ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Συγκριτική Μελέτη Αλγορίθμων Ομαδοποίησης με Εφαρμογή σε Ιατρικά Δεδομένα

Διπλωματική Εργασία

Νικολαΐδης Γεώργιος
ge12039

Επιβλέπων: Στεφανέας Πέτρος, Καθηγητής ΕΜΠ

Φεβρουάριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Συγκριτική Μελέτη Αλγορίθμων Ομαδοποίησης με Εφαρμογή σε Ιατρικά Δεδομένα

Διπλωματική Εργασία
Νικολαΐδης Γεώργιος

Επιβλέπων: Στεφανέας Πέτρος, Καθηγητής ΕΜΠ

Περίληψη

Στα πλαίσια της παρούσας διπλωματικής εργασίας εξετάζονται 3 αλγόριθμοι ομαδοποίησης-χωρισμού δεδομένων σε συστάδες, καθώς επίσης και η αποτελεσματικότητα της εφαρμογής τους σε ιατρικά δεδομένα (ασθενείς με καρκίνο του μαστού). Σκοπός της εφαρμογής των αλγορίθμων είναι η επιλογή εκείνου με τα βέλτιστα αποτελέσματα, καθώς επίσης και η απάντηση στο ερώτημα για το εάν μπορούν να χρησιμοποιηθούν για την πρόβλεψη του καρκίνου του μαστού.

Αρχικά γίνεται αναφορά σε 3 διαφορετικούς αλγόριθμους (K-means, Single Link, DBSCAN) και σε βασικές έννοιες όπως ο τρόπος λειτουργίας τους, η ορθότητα, οι διάφορες τεχνικές παραμετροποίησης, καθώς και η πολυπλοκότητα του χρόνου και του χώρου. Επιπλέον γίνεται αναφορά στα πλεονεκτήματα αλλά και τα μειονεκτήματα του καθενός έναντι των άλλων. Προκειμένου να μπορούν να συγκριθούν οι παραπάνω μέθοδοι, γίνεται αναφορά σε δείκτες αξιολόγησης, η εφαρμογή των οποίων πραγματοποιήθηκε με τη χρήση του λογισμικού πακέτου R.

Τέλος, ακολουθεί γραφική παρουσίαση δυσδιάστατων και πολυδιάστατων δεδομένων καθώς και εφαρμογή των αλγορίθμων με τη βοήθεια του λογισμικού πακέτου R. Πρέπει να σημειωθεί ότι η διαδικασία βελτίωσης της διακριτικής ικανότητας των αλγορίθμων επιτυγχάνεται τόσο με τη βοήθεια των προαναφερθέντων τεχνικών, όσο και με δοκιμές και συνεχείς εκτελέσεις του αντίστοιχου λογισμικού, αφού στόχος της παρούσας διπλωματικής εργασίας είναι η σύγκριση των αποτελεσμάτων κάτω από συνθήκες βέλτιστης παραμετροποίησης.

NATIONAL TECHNICAL
UNIVERSITY OF ATHENS
SCHOOL OF APPLIED MATHEMATICS
AND PHYSICAL SCIENCES
DEPARTMENT OF MATHEMATICS

Comparative Study of Clustering Algorithms with Application to Medical Data

Diploma Thesis

Nikolaidis Georgios

Advisor: Stefaneas Petros, Professor NTUA

Abstract

In the context of this thesis 3 clustering algorithms are examined as well as their effectiveness with the use of medical data (patients with breast cancer). Purpose of applying these algorithms is the choice of the one with the best results as well as the answer to the question of whether they can be used to predict breast cancer.

Initially reference is made to the 3 different algorithms (K - means, Single Link, DBSCAN) and to basic concepts such as how they function, correctness, various parameterization techniques as well as time and space complexity. In addition, the advantages and disadvantages of each one over the others are mentioned. In order to compare the above methods, reference is made to evaluation indicators implemented using the R package software.

Finally a graphical presentation of 2d, multivariate data and application of the algorithms is followed with the help of R package software. It should be noted that the process of improving the performance of the algorithms is achieved both with the help of the aforementioned techniques as well as with the testing and continuous execution of the respective software, since the aim of this thesis was to compare the results under optimal parameterization conditions.

Ευχαριστίες

Επιθυμώ να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Πέτρο Στεφανέα για την επίβλεψη και καθοδήγηση που μου προσέφερε καθ' όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Ακόμα θα ήθελα να ευχαριστήσω την οικογένεια μου για την ψυχολογική και οικονομική υποστήριξη που μου προσέφερε και μου παρείχε τη δυνατότητα να είμαι σε θέση να ολοκληρώσω τις σπουδές μου. Τέλος θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που ήταν δίπλα μου όλα αυτά τα χρόνια.

Αφιερώνω
την εργασία αυτή
στον αδερφό μου

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ 3

ABSTRACT 4

ΠΕΡΙΕΧΟΜΕΝΑ 7

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ 9

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ 11

ΚΕΦΑΛΑΙΟ 1: **Εισαγωγή** 12

ΚΕΦΑΛΑΙΟ 2: Αλγόριθμος K means

2.1 Παρουσίαση Αλγορίθμου K means 14
2.2 Πρόσθετα θέματα 18
2.3 Προσδιορισμός του αριθμού K 21
2.4 Γενίκευση αλγορίθμου 27
2.5 Πλεονεκτήματα και Μειονεκτήματα 30

ΚΕΦΑΛΑΙΟ 3: Αλγόριθμος Single link Hierarchical Clustering

3.1 Ιεραρχικό Clustering 32
3.2 Η μέθοδος Single link 33
3.3 Τρόποι αναπαράστασης ιεραρχίας 35
3.4 Αξιολόγηση Ιεραρχίας 38
3.5 Προσδιορισμός Διαμέρισης 39
3.6 Πλεονεκτήματα και Μειονεκτήματα 41
3.7 Single Link και ελάχιστα διασυνδετικό δέντρο 43

ΚΕΦΑΛΑΙΟ 4: Αλγόριθμος DBSCAN

4.1 Συμπλέγματα με βάση την πυκνότητα 45
4.2 DBSCAN 49
4.3 Προσδιορισμός παραμέτρων 53
4.4 Πλεονεκτήματα και Μειονεκτήματα 55

ΚΕΦΑΛΑΙΟ 5: Δείκτες Αξιολόγησης

5.1 Αξιολόγηση Συμπλεγμάτων 58
5.2 Εξωτερικοί Δείκτες 59
5.3 Προσανατολισμένοι στην ταξινόμηση 59
5.4 Προσανατολισμένοι στην ομοιότητα 61

ΚΕΦΑΛΑΙΟ 6: Εφαρμογή αλγορίθμων σε δυσδιάστατα δεδομένα

6.1 Παρουσίαση δεδομένων 63
6.2 Εφαρμογή αλγορίθμου K means 64
6.3 Εφαρμοφή αλγορίθμου Single Link 67
6.4 Εφαμορφή αλγορίθμου DBSCAN 69

ΚΕΦΑΛΑΙΟ 7: Εφαρμογή αλγορίθμων σε πολυδιάστατα - ιατρικά δεδομένα

7.1 Παρουσίαση ιατρικών δεδομένων 71
7.2 Μείωση διαστάσεων 75
7.3 Εφαρμογή K - means 77
7.4 Εφαρμογή Single link hierarchical clustering 82
7.5 Εφαρμογή DBSCAN 85
7.6 Σύγκριση αποτελεσμάτων 89

ΚΕΦΑΛΑΙΟ 8: Ανακεφαλαίωση - Συμπεράσματα 92

Βιβλιογραφία 93

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

2.1 Ομαδοποίηση δεδομένων μέσω του αλγορίθμου K means 15

2.2 Ολικό ελάχιστο έναντι τοπικά ελαχίστου για K=3 17

2.3 Κακή επιλογή των αρχικών μέσων για K=3 18

2.4 2 ζεύγη με συμπλέγματα με 2 αρχικούς μέσους σε κάθε ζεύγος 19

2.5 2 ζεύγη με συμπλέγματα με 3 αρχικούς μέσους στο ένα ζεύγος 20

2.6 Μέθοδος του Αγκώνα 22

2.7 Μέθοδος του Αγκώνα 23

2.8 Silhouette μέθοδος 25

2.9 Gap statistic μέθοδος 26

2.10 Συμπλέγματα διαφορετικού μεγέθους 31

2.11 Συμπλέγματα διαφορετικής πυκνότητας 32

2.12 Συμπλέγματα μη σφαιρικού σχήματος 33

3.1 Συσσωρευτική και διαιρετική ιεραρχική ομαδοποίηση 32

3.2 Απόσταση μεταξύ 2 συμπλεγμάτων 34

3.3 Δενδρογράφημα συνόλου 5 παρατηρήσεων 36

3.4 Δενδρογράφημα συνόλου 6 παρατηρήσεων 37

3.5 Δενδρογράφημα συνόλου 17 παρατηρήσεων 39

3.6 Διάγραμμα του δείκτη Calin´ski and Harabasz για διάφορες τιμές του k 40

3.7 Διαμέριση παρατηρήσεων με τη μέθοδο single-link και φαινόμενο αλυσίδας . . . 42

3.8 α) Ελάχιστο διασυνδετικό δέντρο των παρατηρήσεων β) Παραγόμενη διαμέριση. . 44

4.1 3 διαφορετικά σύνολα παρατηρήσεων 45

4.2 Eps-γειτονιά για ευκλείδεια μετρική στον R^2 46

4.3 C: ακραία , B: συνοριακή, A: κεντρική παρατήρηση 47

4.4 (α) p πυκνά - προσβάσιμη από q, (β) p,q πυκνά συνδεδεμένες 48

4.5 Η γειτονιά της πρώτης κεντρικής παρατήρησης εισάγεται στο σύμπλεγμα 51

4.6 Η μεταγενέστερη αντιστοίχιση πυκνά προσβάσιμων παρατηρήσεων αποτελεί την πρώτη συστάδα. Το αρχικό σύνολο Σ καθορίζεται για το δεύτερο σύμπλεγμα . . 51

4.7 Το δεύτερο σύμπλεγμα φτάνει στο μέγιστο του μέγεθος. Το αρχικό σύνολο Σ καθορίζεται για το τρίτο σύμπλεγμα 52

4.8 Το τρίτο σύμπλεγμα φτάνει στο μέγιστο του μέγεθος. Το αρχικό σύνολο Σ καθορίζεται για το τέταρτο σύμπλεγμα 52

4.9 Το τελικό αποτέλεσμα ομαδοποίησης με DBSCAN 52
4.10 Θόρυβος (κενές τελείες) 52
4.11 Διάγραμμα 4-dist για ένα δυσδιάστατο σύνολο παρατηρήσεων 54
4.12 4 συμπλέγματα διαφορετικής πυκνότητας ενσωματωμένα σε θόρυβο 55
4.13 Συμπλέγματα που προέκυψαν από εφαρμογή του DBSCAN 57
6.1 Γραφική αναπαράσταση δεδομένων σε 2 διαστάσεις 63
6.2 Η μέθοδος του Αγκώνα 64
6.3 Η μέθοδος Gap Statistic 65
6.4 Η μέθοδος Silhouette 65
6.5 Ομαδοποίηση με τη μέθοδο K means 66
6.6 Διάγραμμα του δείκτη Calin´ski and Harabasz για διάφορες τιμές του k 67
6.7 Ομαδοποίηση με τη μέθοδο Single - link 68
6.8 Διάγραμμα 4-NN σε αύξουσα σειρά 69
6.9 Ομαδοποίηση με τη μέθοδο DBSCAN 70
7.1 Κατανομή παρατηρήσεων στις 2 φυσικές κλάσεις 71
7.2 Εκτίμηση κατανομών, διαγράμματα διασποράς και γραμμική συσχέτιση . . . 72
7.3 Ευθεία ελαχίστων τετραγώνων για Glucose-HOMA και Insuling-HOMA . . . 73
7.4 Εκτίμηση κατανομών κάθε μεταβλητής σε κάθε κλάση 74
7.5 Θυκόγραμμα κάθε μεταβλητής σε κάθε κλάση 74
7.6 Ποσοστό μεταβλητότητας για κάθε βασική συνιστώσα 75
7.**7** Προβολές στις βασικές συνιστώσες (Patient= Κόκκινο , Healthy= Μαύρο) . . 76
7.8 Αθροιστικό ποσοστό έκφρασης της μεταβλητότητας 77
7.**9** Αναπαράσταση ευκλείδειων αποστάσεων 78
7.**10** Μέθοδος του Αγκώνα 79
7.**11** Average Silhouette μέθοδος 79
7.**12** Gap statistic μέθοδος 80
7.**13** Συμπλέγματα μέσω του αλγορίθμου K means 81
7.14 Παραγόμενο δενδρογράφημα του αλγορίθμου Single-link 82
7.15 Παραγόμενα συμπλέγματα (Κόκκινο = 1 , Πράσινο= 2) με μέθοδο Single-link … 83
7.**16** Προβολή παραγόμενων συμπλεγμάτων στις 2 πρώτες κύριες συνιστώσες 84
7.**17** Δείκτης Calin´ski και Harabasz για κάθε k 85
7.18 Γράφημα 7-NN κάθε παρατήρησης 86
7.**19** Συμπλέγματα DBSCAN για Nmin=7 ,Eps=3 86
7.20 Συμπλέγματα DBSCAN για Nmin=7 ,Eps=1.9 και φυσικές κλάσεις 87

