ορίσει, είναι μια διαδικασία διαχωρισμού προτύπων σε ομάδες έτσι ώστε τα πρότυπα που ανήκουν στις ίδιες ομάδες να είναι «κοντινά» (όμοια) και όσο το δυνατόν πιο «μακρινά» (λιγότερο όμοια) μεταξύ διαφορετικών ομάδων. Αυτός ο διαχωρισμός των ομάδων μπορεί να οριστεί ως ένα αποτέλεσμα ομαδοποίησης. Η βέλτιστη ομαδοποίηση ορίζεται ως η έκβαση από την διαλογή μεταξύ όλων των δυνατών συνδυασμών ομαδοποίησης για τα πρότυπα που έχουμε και ουσιαστικά παρουσιάζει ένα αποτέλεσμα συσχετίσεων μεταξύ των προτύπων το οποίο έχει «περισσότερο» νόημα από τα υπόλοιπα. Αν και ο ορισμός του ποια είναι η ιδανική ομαδοποίηση για τα δεδομένα μας είναι αρκετά απλός διαισθητικά, η αποτίμηση ενός αποτελέσματος ομαδοποίησης είναι ένα θεμελιώδες και παράλληλα δύσκολο πρόβλημα. Ένας λόγος είναι πως η ομαδοποίηση θα πρέπει να πραγματοποιηθεί με πρότερη (a priori) γνώση της δομής του συνόλου των δεδομένων. Επιπροσθέτως, είναι αδύνατον να οριστεί ποια κατανομή των ομάδων είναι η καλύτερη για ένα σύνολο μετρήσεων (προτύπων) χωρίς κάποιο αντικειμενικό μέτρο για την βέλτιστη ομαδοποίηση. Πάνω σε αυτό το τομέα, έχουν υπάρξει πάρα πολλές προσπάθειες να

οριστεί μια μετρική για την εύρεση της βέλτιστης ομαδοποίησης. Παρόλα αυτά, λίγες από αυτές τις μετρικές είναι εύκολο να γίνουν κατανοητές τόσο σε μαθηματικό επίπεδο όσο και διαισθητικό [1, 2]. Αυτό έχει ως αποτέλεσμα οι εκατοντάδες συναρτήσεις που έχουν προταθεί στη βιβλιογραφία ως κριτήρια να είναι παραλλαγές του ίδιου κριτηρίου.

Ακόμα και αν δοθεί ένα αντικειμενικό κριτήριο, η δυσκολία του ορισμού της βέλτιστης ομαδοποίησης προέρχεται από τον εκπληκτικά μεγάλο αριθμό πιθανών συνδυασμών ομαδοποίησης. Το πλήθος των συνδυασμών για να παράγουμε k ομάδες από n πρότυπα που υπάρχουν στο σύνολο των μετρήσεών μας είναι ένας Stirling αριθμός δεύτερης τάξης [3]:

$$S_k^{(n)} = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{(k-i)}\binom{k}{i}i^n$$

Συνήθως, το μεγάλο πλήθος καθώς και οι πολλές διαστάσεις των προτύπων αυξάνουν την πολυπλοκότητα για την επίτευξη ενός αποδοτικού μέτρου για την ιδανική ομαδοποίηση. Επιπροσθέτως, είναι δύσκολη η επιλογή ενός κριτηρίου το οποίο να «μεταφράζεται» σε ένα λογικό διαισθητικά ορισμό του τι είναι ομάδα από μια μαθηματική φόρμουλα. Επομένως, από τη στιγμή που δεν υπήρξε κάποια βέλτιστη λύση για το πρόβλημα της ιδανικότερης ομαδοποίησης βάσει της πρόσφατης έρευνας που διεξήχθη, οι αλγόριθμοι που υπάρχουν στοχεύουν κυρίως στην αποδοτικότητα και την επεκτασιμότητα έτσι ώστε να μειώσουν το υπολογιστικό κόστος και να αυξήσουν την επεξεργαστική ικανότητα αντίστοιχα. Μπορεί να είναι πιθανό να παραχθεί ένα αποτέλεσμα ομαδοποίησης πολύ γρήγορα και να επεξεργαστεί μια τεράστια ποσότητα από δεδομένα κατευθείαν. Ωστόσο, συνήθως δεν υπάρχει καμία εγγύηση για την επίτευξη μιας βέλτιστης, ή έστω μιας κοντά στη βέλτιστη, λύσης για το σχηματισμό των ομάδων.

Έτσι, στο [18] προτείνεται μια μέθοδος για τη ποσοτική μέτρηση της βελτιστοποίησης της ομαδοποίησης με σκοπό να χρησιμοποιηθεί για το προσδιορισμό του βέλτιστου αριθμού ομάδων σε διάφορους αλγορίθμους ομαδοποίησης. Αυτό φυσικά έχει ιδιαίτερη απήχηση και στους ιεραρχικούς αλγορίθμους. Η συγκεκριμένη μέθοδος σχεδιάστηκε βασιζόμενη στην υπόθεση πως η βέλτιστη ομαδοποίηση μπορεί να ερμηνευτεί διαισθητικά αλλά και αντικειμενικά από έναν άνθρωπο. Από τη στιγμή που η διαισθητική επικύρωση της ποιότητας της ομαδοποίησης μεγιστοποιείται στο δισδιάστατο χώρο, είναι χρήσιμο να θεωρήσουμε ένα δισδιάστατο Ευκλείδειο χώρο έτσι ώστε να έχουμε μια όσο το δυνατόν καλύτερη αντικειμενικά απόφαση.

Για την ποσοτικοποίηση της έννοιας της βέλτιστης ομαδοποίησης, εισάγεται ο όρος του «**κέρδους ομαδοποίησης**» (clustering gain), το οποίο σαν μέτρο έχει σχεδιαστεί έτσι ώστε να παίρνει την μέγιστη τιμή του όταν η ομοιότητα εντός των ομάδων (intra-cluster) μεγιστοποιείται και αντιστοίχως η ομοιότητα μεταξύ των ομάδων (inter-cluster) ελαχιστοποιείται. Οπότε, ο βέλτιστος αριθμός των ομάδων για το τρέχον σύνολο μετρήσεων μπορεί να βρεθεί από το μέγιστο της καμπύλης του κέρδους ομαδοποίησης. Αυτό το μέτρο μπορεί να χρησιμοποιηθεί απευθείας για να βρεθεί μια βέλτιστη ομαδοποίηση σε όλους του ιεραρχικούς αλγορίθμους καθώς αυτοί προχωρούν. Επίσης, μπορεί να χρησιμοποιηθεί και ως μετρική σύγκρισης αποδοτικότητας μεταξύ αλγορίθμων ομαδοποίησης από τη στιγμή που η απόδοση της ομαδοποίησης αποτιμάται με το κέρδος ομαδοποίησης. Παρακάτω θα παρουσιαστεί πως ο επιθυμητός αριθμός των ομάδων μπορεί να εκτιμηθεί βάσει των δεδομένων και μόνο, χρησιμοποιώντας τη μετρική του κέρδους ομαδοποίησης. Βάσει των πειραμάτων που έχουν γίνει γι' αυτή τη μετρική, οι πιο διαδεδομένοι ιεραρχικοί αλγόριθμοι μπορούν να παράξουν πολύ λογικά αποτελέσματα ομαδοποίησης βασιζόμενοι σε αυτή για την εύρεση των ομάδων σε δεδομένα στα οποία πράγματι εμφανίζονται ομάδες καλά διαχωρίσιμες μεταξύ τους. Αυτό δεν είναι κάτι το οποίο αποτελεί εμπόδιο για εμάς καθώς οι ομάδες κινούμενων αντικειμένων (και δει ανθρώπων που κινούνται σε έναν χώρο με εκθέματα) είναι συνήθως καλά διαχωρίσιμες μεταξύ τους.

Γίνεται λοιπόν σαφές, πως με την βοήθεια της συγκεκριμένης μετρικής δεν χρειάζεται καμία παρέμβαση για τον εντοπισμό ομάδων κινητών χρηστών που βρίσκονται μέσα στον ίδιο χώρο. Αυτό επιτυγχάνεται με την εφαρμογή ενός συγχωνευτικού ιεραρχικού αλγορίθμου ο οποίος κατά τη διάρκεια που θα τρέχει θα υπολογίζει για κάθε ομαδοποίηση που παράγει και το αντίστοιχο κέρδος ομαδοποίησης. Στο υπόλοιπο αυτής της ενότητας θα δοθεί η μαθηματική αποτύπωση των όσων ειπώθηκαν προηγουμένως.

3.3.1 Ισορροπία Ομαδοποίησης (Clustering Balance)

Το πρόβλημα της ομαδοποίησης είναι ο διαχωρισμός της δοσμένης εισόδου από πρότυπα σε ένα συγκεκριμένο πλήθος ομάδων έτσι ώστε η ομοιότητα εντός των ομάδων να μεγιστοποιείται και αντιστοίχως η ομοιότητα μεταξύ των ομάδων να ελαχιστοποιείται σε ένα συγκεκριμένο μετρικό χώρο. Από δω και κάτω θα χρησιμοποιηθούν οι ακόλουθοι συμβολισμοί:

- Το πρότυπο i είναι ένα διάνυσμα χαρακτηριστικών στον m –διάστατο χώρο το οποίο συμβολίζεται ως:

$$p_i = [p_{i1}, p_{i2}, \dots \dots, p_{im}]^T$$

- Η ομάδα C_j είναι ένα σετ από πρότυπα τα οποία έχουν ορισθεί να ανήκουν στην ίδια ομάδα από κάποιον αλγόριθμο ομαδοποίησης και εκφράζεται ως:

$$C_j = \{p_1^{(j)}, p_2^{(j)}, \dots \dots, p_{n_j}^{(j)}\},$$

όπου n_j είναι ο αριθμός των προτύπων (κόμβων) που ανήκουν στην ομάδα C_j.

- Θα υποθέσουμε ότι συνολικά υπάρχουν n διανύσματα τα οποία θα πρέπει να ομαδοποιηθούν και πως ο συνολικός αριθμός των ομάδων είναι k. Επομένως:

$$\sum_{i=1}^{k} n_i = n$$

- Επιπροσθέτως, το $p_0^{(j)}$ συμβολίζει το κεντροειδές της ομάδας j το οποίο και ορίζεται ως εξής:

$$p_0^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} p_i^{(j)}$$

Το κεντροειδές, αποτελώντας το μέσο διάνυσμα της ομάδας, προσφέρει μια συμπιεσμένη αναπαράσταση της ομάδας σε μια απλούστερη μορφή ενώ αρκετά συχνά χρησιμοποιείται για συμπίεση των δεδομένων μιας ομάδας.

Το θέμα της βέλτιστης ομαδοποίησης είναι να βρεθεί μια μορφή ομαδοποίησης η οποία μεγιστοποιεί κάποιο κριτηρίο εκτίμησης της ποιότητας. Ωστόσο, όπως αναφέρθηκε και προηγουμένως, οι πιθανοί συνδυασμοί ομαδοποίησης n προτύπων σε k ομάδες είναι πάρα πολλοί και απαιτείται υψηλή πολυπλοκότητα για την εύρεση όλων. Στη πραγματικότητα, μια συνδυαστική αναζήτηση του συνόλου των πιθανών μορφών ομαδοποίησης είναι υπολογιστικά απαγορευτική και χαρακτηρίζεται ως NP-complete πρόβλημα. Οι συγχωνευτικοί ιεραρχικοί αλγόριθμοι που χρησιμοποιούνται σήμερα υιοθετούν μια προσεγγιστική λογική με το να συγχωνεύουν πιο όμοια πρότυπα (ή και ομάδες σε αργότερα βήματα) πριν την ομαδοποίηση λιγότερο όμοιων προτύπων έτσι ώστε να κατασκευαστεί μια ιεραρχία ομαδοποιήσεων. Αξίζει να αναφερθεί πως η

μετρική της ομοιότητας μεταξύ δύο προτύπων που προέρχονται από τον ίδιο χώρο χαρακτηριστικών παίζει πολύ σημαντικό ρόλο στις διαδικασίες ομαδοποίησης.

Η πιο δημοφιλής μετρική για τη μέτρηση της ομοιότητας μεταξύ προτύπων είναι η Ευκλείδεια απόσταση από τη στιγμή που είναι πιο κατανοητή διαισθητικά και εύκολα εφαρμόσιμη, ειδικά στο δισδιάστατο χώρο χαρακτηριστικών. Η πιο συχνά χρησιμοποιούμενη συνάρτηση κριτηρίου στις τεχνικές ομαδοποίησης είναι το κριτήριο του τετραγωνικού σφάλματος το οποίο ορίζεται ως το άθροισμα των τετραγωνικών αποστάσεων του κεντροειδούς μιας ομάδας από όλα τα μέλη-πρότυπα της ομάδας αυτής οι οποίες και μπορούν να υπολογιστούν με τη χρήση της Ευκλείδειας απόστασης.

Το συνολικό σφάλμα εντός των ομάδων, Λ, ορίζεται με τη βοήθεια του τετραγωνικού σφάλματος e ως:

$$\Lambda = \sum_{j=1}^{k} \sum_{i=1}^{n_j} e(p_i^{(j)}, p_0^{(j)})$$

το οποίο, με την χρήση της Ευκλείδειας απόστασης μπορεί εν τέλει να εκφραστεί ως:

$$\Lambda = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0^{(j)} \right\|^2$$

Το οποίο είναι γνωστό επίσης ως ενδο-ομαδικό σφάλμα.

Το συνολικό σφάλμα μεταξύ των ομάδων λαμβάνει υπόψη του τα σφάλματα μεταξύ των ομάδων θεωρώντας τη συλλογή των κεντροειδών των ομάδων ως ένα καθολικό πρότυπο το οποίο έχει επίσης ένα καθολικό κεντροειδές. Το συνολικό σφάλμα μεταξύ των ομάδων, στη περίπτωση του Ευκλείδειου χώρου, ορίζεται ως εξής:

$$\Gamma = \sum_{j=1}^{k} e(p_0^{(j)}, p_0) = \sum_{j=1}^{k} \left\| p_0^{(j)} - p_0 \right\|^2$$

όπου p_0 είναι το καθολικό κεντροειδές και ορίζεται ως εξής:

$$p_0 = \frac{1}{n} \sum_{i=1}^{n} p_i$$

Τώρα θα παρουσιαστούν κάποια χαρακτηριστικά από αυτά τα δύο αντικρουόμενα συνολικά σφάλματα, για να χρησιμοποιηθούν στο σχεδιασμό ενός μέτρου για την εύρεση μιας βέλτιστης ομαδοποίησης καθώς και ενός κριτηρίου διακοπής για τους ιεραρχικούς αλγορίθμους ομαδοποίησης. Υποθέτουμε πως ο ιεραρχικός αλγόριθμος

στον οποίο αναφερόμαστε είναι συγχωνευτικός (από κάτω προς τα πάνω). Στη περίπτωση των διαιρετικών (από πάνω προς τα κάτω) ανάλογες αλλά ακριβώς αντίθετες τάσεις μπορούν να αποδειχθούν. Όπως έχει παρουσιαστεί και προηγουμένως, στην αρχική φάση του συγχωνευτικού ιεραρχικού αλγορίθμου, κάθε πρότυπο-κόμβος αποτελεί μια ομάδα της οποίας μέλος είναι μόνο αυτός. Είναι ξεκάθαρο πως οι μονήρεις ομάδες δε συνεισφέρουν στο άθροισμα σφαλμάτων Λ εντός της ομάδας και πως η ελάχιστη τιμή που μπορεί να πάρει το Λ είναι μηδέν. Από την άλλη, το Λ μεγιστοποιείται όταν υπάρχει μόνο μια ομάδα η οποία και περιέχει όλα τα πρότυπα (κόμβους). Το πιο ενδιαφέρον γεγονός είναι πως καθώς ο αλγόριθμος ομαδοποίησης τρέχει, η τιμή της μετρικής Λ δε μειώνεται.

Έστω πως δύο ομάδες C_i και C_j συγχωνεύονται σε ένα από τα βήματα ενός συγχωνευτικού ιεραρχικού αλγορίθμου. Αν C_{ij} είναι η ομάδα που προκύπτει από τη συγχώνευση των ομάδων C_i και C_j τότε το κεντροειδές c_{ij} της ομάδας αυτής υπολογίζεται από την σχέση:

$$c_{ij} = \frac{n_i p_0^{(i)} + n_j p_0^{(j)}}{n_i + n_j}$$

Αν Λ_b και Λ_a είναι τα αθροίσματα σφάλματος εντός των ομάδων για τα στοιχεία τα οποία απαρτίζουν τις ομάδες C_i και C_j μόνο, πριν και μετά τη συγχώνευση αντιστοίχως, τότε:

$$\Lambda_b = \sum_{l=1}^{n_i} \left\| p_l^{(i)} - p_0^{(i)} \right\|^2 + \sum_{l=1}^{n_j} \left\| p_l^{(j)} - p_0^{(j)} \right\|^2$$

και

$$\Lambda_a = \sum_{l=1}^{n_i} \left\| p_l^{(i)} - c_{ij} \right\|^2 + \sum_{l=1}^{n_j} \left\| p_l^{(j)} - c_{ij} \right\|^2$$

Από τη στιγμή που κατά τη διάρκεια εκτέλεσης ενός συγχωνευτικού ιεραρχικού αλγορίθμου δεν υπάρχουν διασπάσεις ομάδων, η μετρική του αθροίσματος σφάλματος εντός της ομάδας δε φθίνει όσο η διαδικασία της ομαδοποίησης προχωρά αν $\Lambda_a - \Lambda_b \geq 0$. Έχουμε επομένως:

$$\Lambda_a - \Lambda_b = \sum_{l=1}^{n_i} \left\| p_l^{(i)} \right\|^2 - 2c_{ij}^T \sum_{l=1}^{n_i} p_l^{(i)} + n_i c_{ij}^T c_{ij} + \sum_{l=1}^{n_j} \left\| p_l^{(j)} \right\|^2 - 2c_{ij}^T \sum_{l=1}^{n_j} p_l^{(j)} + n_j c_{ij}^T c_{ij}$$

$$- \left[\sum_{l=1}^{n_i} \left\| p_l^{(i)} \right\|^2 - 2\left(p_0^{(i)}\right)^T \sum_{l=1}^{n_i} p_l^{(i)} + n_i \left\| p_0^{(i)} \right\|^2 + \sum_{l=1}^{n_j} \left\| p_l^{(j)} \right\|^2 - 2\left(p_0^{(j)}\right)^T \sum_{l=1}^{n_j} p_l^{(j)} \right.$$

$$\left. + n_j \left\| p_0^{(j)} \right\|^2 \right]$$

Εφόσον όμως $\sum_{l=1}^{n_i} p_l^{(i)} = n_i p_0^{(i)}$ και $\sum_{l=1}^{n_j} p_l^{(j)} = n_j p_0^{(j)}$, προκύπτει εν τέλει το επιθυμητό αποτέλεσμα:

$$\Lambda_a - \Lambda_b = 2n_i \left\| p_0^{(i)} \right\|^2 - n_i \left\| p_0^{(i)} \right\|^2 - 2n_i \left(p_0^{(i)}\right)^T c_{ij} + n_i \left\| c_{ij} \right\|^2$$

$$+2n_j \left\| p_0^{(j)} \right\|^2 - n_j \left\| p_0^{(j)} \right\|^2 - 2n_j \left(p_0^{(j)}\right)^T c_{ij} + n_j \left\| c_{ij} \right\|^2$$

$$= n_i \left\| p_0^{(i)} - c_{ij} \right\|^2 + n_j \left\| p_0^{(j)} - c_{ij} \right\|^2 \geq 0$$

Ομοίως, το συνολικό σφάλμα μεταξύ των ομάδων ικανοποιεί εκείνα τα χαρακτηριστικά που ακολουθούν την αντίθετη συλλογιστική από εκείνα για το συνολικό σφάλμα εντός των ομάδων. Πρέπει να σημειωθεί πως το καθολικό κεντροειδές (που αφορά όλο το σύνολο των δεδομένων) p_0 δεν αλλάζει κατά τη διάρκεια της ομαδοποίησης. Το άθροισμα των σφαλμάτων μεταξύ των ομάδων Γ μεγιστοποιείται όταν υπάρχουν n μονήρεις ομάδες, κάτι το οποίο συμβαίνει στην αρχή της διαδικασίας ομαδοποίησης. Επίσης, το Γ ελαχιστοποιείται όταν όλα τα πρότυπα n ανήκουν σε μία και μόνο ομάδα, κάτι το οποίο παρατηρείται πάντα στο τέλος ενός συγχωνευτικού ιεραρχικού αλγορίθμου. Είναι εύκολο να αποδειχτεί, όπως και με το Λ, πως η τιμή του Γ δεν αυξάνει όσο ο αλγόριθμος ομαδοποίησης προχωράει, με τη βοήθεια της Ευκλείδειας απόστασης.

Το προτεινόμενο σχήμα, βασίζεται στο γεγονός πως η μετρική συνολικού σφάλματος εντός των ομάδων δεν είναι φθίνουσα και η μετρική συνολικού σφάλματος μεταξύ των ομάδων δεν είναι αύξουσα όσο ο συγχωνευτικός ιεραρχικός αλγόριθμος προχωρά. Όταν ο αλγόριθμος ομαδοποίησης είναι διαιρετικός, τότε ισχύουν τα ακριβώς αντίθετα, ότι δηλαδή η μετρική συνολικού σφάλματος εντός των ομάδων δεν αυξάνει και η μετρική

συνολικού σφάλματος μεταξύ των ομάδων δεν φθίνει όσο ο διαιρετικός αλγόριθμος ομαδοποίησης προχωρά.