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

2.1 K-means: Κοινές επιλογές μετρικής,μέσου και αντικειμενικής συνάρτησης 29

5.1 Πίνακας που προσδιορίζει εάν τα ζεύγη παρατηρήσεων είναι της ίδιας κλάσης και του ίδιου συμπλέγματος 61

6.1 Αριθμητικοί μέσοι παραγόμενων συμπλεγμάτων μέσω του αλγορίθμου K means . 66

6.2 Κατανομή σημείων στις 3 κλάσεις με τις μεθόδους K means και Single link 68

6.3 Ομαδοποίηση σημείων με τη μέθοδο DBSCAN 69

7.1 Υποσύνολο ιατρικών δεδομένων 71

7.2 Ποσοστό μεταβλητότητας που εκφράζεται από κάθε κύρια συνιστώσα 76

7.3 Προβλεπόμενοι αριθμητικοί μέσοι με τη μέθοδο K - means 81

7.4 Πραγματικές τιμές των αριθμητικών μέσων κάθε συμπλέγματος 81

7.5 Πλήθος ορθών και λανθασμένων ταξινομήσεων με τη μέθοδο K - means 82

7.6 Πλήθος ορθών και λανθασμένων ταξινομήσεων με τη μέθοδο Single - link 84

7.7 Προβλεπόμενοι αριθμητικοί μέσοι με τη μέθοδο DBSCAN 88

7.8 Πραγματικές τιμές των αριθμητικών μέσων κάθε συμπλέγματος 88

7.9 Πλήθος ορθών και λανθασμένων ταξινομήσεων με τη μέθοδο DBSCAN 88

7.10 Δείκτες Entropy και Purity για K - means και DBSCAN 89

7.11 Δείκτης Precision για K - means και DBSCAN 90

7.12 Δείκτης Recall για K - means και DBSCAN 90

7.13 Δείκτης F για K - means και DBSCAN 91

7.14 Δείκτες Rand, Jacard και FM για K - means και DBSCAN 91

Κεφάλαιο 1

Εισαγωγή

Οι προηγμένες τεχνολογίες ανίχνευσης και αποθήκευσης καθώς και η δραματική ανάπτυξη σε εφαρμογές όπως η αναζήτηση στο Internet, η ψηφιακή απεικόνιση και η επιτήρηση μέσω βίντεο, έχουν δημιουργήσει σύνολα δεδομένων μεγάλου όγκου και διάστασης. Τα περισσότερα από αυτά τα δεδομένα αποθηκεύονται ψηφιακά σε ηλεκτρονικά μέσα, παρέχοντας έτσι τεράστιες δυνατότητες για την ανάπτυξη μεθόδων αυτόματης ανάλυσης δεδομένων, ταξινόμησης και ανάκτησης. Εκτός από την αύξηση του όγκου των δεδομένων, αυξήθηκε επίσης η ποικιλία των διαθέσιμων δεδομένων (κείμενο, εικόνα και βίντεο). Αυτή η αύξηση τόσο στον όγκο, όσο και στην ποικιλία των δεδομένων, απαιτεί εξέλιξη στη μεθοδολογία για την αυτόματη κατανόηση, επεξεργασία και συνοπτική παρουσίαση των δεδομένων αυτών.

Η ανάλυση συστάδων επικρατεί σε οποιοδήποτε κλάδο που περιλαμβάνει την ανάλυση πολυπαραγοντικών δεδομένων. Έχει διάφορους στόχους, που όλοι σχετίζονται με την ομαδοποίηση ή την κατάτμηση μιας συλλογής αντικειμένων σε υποσύνολα ή "συστάδες", έτσι ώστε όσα βρίσκονται μέσα σε κάθε συστάδα να είναι πιο στενά συνδεδεμένα μεταξύ τους από τα αντικείμενα που αντιστοιχούν σε διαφορετικές συστάδες. Είναι γεγονός πώς είναι δύσκολο να απαριθμηθούν επακριβώς τα πολυάριθμα επιστημονικά πεδία και οι εφαρμογές που έχουν χρησιμοποιήσει τεχνικές ομαδοποίησης. Για παράδειγμα, ο κατακερματισμός της εικόνας, ένα σημαντικό πρόβλημα στην όραση του υπολογιστή, μπορεί να διατυπωθεί ως πρόβλημα ομαδοποίησης (Frigui & Krishnapuram, 1999, Jain & Flynn, 1996, Shi & Malik, 2000). Αντίστοιχα, τα έγγραφα μπορούν να ομαδοποιηθούν (Iwayama & Tokunaga, 1995) για τη δημιουργία τοπικών ιεραρχιών στοχεύοντας στην αποτελεσματική πρόσβαση στις πληροφορίες (Sahami, 1998) ή στην ανάκτηση (Bhatia & Deogun, 1998). Η ομαδοποίηση χρησιμοποιείται επίσης για να κατηγοριοποιήσει τους πελάτες σε διαφορετικούς τύπους με σκοπό το αποτελεσματικό μάρκετινγκ (Arabie & Hubert, 1994). Ομαδοποιούνται επίσης οι υποχρεώσεις παροχής υπηρεσιών για τη διαχείριση και το σχεδιασμό του εργατικού δυναμικού (Hu et al. , 2007), καθώς και για τη μελέτη δεδομένων γονιδιώματος στη βιολογία (Baldi & Hatfield, 2002).

Στα πλαίσια της παρούσας διπλωματικής εργασίας εξετάζονται 3 αλγόριθμοι ομαδοποίησης-χωρισμού δεδομένων σε συστάδες, καθώς επίσης και η αποτελεσματικότητά της εφαρμογής τους σε ιατρικά δεδομένα (ασθενείς με καρκίνο του μαστού). Σκοπός της εφαρμογής των αλγορίθμων είναι η επιλογή εκείνου με τα βέλτιστα αποτελέσματα, καθώς επίσης και η απάντηση στο ερώτημα για το εάν μπορούν να χρησιμοποιηθούν για την αποτελεσματική πρόβλεψη του καρκίνου του μαστού.

Προκειμένου να δωθούν απαντήσεις στα παραπάνω ερωτήματα, αρχικά γίνεται αναφορά στους 3 διαφορετικούς αλγόριθμους (K - means, Single Link, DBSCAN) που θα χρησιμοποιηθούν. Εξετάζονται οι μαθηματικές τους ιδιότητες, πλεονεκτήματα, μειονεκτήματα, αλλά και η αλγοριθμική τους πολυπλοκότητα. Επιπλέον για να μπορούν να συγκριθούν οι παραπάνω μέθοδοι, γίνεται αναφορά σε δείκτες αξιολόγησης. Με βάση τα παραπάνω, ακολουθεί γραφική παρουσίαση των δεδομένων και εφαρμογή των αλγορίθμων με τη βοήθεια του λογισμικού πακέτου R, οδηγώντας μας στην εξαγωγή συμπερασμάτων.

Κεφάλαιο 2

2.1 Παρουσίαση Αλγορίθμου K means

Ο συμβατικός K-means αλγόριθμος είναι ένας από τους πιο χρησιμοποιημένους αλγορίθμους για την ανάλυση συστάδων (συμπλεγμάτων, συγκροτημάτων, clusters) ο οποίος περιγράφηκε για πρώτη φορά από τον Macqueen (1967). Σχεδιάστηκε προκειμένου να ομαδοποιεί αριθμητικά δεδομένα, έτσι ώστε κάθε συστάδα να έχει ένα κέντρο το οποίο καλείται μέσος. Ο αλγόριθμος K-means αποτελεί μέθοδο μη ιεραρχικής ομαδοποίησης σε συστάδες (Jain και Dubes, 1988). Αποτελεί βασική προϋπόθεση ότι ο αριθμός των συστάδων K θεωρείται γνωστός.

Υπάρχει μια συνάρτηση σφάλματος σε αυτόν τον αλγόριθμο με βάση την οποία για ένα δεδομένο αρχικό σύνολο συστάδων K, επανατοποθετεί τα δεδομένα στα πλησιέστερα συμπλέγματα, και στη συνέχεια, επαναλαμβανόμενα αλλάζει το περιεχόμενο των συστάδων, έως ότου η συνάρτηση σφάλματος δεν μεταβληθεί σημαντικά ή το περιεχόμενο των συστάδων δεν αλλάζει πλέον. Ο συμβατικός αλγόριθμος K-means αναπτύχθηκε από τους Hartigan και Wong (1979).

Η συνάρτηση σφάλματος ορίζεται ως εξής :

$$f_{\Sigma}(C) := \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{m} (x_{ij} - \hat{z}_{kj})^2 , \qquad (1)$$

η οποία μετράει την απόσταση κάθε παρατήρησης από το μέσο $\hat{z}_k$ $(k = 1,...,K)$ της συστάδας στην οποία ανήκει. Ο αλγόριθμος K-means ελαχιστοποιεί την f_{Σ} για δεδομένο αριθμό K και διαμέριση $C = (C_1, \ldots, C_K)$ των n παρατηρήσεων $\Pi = [x_1, \ldots, x_n]$ σε K συστάδες. Ο μέσος κάθε συμπλέγματος ορίζεται ως εξής:

$$\hat{z}_{ki} := \frac{1}{|C_k|} \sum_{l=1}^{|C_k|} x_{li} \quad \text{για} \quad i = 1,..,m \ \text{και} \ k = 1,...,K \qquad (2)$$

Επιπλέον έχουμε :

$$n_k := |C_k|, \qquad C_k = [x_{1k}, \ldots, x_{n_k k}]$$

και

$$d(x_i, x_j) := \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

είναι η ευκλείδεια απόσταση (παρατηρούμε ότι στην (1) δεν έχουμε την τετραγωνική ρίζα), ενώ m ο αριθμός των χαρακτηριστικών για κάθε παρατήρηση. Το πρόβλημα της εύρεσης μιας τέτοιας διαμέρισης C όλων των παρατηρήσεων σε K συστάδες, έχει αποδειχθεί ότι είναι NP-hard. Γι αυτό το λόγο, ο αλγόριθμος K-means αποτελεί έναν ευρετικό αλγόριθμο, ο οποίος μας δίνει καλές (όχι απαραίτητα βέλτιστες) λύσεις σε αποδεκτό χρόνο. Μια γενική τεχνική αλγοριθμικού διαχωρισμού για αυτό το πρόβλημα μπορεί να δηλωθεί ως εξής:

Βήμα 1. Θέτουμε τον αριθμό K για το πλήθος των συστάδων ίσο με έναν ακέραιο και επιλέγουμε το μέγιστο αριθμό επαναλήψεων. Επιπλέον, επιλέγουμε με τυχαίο τρόπο μια αρχική διαμέριση και υπολογίζουμε τους μέσους με βάση τη σχέση (2).