Μετατράπηκε επομένως το πρόβλημα βέλτιστης ομαδοποίησης σε ένα πρόβλημα εύρεσης εκείνου του σημείου όπου οι δύο παραπάνω ομοιότητες είναι ισορροπημένες, εκφράζοντάς τις μέσω του αθροίσματος του τετραγωνικού σφάλματος στον Ευκλείδειο χώρο. Ορίζεται λοιπόν ως **ισορροπία ομαδοποίησης**:

$$\mathcal{E}(\chi) = \alpha\Lambda + (1-\alpha)\Gamma$$

όπου, τα Λ και Γ αντιστοιχούν στο άθροισμα σφάλματος εντός και μεταξύ των ομάδων για μία συγκεκριμένη διάταξη ομαδοποίησης χ και $0 \leq \alpha \leq 1$ είναι ένα βαθμωτό μέγεθος το οποίο προσδιορίζει το βάρος μεταξύ αυτών των δύο αθροισμάτων. Η ισορροπία ομαδοποίησης $\mathcal{E}(\chi)$ διατυπώθηκε βάσει της ιδέας πως διαισθητικά η βέλτιστη ομαδοποίηση επιτυγχάνεται όταν αυτά τα δύο αθροίσματα σφαλμάτων έχουν φτάσει στο ισοζύγιό τους. Επικεντρωνόμαστε όμως στην ειδική περίπτωση όπου,

$$\alpha = {}^{1}/_{2}$$

το οποίο προσφέρει μια ισορροπία μεταξύ των σφαλμάτων και προκύπτει πλέον ότι:

$$\mathcal{E}(\chi) = \Lambda + \Gamma$$

Συμπερασματικά, η συμπεριφορά της ομαδοποίησης μπορεί να ερμηνευθεί ως μια διαδικασία αναζήτησης του καθολικού ελαχίστου για την τιμή της μετρικής της ισορροπία ομαδοποίησης. Με την ισορροπία της ομαδοποίησης να βασίζεται σε αυτά τα δύο συνολικά σφάλματα, αυτό μας προϊδεάζει πως θα χρησιμοποιηθεί το κέρδος μεταξύ της ισορροπίας εντός και μεταξύ των ομάδων έτσι ώστε να οριστεί ένα μέτρο για το βέλτιστο αποτέλεσμα ομαδοποίησης.

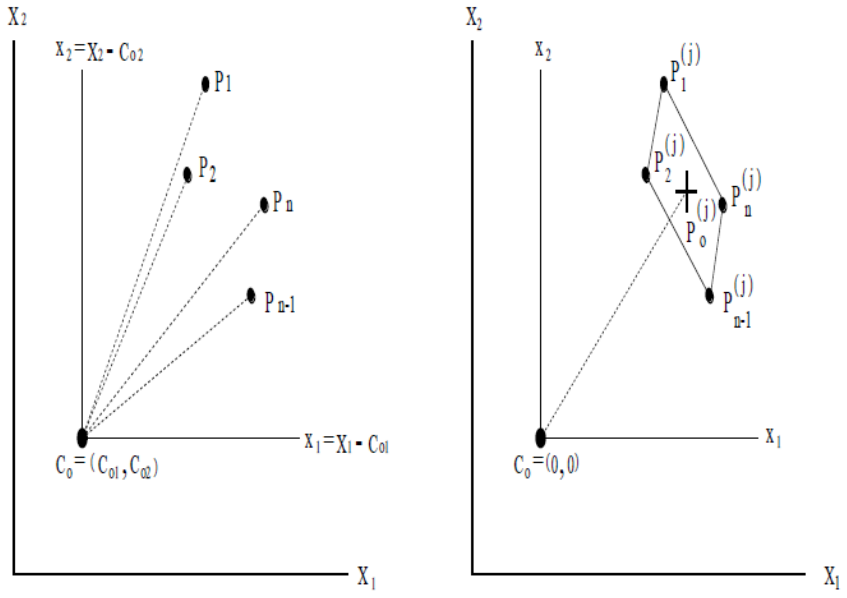
3.3.2 Κέρδος Ομαδοποίησης (Clustering Gain)

Η ισορροπία ομαδοποίησης πρέπει να υπολογίζεται σε κάθε βήμα ενός ιεραρχικού αλγορίθμου έτσι ώστε να προσδιορίσει το βέλτιστο αριθμό των ομάδων. Εντούτοις, ένα σημαντικό μειονέκτημα είναι το υψηλό κόστος για τον υπολογισμό αυτής της μετρικής. Έτσι λοιπόν, παρουσιάζεται το **κέρδος ομαδοποίησης** το οποίο έχει μια ενδιαφέρουσα συσχέτιση με την ισορροπία ομαδοποίησης. Επιπροσθέτως, το κέρδος ομαδοποίησης είναι αρκετά φθηνό υπολογιστικά. Οπότε, μπορεί να υπολογίζεται σε κάθε βήμα της διαδικασίας ομαδοποίησης για τον προσδιορισμό του ιδανικότερου αριθμού ομάδων, χωρίς να αυξάνει την υπολογιστική πολυπλοκότητα του αλγορίθμου ομαδοποίησης.

Το κέρδος ομαδοποίησης Δ_j για την ομάδα C_j ορίζεται ως η διαφορά μεταξύ του μειωμένου αθροίσματος σφάλματος μεταξύ των ομάδων γ_j, συγκριτικά με την αρχική κατάσταση, και του αυξημένου αθροίσματος σφάλματος εντός των ομάδων λ_j ,συγκριτικά με την αρχική κατάσταση. Πιο συγκεκριμένα, το κέρδος ομαδοποίησης για την ομάδα C_j ορίζεται ως:

$$\Delta_j = \gamma_j - \lambda_j$$

Στην παραπάνω εξίσωση, ένας παράγοντας βάρους ίσος με ένα, έχει δοθεί και στα δύο αθροίσματα σφαλμάτων. Το κέρδος ομαδοποίησης απεικονίζεται γραφικά στην ακόλουθη εικόνα:



(α) Χωρίς κάποια ομαδοποίηση (b) Η ομαδοποίηση έχει ολοκληρωθεί

Εικόνα 14: Το κέρδος ομαδοποίησης ορίζεται ως η διαφορά μεταξύ των αθροισμάτων σφάλματος. (a) Αρχική κατάσταση. (b) Τελική κατάσταση της ομάδας C_j

Γενικά, ο αριθμός των προτύπων (κόμβων) της τελική διάταξης/μορφής της ομάδας C_j μπορεί να κινείται από το 1 μέχρι και το n. Η μειωμένη τιμή του αθροίσματος σφάλματος μεταξύ των ομάδων συγκριτικά με την αρχική κατάσταση προσδιορίζεται από τη σχέση:

$$\gamma_j = \sum_{i=1}^{n_j} e\left(p_i^{(j)}, p_0\right) - e\left(p_0^{(j)}, p_0\right)$$

$$= \sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0 \right\|^2 - \left\| p_0^{(j)} - p_0 \right\|^2$$

Ομοίως, η αυξημένη τιμή του αθροίσματος σφάλματος εντός των ομάδων συγκριτικά με την αρχική κατάσταση προσδιορίζει από τη σχέση:

$$\lambda_j = \sum_{i=1}^{n_j} e\left(p_i^{(j)}, p_0^{(j)}\right) = \lambda_j = \sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0^{(j)} \right\|^2$$

Επεκτείνοντας την έκφραση για τον προσδιορισμού του κέρδους για την ομάδα C_j έχουμε επομένως:

$$\begin{aligned} \Delta_j &= \gamma_j - \lambda_j \\ &= \sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0 \right\|^2 - \left\| p_i^{(j)} - p_0 \right\|^2 - \sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0^{(j)} \right\|^2 \\ &= \left(n_j - 1\right) \left\| p_0 - p_0^{(j)} \right\|^2 \end{aligned}$$

εφόσον $\sum_{i=1}^{n_j} p_i^{(j)} = p_0^{(j)} n_j$.

Βάσει των παραπάνω, το ολικό κέρδος ομαδοποίησης για το τρέχον βήμα του αλγορίθμου υπολογίζεται από την ακόλουθη σχέση:

$$\boldsymbol{\Delta = \sum_{j=1}^{k} \left(n_j - 1 \right) \left\| p_0 - p_0^{(j)} \right\|^2} \qquad \textbf{(3.4)}$$

Πρέπει να δοθεί ιδιαίτερη έμφαση στο γεγονός πως το κέρδος ομαδοποίησης είναι πολύ φθηνό υπολογιστικά από τη στιγμή που εμπλέκει μόνο τα κεντροειδή των ομάδων και το καθολικό κεντροειδές και όχι τα στοιχεία των δεδομένων ένα προς ένα, όπως φαίνεται και στη Σχέση (3.4). Το κέρδος ομαδοποίησης Δ_j είναι πάντοτε μεγαλύτερο ή το πολύ ίσο με το μηδέν. Επομένως, το συνολικό κέρδος ομαδοποίησης για κάθε βήμα του αλγορίθμου θα είναι πάντα θετικό, δεδομένου πως δεν έχουμε ως βέλτιστη ομαδοποίηση την αρχική κατάσταση όπου έχουμε μονήρεις ομάδες και το Δ_j τότε είναι ίσο με το μηδέν.

Γίνεται επομένως εύκολα αντιληπτό πως η βέλτιστη σύνθεση ομαδοποίησης που εντοπίζεται από έναν ιεραρχικό αλγόριθμο ομαδοποίησης έχει τη μέγιστη τιμή για το κέρδος ομαδοποίησης, από εκείνες που προέκυψαν κατά το τρέξιμο του αλγορίθμου για άλλες συνθέσεις. Από τη στιγμή που το κέρδος ομαδοποίησης είναι ελάχιστο στο αρχικό (όλες οι ομάδες είναι μονήρεις) και στο τελικό (έχουμε μια ομάδα που περιέχει όλα τα πρότυπα) βήμα του αλγορίθμου ομαδοποίησης, η βέλτιστη σύνθεση ομαδοποίησης θα πρέπει να εντοπιστεί στα ενδιάμεσα βήματα της διαδικασίας. Για να προσδιοριστεί το μέγιστο κέρδος ομαδοποίησης κατά τη διάρκεια των ενδιάμεσων βημάτων του

αλγορίθμου, προτείνεται το κέρδος ομαδοποίησης ως ένα αποδοτικό κριτήριο. Αξίζει να αναφερθεί πως το κέρδος ομαδοποίησης είναι ανάλογο στη μετρική E που προτάθηκε στο [4], για μέτρηση της αποδοτικότητας μιας σύνθεσης ομαδοποίησης.

Ένα ακόμα ενδιαφέρον κομμάτι είναι πως το άθροισμα των μετρικών ισορροπίας ομαδοποίησης και κέρδους ομαδοποίησης είναι μια σταθερή τιμή για κάθε σύνολο δεδομένων από τη στιγμή που:

$$\Omega = \mathcal{E} + \Delta = (\Lambda + \Gamma) + \Delta$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0^{(j)} \right\|^2 + \sum_{j=1}^{k} \left\| p_0^{(j)} - p_0 \right\|^2$$

$$+ \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0 \right\|^2 - \left\| p_0^{(j)} - p_0 \right\|^2 - \sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0^{(j)} \right\|^2 \right)$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left\| p_i^{(j)} - p_0 \right\|^2$$

το οποίο και προσδιορίζεται βασιζόμενο εξολοκλήρου πάνω στα δεδομένα και δεν αλλάζει βάσει του αποτελέσματος της ομαδοποίησης. Για το λόγο αυτό, η ισορροπία ομαδοποίησης μπορεί εναλλακτικά να εκφραστεί μέσω του κέρδους ομαδοποίησης ως εξής:

$$\mathcal{E} = \Lambda + \Gamma = \Omega - \Delta$$

όπου $0 \leq \Lambda, \Gamma, \Delta \leq \Omega$. Πλέον, είμαστε σε θέση να βρούμε τη βέλτιστη σύνθεση ομαδοποίησης καταγράφοντας τις τιμές του κέρδους ομαδοποίησης αντί της ισορροπίας ομαδοποίησης.

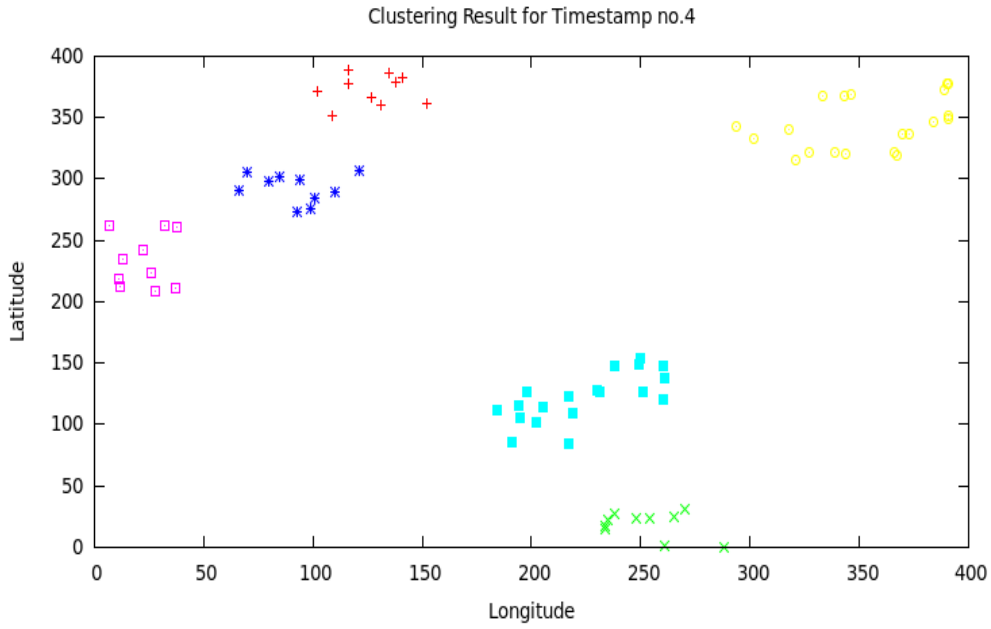
Στον *Αλγόριθμο Εντοπισμού* (Tracking Algorithm) που ακολουθεί συνοψίζεται η διαδικασία για το πως μπορεί να εντοπιστεί η βέλτιστη σύνθεση ομαδοποίησης, έχοντας ένα συγχωνευτικό ιεραρχικό αλγόριθμο, μέσω της παρακολούθησης της τιμή του κέρδους ομαδοποίησης. Πρέπει να κρατούνται οι τιμές του κέρδους ομαδοποίησης $\Delta(t)$ για κάθε βήμα του αλγορίθμου καθώς η μέγιστη τιμή για το κέρδος ομαδοποίησης μπορεί να εντοπιστεί μόνο εφόσον η διαδικασία του αλγορίθμου ομαδοποίησης ολοκληρωθεί. Δηλαδή, ως αρνητικό για τον εντοπισμό της βέλτιστης ομαδοποίησης από τους ιεραρχικούς αλγορίθμους μπορούμε να θεωρήσουμε το γεγονός πως θα πρέπει να τρέξει ο αλγόριθμος μέχρι τέλους έτσι ώστε να αποφανθεί το σύστημα για την καλύτερη

ομαδοποίηση. Φυσικά, το θετικό είναι πως πλέον υπάρχει ένας αυτόματος τρόπος να εντοπίζονται οι ομάδες μέσα σε μια περιοχή παρακολούθησης.

Αλγόριθμος Εντοπισμού (Tracking Algorithm)

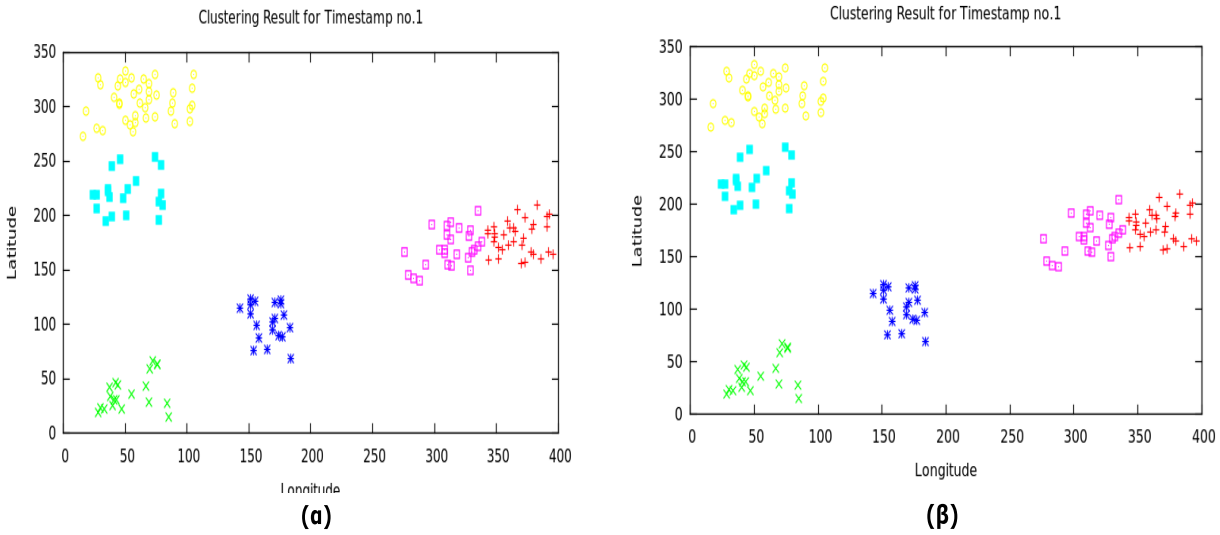
- Αρχικοποίηση:
  - Επιλογή του συγχωνευτικού ιεραρχικού αλγορίθμου που θα χρησιμοποιηθεί
  - $t = 0$
- Επανέλαβε:
  - $t = t + 1$
  - Εντόπισε και συγχώνευσε τις δύο κοντινότερες ομάδες βάσει του μέτρου σύνδεσης
  - Υπολόγισε την τιμή του κέρδους ομαδοποίησης $\Delta(t)$
- Μέχρι να ανήκουν όλα τα στοιχεία του συνόλου N σε μία ομάδα
- Εντόπισε και επέστρεψε το επίπεδο t^* των βημάτων του αλγορίθμου για το οποίο έχουμε το $\max_{1 \le t \le N-1}\{\Delta(t)\}$

Ένα αποτέλεσμα της εφαρμογής της παραπάνω μεθόδου, για μια χρονική στιγμή από τη περίοδο της παρακολούθησης που έχουμε τις θέσεις των χρηστών στο χώρο εποπτείας μας, μπορεί να φανεί στην Εικόνα 15.



Εικόνα 15: Αποτέλεσμα εφαρμογής του αλγορίθμου εύρεσης της βέλτιστης ομαδοποίησης για ένα σύνολο δεδομένων

Στη συγκεκριμένη εικόνα, στον άξονα των x αναπαρίστανται οι τιμές για το γεωγραφικό μήκος της θέσης του κάθε χρήστη ενώ στον άξονα των y οι τιμές για το γεωγραφικό πλάτος του. Παρατηρούμε πως με την ενσωμάτωση του υπολογισμού του κέρδους ομαδοποίησης σε κάθε βήμα του συγχωνευτικού ιεραρχικού αλγορίθμου και έπειτα με τη χρήση του αλγορίθμου εντοπισμού της μεγαλύτερης τιμής του «Δ» πράγματι εντοπίζονται αυτόματα, ομάδες στο χώρο εποπτείας οι οποίες είναι πολύ «λογικές». Φαίνεται και οπτικά πως επιτυγχάνεται με την χρήση αυτού του μέτρου ο ορισμός της «καλής» ομαδοποίησης. Δηλαδή εντοπίζει ομάδες στο χώρο που είναι αρκετά συνεκτικές εσωτερικά και ταυτόχρονα διαφέρουν πολύ μεταξύ τους έτσι ώστε να είναι πολύ καλά διαχωρίσιμες στο χώρο. Εφόσον πλέον έχουμε ένα αποδοτικό μέτρο εντοπισμού της βέλτιστης ομαδοποίησης των συγχωνευτικών ιεραρχικών αλγορίθμων, είμαστε σε θέση να παρουσιάσουμε και ένα άλλο θετικό της συγκεκριμένης κατηγορίας αλγορίθμων. Πιο συγκεκριμένα, σε αντίθεση με τους αλγορίθμους κατάτμησης, οι ιεραρχικοί αλγόριθμοι παράγουν την ίδια ιεραρχία ομαδοποιήσεων, ανεξαρτήτου της σειράς με την οποία τους δίνονται τα δεδομένα. Ένας πολύ γνωστός αλγόριθμος από τους αλγορίθμους κατάτμησης που «πάσχει» από το συγκεκριμένο πρόβλημα είναι ο $k-means$, όπου χρησιμοποιώντας τους ίδιους αρχικούς εκπροσώπους για τις ομάδες που πρέπει να εντοπίσει και δίνοντάς του με διαφορετική σειρά τα δεδομένα, παράγει εντελώς διαφορετικά αποτελέσματα. Αντιθέτως, σε δοκιμές που έγιναν, με διαφορετική σειρά των στοιχείων, στα σύνολα των δεδομένων που χρησιμοποιήθηκαν για την αξιολόγηση του συστήματός μας και θα παρουσιαστούν στο Κεφάλαιο 4, παρατηρήθηκε πως το προτεινόμενο σχήμα αυτόματης ομαδοποίησης δίνει τα ίδια ακριβώς αποτελέσματα, όπως φαίνεται και στην εικόνα που ακολουθεί. Το αποτέλεσμα (α) της Εικόνας 16 φαίνεται πως είναι ακριβώς ίδιο με το (β), κάτι το οποίο επαληθεύει τις παρατηρήσεις που έγιναν παραπάνω σχετικά με την ανεπηρέαστη απόδοση του αλγορίθμου για διαφορετική σειρά παρουσίαση των δεδομένων.



Εικόνα 16: Αποτέλεσμα ομαδοποίησης του ίδιου συνόλου δεδομένων δίνοντας με διαφορετική σειρά τα στοιχεία

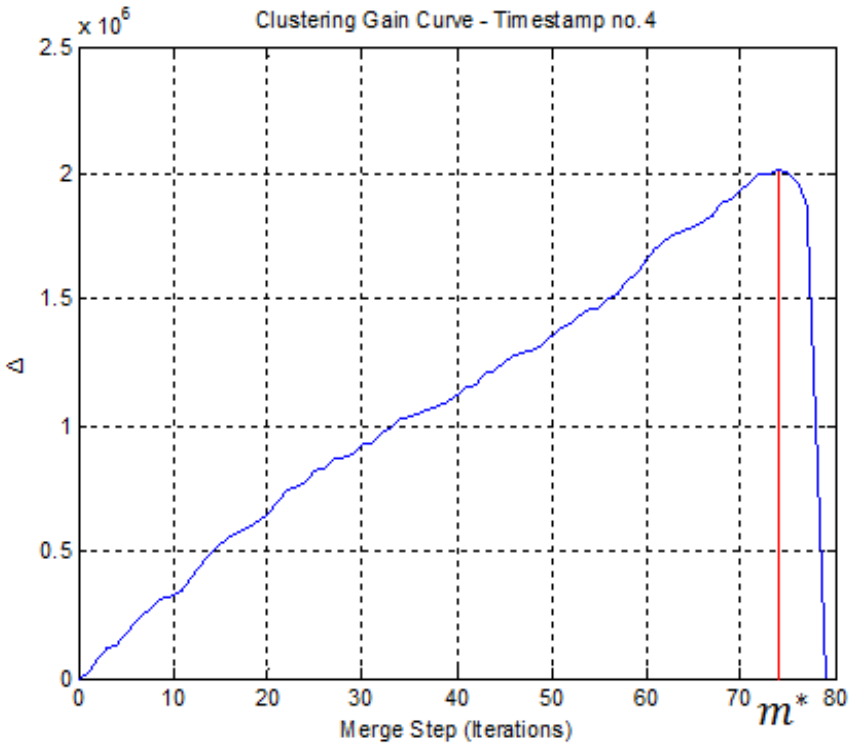
Ας δούμε σε αυτό το σημείο στην Εικόνα 17 και το διάγραμμα της μεταβολής των τιμών του κέρδους ομαδοποίησης για κάθε βήμα του αλγορίθμου έτσι ώστε να κατανοήσουμε καλύτερα τα όσα έχουν αποτυπωθεί με μαθηματικούς τύπους και ορισμούς νωρίτερα. Σε αυτό το διάγραμμα, στον άξονα των x είναι τα βήμα του συγχωνευτικού ιεραρχικού αλγορίθμου ενώ στον άξονα των y έχουμε τις τιμές του κέρδους ομαδοποίησης για κάθε βήμα συγχώνευσης. Με m^* συμβολίζεται το βήμα του αλγορίθμου που μας δίνει τη μεγαλύτερη τιμή του «Δ», δηλαδή:

$$m^* = \underset{1 \le m \le N-1}{\operatorname{argmax}} \{\Delta_m\}$$

και είναι το βήμα στο οποίο βάσει του συγκεκριμένου κριτηρίου έχουμε και τη καλύτερη ομαδοποίηση. Η κόκκινη γραμμή στο διάγραμμα δείχνει σε ποιο βήμα συγχώνευσης του συγχωνευτικού ιεραρχικού αλγορίθμου εντοπίστηκε η τιμή αυτή και προκύπτει εύκολα πως

$$Cluster\ \# = N - m^*$$

όπου "$Cluster\ \#$" είναι το σύνολο των ομάδων της βέλτιστης ομαδοποίησης και N το πλήθος των χρηστών που παρακολουθεί το σύστημα στο χώρο εποπτείας του. Η συγκεκριμένη καμπύλη είναι αυτή που εξετάστηκε για να πάρουμε το αποτέλεσμα της Εικόνας 15. Βλέπουμε επομένως πως πράγματι το πλήθος των ομάδων στη καλύτερη ομαδοποίηση είναι 6 καθώς $N = 80$ και $m^* = 74$. Επίσης, φαίνεται πως η τιμή του κέρδους ομαδοποίησης είναι μηδέν όταν το σύνολο των ομάδων είναι ίσο με N(βήμα 0) ή 1(βήμα $N - 1$).



Εικόνα 17: Καμπύλη μεταβολής του κέρδους ομαδοποίησης «Δ» σε σχέση με τα βήματα συγχωνεύσεων του συγχωνευτικού ιεραρχικού αλγορίθμου

Η εφαρμογή του συγκεκριμένου μέτρου σε συγχωνευτικούς και διαιρετικούς ιεραρχικούς αλγορίθμους είναι αρκετά απλή και γι' αυτό το λόγο επιλέχθηκε και για την περίπτωση που εξετάζουμε. Με την επιλογή μάλιστα και ως μεθόδου ορισμού της σύνδεσης μεταξύ των ομάδων, τη μέθοδο σύνδεσης κεντροειδών, ο μόνος επιπλέον υπολογισμός που χρειάζεται είναι η εύρεση του καθολικού κεντροειδούς του συνόλου των δεδομένων, κάτι το οποίο γίνεται μόνο μια φορά στην αρχή του αλγορίθμου, καθώς και ο υπολογισμός των Ευκλείδειων αποστάσεων των κεντροειδών των ομάδων από το καθολικό κεντροειδές. Το μόνο αρνητικό που έχει σα μεθοδολογία είναι πως για να βρεθεί η βέλτιστη τιμή του κέρδους ομαδοποίησης, άρα και η λογικότερη ομαδοποίηση, απαιτείται το εξαντλητικό τρέξιμο του ιεραρχικού αλγορίθμου. Εντούτοις, ακόμα κι έτσι, από άποψη πολυπλοκότητας είναι πολύ αποδοτικότερη σαν μέθοδος από το να έπρεπε να παράξουμε όλους τους δυνατούς συνδυασμούς ομαδοποίησης n προτύπων σε k ομάδες έτσι ώστε να βρεθεί η καλύτερη, όπως παρουσιάστηκε και στην αρχή αυτού του κεφαλαίου. Συμπερασματικά, μπορούμε να πούμε πως έχουμε ένα αρκετά αποδοτικό μέτρο για τον εντοπισμό της καλύτερης ομαδοποίησης από την αλληλουχία των ομαδοποιήσεων που προκύπτει με το τρέξιμο ενός ιεραρχικού αλγορίθμου.

3.4 Μέθοδος Κατάδειξης «Ύποπτων» Ομάδων

Μέχρι αυτό το σημείο, έχουμε ένα σύστημα το οποίο με τη βοήθεια ενός συγχωνευτικού ιεραρχικού αλγορίθμου καθώς και της μετρικής του κέρδους ομαδοποίησης είναι σε θέση να μας δώσει αυτόματα την κατάτμηση των χρηστών στο χώρο, αν γνωρίζουμε τη θέση τους κάθε χρονική στιγμή. Με αυτό το τρόπο, θα μπορούσε η υπηρεσία, εφόσον οριστούν εκπρόσωποι των ομάδων που εντοπίστηκαν, αντί να στέλνει τη πληροφορία σε όλους τους χρήστες, να τη στέλνει στους εκπροσώπους των ομάδων και εκείνοι με τη σειρά τους να αναλαμβάνουν την αποστολή της πληροφορίας εσωτερικά στην ομάδα που εκπροσωπούν. Έτσι λοιπόν, κάθε χρονική στιγμή, όλοι οι χρήστες θα ζητούσαν πληροφορία άμεσα συσχετισμένη με την θέση τους από το κεντρικοποιημένο σύστημα και αυτό αναγνωρίζοντας τη θέση τους και εφαρμόζοντας τον αλγόριθμο που έχει παρουσιαστεί μέχρι στιγμής θα έστελνε το μήνυμα με τη σχετική πληροφορία στους εκπροσώπους των ομάδων, κερδίζοντας έτσι μεγάλο ποσοστό στο φόρτο του δικτύου. Όπως έχει αναφερθεί και προηγουμένως όμως, το να εκτελείται αυτός ο αλγόριθμος κάθε χρονική στιγμή είναι υπολογιστικά ασύμφορο καθώς έχει μεγάλη πολυπλοκότητα. Το θετικό του είναι πως μπορεί να εντοπίζει αυτόματα τις ομάδες, αλλά δεν προτείνεται για να χρησιμοποιείται σε κάθε χρονική στιγμή.

Εδώ, επομένως, πρέπει να εισάγουμε και την έννοια της παρακολούθησης κινούμενων ομάδων που αναφέραμε στο εισαγωγικό κεφάλαιο αυτής της μελέτης. Πιο συγκεκριμένα, θα ήταν αποδοτικότερο, αν εφόσον κάποια χρονική στιγμή εκτελούσαμε τον παραπάνω αλγόριθμο, βρίσκαμε μια μέθοδο έτσι ώστε να παρακολουθούνται με ένα μεγάλο βαθμό βεβαιότητας οι ομάδες που έχουν προκύψει για κάποια περίοδο, από το να έχουμε ομαδοποίηση σε κάθε χρονική στιγμή. Με αυτό τον τρόπο θα είχαμε και μείωση στο φόρτο του δικτύου όσον αφορά την αποστολή μηνυμάτων αλλά και αιτημάτων από τους χρήστες, καθώς και μια περίοδο στην οποία δεν χρειάζεται να εκτελείται ο «βαρύς» υπολογιστικά αλγόριθμος ομαδοποίησης. Αποφασίζοντας λοιπόν να ακολουθήσουμε τη συγκεκριμένη προσέγγιση έπρεπε να εξετάσουμε το γεγονός του να χρειάζεται κάποιες από τις ομάδες που προέκυψαν να παρακολουθούνται ολόκληρες, ενώ κάποιες άλλες να ελέγχονται μόνο βάσει των εκπροσώπων τους.

Στην εισαγωγή για τους ιεραρχικούς αλγορίθμους και πιο συγκεκριμένα για τους συγχωνευτικούς ιεραρχικούς αλγορίθμους αναφέραμε πως σε αυτές τις μεθόδους ομαδοποίησης όποια σύνδεση μεταξύ ομάδων πραγματοποιηθεί στα βήματα των συγχωνεύσεων δεν μπορεί να αναιρεθεί. Αυτό σημαίνει πως αν αυτή η σύνδεση είναι «κακής» ποιότητας τότε αυτό θα επηρεάσει και το τελικό αποτέλεσμα ομαδοποίησης.

Έπρεπε επομένως να βρεθεί ένας τρόπος κατάδειξης αυτών των «ύποπτων» συγχωνεύσεων στη ομαδοποίηση που μας δίνει ως αποτέλεσμα ο αλγόριθμος που έχει παρουσιαστεί μέχρι στιγμής. Σε αυτό βοήθησε ιδιαίτερα η ιδιότητα της εμφώλευσης των συγχωνευτικών ιεραρχικών αλγορίθμων βάσει της οποίας αν δύο ομάδες συγχωνευτούν σε μια νέα στο επίπεδο t των βημάτων του αλγορίθμου, θα παραμείνουν σε αυτή την ομάδα μέχρι το τέλος της ιεραρχίας των συγχωνεύσεων. Δηλαδή, αν βρεθεί ένας τρόπος να εντοπίζονται αυτές οι συγχωνεύσεις που οδηγούν σε χειρότερο αποτέλεσμα, είναι εύκολο να καταδειχτεί η ομάδα που τελικά τις περιέχει.

Στην ενότητα 3.3 είδαμε έναν αλγόριθμο για τον εντοπισμό της καλύτερης ομαδοποίησης για το σύνολο των δεδομένων μας, μέσα από την ιεραρχία των ομαδοποιήσεων που δίνει ένας συγχωνευτικός ιεραρχικός αλγόριθμος. Ως μέτρο ελέγχου της ποιότητας μιας ομαδοποίησης ορίστηκε το κέρδος ομαδοποίησης το οποίο υπολογίζεται από τη σχέση (3.3)

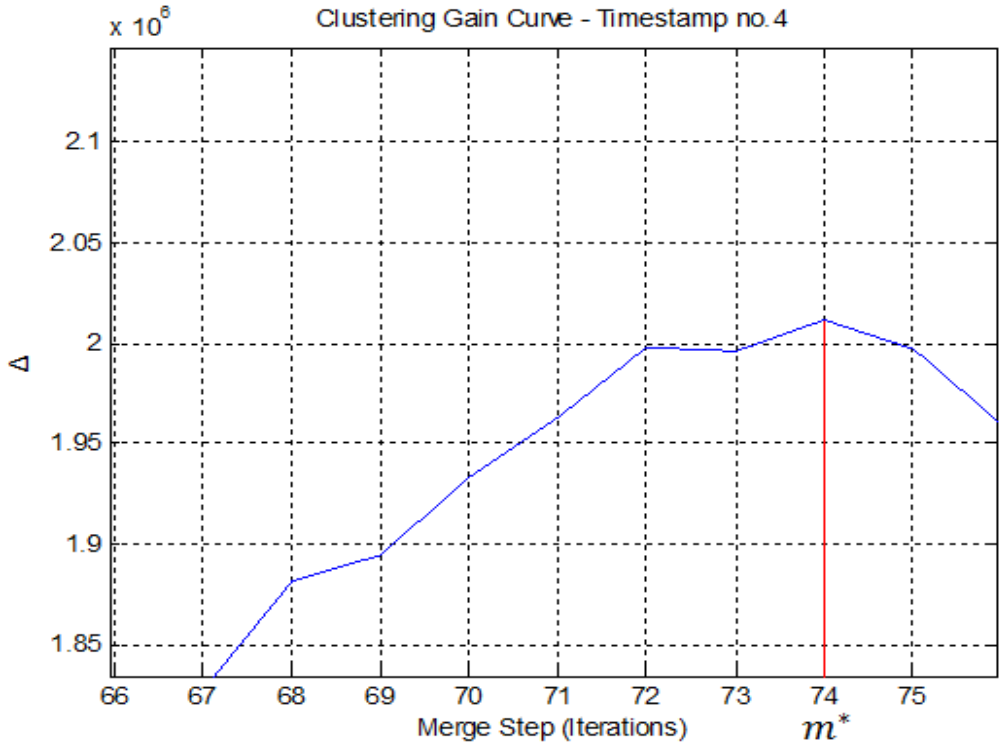
$$\Delta = \sum_{j=1}^{k} (n_j - 1) \left\| p_0 - p_0^{(j)} \right\|^2$$

Το κέρδος ομαδοποίησης δίνει μια τιμή για το πόσο μοιάζουν οι ομάδες που έχουν προκύψει από την ομαδοποίηση του τρέχοντος βήματος του αλγορίθμου καθώς και το πόσο διαφέρουν μεταξύ τους. Δηλαδή, όσο πιο μεγάλη είναι η συνολική εσωτερική ομοιότητα των ομάδων και η εξωτερική ανομοιότητα μεταξύ τους, τόσο μεγαλύτερη είναι και η τιμή του κέρδους ομαδοποίησης «Δ». Προκύπτει και από τον υπολογισμό του επομένως πως πρόκειται για ένα μέτρο ποιότητας, για το λόγο αυτό οι τιμές που μπορεί να πάρει είναι πάντοτε μη αρνητικές. Μπορεί επομένως να έχει τιμή είτε θετική, είτε ίση με το μηδέν, όπου για την περίπτωση των συγχωνευτικών ιεραρχικών αλγορίθμων αυτό συμβαίνει μόνο πριν ξεκινήσουν τα βήματα των συγχωνεύσεων όπου έχουμε N ομάδες με ένα μέλος η κάθε μία καθώς και στο τελευταίο επίπεδο των συγχωνεύσεων ($t = N - 1$) όπου έχουμε μια ομάδα με N μέλη.

Βάσει λοιπόν των παραπάνω ορισμών σκεφτήκαμε πως θα είχε νόημα εκτός από το να υπολογίζουμε την τιμή του κέρδους ομαδοποίησης, να ελέγχουμε και τις διαφορές μεταξύ των τιμών του μέτρου αυτού μεταξύ δύο διαδοχικών ομαδοποιήσεων του αλγορίθμου. Με αυτό τον τρόπο δηλαδή, θα έχουμε πλέον ένα «εργαλείο» για την εύρεση των κακών συγχωνεύσεων. Διαισθητικά, μια κακή συγχώνευση σημαίνει πως προέκυψε μια ομάδα που στο προσεχές μέλλον έχει μεγάλη πιθανότητα να διασπαστεί καθώς προήλθε από την συνένωση δύο όχι και τόσο όμοιων τελικά ομάδων. Από τη

στιγμή που οι τιμές του κέρδους ομαδοποίησης είναι μη αρνητικές, αρνητική διαφορά της τιμής του «Δ» μεταξύ δύο διαδοχικών αποτελεσμάτων ομαδοποίησης, σημαίνει πως οι ομάδες που συγχωνεύτηκαν στο τελευταίο βήμα είχαν αρνητική επίπτωση στο συνολικό κέρδος ομαδοποίησης, καθώς στο προηγούμενο επίπεδο του συγχωνευτικού ιεραρχικού αλγορίθμου είχαμε ένα αποτέλεσμα ομαδοποίησης το οποίο ήταν ποιοτικά καλύτερο. Επομένως, κάνοντας χρήση της ιδιότητας της εμφώλευσης που παρουσιάζουν οι ιεραρχικοί αλγόριθμοι, μπορούμε να καταδείξουμε στη βέλτιστη ομαδοποίηση, αν έχει εντοπιστεί κάποια (ή κάποιες) τέτοια αρνητική μετάβαση, ποια (ή ποιες) από τις ομάδες που έχουν προκύψει πρέπει να τις επισημάνουμε ως «ύποπτες» για το σύστημά μας, έτσι ώστε αυτό στη φάση της παρακολούθησης να την ελέγχει συνεχώς και να μη βασίζεται στον εκπρόσωπό της για την αποστολή τη συσχετισμένης με τη θέση πληροφορίας.

Αν κάναμε μια μεγέθυνση της καμπύλης για το κέρδος ομαδοποίησης της Εικόνας 17 θα μπορούσαμε να παρατηρήσουμε κάτι που θα έμοιαζε με αυτό που αποτυπώνεται στο διάγραμμα της Εικόνας 18. Βλέπουμε δηλαδή πως αν και η καμπύλη των τιμών



Εικόνα 18: Μεγέθυνση της καμπύλης του κέρδους ομαδοποίησης στην Εικόνα 17 για τον εντοπισμό αρνητικής μετάβασης για τιμές του «Δ»

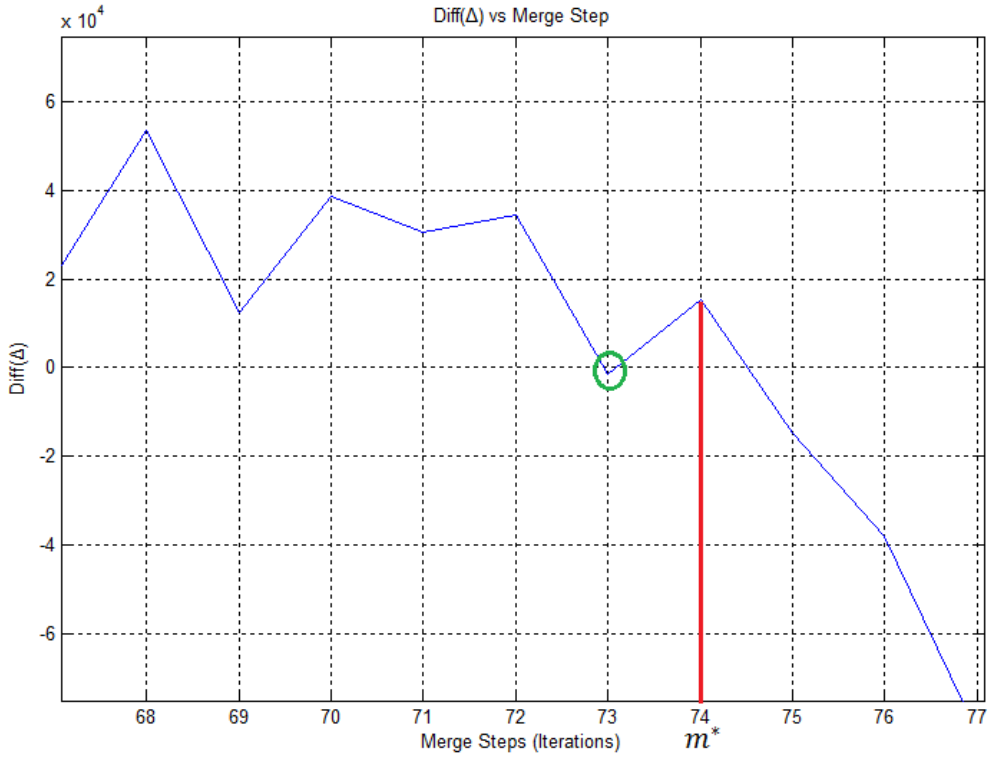
του κέρδους ομαδοποίησης για τα επίπεδα των συγχωνεύσεων του αλγορίθμου μοιάζει συνεχώς αύξουσα μέχρι το σημείο που εντοπίζεται η μεγαλύτερη τιμή του «Δ»,

εντούτοις δεν είναι κάτι το οποίο ισχύει. Όπως αναφέρθηκε, υπάρχουν και περιπτώσεις ομαδοποιήσεων που τελικά δεν είναι και τόσο ποιοτικές και όταν αυτό συμβαίνει θα πρέπει να αναδεικνύεται από το σύστημά μας. Φαίνεται λοιπόν πως η συγχώνευση που συνέβη στο επίπεδο $t = 73$ των βημάτων του αλγορίθμου οδήγησε σε τιμή για το κέρδος ομαδοποίησης χαμηλότερη από την τιμή που είχε για την ομαδοποίηση που προέκυψε μετά τη συγχώνευση που έγινε στο επίπεδο $t = 72$. Άρα το σύστημά μας θα πρέπει να εντοπίσει σε ποια ομάδα τελικά βρίσκεται η ομάδα που προέκυψε στο επίπεδο $t = 73$ και να την υποδείξει ως «ύποπτη» για παρακολούθηση.