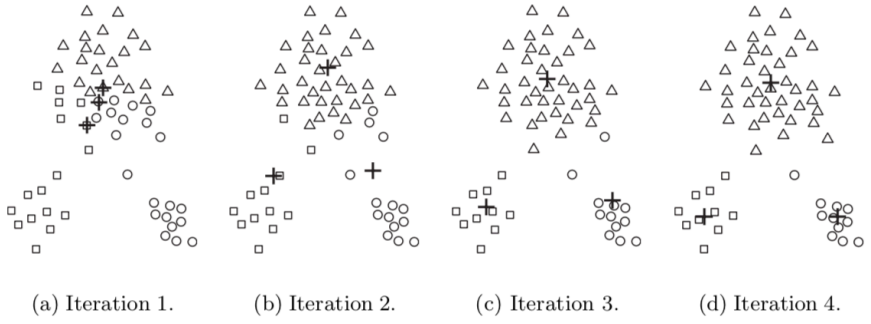
Βήμα 2. Φτιάχνουμε μία νέα διαμέριση, τοποθετώντας κάθε παρατήρηση στη συστάδα με το κοντινότερο μέσο, με βάση πάντα την ευκλείδεια απόσταση.

Βήμα 3. Υπολογίζουμε τους νέους μέσους μέσω της σχέσης (2) και της νέας διαμέρισης.

Βήμα 4. Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρις ότου η συνάρτηση σφάλματος (σχέση (1)) να μην βελτιώνεται - μειώνεται άλλο ή υπερβούμε το μέγιστο αριθμό επαναληψεων.

Αλγόριθμος 2.1 K- means
Δεδομένα: Σύνολο των παρατηρήσεων Π, αριθμός συστάδων K, διαστάσεις m
Αποτέλεσμα: Διαμέριση $(C'_1, C'_2, \ldots, C'_k)$ του συνόλου Π
{ C_i είναι η i-οστή συστάδα }
1: Έστω $(C_1, C_2, \ldots, C_k)$ μια τυχαία αρχική διαμέριση του Π
2: Επανάλαβε
3: d_{ij} = Απόσταση παρατήρησης i απο το μέσο της συστάδας j
4: $n_i = argmin_{1 \leqslant j \leqslant K} d_{ij}$
5: Τοποθέτησε την παρατήρηση i στη συστάδα n_i
6: Υπολόγισε ξανά τους μέσους κάθε συστάδας
7: Μέχρι να μην συμβεί καμία αλλαγή στις συστάδες σε μία πλήρη επανάληψη
8: Επιστροφή του αποτελέσματος, δηλαδή της διαμέρισης $(C'_1, C'_2, \ldots, C'_k)$

Τα βήματα του παραπάνω αλγορίθμου μπορούμε να τα δούμε και με τη βοήθεια της εικόνας 2.1 που ακολουθεί:



Εικόνα 2.1: Ομαδοποίηση δεδομένων μέσω του αλγορίθμου K means

Στο σχήμα (α) παρατηρούμε το πρώτο βήμα του αλγορίθμου, όπου οι παρατηρήσεις τοποθετούνται με τυχαίο τρόπο στις συστάδες και στη συνέχεια βρίσκουμε του μέσους οι οποίοι συμβολίζονται με ένα σταυρό. Στη συνέχεια στα σχήματα (b), (c) υπολογίζονται οι αποστάσεις των παρατηρήσεων από τους μέσους και τοποθετούνται στη συστάδα με την μικρότερη απόσταση από τον αντίστοιχο μέσο. Αφού γίνει αυτό υπολογίζονται οι καινούργιοι μέσοι. Τέλος, στο σχήμα (d) δεν θα έχουμε άλλη μετακίνηση παρατηρήσεων, επομένως ο αλγόριθμος μας επιστρέφει την διαμέριση που βλέπουμε.

Θεώρημα. Ο αλγόριθμος K means μειώνει μονοτονικά την ποσότητα $f_{\Sigma}(C)$

Απόδειξη

Συμβολίζουμε με $A_{(x_i)}^{(t)}$ τη συστάδα στην οποία έχει τοποθετηθεί η παρατήρηση x_i μετά απο t βήματα και $C^{(t)}$ η διαμέριση μετά απο t βήματα.

$$f_{\Sigma}(C^{(t)}) \geq \sum_{j=1}^{K} \sum_{x_i \varepsilon C_j^{(t)}} ||x_i, z_{A_{(x_i)}^{(t+1)}}^{(t)}||^2 \geq \sum_{j=1}^{K} \sum_{x_i \varepsilon C_j^{(t)}} ||x_i, z^{(t+1)}||^2 \geq f_{\Sigma}(C^{(t+1)})$$

□

Συνέπεια. Ο αλγόριθμος K means σταματάει μετά απο πεπερασμένο αριθμό βημάτων

Απόδειξη

Δεδομένου ότι όλες οι πιθανές διαμερίσεις των παρατηρήσεων είναι πεπερασμένου πλήθους $\binom{n}{k}$, η φθίνουσα ακολουθία $f_{\Sigma}(C^{(t)})_{t \varepsilon R}$ έχει ένα πεπερασμένο αριθμό τιμών.
Από το παραπάνω συμπεραίνουμε ότι υπάρχει t τέτοιο ώστε $f_{\Sigma}(C^{(t)}) = f_{\Sigma}(C^{(t+1)})$. Αν δεν υπήρχε η ακολουθία θα ήταν άπειρη, γεγονός που είναι άτοπο. Επομένως έχουμε μία φθίνουσα ακολουθία πραγματικών αριθμών, η οποία είναι κάτω φραγμένη από το 0 και έτσι θα συγκλίνει σύμφωνα με γνωστό λήμμα της πραγματικής ανάλυσης. Επιπλέον θα φτάσει στη σύγκλιση σε πεπερασμένο αριθμό βημάτων, καθώς όλες οι δυνατές τιμές που μπορεί να πάρει είναι πεπερασμένες.

□

Παρατηρήσεις

- Η παραπάνω συνέπεια μας λέει ότι ο αλγόριθμος συγκλίνει, όχι πόσο γρήγορα θα συγκλίνει καθώς έχουμε μόνο ένα εκθετικό φράγμα $\binom{n}{k}$ για τον αριθμό των βημάτων. Ο χρόνος σύγκλισης εξαρτάται απο την αρχική διαμέριση των παρατηρήσεων.

- Η λύση που προκύπτει από τον αλγόριθμο αποτελεί τοπικό ελάχιστο και όχι απαραίτητα ολικό, καθώς εξαρτάται από την αρχική διαμέριση. Γι αυτό το λόγο θα ήταν χρήσιμο να τρέξουμε τον αλγόριθμο μερικές φορές και να διαλέξουμε την καλύτερη διαμέριση ως τελική απάντηση.

Εικόνα 2.2: Ολικό ελάχιστο έναντι τοπικού ελαχίστου για K=3

(a) Optimal clustering. (b) Suboptimal clustering.

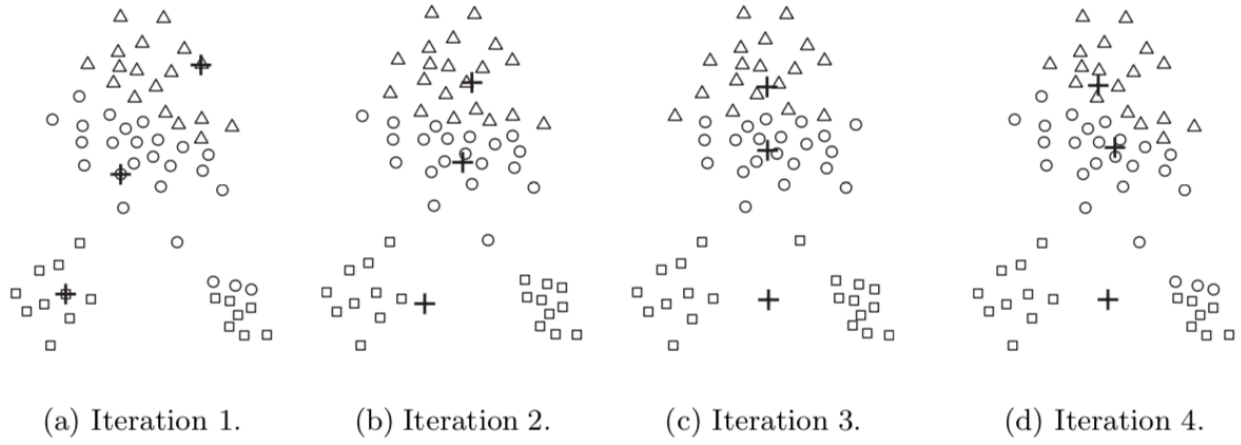
Ανάλυση πολυπλοκότητας χρόνου και χώρου

Προκειμένου να δούμε πόσο χρόνο και χώρο απαιτεί ο αλγόριθμος θα κάνουμε χρήση του ψευδοκώδικα (αλγόριμος 2.1). Στο βήμα 1 έχουμε την αρχικοποίηση, ορίζοντας μία αρχική τυχαία διαμέριση. Η παραπάνω διαδικασία μπορεί να γίνει με πολλούς τρόπους, αν θεωρήσουμε όμως ότι σε κάθε παρατήρηση αντιστοιχούμε τυχαία έναν αριθμό από το 1 έως το K, τότε θα έχουμε O(n), διότι η εύρεση ενός τυχαίου αριθμού απαιτεί O(1). Έστω ότι το βήμα 2 τρέχει I φορές. Επιπλέον, δεδομένου ότι είμαστε στις m διαστάσεις, ο υπολογισμός της απόστασης μεταξύ 2 σημείων απαιτεί m χρόνο και έτσι κάθε επανάληψη είναι πολυπλοκότητας O(Knm). Επομένως, αν I ο αριθμός των επαναλήψεων, τότε η συνολική πολυπλοκότητα θα είναι O(IKnm). Το I είναι συχνά μικρό και συνήθως μπορούμε να το φράξουμε, καθώς οι περισσότερες αλλαγές συμβαίνουν στις πρώτες επαναλήψεις. Δεδομένου ότι K και m είναι μικρότερα σε σχέση με το πλήθος των παρατηρήσεων n, βλέπουμε ότι ο αλγόριθμος K means είναι γραμμικός στο πλήθος των παρατηρήσεων. Όσον αφορά τη χωρική πολυπλοκότητα, τα μόνα που αποθηκεύονται είναι οι παρατηρήσεις και οι μέσοι, επομένως θα έχουμε O((n + K)m). Με βάση τις παραπάνω πολυπλοκότητες μπορούμε να συμπεράνουμε ότι ο αλγόριθμος K means είναι αρκετά αποδοτικός για μεγάλα σύνολα δεδομένων.

2.2 Πρόσθετα θέματα

Επιλογή αρχικής διαμέρισης - αρχικών μέσων

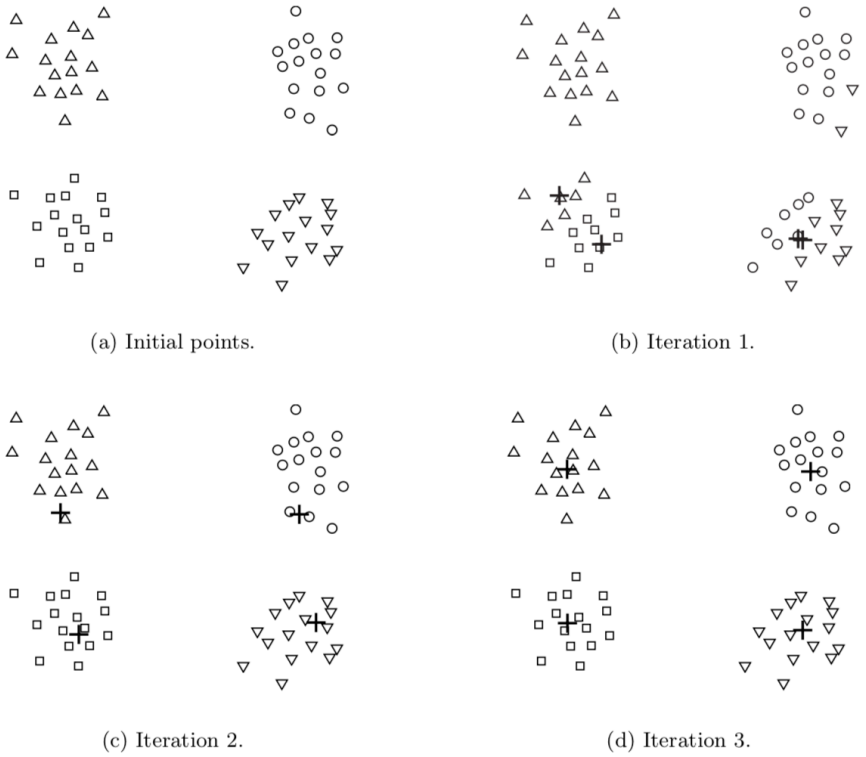
Η αρχική διαμέριση των παρατηρήσεων μπορεί να επηρεάσει αισθητά το αποτέλεσμα του αλγορίθμου, καθώς μία κακή επιλογή των αρχικών μέσων μπορεί να μας οδηγήσει σε αποτέλεσμα που είναι τοπικό ελάχιστο και όχι ολικό (εικόνα 2.3)



Εικόνα 2.3: Κακή επιλογή των αρχικών μέσων για K=3

Με βάση τις εικόνες 2.1 και 2.3 βλέπουμε ότι στην πρώτη περίπτωση η τυχαία επιλογή μας έδωσε το σωστό αποτέλεσμα, εν αντιθέσει με την δεύτερη περίπτωση όπου καταλήξαμε σε ένα τοπικό ελάχιστο.

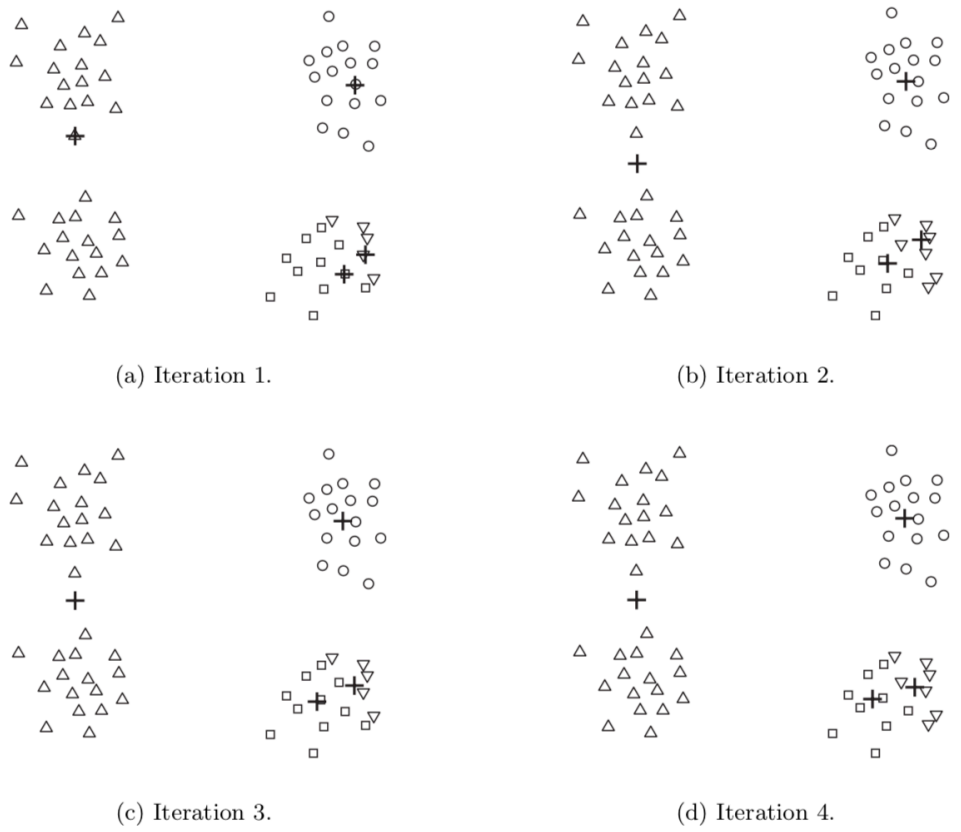
Ένας απλός τρόπος αντιμετώπισης του παραπάνω προβλήματος είναι να τρέξουμε τον αλγόριθμο πολλές φορές, έχοντας διαφορετική αρχικοποίηση κάθε φορά και στο τέλος να διαλέξουμε τη διαμέριση με τη μικρότερη τιμή $f_{\Sigma}(C)$. Η παραπάνω τεχνική αν και είναι σχετικά εύκολη μπορεί να μην δουλεύει πάντα, καθώς εξαρτάται από το σύνολο των παρατηρήσεων καθώς και τον αριθμό K για το πλήθος των συμπλεγμάτων. Παρακάτω μπορούμε να δούμε ένα παράδειγμα μη αποδοτικής λειγουργίας της παραπάνω μεθόδου.



Εικόνα 2.4: 2 ζεύγη με συμπλέγματα, με 2 αρχικούς μέσους σε κάθε ζεύγος

Με βάση την εικόνα 2.4 παρατηρούμε ότι επειδή τα συμπλέγματα σε κάθε ζεύγος βρίσκονται πιο κοντά μεταξύ τους, άμα ξεκινήσουμε βάζοντας από 2 μέσους σε κάθε ζευγάρι θα καταλήξουμε στο ολικό ελάχιστο. Αυτό συμβαίνει ακόμα και αν οι αρχικοί μέσοι βρίσκονται στο ίδιο σύμπλεγμα σε κάθε ζεύγος. Αντίθετα, μια προσέγγιση όπως αυτή της εικόνας 2.5, όπου δηλαδή θα έχουμε ένα μέσο στο πρώτο ζεύγος και 3 στο άλλο, δεν θα καταλήξει στη βέλτιστη λύση αλλά σε ένα τοπικό ελάχιστο. Επομένως παρατηρούμε ότι μια βέλτιστη διαμέριση θα προκύψει μόνο όταν σε κάθε ζεύγος έχουμε από 2 αρχικούς μέσους σε κάθε ζεύγος συμπλεγμάτων. Δυστυχώς καθώς το πλήθος των συμπλεγμάτων αυξάνεται, είναι πολύ πιθανό κάθε φορά που επιλέγουμε τυχαία τους αρχικούς μέσους να υπάρχει κάποιο ζευγάρι συμπλεγμάτων στο οποίο να έχουμε μόνο ένα μέσο. Συνεπώς, επειδή οι αποστάσεις μεταξύ των ζευγαριών είναι πολύ μεγαλύτερες από τις αποστάσεις μεταξύ των συμπλεγμάτων εντός κάθε ζεύγους, ο αλγόριθμος δεν θα μας επιστρέψει μία βέλτιστη διαμέριση, αλλά ένα τοπικό ελάχιστο.

Δεδομένου ότι η παραπάνω μέθοδος ενδεχομένως να μη μας δώσει καλά αποτελέσματα, μια διαφορετική προσέγγιση θα ήταν να πάρουμε ένα δείγμα από το σύνολο των παρατηρήσεων και να παράξουμε μία διαμέρισή του χρησιμοποιώντας τεχνικές ιεραρχικής ομαδοποίησης. Στη συνέχεια, παίρνουμε τους μέσους των K συμπλεγμάτων και τους χρησιμοποιούμε σαν αρχικούς μέσους για τον αλγόριθμο K means. Η παραπάνω προσέγγιση συχνά λειτουργεί καλά αλλά είναι πρακτική μόνο όταν το μέγεθος του δείγματος είναι σχετικά μικρό (καθώς οι τεχνικές ιεραρχικής ομαδοποίησης είναι αρκετά χρονοβόρες), και όταν ο αριθμός K για το πλήθος των συμπλεγμάτων είναι αρκετά μικρός, συγκριτικά με το μέγεθος του δείγματος.



Εικόνα 2.5: 2 ζεύγη με συμπλέγματα, με 3 αρχικούς μέσους στο ένα ζεύγος

Μία εναλλακτική μέθοδος θα ήταν να επιλέξουμε τυχαία μία παρατήρηση ή να πάρουμε το μέσο όλων των παρατηρήσεων και να το θέσουμε ως ένα από τους αρχικούς μέσους. Στη συνέχεια, για κάθε επιλογή αρχικού μέσου διαλέγουμε εκείνο το σημείο που έχει την μεγαλύτερη απόσταση απ' όλους όσους έχουμε διαλέξει μέχρι στιγμής. Με αυτόν τον τρόπο οι αρχικοί μέσοι που θα έχουμε επιλέξει εκτός από το γεγονός ότι θα είναι τυχαίοι, θα είναι και σχετικά μακριά μεταξύ τους. Μια τέτοια προσέγγιση εκτός του ότι θα μπορούσε να μας οδηγήσει στο να επιλέξουμε ακραίες τιμές σαν αρχικούς μέσους, αποτελεί και μια αρκετά χρονοβόρα διαδικασία. Προκειμένου να το αποφύγουμε, επιλέγουμε ένα τυχαίο δείγμα για να πραγματοποιήσουμε την παραπάνω διαδικασία. Δεδομένου ότι οι ακραίες παρατηρήσεις θα είναι λίγες, η πιθανότητα να επιλεχθεί κάποια από αυτές στο τυχαίο δείγμα θα είναι αρκετά μικρή, σε αντίθεση με το να επιλέξουμε παρατηρήσεις από κάθε πυκνή περιοχή. Επιπλέον, το μέγεθος του δείγματος θα είναι συγκριτικά πιο μικρό από το σύνολο των παρατηρήσεων, με αποτέλεσμα η διαδικασία να είναι λιγότερο χρονοβόρα.

Ακραίες τιμές

Επειδή μας ενδιαφέρει η ελαχιστοποίηση της ποσότητας $f_{\Sigma}(C)$, η οποία έτσι όπως έχει οριστεί από τον τύπο (1) περιλαμβάνει την ευκλείδεια απόσταση, ακραίες παρατηρήσεις θα ασκούν μεγάλη επιρροή στην τελική διαμόρφωση των μέσων. Αρκετά συχνά κρίνεται αναγκαίος ο εντοπισμός τέτοιων παρατηρήσεων και η αφαίρεσή τους, πριν τρέξουμε τον αλγόριθμο. Παρ΄ όλα αυτά

υπάρχουν μερικές εφαρμογές στις οποίες οι ακραίες παρατηρήσεις δεν πρέπει να αφαιρεθούν, όπως στη συμπίεση δεδομένων ή σε εφαμοργές σε οικονομικά δεδομένα (π.χ ασυνήθιστα κερδοφόροι πελάτες), όπου παρουσιάζουν ιδιαίτερο ενδιαφέρον. Για την αντιμετώπιση του παραπάνω ζητήματος έχουν αναπτυχθεί τροποποιήσεις του αλγορίθμου K - means αλλά και άλλοι αλγόριθμοι (π.χ DBSCAN), που αποδίδουν αρκετά καλά σε τέτοιες περιπτώσεις, λόγω της διαφορετικής λογικής στην οποία βασίσονται.

Άδεια Συμπλέγματα

Ένα από τα προβλήματα του αλγορίθμου K means είναι ότι μπορεί να επιστραφούν άδεια συμπλέγματα, όταν κατά τη διάρκεια της αρχικοποίησης δεν τοποθετηθεί καμία παρατήρηση σε κάποιο από τα συμπλέγματα. Αυτό θα έχει ως αποτέλεσμα η ποσότητα $f_{\Sigma}(C)$ της διαμέρισης να είναι πιο μεγάλη απ΄ ότι χρειάζεται. Ένας τρόπος αντιμετώπισης του παραπάνω ζητήματος θα ήταν να αντικαταστήσουμε το συγκεκριμένο αρχικό μέσο με την παρατήρηση που βρίσκεται πιο μακριά απ όλους τους άλλους μέσους, μειώνοντας την ποσότητα $f_{\Sigma}(C)$, αποκλείοντας με αυτόν τον τρόπο την παρατήρηση που συνεισφέρει περισσότερο. Μία εναλλατική προσέγγιση θα αποτελούσε να επιλεχθεί ένα σημείο από το σύμπλεγμα, που συνεισφέρει περισσότερο στην ποσότητα $f_{\Sigma}(C)$, χωρίζοντας έτσι το σύμπλεγμα σε 2 μέρη και μειώνοντας τη συνολική συνεισφορά. Η παραπάνω διαδικασία μπορεί να γίνει παραπάνω από μία φορές, αν τα αρχικά συμπλέγματα που είναι άδεια είναι περισσότερα από ένα.

2.3 Προσδιορισμός του αριθμού K

Βασική προϋπόθεση για να εκτελέσουμε τον αλγόριθμο k means αποτελεί ο προσδιορισμός του αριθμού των συμπλεγμάτων K, που θεωρούμε ότι υπάρχουν στα δεδομένα μας. Το πρόβλημα της επιλογής του αριθμού των συμπλεγμάτων αποτελεί από μόνο του ένα αρκετά δύσκολο πρόβλημα. Στη συνέχεια θα αναφερθούμε σε μερικούς τρόπους εκτίμησης της παραπάνω ποσότητας.