Ορίζοντας και μαθηματικά τα παραπάνω, αν Δ_i είναι η τιμή του κέρδους ομαδοποίησης για το επίπεδο $t = i$ του αλγορίθμου και Δ_j η αντίστοιχη τιμή του για το επίπεδο $t = j$, με $i > j$, τότε η ομάδα που δημιουργήθηκε στο επίπεδο $t = i$ πρέπει να θεωρηθεί ως ύποπτη αν ισχύει

$$\Delta_i - \Delta_j < 0$$

Διαισθητικά, θα μπορούσαμε επίσης να πούμε πως περίπτωση χειρότερου αποτελέσματος ομαδοποίησης θα έχουμε, αν αντιμετωπίσουμε το κέρδος ομαδοποίησης «Δ» ως μια ανεξάρτητη μεταβλητή, όταν η διαφορά δύο διαδοχικών τιμών της είναι αρνητική, κάτι το οποίο φαίνεται στο διάγραμμα της Εικόνας 19 που ακολουθεί.



Εικόνα 19: Διάγραμμα διαφορών για τις τιμές του κέρδους ομαδοποίησης «Δ»

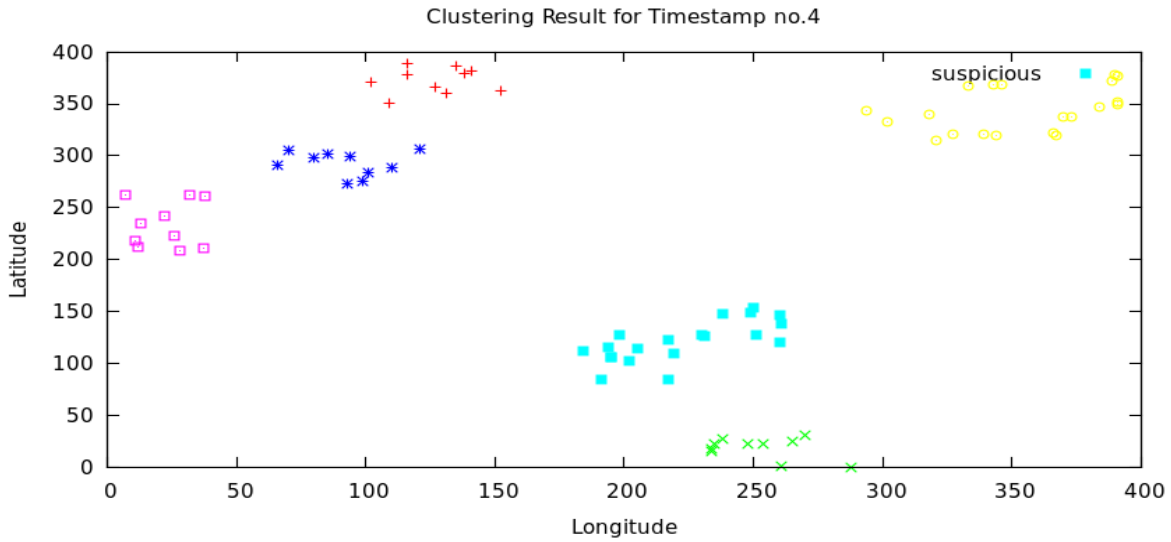
Ενισχύεται επομένως, μέσω και της Εικόνας 19, το γεγονός πως η διαφορά μεταξύ των τιμών του κέρδους ομαδοποίησης για τα επίπεδα 72 και 73 είναι αρνητική. Σε αυτό το σημείο μπορούμε να δώσουμε τον Αλγόριθμο Αυτόματης Ομαδοποίησης και Εντοπισμού «Ύποπτων» Ομάδων.

Αλγόριθμος Αυτόματης Ομαδοποίησης και Εντοπισμού Ύποπτων Ομάδων (Automated Hierarchical Clustering Algorithm with Suspicious Clusters Annotation Mechanism - AHCASCAM)

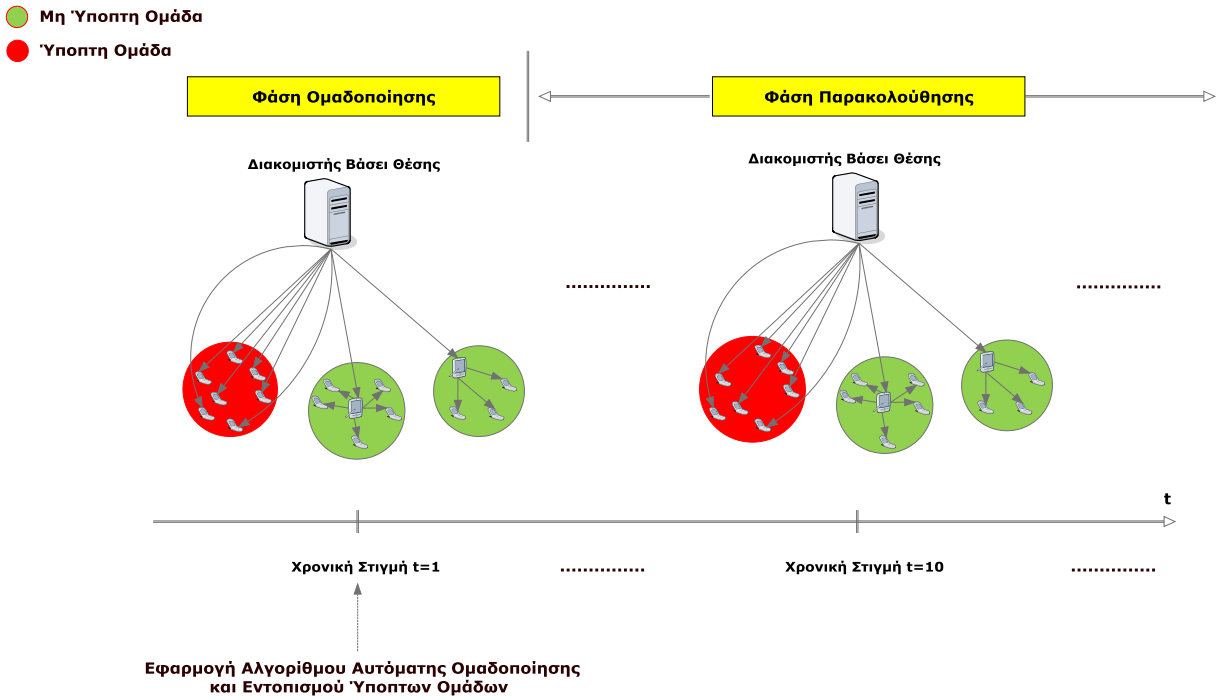
- Αρχικοποίηση:
  - Επιλογή του συγχωνευτικού ιεραρχικού αλγορίθμου που θα χρησιμοποιηθεί
  - $t = 0$
- Επανέλαβε:
  - $t = t + 1$
  - Εντόπισε και συγχώνευσε στην ομάδα C_q τις δύο κοντινότερες ομάδες βάσει του μέτρου σύνδεσης
  - Υπολόγισε την τιμή του κέρδους ομαδοποίησης $\Delta(t)$
  - Αν $t > 1$
    - Υπολόγισε τη διαφορά $diff = \Delta(t) - \Delta(t-1)$
    - Αν $diff < 0$ τότε $C_q.suspicious = true$
- Μέχρι να ανήκουν όλα τα στοιχεία του συνόλου N σε μία ομάδα
- Εντόπισε και επέστρεψε το επίπεδο t^* των βημάτων του αλγορίθμου για το οποίο έχουμε το $\max_{1 \le t \le N-1}\{\Delta(t)\}$
- Με τη βοήθεια της ιεραρχικής δομής των ομάδων του επιπέδου t^* εντόπισε αν υπάρχουν ομάδες που έχουν ύποπτες υποομάδες ή που είναι ύποπτες (από προηγούμενη συγχώνευση) οι ίδιες και σημείωσέ τες

Δημιουργήσαμε βήμα-βήμα επομένως ένα σύστημα το οποίο είναι σε θέση να ομαδοποιεί αυτόματα ένα σύνολο N χρηστών σε ένα χώρο εποπτείας και να καταδεικνύει στην ομαδοποίηση που προκύπτει τις ομάδες που χρειάζονται συνεχή παρακολούθηση (αν υπάρχουν βεβαίως), όπως φαίνεται και στην Εικόνα 20. Εφόσον το σύστημα έχει πλέον πληροφορία σχετικά με το ποιες ομάδες είναι ύποπτες και ποιες όχι (φάση ομαδοποίησης), μπορεί στις επόμενες χρονικές στιγμές να περάσει στη φάση της παρακολούθησης, όπου θα χρειάζεται για τις ύποπτες ομάδες να λαμβάνει συνεχώς

τη θέση των μελών τους και να τους στέλνει τα μηνύματα που σχετίζονται με την περιοχή που βρίσκονται ενώ για τις ομάδες που δεν έχουν επισημανθεί ως ύποπτες μπορεί να ελέγχει μόνο τους εκπροσώπους των ομάδων τους κάθε χρονική στιγμή και να αποστέλλει σε εκείνους το μήνυμα που αφορά την ομάδα. Οι εκπρόσωποι μετά είναι υπεύθυνοι για τη διανομή της πληροφορίας, που τους μεταφέρθηκε από την υπηρεσία, εντός της ομάδας που αντιπροσωπεύουν, όπως φαίνεται στην Εικόνα 21.



Εικόνα 20: Αποτέλεσμα εφαρμογής του αλγορίθμου αυτόματης ομαδοποίησης και εντοπισμού ύποπτων ομάδων σε ένα σύνολο δεδομένων



Εικόνα 21: Περιγραφή Λογικής Προτεινόμενου Συστήματος

Συμπερασματικά, το κέρδος με την ενσωμάτωση του προτεινόμενου σχήματος είναι άμεσα συσχετιζόμενο με τις χρονικές στιγμές που το σύστημα παρακολουθεί και δεν χρειάζεται να στέλνει σε όλους τους χρήστες την πληροφορία, καθώς δεν είναι απαραίτητο σε κάθε στιγμή να κάνει την ομαδοποίηση των δεδομένων. Δηλαδή, η κατηγοριοποίηση των ομάδων σε ύποπτες (για να διασπαστούν στο εγγύς μέλλον) και μη ωφελεί στην δημιουργίας μιας σιγουριάς στο σύστημα έτσι ώστε να μην χρειάζεται να πραγματοποιεί σε κάθε χρονική στιγμή τον αλγόριθμο αυτόματης ομαδοποίησης που είναι ακριβός υπολογιστικά, αλλά για κάποια περίοδο απλά να παρακολουθεί το αποτέλεσμα που προέκυψε από τη στιγμή που εκτελέστηκε ο Αλγόριθμος Αυτόματης Ομαδοποίησης και Εντοπισμού Ύποπτων Ομάδων. Φυσικά, βέλτιστη λύση για τέτοιου είδους προβλήματα δεν υπάρχει, γι' αυτό και ο προτεινόμενος αλγόριθμος κατασκευάστηκε για να χρησιμοποιηθεί από συστήματα που προσφέρουν υπηρεσίες πληροφοριών βάσει θέσης, με στόχο τη μείωση της αποστολής μηνυμάτων εντός του δικτύου, διατηρώντας το ποσοστό απώλειας πληροφορίας (δηλαδή, μη ενημερωμένων χρηστών) και τις απαιτήσεις σε επεξεργαστική ισχύ σε ανεκτά επίπεδα, όπως θα δείξουμε και στο 4^{ο} Κεφάλαιο που περιλαμβάνει την αποτίμηση της παραπάνω λογικής μέσω πειραματικών δοκιμών.

3.4.1 Μέθοδος Επιλογής Κόμβου Κεφαλής

Έχει αναφερθεί πολλές φορές μέχρι στιγμής πως από τη στιγμή που θα γίνει η κατηγοριοποίηση των ομάδων από τον αλγόριθμό μας σε ύποπτες (αν υπάρχουν φυσικά τέτοιες) και μη, έπειτα έχουμε τη φάση της παρακολούθησης, όπου στις μη ύποπτες ομάδες, η πληροφορία στέλνεται από την υπηρεσία στους εκπροσώπους αυτών των ομάδων. Ο εκπρόσωπος μιας ομάδας πολύ συχνά στη βιβλιογραφία αναφέρεται και ως κόμβος κεφαλή (cluster head). Πώς όμως τελικά προκύπτει αυτός ο κόμβος; Σε άλλες μορφές δικτύων (όπως δίκτυα αισθητήρων), όπου η ενέργεια σε κάθε κόμβο τους είναι περιορισμένη, η «ετικέτα» του κόμβου κεφαλή για μια ομάδα από κόμβους, εναλλάσσεται μεταξύ των μελών της ανά τακτά χρονικά διαστήματα έτσι ώστε η κατανάλωση της ενέργειας να είναι ίση όσο το δυνατόν για τους κόμβους της ίδιας ομάδας. Ας δούμε για παράδειγμα τη μέθοδο που ακολουθείται από την αρχιτεκτονική LEACH για την επιλογή του εκπροσώπου μιας ομάδας κόμβων [19].

❖ *Περιγραφή Αλγορίθμου Επιλογής Κόμβου Κεφαλής του σχήματος LEACH*

Η ακόλουθη τεχνική επιλογής του κατάλληλου κόμβου κεφαλής βασίζεται στον λόγο ενέργειας και απόστασης όπως αυτός περιγράφεται παρακάτω.

$$Ratio = \frac{Node\ Energy}{Distance\ of\ node\ from\ BS}$$

ή

$$R_i = \frac{E_i}{D_i}$$

Όπου $D_i = distance(N_i, BS)$ και R_i είναι ο λόγος Ενέργειας και Απόστασης του κάθε κόμβου ο οποίος υπολογίζεται ως ακολούθως:

$$R_1 = \frac{E_1}{D_1}$$

$$R_2 = \frac{E_2}{D_2}$$

.
.
.

$$R_n = \frac{E_n}{D_n}$$

Θεωρούμε ως κόμβο κεφαλή της κάθε ομάδας εκείνο τον κόμβο ο οποίος παρουσιάζει το μεγαλύτερο R_j την τρέχουσα χρονική στιγμή.

$$R_j = \max\{R_1, R_2, \dots\dots\dots, R_n\}$$

Αν 2 κόμβοι στην ίδια ομάδα παρουσιάζουν την ίδια τιμή $R_i = R_j$, η οποία είναι ταυτόχρονα και η μέγιστη για την ομάδα, τότε επιλέγουμε ως κόμβο κεφαλής της ομάδας εκείνον που έχει τα μεγαλύτερα αποθέματα ενέργειας.

$$R_i = R_j = \max\{E_1, E_2, \dots\dots\dots, E_n\}$$

Και σε μορφή ψευδοκώδικα τα παραπάνω συνοψίζονται ως εξής:

- Επανέλαβε για κάθε ομάδα C_i
  - Επανέλαβε για κάθε κόμβο N_i που ανήκει στην ομάδα C_i
    - Υπολόγισε την απόσταση D_i of N_i from Base Station (BS)
    - Find Energy E_i of N_i
    - Calculate ratio: $R_i = E_i/D_i$
  - Μέχρι να ελεγχθούν όλοι οι κόμβοι
  - Επέλεξε το max(R_i)
  - Αν length(max(R_i)) > 1, επέλεξε ως εκπρόσωπο της ομάδας C_i τον κόμβο N_i που έχει max(E_i)
- Μέχρι να ελεγχθούν όλες οι ομάδες

❖ *Περιγραφή Αλγορίθμου Επιλογής Κόμβου Κεφαλής Βάσει του Ιδεατού Κεντροειδούς*

Μια πιο απλουστευμένη μέθοδος επιλογής του κόμβου κεφαλή μιας ομάδας κόμβων, η οποία είναι και πιο άμεσα εφαρμόσιμη στο πρόβλημα που προσπαθούμε να αντιμετωπίσουμε, είναι αυτή της επιλογής ως εκπροσώπου της ομάδας του κόμβου που βρίσκεται κοντινότερα στο ιδεατό κεντροειδές αυτής της ομάδας. Όπως γίνεται κατανοητό, σε αυτή τη μέθοδο δε λαμβάνεται υπόψη το απόθεμα σε ενέργεια του εκάστοτε κόμβου. Το μόνο κριτήριο είναι το πόσο κοντά βρίσκεται ο κόμβος αυτός. Οπότε, αν με CC_i συμβολίσουμε το ιδεατό κεντροειδές της ομάδας C_i, το οποίο και υπολογίζεται μετά την εύρεση των μελών της κάθε ομάδας στη φάση της ομαδοποίησης ως ακολούθως:

$$CC_i = \frac{1}{n_{C_i}} \sum_{k=1}^{n_{C_i}} N_k$$

και με N_j συμβολίσουμε το κόμβο j της ομάδας αυτής τότε, για κάθε ομάδα C_i υπολογίζουμε το

$$\min \{D_j, \forall j \in C_i\}$$

με

$$D_j = distance(N_j, CC_i)$$

και ορίζουμε τελικά ως εκπρόσωπο της ομάδας C_i τον κόμβο i, που είχε τη μικρότερη απόσταση από το ιδεατό κεντροειδές CC_i της ομάδας.

Και σε μορφή ψευδοκώδικα τα παραπάνω συνοψίζονται ως εξής:

- Επανέλαβε για κάθε ομάδα C_i
  - Επανέλαβε για κάθε μέλος N_j της ομάδας C_i
    - Υπολόγισε την απόσταση D_j του N_j από το ιδεατό κεντροειδές CC_i της ομάδας C_i
  - Μέχρι να ελεγχθούν όλα τα μέλη
  - Επέλεξε ως εκπρόσωπο της ομάδας C_i τον κόμβο N_j που έχει min(D_j)
- Μέχρι να ελεγχθούν όλες οι ομάδες

Λόγω της απλότητας της παραπάνω μεθόδου, είναι αυτή η οποία τελικά χρησιμοποιήθηκε στα πειράματά μας καθώς είναι και φθηνότερη υπολογιστικά έναντι αυτής που χρησιμοποιείται στην αρχιτεκτονική LEACH. Έχουμε πλέον δημιουργήσει ένα σύστημα το οποίο λαμβάνοντας ως είσοδο τις πληροφορίες θέσης των χρηστών σε μια περιοχή ελέγχουν είναι σε θέση να τις ομαδοποιεί αυτόματα, ορίζοντας τους εκπροσώπους τους και εντοπίζοντας πιθανές ομάδες για να διασπαστούν στο προσεχές μέλλον. Αυτό το οποίο μένει είναι να δούμε κατά πόσο είναι ισχυρή μια τέτοια μέθοδος, κάτι το οποίο θα παρουσιαστεί το Κεφάλαιο 4 που ακολουθεί.

ΚΕΦΑΛΑΙΟ 4

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ ΣΥΣΤΗΜΑΤΟΣ

Το προτεινόμενο σύστημα αυτόματης ομαδοποίησης και εντοπισμού ύποπτων ομάδων στα πλαίσια των υπηρεσιών βάσει πληροφορίας θέσης αξιολογήθηκε ποιοτικά και ποσοτικά έτσι ώστε να αποτιμηθεί η αξία του και το ποσοστό εκπλήρωσης των αρχικών απαιτήσεων από αυτό. Στο συγκεκριμένο κεφάλαιο θα περιγραφεί το σενάριο βάσει του οποίου έγινε η αξιολόγηση του προτεινόμενου συστήματος, τα αποτελέσματα των πειραματικών δοκιμών καθώς και κάποια συμπεράσματα πάνω σε αυτά. Η παρούσα αξιολόγηση είχε ως στόχο τόσο τη μελέτη των επιδόσεων του συστήματος όσο και την επικύρωση της ορθής λειτουργίας του.