Μέθοδος του Αγκώνα (Elbow method)

Μία πολύ βασική μέθοδος προσδιορισμού του αριθμού K για ένα σύνολο παρατηρήσεων αποτελεί η μέθοδος του αγκώνα. Η βασική ιδέα της μεθόδου είναι η εξής: Ξεκινάμε με K=1 και συνεχώς το αυξάνουμε, αξιολογώντας παράλληλα τη διαμέριση που προκύπτει για κάθε διαφορετική τιμή του K. Προκειμένου να αξιολογήσουμε τα αποτελέσματα για κάθε τιμή του K, θεωρούμε ως εσω-συμπλεγματική απόκλιση για κάθε σύμπλεγμα της διαμέρισης την ποσότητα:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - z_k)^2 \ ,$$

και συνολική εσω-συμπλεγματική απόκλιση την ποσότητα:

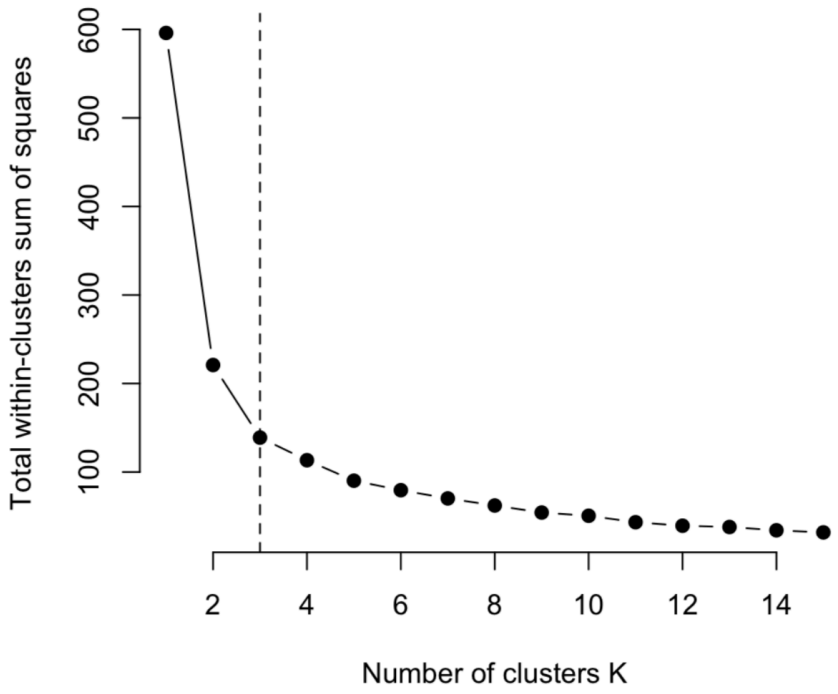
$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - z_k)^2 \quad (3).$$

Η παραπάνω τιμή βλέπουμε ότι ταυτίζεται με αυτήν της σχέσης (1), την οποία γνωρίζουμε ότι ο αλγόριθμος θέλει να ελαχιστοποιήσει. Όσο μικρότερη είναι η παραπάνω τιμή, τόσο καλύτερη είναι και η διαμέριση. Αν για κάποια τιμή του K η τιμή που προκύπτει από τη σχέση (3) μειωθεί δραματικά, τότε αυτή θα είναι και η πραγματική τιμή του K. Άν γράψουμε την παραπάνω διαδικασία με μορφή ψευδοκώδικα θα έχουμε:

Αλγόριθμος 2.2 Elbow method

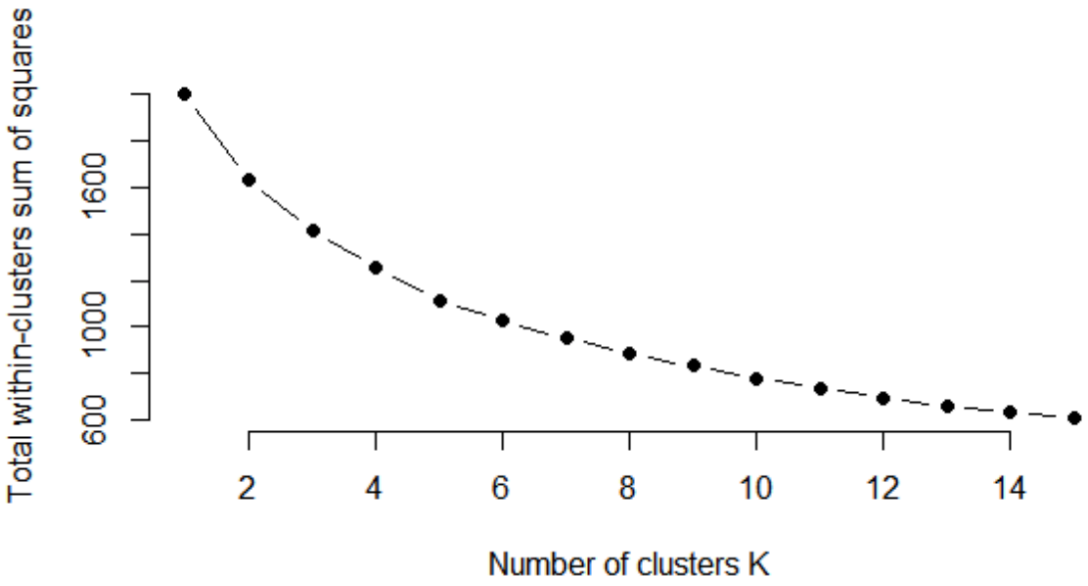
Δεδομένα: Σύνολο των παρατηρήσεων Π, διαστάσεις m
Αποτέλεσμα: Προτεινόμενο K

1. Εφαρμογή του K-means για τιμές του K σε κάποιο διάστημα π.χ. από 1 έως 15
2. Για κάθε τιμή του K, υπολογισμός της ποσότητας WSS από τη σχέση (3)
3. Σχεδιασμός γραφικής παράστασης WSS-K
4. Το σημείο στην καμπύλη που κάνει απότομη στροφή και μοιάζει με αγκώνα συνήθως θεωρείται ένδειξη υποψήφιου αριθμού K για τον αριθμό των συμπλεγμάτων.



Εικόνα 2.6 Μέθοδος του Αγκώνα

Στην εικόνα 2.6 παρατηρούμε την ποσότητα WSS στον Y άξονα, και τον αριθμό των συμπλεγμάτων K στον X άξονα. Με βάση το γράφημα παρατηρούμε ότι μετά την τιμή K=3 παρουσιάζεται σχετικά μικρή πτώση στην τιμή WSS, με αποτέλεσμα να θεωρήσουμε ως αποδεκτή την τιμή K=3. Παρ' όλη την πρακτικότητα της παραπάνω μεθόδου, είναι αρκετά πιθανό να υπάρχουν περιπτώσεις όπου το σημείο επιλογής δεν είναι αρκετά εμφανές.



Εικόνα 2.7 Μέθοδος του Αγκώνα

Στην εικόνα 2.7 παρατηρούμε ότι το σημείο επιλογής θα μπορούσε να είναι το K=4 αλλά και το K=5. Προκειμένου να αντιμετωπιστούν τέτοιες περιπτώσεις μπορούμε να εφαρμόσουμε άλλες τεχνικές, όπως θα δούμε στη συνέχεια.

Average Silhouette μέθοδος

Η μέθοδος Silhouette αποτελεί ένδειξη καλής διαμέρισης των παρατηρήσεων και μας παρέχει μια γραφική αναπαράσταση του πόσο καλά είναι τοποθετημένη κάθε παρατήρηση στο σύμπλεγμα στο οποίο έχει τοποθετηθεί. Περιγράφεται για πρώτη φορά από το Peter J. Rousseeuw (1987). Στη συνέχεια οι Kaufman και Rousseeuw πρότειναν το δείκτη Silhouette, ως εκτιμήτρια του βελτιστού αριθμού συμπλεγμάτων K. Έχει αρκετές ομοιότητες με τη Elbow method και σε μορφή ψευδοκώδικα θα είναι ως εξής:

Αλγόριθμος 2.3 Silhouette

Δεδομένα: Σύνολο των παρατηρήσεων Π , διαστάσεις m
Αποτέλεσμα: Προτεινόμενο K

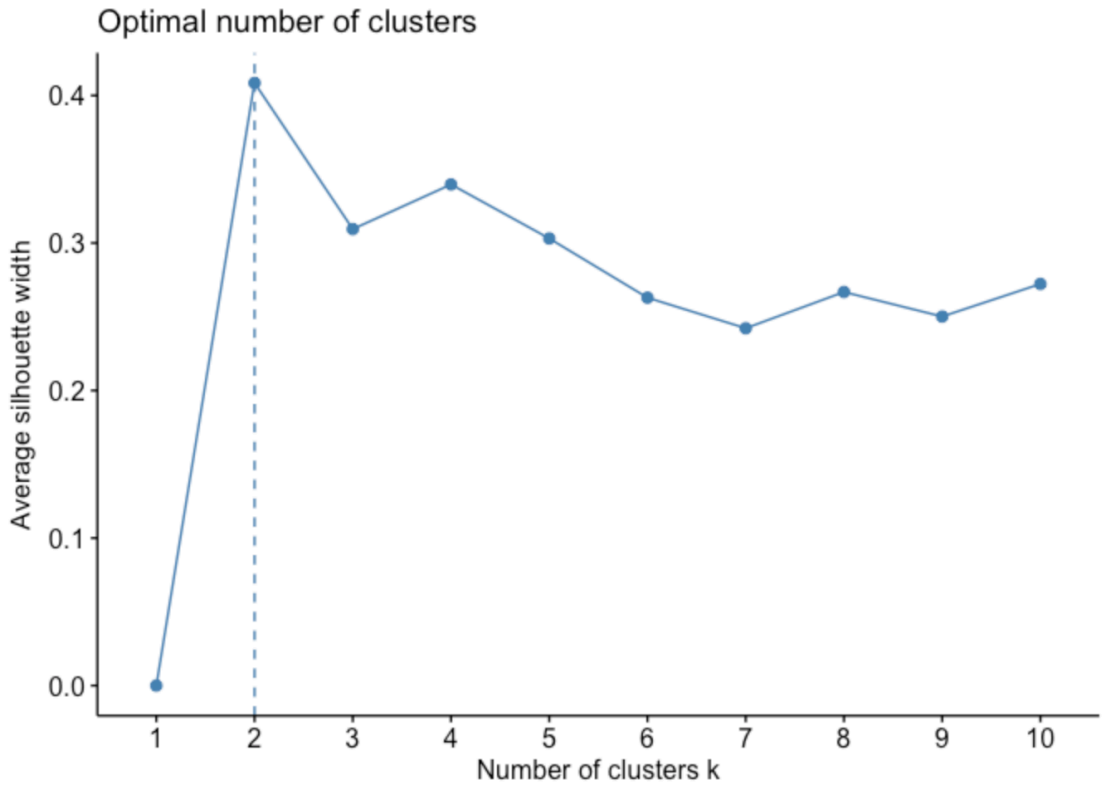
1. Εφαρμογή του K-means για τιμές του K σε κάποιο διάστημα, π.χ. από 1 έως 10
2. Για κάθε τιμή του k υπολογίζουμε τη μέση silhouette των παρατηρήσεων (avg.sil)
3. Σχεδιασμός γραφικής παράστασης avg.sil-k
4. Το σημείο στο οποίο εμφανίζεται μέγιστο θεωρείται ως ο ιδανικός αριθμός συμπλεγμάτων K

Για κάθε παρατήρηση i ο δείκτης αυτός υπολογίζεται ως εξής:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \qquad , -1 \leq s(i) \leq 1,$$

όπου $a(i)$ η μέση απόσταση της i παρατήρησης από όλες τις παρατηρήσεις στο ίδιο σύμπλεγμα. Όσο πιο μικρή είναι η τιμή $a(i)$, τόσο πιο καλά θεωρούμε ότι έχει τοποθετηθεί η παρατήρηση i. Στη συνέχεια ορίζουμε τη μέση ανομοιομορφία ενός σημείου i με ένα σύμπλεγμα C_k, ως την μέση τιμή της απόστασης του i απ' όλες τις παρατηρήσεις στο σύμπλεγμα C_k. Υπολογίζουμε την παραπάνω τιμή για κάθε σύμπλεγμα στο οποίο δεν ανήκει η παρατήρηση i και ορίζουμε τη μικρότερη από αυτές ως $b(i)$. Το σύμπλεγμα από το οποίο προέκυψε η τιμή $b(i)$ ονομάζεται γειτονικό του i. Επιπλέον ο συντελεστής $s(i)$ παίρνει τιμές από το -1 μέχρι το 1, και όσο πιο κοντά στο 1 είναι, τόσο πιο καλά τοποθετημένη είναι η παρατήρηση i. Αν η τιμή είναι κοντά στο -1 σημαίνει ότι θα ήταν καλύτερο να είχε τοποθετηθεί στο γειτονικό σύμπλεγμα, ενώ για τιμές κοντά στο 0 σημαίνει ότι η παρατήρηση βρίσκεται στο όριο μεταξύ συμπλεγμάτων. Επομένως, υπολογίζοντας την τιμή αυτή για κάθε παρατήρηση και παίρνοντας τη μέση τιμή τους, προκύπτει η τιμή avg.sil, η οποία ανάλογα με το πόσο κοντά στο 1 βρίσκεται μπορούμε να κρίνουμε αν η ομαδοποίηση που προέκυψε είναι καλή η όχι. Οι παραπάνω τιμές μπορούν να υπολογιστούν και με τη χρήση άλλων μετρικών πέρα από την ευκλείδεια.