Αρχικά θα παρουσιαστούν οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος, τα σύνολα των δεδομένων καθώς και η τεχνική που χρησιμοποιήθηκε για τα πειράματά μας ενώ στη συνέχεια θα δοθούν τα αποτελέσματα των πειραμάτων μας.

4.1 Σενάριο Αξιολόγησης

4.1.1 Μετρικές

Στη συγκεκριμένη παράγραφο θα γίνει μια παρουσίαση των μετρικών που αποφασίσαμε να χρησιμοποιήσουμε για την αποτίμηση του συστήματός μας ως προς το φόρτο του δικτύου καθώς και το πόσο καλή είναι η κατάδειξη των ύποπτων ομάδων. Έχουμε τις ακόλουθες μετρικές λοιπόν:

- *False Negative percentage (FN)*

  Θα μπορούσαμε να το περιγράψουμε ως: Λανθασμένη μη κατάδειξη ενός χρήστη ως ύποπτο. Δηλαδή, το να έχουμε ορίσει έναν κόμβο τη χρονική στιγμή t_i ως μη-ύποπτο αλλά σε κάποια επόμενη χρονική στιγμή t_j να έχει αλλάξει ομάδα. Μπορεί να φαίνεται επομένως πως είχαμε κέρδος, καθώς από αυτή την ομάδα στην οποία ανήκε ο χρήστης παρακολουθούσαμε μόνο τον εκπρόσωπό της, εντούτοις, λόγω του ότι ο χρήστης έφυγε από αυτήν χωρίς να ενημερωθεί το σύστημα, τελικά έχουμε απώλεια πληροφορίας. Τέλος, από τη στιγμή που μιλάμε

για ποσοστό, η τιμή της μετρικής αυτής, κατά τη διάρκεια των πειραμάτων υπολογίστηκε από τη ακόλουθη σχέση:

$$FN = \frac{total\_FN\_hits}{total\_checks} \cdot 100\% \qquad (\mathbf{4.1})$$

όπου, με $total\_FN\_hits$ συμβολίζουμε τα συνολικά περιστατικά λανθασμένης μη κατάδειξης χρηστών ως ύποπτους και με $total\_checks$ το σύνολο των ελέγχων που έγιναν κατά τη διάρκεια των πειραμάτων στη φάση της προσομοίωσης. Ο τρόπος υπολογισμού των συνολικών ελέγχων θα περιγραφεί αργότερα στην ενότητα που αφορά τις μεθόδους αξιολόγησης.

- Σε αυτό το σημείο πρέπει να αναφερθεί πως η αλλαγή ομάδας για έναν χρήστη ορίζεται ως το να βρίσκεται σε μια ομάδα η οποία εκπροσωπείται από διαφορετικό χρήστη, σε σχέση με αυτόν που εκπροσωπούσε την παλιά ομάδα του (τη χρονική στιγμή της ομαδοποίησης).

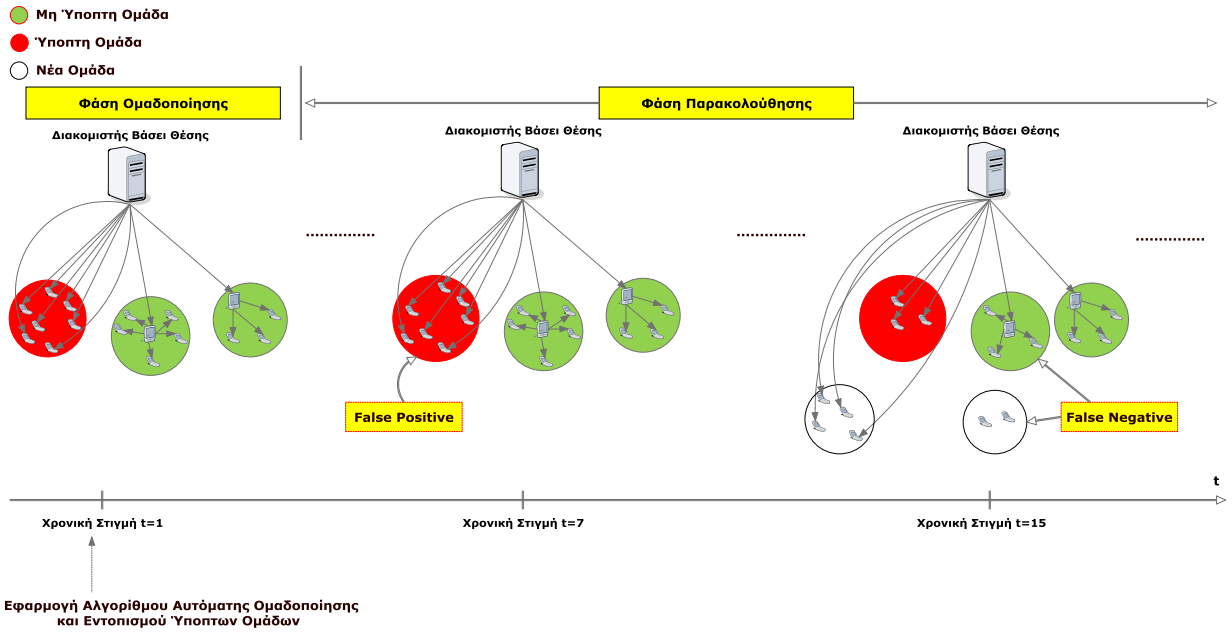
❖ *False Positive percentage (FP)*

Θα μπορούσαμε να το περιγράψουμε ως: Λανθασμένη υπόδειξη ενός χρήστη ως ύποπτο. Δηλαδή, το να έχουμε ορίσει έναν κόμβο τη χρονική στιγμή t_i ως ύποπτο και σε κάποια επόμενη χρονική στιγμή t_j να παραμένει στην ίδια ομάδα. Στην περίπτωση αυτή, είμαστε ζημιωμένοι όσον αφορά το φόρτο του δικτύου μας, καθώς του έχουμε δώσει έναν χρήστη να τον παρακολουθεί συνέχεια ως ύποπτο αλλά αυτός τελικά δεν άλλαξε ομάδα. Εντούτοις, σε αυτή την περίπτωση είναι δεν έχουμε απώλεια πληροφορίας. Τέλος, από τη στιγμή που μιλάμε για ποσοστό, η τιμή της μετρικής αυτής, κατά τη διάρκεια των πειραμάτων υπολογίστηκε από τη ακόλουθη σχέση:

$$FP = \frac{total\_FP\_hits}{total\_checks} \cdot 100\% \qquad (\mathbf{4.2})$$

όπου, με $total\_FP\_hits$ συμβολίζουμε τα συνολικά περιστατικά λανθασμένης κατάδειξης χρηστών ως ύποπτους και με $total\_checks$ και εδώ το σύνολο των ελέγχων που έγιναν κατά τη διάρκεια των πειραμάτων στη φάση της προσομοίωσης.

Η λογική των False Negative και False Positive μετρικών μπορεί να φανεί καλύτερα και μέσω της εικόνας που ακολουθεί.



Εικόνα 22: Αναπαράσταση των μετρικών False Positive (FP) και False Negative (FN) κατά τη διάρκεια μιας προσομοίωσης

❖ *Ποσοστό Φόρτου Δικτύου (Network Load percentage - NL)*

Ως φόρτο δικτύου για μια χρονική στιγμή t ορίζεται ο λόγος του συνολικού πλήθους των χρηστών που χρειάζεται να στείλουμε το μήνυμα με την πληροφορία προς το σύνολο των χρηστών του συνόλου δεδομένων. Το πηλίκο της παραπάνω μετρικής θα μπορούσε να περιγραφεί και ως η χείριστη των περιπτώσεων για τέτοιου είδους συστήματα, όπου δεν υπάρχει κάποιος μηχανισμός ομαδοποίησης. Για την ποσοστιαία τιμή της μετρικής αυτής λοιπόν έχουμε

$$network\_load(t) = \frac{\sum_{i=1}^{k} N_{C_i} + l}{\sum_{i=1}^{k+l} N_{C_i}} \cdot 100\% \qquad (\mathbf{4.3})$$

όπου, k είναι το πλήθος των ύποπτων ομάδων της βέλτιστης ομαδοποίησης για τη χρονική στιγμή t, l το πλήθος των μη-ύποπτων ομάδων και $\boldsymbol{N_{C_i}}$ το πλήθος των χρηστών που ανήκουν στην ομάδα $\boldsymbol{C_i}$ τη χρονική στιγμή t. Από τη Σχέση (4.1) επομένως μπορεί να προκύψει πολύ εύκολα ποια είναι η μέγιστη και ποια η ελάχιστη τιμή για τη συγκεκριμένη μετρική, οπότε

- $NL(t) = NL\_\max(\mathrm{t}) = \frac{\sum_{i=1}^{k} N_{C_i}}{\sum_{i=1}^{k+l} N_{C_i}} \cdot 100\% = 100\%$, όταν $l = 0$ (όλες οι ομάδες ύποπτες ή απώλεια μηχανισμού ομαδοποίησης)
- $NL(t) = NL\_\min(\mathrm{t}) = \frac{l}{\sum_{i=1}^{k+l} N_{C_i}} \cdot 100\%$, όταν $k = 0$ (όλες οι ομάδες είναι μη ύποπτες ή χρησιμοποιείται απλά μηχανισμός ομαδοποίησης, χωρίς μέθοδο κατάδειξης ύποπτων ομάδων)

Είναι φανερό πως για μια προσομοίωση S, με N_s χρονικές στιγμές, το *μέσο ποσοστό φόρτου δικτύου (Average Network Load - ANL)*, αν θέλουμε μια γενικότερη εικόνα για την βελτίωση που προσφέρει το σύστημά μας, υπολογίζεται μέσω της ακόλουθης σχέσης

$$ANL = \frac{\sum_{t=1}^{N_s} NL(t)}{N_s} \qquad (\mathbf{4.4})$$

Πρέπει να αναφερθεί σε αυτό το σημείο πως μπορεί σε κάποια από το πειράματα να εμφανιστεί και η μετρική του ποσοστού κέρδους του δικτύου (Network Gain percentage - NG) το οποίο όμως κρίνεται πως δεν είναι απαραίτητο να οριστεί χωριστά καθώς είναι το ακριβός αντίθετο από το ποσοστό φόρτου του δικτύου. Δηλαδή, το κέρδος του δικτύου είναι το πόσα μηνύματα τελικά δεν χρειάζεται να σταλούν και ισχύει

$$NG(t) = 100\% - NL(t)$$

καθώς και για τις ακραίες τιμές του

- $NG\_\max(t) = 100\% - NL\_\min(t)$
- $NG\_\min(t) = 100\% - NL\_\max(t) = 0\%$

Τέλος, για το *μέσο ποσοστό κέρδους δικτύου (Average Network Gain - ANG)* σε μια προσομοίωση S, προκύπτει ομοίως με τα παραπάνω πως

$$ANG = 100\% - ANL$$

❖ *Μέσος Χρόνος Εκτέλεσης (Average Execution Time - AET)*

Ως χρόνος εκτέλεσης του αλγορίθμου ορίζεται το χρονικό διάστημα που μεσολαβεί από τη στιγμή που θα ξεκινήσει η ομαδοποίηση των δεδομένων μέχρι να βρεθεί η βέλτιστη ομαδοποίηση και να καταδειχθούν οι ύποπτες ομάδες, αν

τελικά υπάρχουν τέτοιες στο τελικό αποτέλεσμα. Έτσι λοιπόν, για τη χρονική στιγμή t ισχύει

$$ET(t) = end\_time(t) - start\_time(t) \qquad (\mathbf{4.5})$$

ενώ για το μέσο χρόνο εκτέλεσης, ο οποίος ορίζεται για την περίοδο που διαρκεί μια προσομοίωση, έχουμε

$$AET = \frac{\sum_{t=1}^{N_s} ET(t)}{N_s} \qquad (\mathbf{4.6})$$

όπου, με N_s συμβολίζουμε το πλήθος των χρονικών στιγμών που περιλαμβάνει η προσομοίωση S που εκτελείται.

- *Μέση Περίοδος Επιτυχημένης Κατάδειξης (Average Period for Successive Annotation - APSA)*

Ένα πρωτοποριακό στοιχείο της προτεινόμενης μεθόδου, είναι πως μπορεί να καταδείξει αν κάποια (ή κάποιες) από τις ομάδες που ανήκουν στη βέλτιστη ομαδοποίηση είναι ύποπτη για διάσπαση στο εγγύς μέλλον. Πόσο επιτυχημένη είναι αυτή η κατάδειξη όμως; Αυτό ορίζεται από τη μετρική της μέσης περιόδου επιτυχημένης κατάδειξης, η οποία και προσδιορίζει το μέσο αριθμό βημάτων που μεσολαβούν από τη κατάδειξη μιας ομάδας ως ύποπτη μέχρι η ομάδα αυτή να διασπαστεί, αλλάζοντας τη δομή της. Φυσικά, δε λαμβάνονται υπόψη στην καταμέτρηση οι περιπτώσεις λανθασμένης κατάδειξης, καθώς αυτό το οποίο θέλουμε να μετρήσουμε είναι το πόσο αποδοτική είναι η μέθοδός μας για τις επιτυχείς καταδείξεις ύποπτων ομάδων. Ο υπολογισμός της τιμής της παραπάνω μετρικής μπορεί να περιγραφεί από την παρακάτω σχέση

$$APSA = \frac{\sum_{i=1}^{N_{sa}} (t_s - t_a)}{N_{sa}} \qquad (\mathbf{4.7})$$

όπου, με N_{sa} συμβολίζουμε το πλήθος των επιτυχών καταδείξεων (successive annotation) ομάδων ως ύποπτες κατά τη διάρκεια της προσομοίωσης S, με t_a τη χρονική στιγμή όπου αυτή η ομάδα εντοπίστηκε ως ύποπτη (annotation) και με t_s τη χρονική στιγμή που τελικά η δομή της άλλαξε (success).

4.1.2 Μέθοδος Αξιολόγησης

Σε αυτό το σημείο θα περιγραφεί η λογική που ακολουθήθηκε για την αξιολόγηση του προτεινόμενου συστήματος και τη διεξαγωγή των πειραμάτων μας. Από τη στιγμή που το σύστημά μας ουσιαστικά ενεργεί μόνο σε επίπεδο χρονικής στιγμής δεν θα μπορούσε να αξιολογηθεί αυστηρώς στο βάθος του χρονικού ορίζοντα μιας προσομοίωσης. Ας δούμε επομένως τη λογική της προσομοίωσης για κάθε μία από τις μετρικές που παρουσιάστηκαν παραπάνω.

- *Προσομοίωση για False Negative και False Positive*

  Πρόκειται για δύο μετρικές οι οποίες υπολογίζονται ταυτόχρονα. Έτσι λοιπόν, αν θεωρήσουμε πως έχουμε τη προσομοίωση S με N_s χρονικές στιγμές τότε για κάθε χρονική στιγμή εκτελούμε τον Αλγόριθμο Αυτόματης Ομαδοποίησης και Εντοπισμού Ύποπτων ομάδων. Στο αποτέλεσμα που προκύπτει, παίρνουμε την ομαδοποίηση κάθε χρονικής στιγμής της προσομοίωσης επαναληπτικά και ελέγχουμε από την επόμενη χρονική στιγμή και μέχρι το πέρας της προσομοίωσης σε ποιες χρονικές στιγμές οι ύποπτες ομάδες που καταδείχθηκαν διατήρησαν τη δομή τους (False Positives), έτσι ώστε να θεωρηθούν όλοι τους οι κόμβοι ως λανθασμένη κατάδειξη, καθώς και σε ποιες οι μη-ύποπτες ομάδες δεν διατήρησαν τη δομή τους (False Negatives). Στην περίπτωση εύρεσης καταστάσεων λανθασμένων μη καταδείξεων, από τους χρήστες εκείνων των ομάδων, ως False Negatives μετράμε ***μόνο*** τους χρήστες που μετακινήθηκαν σε μια νέα ομάδα η οποία δεν έχει τον εκπρόσωπο που το σύστημα νόμιζε πως είχαν αυτοί οι χρήστες. Μας νοιάζει δηλαδή για τα False Positives πόσοι είναι οι χρήστες που τελικά λαμβάνουν μηνύματα πληροφοριών ενώ θα μπορούσαμε να το είχαμε αποφύγει και για τα False Negatives πόσοι είναι οι χρήστες που τελικά δε λαμβάνουν μηνύματα που θα έπρεπε να είχαν λάβει. Και οι δύο τιμές συγκρίνονται με το συνολικό πλήθος των μηνυμάτων που θα στέλνονταν στο δίκτυο αν τελικά δεν είχαμε εφαρμόσει κάποιο σχήμα ομαδοποίησης για τη χρονική στιγμή που ελέγχουμε. Δηλαδή, το $total\_checks$ που εμφανίζεται στις σχέσεις (4.1) και (4.2) μπορεί πλέον να οριστεί ως εξής

$$total\_checks = \sum_{i=1}^{N_s-1} \sum_{j=i+1}^{N_s} \sum_{k=1}^{r_i} N_{C_k} = \cdots = \frac{N(N_s-1)N_s}{2}$$

όπου, με N συμβολίζουμε το πλήθος των χρηστών του συνόλου των δεδομένων μας για κάθε χρονική στιγμή (αυτό είναι σταθερό και δεν αλλάζει), με N_{C_k} το πλήθος των μελών της ομάδας k (από το αποτέλεσμα ομαδοποίησης τη χρονική στιγμή $t = i$) και με r_i το πλήθος των ομάδων που έχουν προκύψει ως βέλτιστη ομαδοποίηση για τη χρονική στιγμή $t = i$ της προσομοίωσης.

- *Προσομοίωση για Ποσοστό Φόρτου Δικτύου*

  Από τη στιγμή που το σύστημα που έχει δημιουργηθεί μπορεί να προσφέρει μια αυτόματη ομαδοποίηση των δεδομένων και κατάδειξη των ομάδων που προκύπτουν σε ύποπτες και μη, ο υπολογισμός της συγκεκριμένης μετρικής θα πρέπει να γίνεται για κάθε μία από τις N_s χρονικές στιγμές της προσομοίωσης t. Δηλαδή, και γι' αυτή τη μετρική τρέχουμε τον αλγόριθμο σε κάθε χρονική στιγμή αλλά για να υπολογίσουμε τη τιμή της δε χρειάζεται να συγκριθεί με μετέπειτα χρονικές στιγμές. Τέλος, για να αποκτήσουμε μια γενικότερη εικόνα για τη βελτίωση που προσφέρει η μέθοδός μας στο φόρτο του δικτύου, θα μπορούσαμε μετά το πέρας της προσομοίωσης και εφόσον θα έχουμε τη τιμή για το φόρτο του δικτύου σε κάθε χρονική στιγμή να υπολογίσουμε και το μέσο ποσοστό φόρτου δικτύου. Μέσω της ίδιας ακριβώς διαδικασία μπορούν να υπολογιστούν και οι μετρικές που αφορούν το κέρδος δικτύου, οι οποίες όπως δείξαμε παραπάνω είναι συμπληρωματικές με τις μετρικές για το φόρτο δικτύου, επομένως αν γνωρίζουμε τις τιμές για τη μια κατηγορία, μέσω μιας απλής αφαίρεσης μπορούμε να μάθουμε τις τιμές και της άλλης.

- *Προσομοίωση για Μέσο Χρόνο Εκτέλεσης*

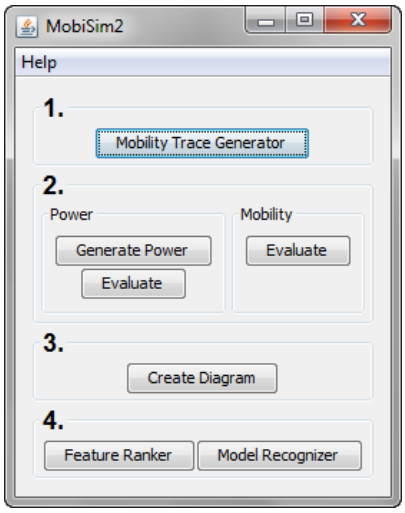
  Και γι' αυτή τη μετρική η μέθοδος της προσομοίωσης που ακολουθείται για τον υπολογισμό της είναι ίδια με αυτή που περιγράφηκε για το Ποσοστό Φόρτου Δικτύου. Έτσι λοιπόν και εδώ για να υπολογίσουμε το μέσο χρόνο εκτέλεσης, υπολογίζουμε το χρόνο εκτέλεσης του αλγορίθμου μας, εφαρμόζοντάς τον στο σύνολο των δεδομένων μας κάθε χρονική στιγμή της προσομοίωσης και στη συνέχεια αθροίζουμε τους χρόνους αυτούς και τους διαιρούμε με τις συνολικές χρονικές στιγμές της προσομοίωσής μας

- *Προσομοίωση για Μέση Περίοδο Επιτυχημένης Κατάδειξης*

  Αναφέρθηκε πως η συγκεκριμένη μετρική λαμβάνει υπόψη της μόνο τις επιτυχείς καταδείξεις ύποπτων ομάδων και στοχεύει στην εξαγωγή ενός μέσου αριθμού βημάτων από τη στιγμή της κατάδειξης μιας ομάδας ως ύποπτης και της στιγμής που τελικά η ομάδα αυτή χαλάει τη δομή της. Σε αντίθεση με τα False Positives, εδώ δε μας ενδιαφέρει μετά τη διάσπαση της ομάδας αν αυτό θα ξανασυμβεί. Επομένως ακολουθείται μια παρόμοια προσέγγιση προσομοίωσης με αυτή του υπολογισμού των False Negative και False Positive δεικτών με κάποιες διαφοροποιήσεις όμως. Έτσι λοιπόν, αν θεωρήσουμε πως έχουμε τη προσομοίωση S, με N_s χρονικές στιγμές τότε για κάθε χρονική στιγμή εκτελούμε τον Αλγόριθμο Αυτόματης Ομαδοποίησης και Εντοπισμού Ύποπτων ομάδων. Στο αποτέλεσμα που προκύπτει, παίρνουμε την ομαδοποίηση κάθε χρονικής στιγμής της προσομοίωσης επαναληπτικά και αν έχουν καταδειχτεί ομάδες στο συγκεκριμένο αποτέλεσμα ως ύποπτες ελέγχουμε από την επόμενη χρονική στιγμή και μέχρι το πέρας της προσομοίωσης αν οι ομάδες αυτές τελικά διασπόνται. Εφόσον διασπόνται, δηλαδή είχα επιτυχή κατάδειξη ύποπτης ομάδας (successive annotation), εντοπίζω σε ποιό από τα επόμενα βήματα της προσομοίωσης σε σχέση με αυτό που είναι το τρέχον, γίνεται αυτή η διάσπαση και κρατάω τη διάρκεια αυτής της χρονικής περιόδου. Στο τέλος, αθροίζονται αυτές οι περίοδοι για τις επιτυχείς καταδείξεις και διαιρούνται με το συνολικό πλήθος των επιτυχών καταδείξεων, δίνοντας έτσι μια διαισθητική ένδειξη για το πόσο ισχυρό είναι το μέτρο κατάδειξης ύποπτων ομάδων του αλγορίθμου μας.