Στην εικόνα 2.8 βλέπουμε την τιμή avg.sil στον Y άξονα για κάθε τιμή K του X άξονα και παρατηρούμε ότι για K=2, η διαμέριση που προέκυψε είναι καλύτερη από αυτήν για τις υπόλοιπες τιμές. Για K=2 έχουμε τη μεγαλύτερη τιμή avg.sil, και συνεπώς θα είναι και πιο κοντά στο 1 σε σχέση με τις υπόλοιπες, γεγονός που αντιστοιχεί σε καλύτερη διαμέριση.



Εικόνα 2.8 Silhouette μέθοδος

Gap Statistic μέθοδος

Η Gap Statistic μέθοδος δημοσιεύτηκε από τους Tibshirani, Walther και Hastie (2001) και μπορεί να εφαρμοστεί με πολλές τεχνικές ομαδοποίησης των παρατηρήσεων. Δεδομένου ότι αυξάνοντας τον αριθμό K είναι αναμενόμενο η τιμή W(C) (σχέση 3) να μειώνεται, χρειαζόμαστε ένα δείκτη που να μας λέει πότε μία τέτοια μείωση είναι σημαντική. Η μέθοδος αυτή συγκρίνει τη συνολική εσω-συμπλεγματική απόκλιση για διαφορετικές τιμές του k με τις αναμενόμενες τιμές τους υπό μηδενική κατανομή αναφοράς των δεδομένων (π.χ. ομοιόμορφη). Τα δεδομένα αναφοράς παράγονται κάνοντας χρήση προσομοίωσης Monte Carlo, δημιουργώντας για κάθε μεταβλητή x_i n τιμές, με βάση την ομοιόμορφη κατανομή στο διάστημα $[min(x_i), max(x_i)]$ για κάθε i=1….m. Με βάση την παραπάνω διαδικασία παράγουμε B δείγματα, και στη συνέχεια για δεδομένο K υπολογίζουμε την ποσότητα log[W(C)] μέσω της σχέσης (3) για το κάθε ένα και παίρνουμε τον μέσο όρο αυτών των ποσοτήτων. Συμβολίζοντας με $logW_{unif}(k)$ την παραπάνω ποσότητα έχουμε την τιμή της συνάρτησης gap:

$$Gap(k) = logW_{unif}(k) + logW(k)\,, \qquad (4)$$

όπου το K βρίσκεται σε ένα διάστημα και $logW(k)$ αναφέρεται στις n αρχικές παρατηρήσεις που θέλουμε να ομαδοποιήσουμε. Η ιδανική τιμή για τον αριθμό των συμπλεγμάτων θα είναι

αυτή που μεγιστοποιεί τη στατιστική συνάρτηση gap (σχέση 4). Αυτό σημαίνει ότι για εκείνη τη τιμή του K η διαμέριση που προέκυψε διαφέρει αρκετά από αυτήν της ομοιόμορφης κατανομής. Άν γράψουμε την παραπάνω διαδικασία με μορφή ψευδοκώδικα θα έχουμε:

Αλγόριθμος 2.4 Gap Statistic

Δεδομένα: Σύνολο των παρατηρήσεων Π , διαστάσεις m

Αποτέλεσμα: Προτεινόμενο K

1. Τοποθετούμε τις παρατηρήσεις σε συμπλέγματα για διάφορες τιμές του k = 1,….,k_{max} και υπολογίζουμε την αντίστοιχη συνολική εσω-συμπλεγματική απόκλιση *W(k)*.

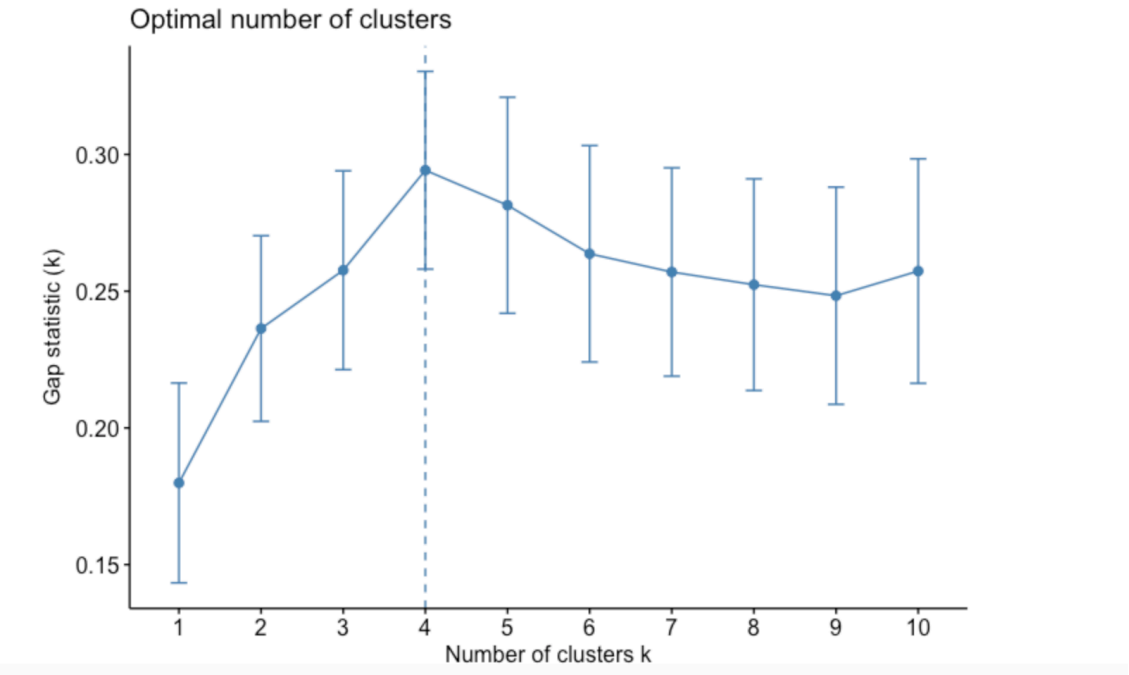
2. Παράγουμε B σύνολα δεδομένων με μία τυχαία ομοιόμορφη κατανομή. Ομαδοποιούμε τα δεδομένα αυτά σε συμπλέγματα για κάθε μία από τιμές του k = 1,....,k_{max} και υπολογίζουμε την αντίστοιχη συνολική εσω-συμπλεγματική απόκλιση W_{kb}.

3. Έστω $\bar{w} = \frac{1}{B}\sum_b log(W_{kb})$, υπολογισμός της τυπικής απόκλισης

$$sd(k) = \sqrt{(1/b)\sum_b (log(W_{kb}) - \bar{w})^2} \quad \text{και} \quad s_k = sd(k) \times \sqrt{1 + 1/B}$$

4. Επιλέγουμε τον μικρότερο αριθμό k για τον οποίο ισχύει:

$$Gap(k) \geq Gap(k+1) - s_{k+1} \qquad (5)$$



Εικόνα 2.9 Gap statistic μέθοδος

Στην εικόνα 2.9 βλέπουμε τις τιμές της συνάρτησης Gap (σχέση 4) για διάφορες τιμές του K, όπου παρατηρούμε ότι η μικρότερη τιμή του K για την οποία ικανοποιείται η ανίσωση 5 είναι η K=4. Για τη δεδομένη τιμή του K η διαφοροποίηση της διαμέρισης από την ομοιόμορφη κατανομή είναι μεγαλύτερη σε σχέση με την τιμή K=5.

2.4 Γενίκευση Αλγορίθμου

Ο αλγόριθμος K-means όπως τον ορίσαμε πριν (αλγόριθμος 2.1), χρησιμοποιεί την έννοια της απόστασης ως ένα μέτρο ομοιότητας μεταξύ 2 παρατηρήσεων. Στην πραγματικότητα ως απόσταση μπορούμε να χρησιμοποιήσουμε οποιαδήποτε συνάρτηση ικανοποιεί τις ιδιότητες μίας μετρικής, δεδομένου ότι μπορούμε να αποδείξουμε ότι ο αλγόριθμος θα συγκλίνει, κάνοντας χρήση αυτής της συνάρτησης στο κριτήριο ελαχιστοποίησης. Μία συνάρτηση

$d : \Pi x \Pi \to \mathbb{R}$ θα καλείται μετρική αν ικανοποιεί τα εξής:

- $d(x, y) = 0 \Leftrightarrow x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(z, y)$

Η επιλογή της κατάλληλης μετρικής εξαρτάται από τη φύση των παρατηρήσεων και το πεδίο εφαρμογής τους. Παραδείγματα μετρικών που χρησιμοποιούνται συχνά είναι:

α) **Μετρική Minkowski** L_q

Η μετρική αυτή μετράει την απόσταση d μεταξύ 2 αντικειμένων x και y, συγκρίνοντας τις τιμές των m χαρακτηριστικών τους. Μπορεί να εφαρμοστεί σε συχνότητες, πιθανότητες και δυαδικές τιμές

$$d(x, y) = L_q(x, y) = \sqrt[q]{\sum_{i=1}^{m} (x_i - y_i)^q}.$$

Ειδικές περιπτώσεις της παραπάνω μετρικής αποτελεί η Ευκλείδεια για q=2 και η Manhattan για q=1

- Manhattan Απόσταση L_1

$$d(x, y) = L_1 = \sum_{i=1}^{m} |x_i - y_i|$$

- Ευκλείδεια Απόσταση L_2

$$d(x, y) = L_2 = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

β) **Μετρική Cosine**

Η $cos(x, y)$ μετράει την ομοιότητα 2 αντικειμένων x και y, υπολογίζοντας το συνημίτονο της γωνίας μεταξύ των χαρακτηριστικών τους διανυσμάτων στον $\mathbb{R}^m$. Παίρνει τιμές μεταξύ του -1 (μεγαλύτερος βαθμός ανομοιομορφίας με μεταξύ τους γωνία 180°) και 1 (μεγαλύτερος βαθμός ομοιομορφίας με μεταξύ τους γωνία 0°). Η παραπάνω μετρική μπορεί να εφαρμοστεί σε συχνότητες, πιθανότητες και δυαδικές τιμές

$$sin(x, y) = cos(x, y) = \frac{\sum_{i=1}^{m} x_i * y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} * \sqrt{\sum_{i=1}^{m} y_i^2}} .$$

γ) **Απόκλιση του Bregman**

Έστω $F : \Omega \rightarrow \mathbb{R}$ απείρως παραγωγίσιμη, κυρτή συνάρτηση, ορισμένη σε ένα κλειστό κυρτό σύνολο Ω. Η απόσταση Bregman που σχετίζεται με την F για σημεία p,q του Ω ορίζεται ως εξής:

$$D_F(p, q) = F(p) - F(q) - < \bigtriangledown F(q), p - q >$$

Η παραπάνω συνάρτηση έτσι όπως έχει οριστεί δεν ικανοποιεί όλες τις ιδιότητες της μετρικής. Για δεδομένες συναρτήσεις F προκύπτουν γνωστές μετρικές. Για παράδειγμα για $F(x) = ||x||^2$ θα είναι η τετραγωνική ευκλείδεια απόσταση. Αποτελεί επομένως μία γενικότερη κλάση συναρτήσεων ομοιότητας. Στην πραγματικότητα εφαργογές του K means με χρήση της Ευκλείδειας αλλά και της μετρικής cosine αποτελούν ειδικές περιπτώσεις ενός γενικότερου αλγορίθμου ομαδοποίησης, που βασίζονται στις αποκλίσεις του Bregman.

Στη συνέχεια έχοντας επιλέξει την κατάλληλη μετρική, θα πρέπει να τροποποιήσουμε ανάλογα την τιμή που θέλουμε να ελαχιστοποιήσουμε:

$$f_{\Sigma}(C) := \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{m} dist(x_{ij}, \hat{z}_{kj}) , \qquad (6)$$

όπου dist( ) η επιλεγμένη μετρική και $\hat{z}_{kj}$ ο μέσος για την κάθε συστάδα, τη μορφή του οποίου μπορούμε να υπολογίσουμε με μία καθαρά μαθηματική διαδικασία. Απαιτώντας η ποσότητα (6) να είναι ελάχιστη, μπορούμε να λύσουμε ως προς $\hat{z}_{kj}$, προσδιορίζοντας την κατάλληλη τιμή για την οποία έχουμε ελαχιστοποίηση. Για παράδειγμα για την Ευκλείδεια απόσταση θα έχουμε :

$$\frac{\partial f_{\Sigma}}{\partial C_k} = \frac{\partial}{\partial C_k} \sum_{i=1}^{K} \sum_{x \varepsilon C_i} (z_i - x)^2 = \sum_{i=1}^{K} \sum_{x \varepsilon C_i} \frac{\partial}{\partial C_k} (z_i - x)^2 =$$

$$= \sum_{x \varepsilon C_k} 2 * (z_k - x_k) = 0 \Rightarrow n_k z_k = \sum_{x \varepsilon C_k} x_k \Rightarrow \hat{z}_k = \frac{1}{n_k} \sum_{x \varepsilon C_k} x_k$$

□

Με παρόμοιο τρόπο λειτουργούμε σε οποιαδήποτε άλλη περίπτωση άμα θεωρήσουμε κάποια άλλη μετρική αντί για την Ευκλείδεια. Δεδομένου μετρικής, μέσου και αντικειμενικής συνάρτησης (σχέση 6) προκύπτει και μία διαφορετική υλοποίηση του αλγορίθμου, η οποία θα πρέπει να σιγουρευτούμε ότι συγκλίνει.

Πίνακας 2.1 K-means: Κοινές επιλογές μετρικής,μέσου και αντικειμενικής συνάρτησης

Μετρική	Μέσος	Αντικειμενική Συνάρτηση
Manhattan (L_1)	Διάμεσος	Ελαχιστοποίηση αθροίσματος της L_1 απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο
Τετραγωνική Ευκλείδεια (L_2^2)	Μέση Τιμή	Ελαχιστοποίηση αθροίσματος της L_2^2 απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο
cosine	Μέση Τιμή	Ελαχιστοποίηση αθροίσματος της cosine απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο
Απόκλιση του Bregman	Μέση Τιμή	Ελαχιστοποίηση αθροίσματος της απόκλισης του Bregman απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο

Στο πίνακα 2.1 παρατηρούμε ότι κάνοντας χρήση της μετρικής Manhattan (L_1) ο μέσος που ελαχιστοποιεί τη σχέση 6 θα είναι η διάμεσος. Επιπλέον, οι μετρικές L_2^2, cosine αποτελούν ειδικές περιπτώσεις της απόκλισης του Bregman, γεγονός που φαίνεται και από την κοινή μορφή του μέσου.

2.5 Πλεονεκτήματα και Μειονεκτήματα

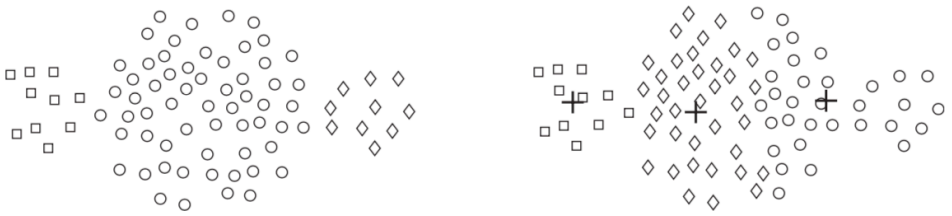
Πλεονεκτήματα k-means

- **Απλότητα**: Αποτελεί ένα αρκετά απλό αλγόριθμο τόσο στην κατανόηση, όσο και στην υλοποίηση, που μπορεί να χρησιμοποιηθεί για μεγάλη ποικιλία δεδομένων.
- **Αποδοτικότητα**: Λόγω της σχεδόν γραμμικής πολυπλοκότητάς του ως προς το σύνολο των παρατηρήσεων, είναι αρκετά αποδοτικός για μεγάλα σύνολα δεδομένων, ακόμα και αν τρέχει πολλές φορές.
- **Σφαιρικά Συμπλέγματα**: Επιστρέφει αρκετά καλά αποτελέσματα όταν η μορφή των συμπλεγμάτων είναι σφαιρική, δηλαδή τα χαρακτηριστικά των παρατηρήσεων εντός κάθε συμπλέγματος έχουν κοινή διασπορά και είναι ανεξάρτητα μεταξύ τους. Εν γένει τα συμπλέγματα που θα παραχθούν θα είναι κυρτά σχήματα, όπως για παράδειγμα η μπάλα στο τρισδιάστατο χώρο (Anderberg 1973), η μορφή των οποίων εξαρτάται από την μετρική που χρησιμοποιείται κάθε φορά.

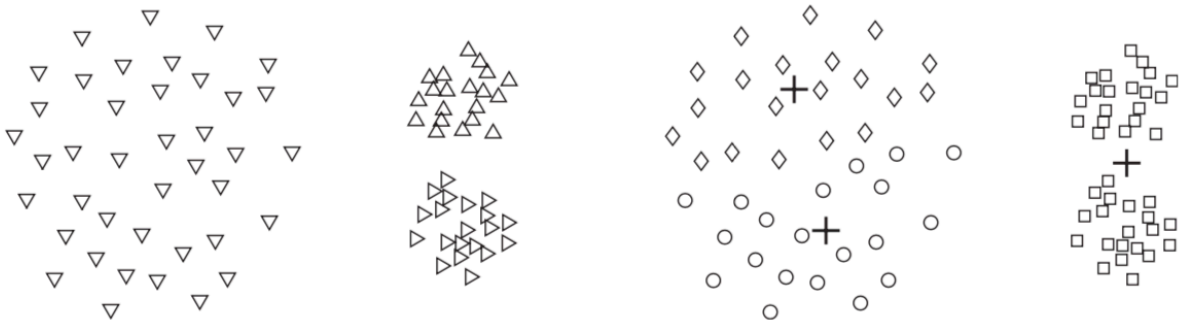
Μειονεκτήματα k-means

- **Αριθμητικά Δεδομένα:** Ο απλός αλγόριθμος K means προϋποθέτει ότι τα χαρακτηριστικά των παρατηρήσεων θα είναι όλα ποσοτικές μεταβλητές προκειμένου να εφαρμοστεί κάποια από τις μετρικές ως μέτρο ομοιογένειας των παρατηρήσεων.
- **Τοπικό Ελάχιστο:** Η διαμέριση που επιστρέφει αποτελεί τοπικό ελάχιστο και εξαρτάται από την αρχικοποίηση των συμπλεγμάτων. Διαφορετικά τρεξίματα του αλγορίθμου επιστρέφουν διαφορετικά αποτελέσματα, ενώ ιδιαίτερη σημασία έχει και η σειρά με την οποία έχουν αποθηκευτεί οι παρατηρήσεις.
- **Ακραίες Παρατηρήσεις - Μέγεθος Κλίμακας:** Το αποτέλεσμα του αλγορίθμου επηρεάζεται αρκετά από ακραίες παρατηρήσεις, καθώς επίσης και από τη διαφορά στην κλίμακα των χαρακτηριστικών που έχει κάθε παρατήρηση. Αποτελεί λογικό συμπέρασμα ότι χαρακτηριστικά που παίρνουν μεγάλες τιμές, έχουν μεγαλύτερη επίδραση στην Ευκλείδεια μετρική.
- **Διαφορετικό μέγεθος,πυκνότητα και μη σφαιρικό σχήμα:** Η ακρίβεια της διαμέρισης δεν είναι ιδιαίτερα καλή όταν τα συμπλέγματα διαφέρουν σε μέγεθός, πυκνότητα αλλά και όταν έχουν σφαιρικό σχήμα. (εικόνες 2.10, 2.11, 2.12).
- **Πολυδιάστατα δεδομένα:** Η απόδοση του αλγορίθμου επηρεάζεται αρνητικά όταν το πλήθος των χαρακτηριστικών m για κάθε παρατήρηση είναι μεγάλο (Keim, Hinneburg, 1999).

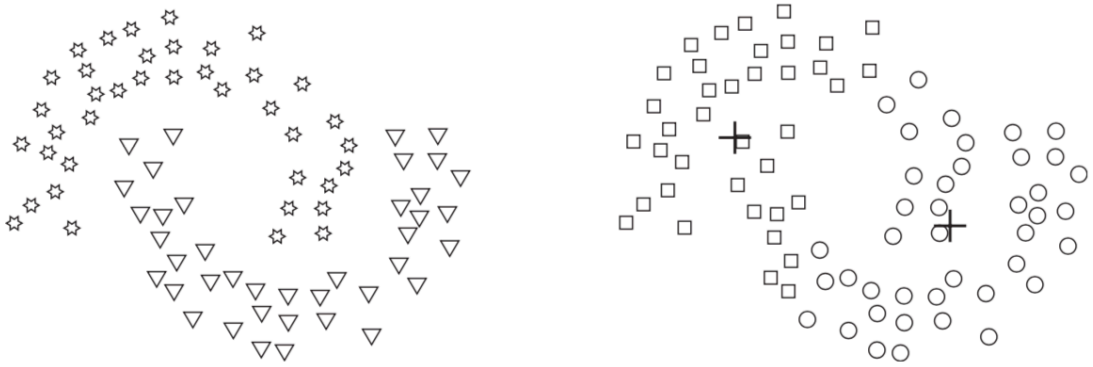
- **Ομοιόμορφο αποτέλεσμα:** Αρκετά συχνά τα συμπλέγματα που προκύπτουν έχουν σχετικά ομοιόμορφο μέγεθος, ακόμα και για σύνολα παρατηρήσεων για τα οποία αυτό δεν ισχύει στη πραγματικότητα.