4.1.3 Σύνολα Δεδομένων (Datasets)

Σε αυτό το σημείο πρέπει να αναφερθεί πως λόγω της έλλειψης αντιπροσωπευτικών συνόλων δεδομένων σε αυτή τη περιοχή που να προέρχονται από μετρήσεις σε πραγματικό πεδίο, προτιμήθηκε η παραγωγή συνθετικών δεδομένων τα οποία και προσομοιώνουν τυχαία κινήσεις ομάδων σε ένα χώρο. Έπειτα από αρκετή μελέτη για το εργαλείο που θα χρησιμοποιούταν για τη παραγωγή των συνόλων δεδομένων για την πειραματική αξιολόγηση του συστήματός μας, καταλήξαμε στη χρήση του MobiSim [20]. Το πρόγραμμα αυτό επιλέχθηκε καθώς ενσωματώνει το μοντέλο κινητικότητας που



Εικόνα 23: Mobility Simulator 2

θέλαμε να ακολουθούν οι ομάδες του συνόλου των δεδομένων μας. Πιο συγκεκριμένα, η κίνηση κάθε ομάδας προσομοιώνει αυτή που ορίζεται από το μοντέλο Reference Point Group Mobility (RPGM) [21]. Στο συγκεκριμένο μοντέλο, κάθε κόμβος μέσα σε μια ομάδα έχει δύο συνιστώσες στο διάνυσμα της κίνησής του: την ατομική συνιστώσα και τη συνιστώσα της ομάδας. Η ατομική συνιστώσα βασίζεται στο Random WayPoint (RWP) μοντέλο κινητικότητας [22]. Βάσει αυτού του μοντέλου, ένας κόμβος επιλέγει τυχαία ένα προορισμό, εντός του πεδίου δράσης της ομάδας που ανήκει, και κινείται προς αυτόν με σταθερή ταχύτητα. Όταν ο κόμβος φτάσει στον προορισμό το, διαλέγει τυχαία πάλι ένα νέο και ξεκινάει να κινείται προς αυτόν μετά από ένα χρονικό διάστημα παύσης στον σημείο που βρίσκεται. Αυτή η συμπεριφορά επαναλαμβάνεται μέχρι το τέλος της προσομοίωσης. Η συνιστώσα κινητικότητας της ομάδας από την άλλη, είναι κοινή για όλα τα μέλη της ίδιας ομάδας και ακολουθεί και εκείνη το RWP μοντέλο. Εδώ όμως ο προορισμός δεν είναι ένα σημείο αλλά μια αυθαίρετη περιοχή εντός τους χώρου της προσομοίωσης.

Μέσω του MobiSim λοιπόν μπορούμε να προσομοιώσουμε την κίνηση μιας ομάδας μέσω του RPGM μοντέλου λαμβάνοντας υπόψη μας πως προσεγγίζει το θέμα με διαφορετικό τρόπο. Πιο συγκεκριμένα, ορίζουμε το πλήθος ομάδων που θέλουμε, καθώς και πόσους κόμβους έχει η κάθε μία, και το πρόγραμμα ορίζει από μόνο του ένα κόμβο ηγέτη-εκπρόσωπο για την ομάδα αυτή, ο οποίος και κινείται ακολουθώντας τη λογική του RWP μοντέλου. Οι υπόλοιποι κόμβοι της ομάδας, ακολουθούν την τροχιά κίνησης που ορίζει ο εκπρόσωπος της ομάδας με μικρές αποκλίσεις όσον αφορά τη ταχύτητα (speed deviation ratio) και την κατεύθυνση (angle deviation ratio) της κίνησης. Για να προσομοιωθεί σωστά η κίνηση ομάδων σε κλειστούς χώρους, η ταχύτητα που

για τους κόμβους των ομάδων των συνόλων δεδομένων μας κινείται πάντοτε στο διάστημα $[0.2, 1.4]$ m/s όπως ορίζεται και στο [23].

Ορίζοντας λοιπόν τις παραμέτρους του MobiSim κατάλληλα, πήραμε τα ακόλουθα σύνολα δεδομένων για τη διεξαγωγή των πειραμάτων μας:

Πίνακας 1: Σύνολα Δεδομένων Πειραμάτων

Σύνολο Δεδομένων	Angle Deviation Ratio (adr)	Speed Deviation Ratio (sdr)	Πλήθος Κόμβων (N)
trace1	0,05	0,05	80
trace2	0,06	0,06	80
trace3	0,08	0,07	80
trace4	0,08	0,08	80
trace5	0,09	0,08	80
trace6	0,05	0,05	160
trace7	0,05	0,05	240
trace8	0,05	0,05	400

4.1.4 Περιβάλλον Προσομοιώσεων

Οι προσομοιώσεις διεξήχθησαν όλες στο ίδιο υπολογιστικό σύστημα με τα ακόλουθα χαρακτηριστικά:

- Επεξεργαστής: Intel Pentium Dual Core T4300 @ 2.1GHz
- Μνήμη RAM: 4GB DDR3 @ 1333MHz
- Λειτουργικό Σύστημα: Windows 7 Home Premium [x64]

ενώ η γλώσσα ανάπτυξης των αλγορίθμων που χρησιμοποιήθηκε είναι η Java σε περιβάλλον Eclipse Indigo. Τέλος, τα διαγράμματα για την αποτίμηση του συστήματος που θα δούμε στη συνέχεια έγιναν με τη βοήθεια του MatLab.

4.2 Αποτελέσματα Αποτίμησης

4.2.1 Αξιολόγηση FN και FP

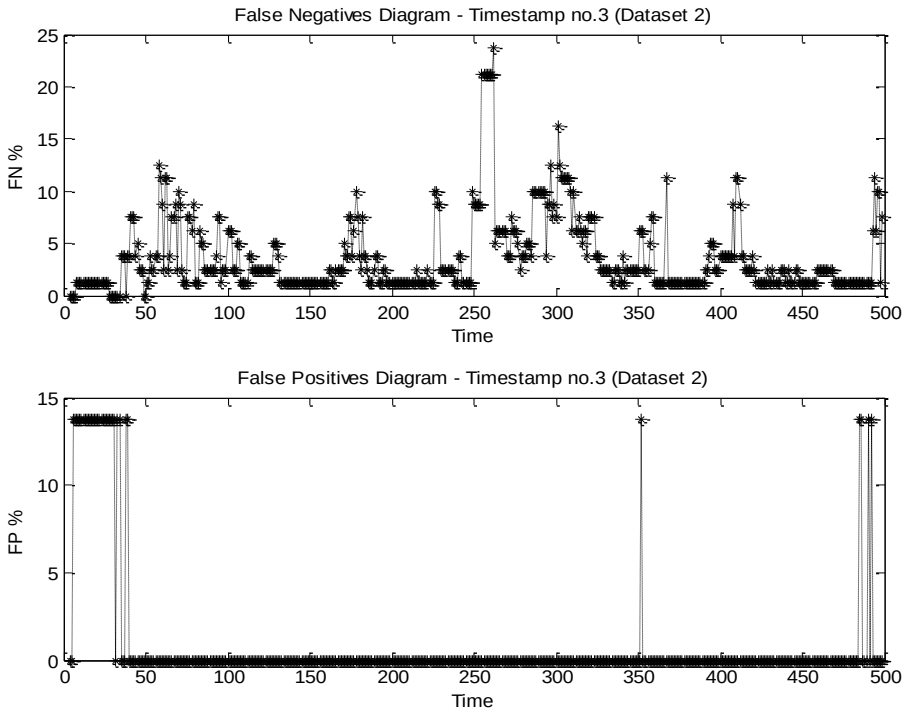
Στη παρούσα παράγραφο θα παρουσιαστούν τα πειραματικά αποτελέσματα για την αποτίμηση του προτεινόμενου Αλγορίθμου Αυτόματης Ομαδοποίησης και Εντοπισμού Ύποπτων Ομάδων όσον αφορά τα ποσοστά των επιτυχών και των μη-επιτυχών καταδείξεων, σε σχέση με τα σύνολα των δεδομένων που χρησιμοποιήθηκαν για τον υπολογισμό αυτών των μετρικών.

Πίνακας 2: Πίνακας Τιμών για FP και FN

Σύνολο Δεδομένων	Angle Deviation Ratio (adr)	Speed Deviation Ratio (sdr)	FP %	FN %
trace1	0,05	0,05	**1,767034**	**12,625872**
trace2	0,06	0,06	**3,7586775**	**13,498987**
trace3	0,08	0,07	**3,19006**	**16,362064**
trace4	0,08	0,08	**2,87506**	**18,282253**
trace5	0,09	0,08	**2,281533**	**20,019588**

Από τα αποτελέσματα των πειραμάτων γίνεται κατανοητό πως σαν μέθοδος κατάδειξης ύποπτων ομάδων, άρα και χρηστών λειτουργεί πολύ καλά καθώς το μεγαλύτερο ποσοστό αποτυχίας (FP) που σημειώθηκε για τα σύνολα δεδομένων που χρησιμοποιήθηκαν (όλα με $N = 80$) ήταν της τάξης του 3,76%. Αυτό μεταφράζεται στο ότι μόνο το 3,76% από το συνολικό πλήθος των μηνυμάτων που εστάλησαν συνολικά στο δίκτυο κατά τη διάρκεια της προσομοίωσης ήταν αχρείαστα να σταλούν, καθώς οι χρήστες βρίσκονταν ακόμα στην ίδια ομάδα με αυτή που τους ανιχνεύσαμε τη στιγμή που εφαρμόστηκε ο αλγόριθμος. Παρατηρείται ακόμη πως και το ποσοστό λανθασμένων μη καταδείξεων δεν είναι απαγορευτικό καθώς το μεγαλύτερο ποσοστό που συναντάμε είναι περίπου 20,02%. Δηλαδή, ένα 20,02% των συνολικών μηνυμάτων τελικά δε κατέληξαν στους χρήστες που θα έπρεπε κατά τη διάρκεια όλης της προσομοίωσης όμως. Και αυτό τονίζεται ιδιαιτέρως καθώς μέχρι στιγμής το σύστημα δεν έχει μηχανισμό παρακολούθησης της κίνησης των ομάδων που προέκυψαν από την ομαδοποίηση. Με την ενσωμάτωση ενός τέτοιου μηχανισμού υπάρχει η πεποίθηση

πως τα ποσοστά των FN θα μειωθούν αρκετά. Ακόμα κι έτσι όμως, αν δούμε και τα ποσοστά του φόρτου του δικτύου που επιτυγχάνονται με αυτή τη μέθοδο, δεν μπορούμε να πούμε πως απέτυχε το στόχο της. Αντιθέτως, εκπληρώνει αυτό για το οποίο κατασκευάστηκε προσφέροντας και πρόσφορο έδαφος για περαιτέρω βελτίωση. Τελειώνοντας, παρατηρείται πως όσο η κινητικότητα του συνόλου των δεδομένων που χρησιμοποιείται για την προσομοίωση αυξάνεται, τόσο αυξάνεται το ποσοστό των False Negatives (FN) και αντιστρόφως μειώνεται το ποσοστό των False Positives (FP). Φυσικά, αυτό είναι κάτι αναμενόμενο καθώς αυξημένη κινητικότητα σημαίνει πως οι ύποπτες ομάδες έχουν μεγάλη πιθανότητα να διασπαστούν στο εγγύς μέλλον, κάτι το οποίο όμως και για τις ομάδες που θεωρήθηκαν πιο «ήσυχες» από τον αλγόριθμό μας και δεν καταδείχτηκαν ως ύποπτες στο αποτέλεσμα της βέλτιστης ομαδοποίησης. Μια στιγμιαία αποτύπωση του πως λειτουργεί ο υπολογισμών των τιμών για τα FN και FP που γίνεται μέσω αυτή της προσομοίωσης δίνεται στην Εικόνα 24. Στη συγκεκριμένη



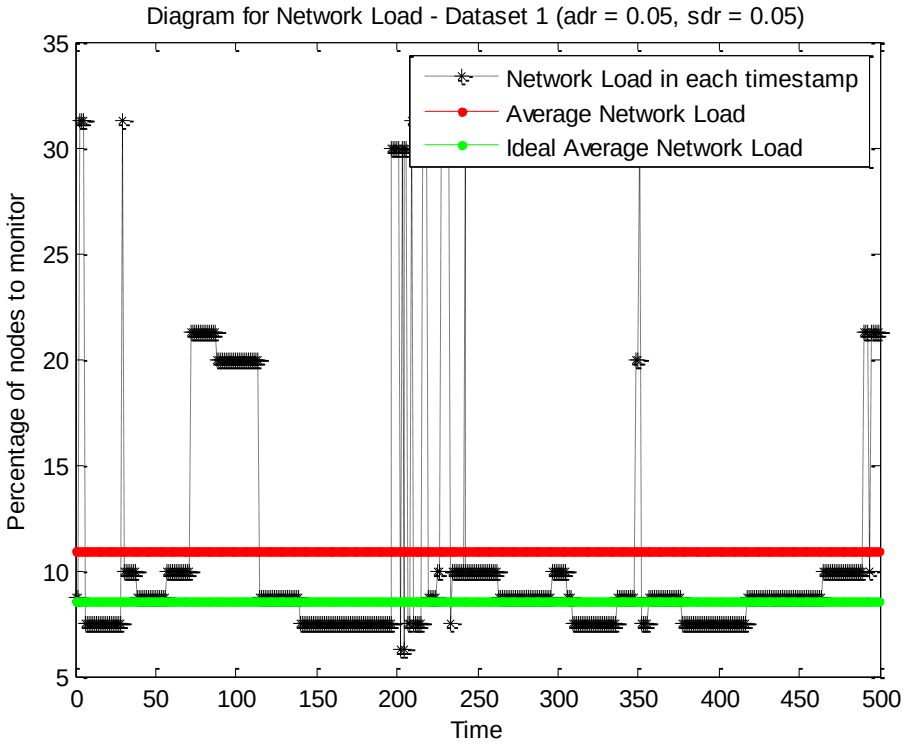
Εικόνα 24: Αποτίμηση των FN και FP για τη χρονική στιγμή t=3 της προσομοίωσης

εικόνα φαίνονται οι γραφικές παραστάσεις για τις τιμές που παίρνουν τα FP και FN αν ελέγχουμε το αποτέλεσμα της ομαδοποίησης που είχαμε τη χρονική στιγμή $t = 3$ της προσομοίωσής μας και το συγκρίνουμε με τα αποτελέσματα των ομαδοποιήσεων των

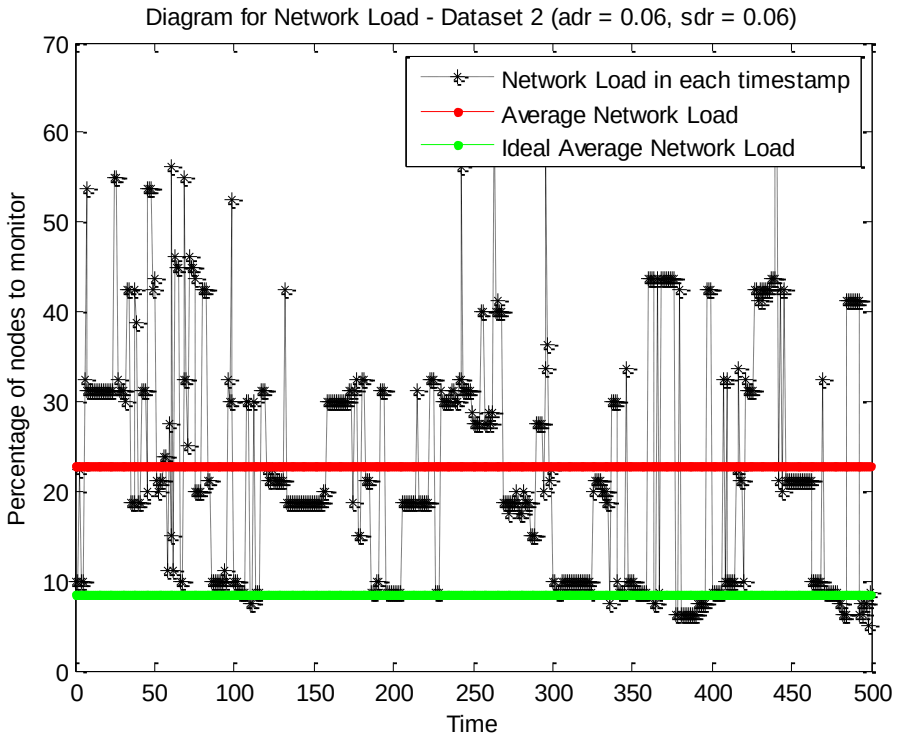
χρονικών στιγμών $t = 4, \dots, N_s$. Στο συγκεκριμένο παράδειγμα το σύνολο των δεδομένων που χρησιμοποιείται είναι το 2ο όπου $\mathrm{adr} = 0.06$ και $\mathrm{sdr} = 0.06$ βάσει του Πίνακα 1. Παρατηρούμε λοιπόν τις μεταβολές των ποσοστών για τα False Negative που κατά κύριο λόγο βρίσκονται κάτω από το 15% καθώς και για τα False Positive που όπως φαίνεται έχουμε πολύ μικρό αριθμό λανθασμένων καταδείξεων και αυτές εμφανίζονται μόνο κάποιες από τις επόμενες χρονικές στιγμές, ενώ την περισσότερη περίοδο φαίνεται πως οι ομάδες που καταδείχτηκαν ως ύποπτες δε διατηρούν τη δομή τους.

4.2.2 Αξιολόγηση Φόρτου Δικτύου

Στην παρούσα παράγραφο θα παρουσιαστούν τα αποτελέσματα της αποτίμησης του προτεινόμενου αλγορίθμου ως προς την επιρροή του στο φόρτο του δικτύου, όπως ορίζεται από τις μετρικές των σχέσεων (4.3) και (4.4). Στα διαγράμματα των Εικόνων 25, 26 και 27 που ακολουθούν, βλέπουμε για κάθε ένα από τα σύνολα δεδομένων που χρησιμοποιήθηκαν, το φόρτο του δικτύου με την εφαρμογή του αλγορίθμου μας σε κάθε χρονική στιγμή της προσομοίωσης S. Ταυτόχρονα, σε κάθε ένα από αυτά τα διαγράμματα, υπάρχει το μέσο ποσοστό φόρτου στο δίκτυο (Average Network Load - ANL) όπως αυτό ορίστηκε από τη σχέση (4.4) και εμφανίζεται με τη βοήθεια της κόκκινης γραμμής. Παρατηρούμε λοιπόν, πως υπολογίζοντας την τιμή του ANL για κάθε ένα από τα σύνολα των δεδομένων μας, έχουμε κέρδος με την εφαρμογή του συγκεκριμένου σχήματος που προτείνεται καθώς χρειάζεται, κατά μέσο όρο, να στέλνουμε ένα ποσοστό μηνυμάτων μικρότερο του 30% όπως φαίνεται. Αυτό όμως ήταν κάτι αναμενόμενο καθώς η ενσωμάτωση ομαδοποίησης στο προτεινόμενο σχήμα θα μείωνε κατά πολύ την αποστολή των μηνυμάτων από την υπηρεσία προς τους χρήστες καθώς κατηγοριοποιώντας τους σε ομάδες χρειάζεται να στείλει μόνο στους εκπροσώπους των ομάδων το μήνυμα και έπειτα εκείνοι αναλαμβάνουν τοπικά τη διανομή του εντός της ομάδας που εκπροσωπούν. Αν λοιπόν ιδανικά μπορούσαμε να εκτελούμε έναν αλγόριθμο αυτόματης ομαδοποίησης σε κάθε χρονική στιγμή, χωρίς αυτός να περιλαμβάνει κάποιο μηχανισμό κατάδειξης ύποπτων ομάδων, το μέσο ποσοστό φόρτου του δικτύου μας θα ήταν το χαμηλότερο δυνατό, όπως φαίνεται και από την πράσινη γραμμή στα διαγράμματα του φόρτου δικτύου για τα σύνολα των δεδομένων μας. Εντούτοις, όπως έχει προαναφερθεί, το να εκτελούμε έναν ιεραρχικό αλγόριθμο αυτόματης ομαδοποίησης κάθε χρονική στιγμή δεν αποτελεί αποδοτική λύση για την παρακολούθηση κινούμενων ομάδων καθώς αυξάνεται αρκετά η υπολογιστική

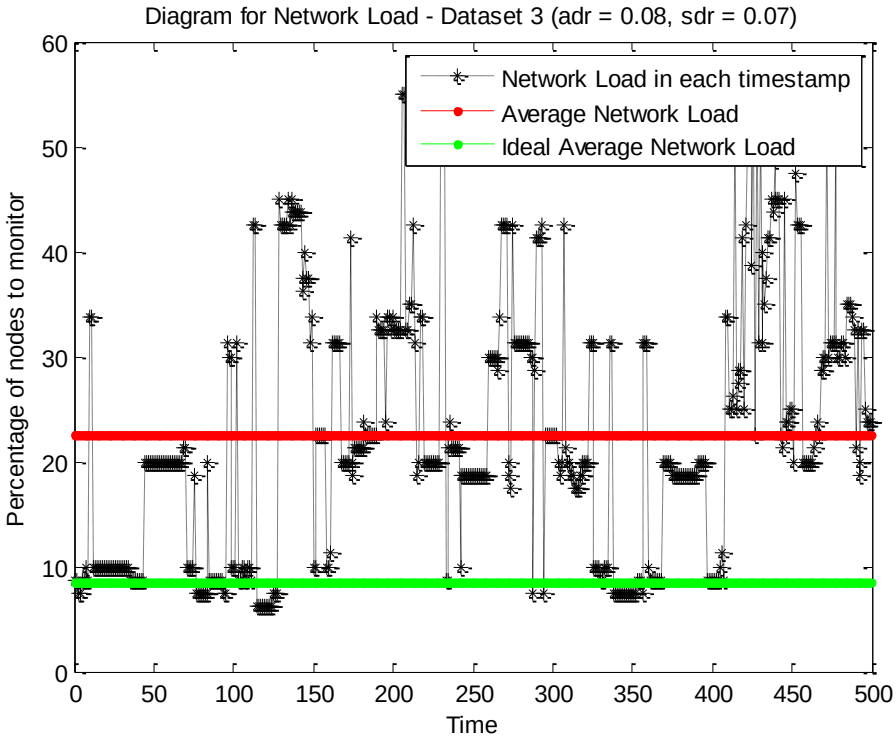


(α)

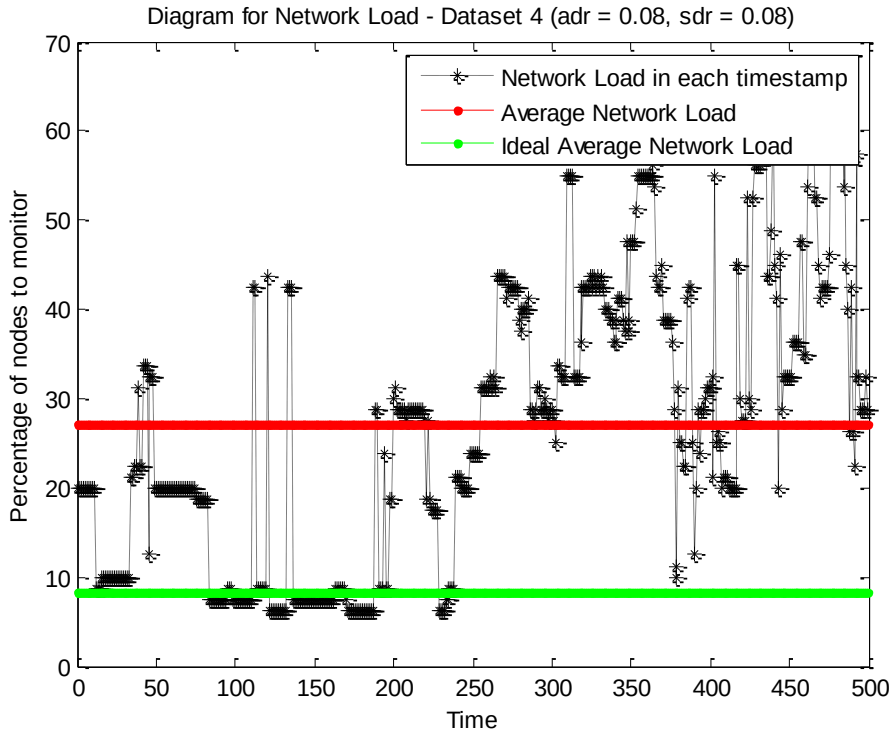


(β)

Εικόνα 25: Διαγράμματα Φόρτου Δικτύου (Μέρος Α)

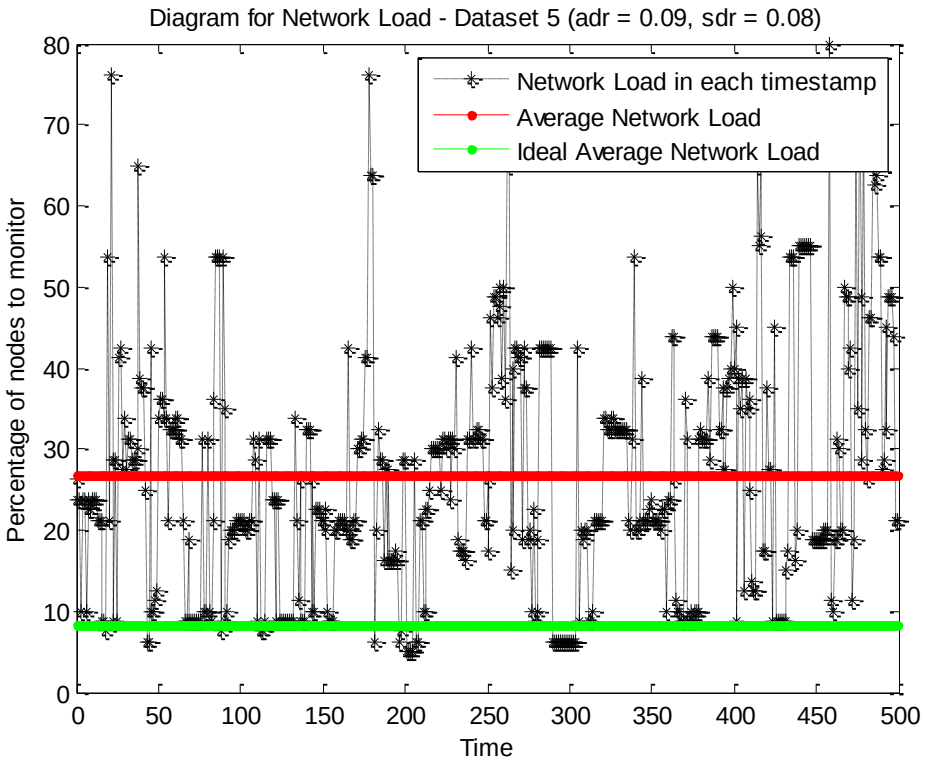


(α)



(β)

Εικόνα 26: Διαγράμματα Φόρτου Δικτύου (Μέρος Β)



Εικόνα 27: Διαγράμματα Φόρτου Δικτύου (Μέρος Γ)

πολυπλοκότητα. Το σχήμα μας δηλαδή, μπορεί να μην πετυχαίνει τη μέγιστη δυνατή μείωση στο φόρτο του δικτύου, αλλά μέσω της κατάδειξης των ύποπτων ομάδων προσφέρει μια πολύ καλύτερη βάση στην οποία θα μπορέσει να πατήσει ο μηχανισμός παρακολούθησης των δημιουργηθέντων ομάδων έτσι ώστε να είναι πιο αποδοτικός όσον αφορά την απώλεια πληροφορίας εντός του δικτύου. Βεβαίως, γνωρίζοντας πως η χειρότερη περίπτωση για το φόρτο του δικτύου είναι να στέλνουμε μήνυμα σε όλους της χρήστες, δηλαδή φόρτο της τάξης του 100%, το κέρδος που επιτυγχάνεται με τον προτεινόμενο σχήμα είναι κάτι παραπάνω από ικανοποιητικό, αν αναλογιστούμε τα θετικά του αλγορίθμου και τα ποσοστά των FP και FN που παρουσιάστηκαν παραπάνω. Έχουμε επομένως μείωση στην αποστολή πληροφορίας διατηρώντας τα ποσοστά απώλειας πληροφορίας σε ανεκτά επίπεδα, ειδικά για ένα σύστημα που δεν έχει κάποιο μηχανισμό παρακολούθησης. Τέλος, μια παρατήρηση που προκύπτει από το διάγραμμα (β) της Εικόνας 26 (που αφορά το σύνολο δεδομένων με τίτλο “trace4”) είναι πως όσο η προσομοίωση προχωράει και το σύνολο των δεδομένων μας γίνεται πιο κινητικό (λόγω της τιμής των adr και sdr) τόσο αυξάνεται και ο φόρτος του δικτύου, κάτι που είναι απολύτως λογικό. Λόγω του χαμηλού φόρτου που έχουμε στην αρχή όμως, τα επίπεδα του ANL διατηρούνται και πάλι κάτω από το 30%.

4.2.3 Αξιολόγηση Περιόδου Επιτυχημένης Κατάδειξης

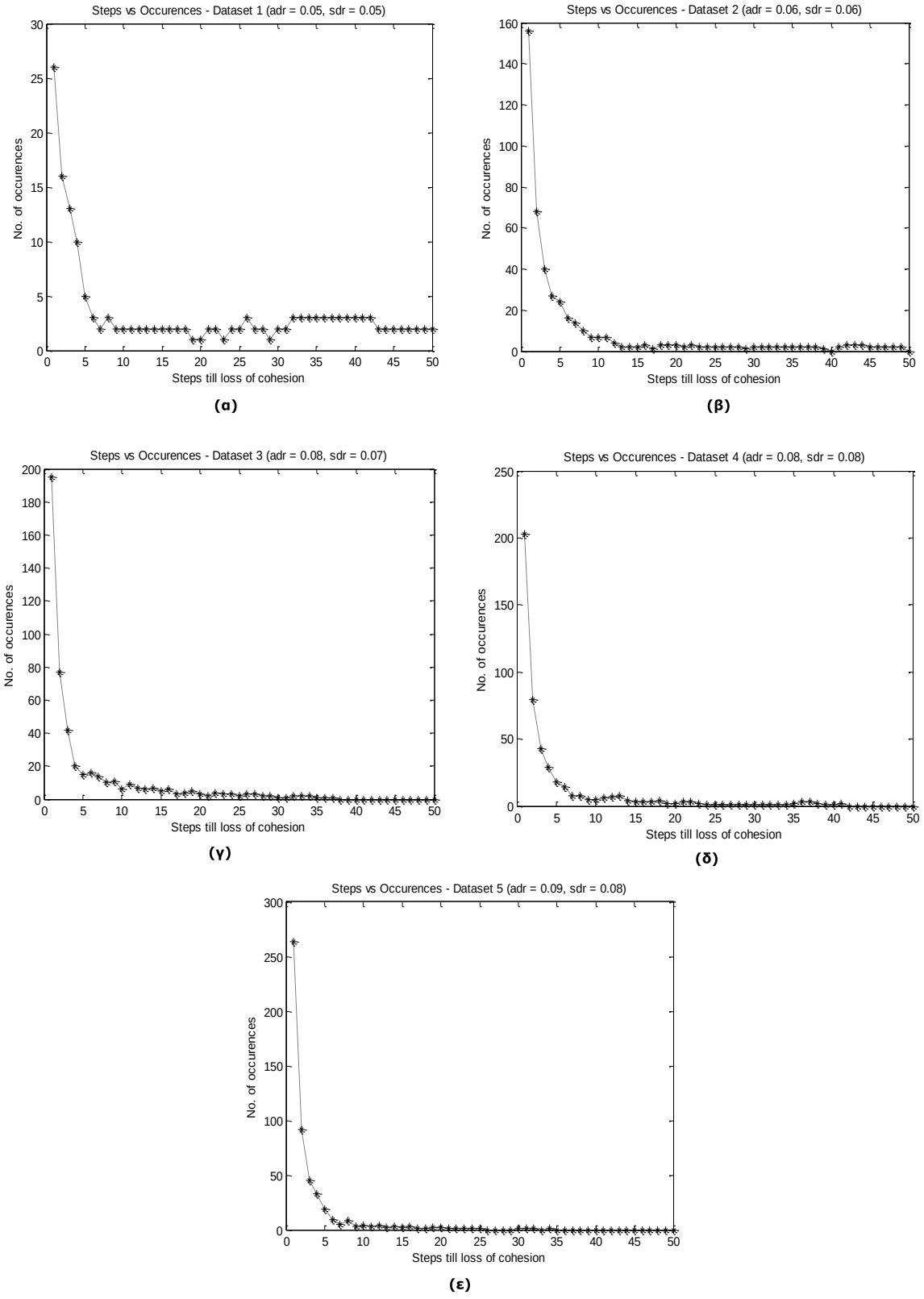
Για να μπορέσουμε να δούμε αν το σύστημά μας επιτυχώς εντοπίζει ύποπτες ομάδες, θα πρέπει να ελέγξουμε σε πόσο χρονικό διάστημα, κατά μέσο όρο, το σύνολο των ύποπτων ομάδων που εντοπίζονται τελικά διασπόνται αλλάζοντας τη δομή τους. Αν η τιμή που θα προκύψει για το APSA (*Average Period for Successive Annotation*) βάση της σχέσης (4.7) είναι μικρή συγκριτικά με τις χρονικές στιγμές N_s της προσομοίωσης S, αυτό δείχνει πως η κατάδειξη που γίνεται είναι αποδοτική καθώς εντοπίζει ομάδες που πράγματι σύντομα διασπόνται, ενώ αν η τιμή που θα προκύψει είναι σχετικά μεγάλη σημαίνει πως η κατάδειξη δεν είναι τόσο ισχυρή και ίσως θα έπρεπε να εξεταστεί η εισαγωγή ενός κατωφλίου για τη διαφορά μεταξύ των τιμών του κέρδους ομαδοποίησης δύο διαφορετικών επιπέδων στην ιεραρχία των ομαδοποιήσεων του συγχωνευτικού ιεραρχικού αλγορίθμου που χρησιμοποιείται ως μέθοδος για την κατάδειξη των ύποπτων συγχωνεύσεων άρα και ομάδων. Έχουμε λοιπόν τον ακόλουθο πίνακα για τις τιμές που λαμβάνει αυτή η μετρική για τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την αποτίμηση του συστήματος ως προς αυτή την παράμετρο.

Πίνακας 3: Πίνακας Τιμών Μέσης Περιόδου Επιτυχούς Κατάδειξης (APSA)

Σύνολο Δεδομένων	Angle Deviation Ratio (adr)	Speed Deviation Ratio (sdr)	*APSA*
trace1	0,05	0,05	**24**
trace2	0,06	0,06	**8**
trace3	0,08	0,07	**7**
trace4	0,08	0,08	**6**
trace5	0,09	0,08	**3**

Πρέπει να σημειώσουμε εδώ πως όλα τα σύνολα δεδομένων που παράχθηκαν μέσω του MobiSim έχουν τιμή $N_s = 500$. Παρατηρούμε λοιπόν πως για τα πιο «ήσυχα» σύνολα δεδομένων η κατάδειξη δεν είναι τόσο ισχυρή (δεν θα μπορούσε όμως να χαρακτηριστεί και ως αποτυχημένη) αλλά όσο η κινητικότητα αυξάνεται, τόσο και η διάρκεια της περιόδου που οι ομάδες παραμένουν συνεκτικές, από τη στιγμή της κατάδειξής τους ως ύποπτες μέχρι και τη διάσπασή τους, μειώνεται. Αυτά που δόθηκαν ως πληροφορία στον Πίνακα 3 μπορούν να αποτυπωθούν και από τα ακόλουθα διαγράμματα της Εικόνας 26 που ουσιαστικά αποτελούν ιστογράμματα της συχνότητας εμφάνισης μιας συγκεκριμένης περιόδου επιτυχούς κατάδειξης, όπως αυτή έχει οριστεί

στην παράγραφο 4.1.1, για την πειραματική αξιολόγηση καθενός από τα παραπάνω σύνολα δεδομένων όσον αφορά τη συγκεκριμένη μετρική. Όπως φαίνεται λοιπόν, τη συχνότερη εμφάνιση έχουμε για καταδείξεις ομάδων όπου την επόμενη χρονική στιγμή χαλάνε τη συνοχή τους. Αυτό ισχύει για όλα τα σύνολα των δεδομένων που εξετάστηκαν και παρατηρείται επίσης το γεγονός πως όσο αυξάνεται η κινητικότητα των ομάδων, τόσο μεγαλύτερο είναι το πλήθος των καταδείξεων ως ύποπτες καθώς και ο αριθμός της εμφάνισης της περιόδου διάρκειας ίσης με μία χρονική στιγμή. Αυτό λοιπόν που μπορούμε να εξάγουμε συμπερασματικά από τα συγκεκριμένα πειράματα είναι πως δημιουργήσαμε μια μέθοδο η οποία, χωρίς να έχει πρότερη γνώση για την κατάσταση των κόμβων των ομάδων (αν πχ κάποιοι από τους κόμβους ήταν για μεγάλη περίοδο στο παρελθόν ύποπτοι), μπορεί και εντοπίζει με μεγάλη επιτυχία και καλή ευαισθησία ομάδες που τελικά στη συνέχεια αλλάζουν τη δομή τους (προφανώς λόγω κάποιας διάσπασης των μελών τους).



Εικόνα 28: Διαγράμματα Μέσης Περιόδου Επιτυχούς Κατάδειξης

4.2.4 Αξιολόγηση Μέσου Χρόνου Εκτέλεσης

Έχουμε μέχρι στιγμής έναν αλγόριθμο, που βάσει των προηγούμενων πειραμάτων επιτυγχάνει τους στόχους του αναφορικά με την επιτυχή κατάδειξη ύποπτων ομάδων καθώς και τη μείωση του φόρτου του δικτύου, κρατώντας σε ανεκτά επίπεδα την απώλεια πληροφορίας στο δίκτυο. Ποιος είναι όμως ο χρόνος εκτέλεσης ενός τέτοιου σχήματος; Γνωρίζουμε πως οι ιεραρχικοί αλγόριθμοι έχουν το αρνητικό της μεγάλης υπολογιστικής πολυπλοκότητας το οποίο αντισταθμίζεται από τα πολλά άλλα θετικά που προσφέρουν ως αλγόριθμοι και τα οποία εκμεταλλευτήκαμε για τη δημιουργία του συστήματός μας. Ας δούμε λοιπόν σε αυτό το σημείο και τους πραγματικούς χρόνους εκτέλεσης που επιτυγχάνει το σύστημά μας για τα σύνολα των δεδομένων που παρουσιάστηκαν στη παράγραφο 4.1.3 έτσι ώστε να μπορέσουμε να εξάγουμε ασφαλή συμπεράσματα για το τι ταχύτητες είναι ικανό να πετύχει το σύστημά μας, βασιζόμενο σε έναν συγχωνευτικό ιεραρχικό αλγόριθμο.

Πίνακας 4: Πίνακας Τιμών Μέσου Χρόνου Εκτέλεσης του Αλγορίθμου (AET)

Σύνολο Δεδομένων	Πλήθος Κόμβων (N)	*AET (ms)*
trace1	80	**14,286**
trace6	160	**104,04**
trace7	240	**349,428**
trace8	400	**2176,186**

Όπως ήταν αναμενόμενο λοιπόν, βάσει των αποτελεσμάτων που παρουσιάζονται στον Πίνακα 4, με την αύξηση του πλήθους των χρηστών, αυξάνεται αρκετά και ο χρόνος εκτέλεσης για τον Αλγόριθμο Αυτόματης Ομαδοποίησης και Εντοπισμού Ύποπτων Ομάδων. Η αύξηση αυτή είναι μάλιστα εκθετική, όπως αναμενόταν και λόγω της πολυπλοκότητας των συγχωνευτικών ιεραρχικών αλγορίθμων. Φαίνεται λοιπόν, πως το προτεινόμενο σύστημα μπορεί να αντιμετωπίσει αρκετά αποδοτικά ακόμα και περιπτώσεις με σχετικά μεγάλο πλήθος κόμβων N. Αν για παράδειγμα αναλογιστούμε ένα σενάριο στο οποίο έχουμε μια υπηρεσία που βασίζεται σε πληροφορία θέσης και θέλει να ενημερώνει τους χρήστες που βρίσκονται σε μια αίθουσα ενός μουσείου, τότε υπό φυσιολογικές συνθήκες, δεν θα χρειαστεί να αντιμετωπίσει καταστάσεις με πλήθος χρηστών μεγαλύτερο των 200. Ακόμα και κάθε χρονική στιγμή να χρειάζεται να εκτελείται το προτεινόμενο σχήμα, θα προλαβαίνει να ομαδοποιήσει τα δεδομένα, να

καταδείξει τις ύποπτες ομάδες και να κάνει την απαραίτητη αποστολή των πληροφοριών. Εν κατακλείδι λοιπόν, μπορούμε να πούμε πως το σύστημά μας είναι αρκετά αποδοτικό και από την οπτική πλευρά του χρόνου εκτέλεσης, όσο το πλήθος των χρηστών διατηρείται σε φυσιολογικά επίπεδα.

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΑΝΟΙΚΤΑ ΘΕΜΑΤΑ

5.1 Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκε η αρχιτεκτονική και οι λεπτομέρειες υλοποίησης ενός σχήματος αυτόματης ομαδοποίησης και εντοπισμού ύποπτων ομάδων στον τομέα των υπηρεσιών βάσει πληροφορίας θέσης. Αρχικά, παρουσιάστηκαν τα γενικά χαρακτηριστικά τέτοιων συστημάτων καθώς και ποιοι είναι οι στόχοι τους, τι προσπαθούν να αντιμετωπίσουν, ενώ στη συνέχεια δόθηκαν πληροφορίες σχετικά με διάφορες προσεγγίσεις που έχουν γίνει από την ερευνητική κοινότητα πάνω σε αυτό το χώρο. Ιδιαίτερη έμφαση δόθηκε στη χρήση της ομαδοποίησης σε τέτοιου είδους υπηρεσίες, οπότε έγινε μια αναφορά στο τι περιλαμβάνει η συγκεκριμένη ερευνητική περιοχή. Λόγω του ότι θέλαμε ένα σύστημα εντελώς αυτόνομο, επιλέχθηκαν οι ιεραρχικοί αλγόριθμοι ομαδοποίησης για την εκπλήρωση αυτού του στόχου για τους οποίους δόθηκαν τα κύρια χαρακτηριστικά τους καθώς και οι ιδιαιτερότητές τους. Ακολούθησε η παρουσίαση ενός κριτηρίου εύρεσης της βέλτιστης ομαδοποίησης για τους συγχωνευτικούς ιεραρχικούς αλγορίθμους, όπως αυτό προτείνεται στη βιβλιογραφία [18], καθώς και μιας μεθόδου εντοπισμού ύποπτων ομάδων στο αποτέλεσμα της ομαδοποίησης. Τέλος, μετά την παρουσίαση της αρχιτεκτονικής του σχήματός μας και των χαρακτηριστικών του ακολούθησε η πειραματική αποτίμησή του.

Η συνεισφορά της συγκεκριμένης διπλωματικής εργασίας συνοψίζεται στα εξής:

- Παρουσίαση ενός ιεραρχικού σχήματος ομαδοποίησης για τη χρήση του στις υπηρεσίες βάσει πληροφορίας θέσης για τη μείωση του φόρτου που εμφανίζεται στο δίκτυο από πλευράς αποστολής μηνυμάτων
- Ενσωμάτωση ενός κριτηρίου εύρεσης της βέλτιστης ομαδοποίησης έτσι ώστε να είναι δυνατός ο αυτόματος εντοπισμός της καλύτερης από την ιεραρχία των ομαδοποιήσεων που προσφέρει ένας συγχωνευτικός ιεραρχικός αλγόριθμος. Με αυτό τον τρόπο το σχήμα μας είναι σε θέση να εντοπίζει αυτόματα τις λογικότερες ομάδες εντός μιας περιοχής που παρακολουθεί, βασιζόμενο στην πληροφορία θέσης των χρηστών
- Αξιοποίηση της φυσικής σημασία του μέτρου «Κέρδος Ομαδοποίησης Δ» για τον εντοπισμό ομάδων στο αποτέλεσμα της βέλτιστης ομαδοποίησης που

θεωρούνται ύποπτες υπό την έννοια πως παρουσιάζουν μεγάλη πιθανότητα στο εγγύς μέλλον να χαλάσουν τη δομή τους, λόγω κινητικότητας των χρηστών τους

- Δημιουργία ποιοτικής πληροφορίας σχετικά με την κατάσταση των ομάδων που εντοπίστηκαν αυτόματα από το συγχωνευτικό ιεραρχικό σχήμα ομαδοποίησης, έτσι ώστε να ενισχυθεί η βάση γνώσης του μηχανισμού παρακολούθησης αυτών των ομάδων, με στόχο την διατήρηση της απώλειας πληροφορίας σε ανεκτά επίπεδα

5.2 Ανοικτά Θέματα

Στα πλαίσια της παρούσας εργασίας παρουσιάστηκαν κάποια ανοικτά θέματα. Πιο συγκεκριμένα, αναφέρθηκε αρκετές φορές πως στις υπηρεσίες βάσει πληροφορίας θέσης, το ιδανικό είναι να υπάρχει ένας μηχανισμός παρακολούθησης των ομάδων έτσι ώστε να μην χρειάζεται συνεχώς η εκτέλεση του αλγορίθμου ομαδοποίησης. Αν και η προσέγγισή μας προσφέρει έναν αυτόματο σχήμα εντοπισμού των ομάδων, με τρόπο που περιγράφηκε στο Κεφάλαιο 3, εντούτοις δεν είναι σε θέση να παρακολουθήσει τις ομάδες που δημιουργούνται, έτσι ώστε να συνεισφέρει και σε αυτό το επίπεδο. Η μόνη προσφορά του σχήματός μας σε σχέση με τη φάση της παρακολούθησης είναι η πληροφορία που δίνει σχετικά με το αν κάποιες ομάδες στο αποτέλεσμα ομαδοποίησης που παράγεται είναι ύποπτες, έτσι ώστε στο στάδιο της παρακολούθησης να στέλνονται συνεχώς οι πληροφορίες σε όλους τους χρήστες αυτών των ομάδων και να ελέγχονται και ως προς τη συνοχή τους για πιθανή διάσπαση. Προτείνεται επομένως η ανάπτυξη ενός μηχανισμού ο οποίος θα παρακολουθεί την κίνηση των ομάδων που προέκυψαν από τον παραπάνω αλγόριθμο για πιθανές διασπάσεις (ομάδες εντοπισμένες ως ύποπτες) ή επικαλύψεις, δημιουργώντας έτσι ένα πλήρως αυτοματοποιημένο σύστημα ομαδοποίησης και παρακολούθησης χρηστών σε κλειστούς χώρους. Σε περίπτωση που ο μηχανισμός ανιχνεύσει κάποια διάσπαση ύποπτης ομάδας ή συγχώνευση (επικάλυψη) ομάδων, τότε θεωρεί πως η ομαδοποίηση που γνωρίζει δεν είναι πια έγκυρη και εκτελεί εκ νέου τον Αλγόριθμο Αυτόματης Ομαδοποίησης και Εντοπισμού Ύποπτων Ομάδων.

Υπάρχουν όμως και περιπτώσεις κατάρρευσης της εγκυρότητας της ισχύουσας ομαδοποίησης που, ακόμα και με την ενσωμάτωση ενός μηχανισμού σαν τον παραπάνω, στη φάση της παρακολούθησης δεν είναι «ορατές». Τέτοια είναι η διάσπαση μιας ομάδας η οποία δεν είχε καταδειχτεί ως ύποπτη από το μηχανισμό μας, που είναι μια περίπτωση False Negative (βλ. Εικόνα 22). Για το λόγο αυτό προτείνεται ο

ορισμός ενός «παραθύρου επαναομαδοποίησης». Πιο συγκεκριμένα, θεωρώντας πως το παράθυρο αυτό συμβολίζεται με T, αν εντός T χρονικών στιγμών παρακολούθησης, μετρώντας από το στάδιο της ομαδοποίησης και μετά, το σύστημα δεν εντοπίσει κάποιο γεγονός το οποίο αλλάζει τη μορφή της ισχύουσας ομαδοποίησης, τότε θα περνάει και πάλι σε φάση ομαδοποίησης για να καλύψει περιπτώσεις σαν αυτή που αναφέρθηκε. Με αυτό τον τρόπο αυξάνεται μεν η πολυπλοκότητα, αλλά θα υπάρχει μείωση των False Negatives συνολικά, άρα και βελτίωση όσον αφορά την απώλεια πληροφορίας.

Τέλος, ένα αρνητικό της κατάδειξης που προσφέρει το σύστημά μας είναι πως το πραγματοποιεί λαμβάνοντας υπόψη του μόνο τη τρέχουσα θέση των ομάδων στη φάση της ομαδοποίησης, χωρίς να κάνει χρήση κάποιου μηχανισμού μνήμης. Θα ήταν καλύτερο η κατάδειξη των ύποπτων ομάδων να γίνεται κάνοντας χρήση πληροφορίας σχετική με του τι είδους ομάδα ήταν στο παρελθόν η ομάδα που πάει να καταδείξει ως ύποπτη το σύστημά μας αυτή τη χρονική στιγμή. Αν είχαμε δηλαδή μια ομάδα η οποία μετά από μια προηγούμενη κατάδειξη όντως έχει αλλάξει δομή, τότε πρέπει να την καταδείξουμε ως ύποπτη και τώρα. Αντιθέτως, αν η ομάδα αυτή έχει επισημανθεί στο παρελθόν ως ύποπτη αλλά τελικά διατηρούσε τη δομή της για αρκετό διάστημα, καλό θα είναι να περιμένει λίγο το σύστημά μας προτού την καταδείξει ως ύποπτη. Να ενσωματωθεί δηλαδή, μια φάση εκπαίδευσης στο σύστημά μας, που θα ενισχύει ακόμα περισσότερο την κατάδειξη που προσφέρει.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος Όρος	Ελληνικός Όρος
A priori	Πρότερη
Additivity	Προσθετικότητα
Agglomerative	Συγχωνευτικός
Annotation	Κατάδειξη
Average Linkage	Σύνδεση Μέσου Όρου
Centroid	Κεντροειδές
Centroid Linkage	Σύνδεση Κεντροειδών
Classification	Κατηγοριοποίηση
Cluster Analysis	Ανάλυση κατά Συστάδες
Cluster Head	Κεντροειδές
Cluster Tendency	Τάση των Ομάδων
Clustering	Ομαδοποίηση
Clustering Balance	Ισορροπία Ομαδοποίησης
Clustering Examples	Ομαδοποίηση Παραδειγμάτων
Clustering Feature	Χαρακτηριστικό Ομαδοποίησης
Clustering Gain	Κέρδος Ομαδοποίησης
Clustering Validity	Εγκυρότητα Ομαδοποίησης
Clustering Variables	Ομαδοποίηση Μεταβλητών
Clusters	Ομάδες
Compactness of Representation	Συμπαγής Αναπαράσταση
Complete Linkage	Πλήρης Σύνδεση
Confirmatory Procedures	Επιβεβαιωτικές Διαδικασίες
Data Abstraction	Αφαίρεση Δεδομένων
Data Analysis	Ανάλυση Δεδομένων
Data Streams	Ροές Δεδομένων
Datasets	Σύνολα Δεδομένων
Dendrogram	Δενδρόγραμμα
Density-Based	Βάσει Πυκνότητας
Dissimilarity Measures	Μέτρα Ανομοιότητας
Divisive	Διαιρετικός
Expansion	Επέκταση
Exploratory Procedures	Διερευνητικές Διαδικασίες
External Assessment of Validity	Εξωτερική Αξιολόγηση Εγκυρότητας

Feature Extraction	Εξαγωγή Χαρακτηριστικού
Feature Selection	Επιλογή Χαρακτηριστικού
Graph Theory	Θεωρία Γράφων
Grid-Based	Βάσει Πλέγματος
Group Leader	Ηγέτης Ομάδας
Hierarchical Algorithms	Ιεραρχικοί Αλγόριθμοι
Hierarchical Clustering	Ιεραρχική Ομαδοποίηση
Identity Matrix	Μήτρα Ταυτότητας
Incrementality	Επαύξηση
Inter-Cluster	Μεταξύ των Ομάδων
Internal Assessment of Validity	Εσωτερική Αξιολόγηση Εγκυρότητας
Intra-Cluster	Εντός της Ομάδας
Label	Ετικέτα
Linkage Metrics	Μετρικές Σύνδεσης
Location Based Services	Εφαρμογές Βάσει Πληροφορίας Θέσης
Micro-Clustering	Μικρο-ομαδοποίηση
Micro-Clusters	Τοπικά Μοντέλα
Mobility	Κινητικότητα
Model-Based	Βάσει Μοντέλου
Monitoring	Παρακολούθηση
Moving Objects	Κινούμενα Αντικείμενα
Network Gain	Κέρδος Δικτύου
Network Load	Φόρτο Δικτύου
Normalized	Κανονικοποιημένα
Outliers	Ακραίες Τιμές
Partitioning Algorithms	Αλγόριθμοι Κατάτμησης
Pattern Proximity	Εγγύτητα Προτύπων
Pattern Representation	Αναπαράσταση Προτύπων
Patterns	Πρότυπα
Probabilistic	Πιθανοτικό
Proximity Matrix	Μήτρα Εγγύτητας
Proximity Measures	Μέτρα Εγγύτητας
Radius	Ακτίνα
Relative Test	Σχετικός Έλεγχος
Similarity Measures	Μέτρα Ομοιότητας
Single Linkage	Μονή Σύνδεση

Supervised Classification	Εποπτευόμενη Κατηγοριοποίηση
Tracking Algorithm	Αλγόριθμος Εντοπισμού
Transpose	Μετατόπιση
Unsupervised Classification	Μη Εποπτευόμενη Κατηγοριοποίηση
User Ids	Ταυτότητες Χρηστών
Valid Group	Έγκυρη Ομάδα

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ - ΑΚΡΩΝΥΜΙΑ

adr	Angle Deviation Ratio
AET	Average Execution Time
AHCASCAM	Automated Hierarchical Clustering Algorithm with Suspicious Clusters Annotation Mechanism
ANG	Average Network Gain
ANL	Average Network Load
APSA	Average Period for Successive Annotation
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
BS	Base Station
C2P	Clustering Based on Closest Pairs
CF	Clustering Feature
CURE	Clustering Using Representatives
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
ET	Execution Time
FN	False Negative
FP	False Positive
GPS	Global Positioning System
KNN	K - Nearest Neighbors
LBS	Location Based Services
LEACH	Low Energy Adaptive Clustering Hierarchy
MinPts	Minimum Points
MobiSim	Mobility Simulator
ms	Milliseconds
MUA	Matrix Updating Algorithm
NG	Network Gain
NL	Network Load
ODAC	Online Divisive and Agglomerative Clustering
OPTICS	Ordering Points to Identify the Clustering Structure
RPGM	Reference Point Group Mobility
RWP	Random Waypoint
sdr	Speed Deviation Ratio
VANETs	Vehicular Ad-Hoc Networks

ΑΕΜ	Αλγόριθμος Ενημέρωσης Μήτρας
ΓΣΣ	Γενικευμένο Συγχωνευτικό Σχήμα
ΕΚΠΑ	Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΑΝΑΦΟΡΕΣ

[1] A.K. Jain and M.N. Murty and P.J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol.31 no.3, pp. 264-323, 1999

[2] A.K. Jain and R.C. Dubes, “Algorithms for Clustering Data”, Prentice Hall, Englewood Cliffs, NJ, 1988

[3] M. Abramowtiz and I.A. Stegun, “Handbook of Mathematical Functions with Formulas, Graphics and Mathematical Table”, US Govt. Printing Office, Washington, D.C., 1968

[4] N. Jardine and C.J. Rijsbergen, “The Use of Hierarchical Clustering in Information Retrieval”, Information Storage and Retrieval, Vol.7, pp. 217-240, 1971

[5] D. Barbara, “Requirements for Clustering Data Streams”, 2002

[6] Chernoff, H., “A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations”, Annals of Mathematical Statistics, Vol. 23, pages 493-509, 1952

[7] Kaufman, L., Rousueeuw, P., “Finding Groups in Data: an Introduction to Cluster Analysis”, John Wiley and Sons (1990)

[8] Ng, R.T., Han, J., “Efficient and effective clustering methods for spatial data mining”, In: Proc. of VLDB (1994)

[9] Zhang, T., Ramakrishnan, R., Linvy, M., “BIRCH: An efficient data clustering method for very large databases”, In: Proc. of ACM SIGMOD (1996)

[10] Guha, S., Rastogi, R., Shim, K., “CURE: An efficient clustering algorithm for large databases”, In: Proc. of ACM SIGMOD (1998)

[11] Nanopoulos, A., Theodoridis, Y., Manolopoulos, Y., “C2P: Clustering based on closest pairs”, In: Proc. of VLDB (2001)

[12] Ankerst, M., Breunig, M., Kriegel, H.P., Sander, J., “OPTICS: ordering points to identify the clustering structure”, In: Proc. of ACM SIGMOD (1999) 49-60

[13] Vlachos, M., Kollios, G., Gunopoulos, D., “Discovering similar multidimensional trajectories”, In: Proc. of ICDE (2002) 673-684

[14] Y. Li, J. Han, J. Yang, “Clustering moving objects”, Proc. tenth ACM SIGKDD international conference of Knowledge discovery and data mining pp. 617-622, 2004

[15] C.S. Jensen, D. Lin, B. Chin Ooi, “Continuous Clustering of Moving Objects”, IEEE trans. Knowledge and Data Engineering, 19(9): 1161-1174, 2007

[16] G. Homans, “The Human Group”, Transaction Pub., New York, Reprint 2001 (orig. 1950)

[17] T. Zhu, Y. Zhang, F. Wang, W. Lv, “A location-based push service architecture with clustering method”, Network Computing and Advanced Information Management (NCM), 6th Intl., Conf., pp.107-112, 2012

[18] Y. Jung, H. Park, and D.Z. Du, “A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering”, Journal of Global Optimization, vol. 25, pp. 91-111, 2003

[19] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy Efficient Communication Protocols for Wireless Microsensor Networks” Proc. Hawaiian International Conference on Systems Science, Jan.2000

[20] S. M. Mousavi, H. R. Rabiee, M. Moshref, and A. Dabirmoghaddam, “MobiSim: A Framework for Simulation of Mobility Models in Mobile Ad-Hoc Networks”, in Proc. IEEE / WiMOB, White Plains, NY, USA, 2007

[21] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, “A Group Mobility Model for Ad hoc Wireless Networks”, In Proceedings of the ACM/IEEE MSWIM’99, pp.53-60, Seattle, WA, August 1999

[22] D.B. Johnson and D.A. Maltz, “Dynamic Source Routing in Ad Hoc Wireless Networks”, In Mobile Computing, edited by T. Imielinski and H. Korth, Chapter 5, pp. 153-181, Kluwer Publishing Company, 1996

[23] N. Aschenbruck, R. Ernst, P. Martini, “Indoor Mobility Modeling”, IEEE, GLOBECOM Workshops (GC Wkshps), pp.1264-1269, 2010

[24] Rodrigues, P.P., Gama, J., Pedroso, J.P., “ODAC: Hierarchical clustering of time series data streams”, In: SDM 2006. Proceedings of the Sixth SIAM International Conference on Data Mining, pp. 499–503. SIAM (April 2006)

[25] Ester M., Kriegel H.-P., Sander J., Xu X., “A density-based algorithm for discovering clusters in large spatial databases with noise”, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, Portland, OR, 1996

[26] Gower J.C., “A comparison of some methods of cluster analysis”, *Biometrics*, Vol. 23, pp. 623–628, 1967

[27] Lance G.N., Williams W.T., “A general theory of classificatory sorting strategies: II. Clustering systems”, *Computer Journal*, Vol. 10, pp. 271–277, 1967

[28] Fisher, L., and Van Ness, J., “Admissible Clustering Procedures”, *Biometrika*, 58 (1), pp. 91–104, 1971

[29] Haversine Formula: http://en.wikipedia.org/wiki/Haversine_formula
[30] Farnstrom, F., J. Lewis, and C. Elkan, “Scalability for clustering algorithms revisited”, SIGKDD Explorations 2 (1), pp.51-57, 2000
[31] Aggarwal, C., J. Han, J. Wang, and P. Yu, “A framework for clustering evolving data streams”, In Proceedings of the International Conference on Very Large Data Bases, Berlin, Germany, pp. 81-92, Morgan Kaufmann, 2003
[32] Barbara, D., P. Chen, “Using the fractal dimension to cluster datasets”, In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, Boston, MA, pp. 260-264, ACM Press, 2000
[33] Hinneburg, A., D. A. Keim, “Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering”, In Very Large Data Bases, Edinburgh, Scotland, pp. 506-517, Morgan Kaufmann, 1999
[34] Fisher, D. H., “Knowledge acquisition via incremental conceptual clustering”, Machine Learning 2, pp. 139-172, 1987
[35] Kaski, S., T. Kohonen, “Winner-take-all networks for physiological models of competitive learning”, Neural Networks 7 (6-7), 973-984, 1994
[36] Spath, H., “Cluster Analysis Algorithms for Data Reduction and Classification”, Ellis Horwood, 1980
[37] Sheikholeslami, G., S. Chatterjee, A. Zhang, “Wavecluster: A multiresolution clustering approach for very large spatial databases”, In Proceedings of the 24rd International Conference on Very Large Data Bases, San Francisco, CA, pp. 428-439, Morgan Kaufmann, 1998
[38] J., Raper, G., Gartner, H., Karimi, C., Rizos, “A critical evaluation of location based services and their potential”, Journal of Location Based Services, Vol.1, pp. 5-45, 2007
[39] D., Mohapatra, “Survey of location based wireless services”, IEEE / ICPWC, pp. 358-362, 2005