Εικόνα 2.10: Συμπλέγματα διαφορετικού μεγέθους



Εικόνα 2.11: Συμπλέγματα διαφορετικής πυκνότητας



Εικόνα 2.12: Συμπλέγματα μη σφαιρικού σχήματος

Κεφάλαιο 3

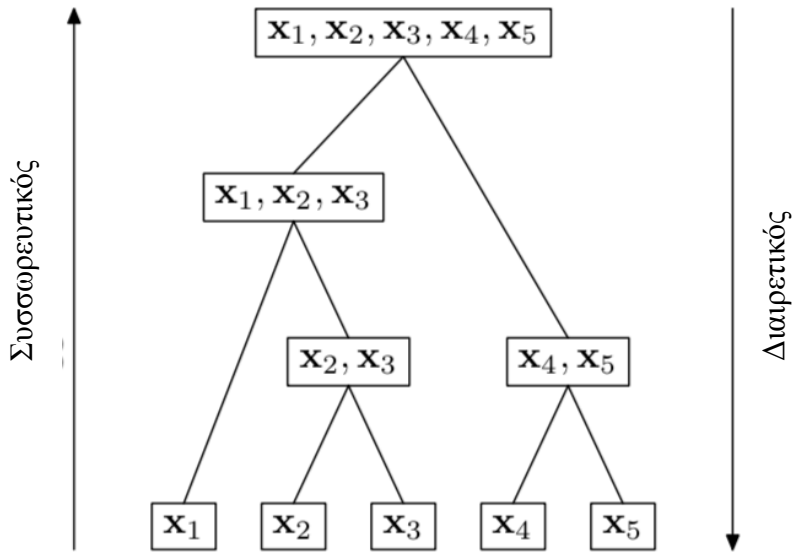
3.1 Hierarchical Clustering

Η τεχνική ομαδοποίησης, στην οποία θα γίνει αναφορά στο συγκεκριμένο κεφάλαιο, ανήκει στην ευρύτερη κατηγορία των ιεραρχικών αλγορίθμων, οι οποίοι παράγουν ιεραρχίες με υποσύνολα του συνόλου των παρατηρήσεων. Οι παραπάνω αλγόριθμοι χωρίζονται σε 2 υποκατηγορίες: α) διαιρετικούς και β) συσσωρευτικούς. Μία διαιρετική μέθοδος ξεκινάει με όλες τις παρατηρήσεις σε ένα σύμπλεγμα το οποίο σταδιακά μειώνεται σε μικρότερα κομμάτια. Αντίθετα οι συσσωρευτικές τεχνικές ξεκινάνε με κάθε παρατήρηση σε διαφορετικό σύμπλεγμα, και στη συνέχεια ενώνουν τα 2 πιο κοντινά συμπλέγματα σε ένα, μειώνοντας έτσι το συνολικό αριθμό σε n-1. Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι όλες οι παρατηρήσεις να τοποθετηθούν σε ένα σύμπλεγμα. Οι ιεραρχικές μέθοδοι γενικά ορίζονται ως εξής:

Ορισμός 3.1 Έστω $\Pi = [x_1, \ldots, x_n]$ ένα σύνολο n παρατηρήσεων. Ένα σύνολο $S = (C_1, C_2, \ldots, C_K)$ από υποσύνολα του Π θα καλείται ιεραρχία του Π, εάν όλα τα σύνολα $C_1, C_2, \ldots, C_K \sqsubseteq \Pi$ είναι διαφορετικά και αν και για 2 οποιαδήποτε σύνολα $C_k, C_l \sqsubseteq \Pi$ με $C_k \neq C_l$ μπορεί να ισχύει μόνο ένα από τα ακόλουθα:

$$C_k \cap C_l = \emptyset \quad \text{ή} \quad C_k \subset C_l \quad \text{ή} \quad C_l \subset C_k$$

Τα σύνολα στο $S = (C_1, C_2, \ldots, C_K)$ είναι γνωστά ως οι κλάσεις του Π.



Εικόνα 3.1 Συσσωρευτική και διαιρετική ιεραρχική ομαδοποίηση

3.2 Η μέθοδος Single-link

Η Single-Link μέθοδος είναι μία από τις πιο απλές συσσωρευτικές ιεραρχικές μεθόδους. Εισήχθη για πρώτη φορά από τους Florek et al. (1951) και στη συνέχεια ανεξάρτητα από τους McQuitty (1957) και Sneath (1957). Προκειμένου να κατανοήσουμε την λειγουργία της μεθόδου, θα αναφερθούμε στους πίνακες εγγύτητας των παρατηρήσεων. Ο πίνακας εγγύτητας (Jain and Dubes, 1988) είναι ένας πίνακας που περιέχει τους ζευγαρώδεις δείκτες εγγύτητας ενός συνόλου παρατηρήσεων. Ο δείκτης εγγύτητας μπορεί να αναφέρεται σε απόσταση μεταξύ 2 παρατηρήσεων μέσω κάποιας μετρικής, όπως έχει οριστεί σε προηγούμενη ενότητα. Επιπλέον μπορεί να είναι κάποια συνάρτηση ομοιότητας $s : \Pi x \Pi \rightarrow \mathbb{R}$ η οποία ικανοποιεί τις εξής ιδιότητες:

1. $0 \leq s(x, y) \leq 1$
2. $s(x, x) = 1$
3. $s(x, y) = s(y, x)$

Επομένως, δεδομένου ενός συνόλου παρατηρήσεων $\Pi = [x_1, x_2, \ldots, x_n]$ ο πίνακας εγγύτητας ορίζεται ως εξής:

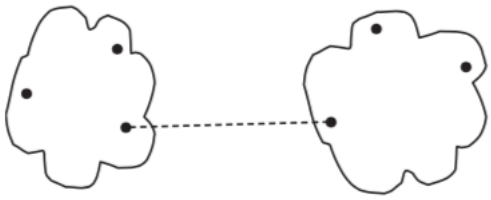
$$M(\Pi) = \begin{pmatrix} f(1,1) & f(1,2) & .. & f(1,n) \\ f(2,1) & . & & . \\ . & & . & . \\ f(n,1) & & & f(n,n) \end{pmatrix} \qquad (1)$$

όπου f(,) θα είναι είτε κάποια μετρική, είτε κάποια συνάρτηση ομοιότητας των παρατηρήσεων. Συνήθως ο παραπάνω πίνακας είναι συμμετρικός.

Ο αλγόριθμος Single-link δημιουργεί ιεραρχίες παρατηρήσεων ξεκινώντας με n συμπλέγματα από 1 παρατήρηση στο κάθε ένα. Στη συνέχεια, βρίσκουμε τα 2 συμπλέγματα τα οποία είναι πιο κοντά (όμοια) μεταξύ τους και τα ενώνουμε, μειώνοντας το συνολικό αριθμό των συμπλεγμάτων σε n-1. Η παραπάνω διαδικασία συνεχίζεται μέχρι όλες οι παρατηρήσεις να έχουν τοποθετηθεί σε ένα σύμπλεγμα. Παρατηρούμε ότι βασική διαδικασία είναι ο υπολογισμός της εγγύτητας μεταξύ 2 συμπλεγμάτων, ο ορισμός της οποίας είναι αυτός που διαφοροποιεί τις διαφορετικές τεχνικές συσσωρευτικής ιεραρχικής ομαδοποίησης. Στη μέθοδο single-link η εγγύτητα μεταξύ 2 συμπλεγμάτων C_i και C_j ορίζεται ως η ελάχιστη απόσταση-ομοιότητα των αντικειμένων των συμπλεγμάτων.

$$d(C_i, C_j) = d_{min}(C_i, C_j) = min_{x \varepsilon C_i, y \varepsilon C_j} f(x, y) \qquad (2)\,,$$

όπου f( ) η συνάρτηση του πίνακα εγγύτητας. Μέσω της παραπάνω σχέσης, υπολογίζονται οι αποστάσεις μεταξύ των συμπλεγμάτων σε κάθε βήμα του αλγορίθμου, κάνοντας την αντίστοιχη ενημέρωση στον πίνακα εγγύτητας M (σχέση 1) μετά από κάθε συγχώνευση.



Εικόνα 3.2 Απόσταση μεταξύ 2 συμπλεγμάτων

Η έννοια της απόστασης όπως ορίστηκε από τη σχέση (2) αποτελεί ειδική περίπτωση της σχέσης Lance-Williams για τον υπολογισμό της εγγύτητας μεταξύ συμπλεγμάτων. Με άλλα λόγια, αν η ένωση του A και του B μας κάνει το σύμπλεγμα R, η εγγύτητα του R με ένα προϋπάρχον σύμπλεγμα Q δίνεται από τη σχέση:

$$d(R,Q) = \alpha_A d(A,Q) + \alpha_B d(B,Q) + \beta d(A,B) + \gamma \,|\, d(A,Q) - d(B,Q) \,|\qquad (3)$$

Αν στην παραπάνω σχέση θέσουμε $\alpha_A = 0.5$, $\alpha_B = 0.5$, $\beta = 0$ και $\gamma = -0.5$ θα προκύψει:

$$\begin{aligned} d(R,Q) &= \frac{1}{2}d(A,Q) + \frac{1}{2}d(B,Q) - \frac{1}{2}\,|\, d(A,Q) - d(B,Q) \,|\\ &= min(d(A,Q), d(B,Q)) \end{aligned}$$

Από τα παραπάνω προκύπτει πολύ εύκολα η απόσταση μεταξύ 2 συμπλεγμάτων, όπως ορίστηκε στη σχέση (2). Για διαφορετικές τιμές των παραμέτρων της σχέσης (3) προκύπτει μία εντελώς διαφορετική μέθοδος, καθώς βασίζεται σε διαφορετικό υπολογισμό της εγγύτητας.

Αλγόριθμος 3.1 Single-link Hierarchical Clustering
Δεδομένα: Σύνολο των παρατηρήσεων $\Pi = (x_1, x_2, \ldots, x_n)$
Αποτέλεσμα: Ιεραρχία συμπλεγμάτων
1. Για κάθε $x_i \varepsilon \Pi$ επανάλαβε
2. όρισε ως σύμπλεγμα $C_i = (x_i)$
3. Έστω $C = (C_1, C_2, \ldots, C_n)$
4. Όσο $|\, C \,| \neq 1$ κάνε
5. Για όλα τα ζευγάρια συμπλεγμάτων $< C_i, C_{j\neq i} > \varepsilon CxC$ επανάλαβε
6. υπολόγισε $d(C_i, C_j)$
7. Όρισε $best(C_i, C_j) = \forall < C_{k\neq i}, C_{l\neq k,j} > \varepsilon CxC : [d(C_i, C_j) \leq d(C_k, C_l)]$
8. Για $best(C_i, C_j)$ κάνε
9. όρισε $C_{ij} = C_i \cup C_j$
10. όρισε $C^{new} = C - (C_i, C_j)$
11. όρισε $C = C^{new} \cup C_{ij}$
12. Τέλος

Στον παραπάνω ψευδοκώδικα συνοψίζεται η διαδικασία του αλγορίθμου που αναφέραμε νωρίτερα. Παρατηρούμε ότι δεν προσδιορίζουμε τον αριθμό των συμπλεγμάτων όπως στον αλγόριθμο k means. Επιπλέον, έχει την ιδιότητα μονοτονίας, δηλαδή η ανομοιογένεια μεταξύ συγχωνευμένων συμπλεγμάτων είναι μονότονη και αυξάνεται με το επίπεδο της συγχώνευσης.

Ανάλυση πολυπλοκότητας χρόνου και χώρου

Αρχικά θα αναλύσουμε την πολυπλοκότητα χώρου, που οφείλεται κυρίως στον πίνακα εγγύτητας και στα συμπλέγματα σε κάθε βήμα του αλγορίθμου. Δεδομένου ότι έχουμε n παρατηρήσεις και ότι ο πίνακας είναι συμμετρικός, θα έχει συνολικά $\frac{1}{2}n^2$ στοιχεία. Επιπλέον για την αποθήκευση των συμπλεγμάτων, αρχικά απαιτούνται n στοιχεία με τον αριθμό τους να μειώνεται κατά 1 μέχρι να τερματίσει ο αλγόριθμος. Επομένως η συνολική πολυπλοκότητα χώρου είναι $O(n^2)$.

Για την ανάλυση της πολυπλοκότητας χρόνου θα θεωρήσουμε ότι για την αποθήκευση του πίνακα εγγύτητας γίνεται χρήση ενός πίνακα με αποτέλεσμα η συμπλήρωση του, καθώς και η εύρεση του βέλτιστου ζεύγους να απαιτούν $O(n^2)$. Τα βήματα 1,2 αποτελούν το στάδιο της αρχικοποίησης και είναι $O(n)$, καθώς το βήμα 1 πραγματοποιείται n φορές. Επιπλέον, το βήμα 4 θα τρέξει n φορές καθώς ξεκινάμε με n συμπλέγματα, τα οποία σε κάθε βήμα του αλγόριθμου μειώνονται κατά 1 μέχρι όλες οι παρατηρήσεις να ανήκουν στο ίδιο σύμπλεγμα. Σύμφωνα με τα παραπάνω η συνολική πολυπλοκότητα χρόνου θα είναι $O(n^3)$. Παρατηρούμε ότι η πολυπλοκότητα χώρου και χρόνου της ιεραρχικής ομαδοποίησης περιορίζει σημαντικά το μέγεθος των συνόλων δεδομένων που μπορούν να επεξεργαστούν, καθώς για πολύ μεγάλες τιμές του n ο αλγόριθμος θα απαιτεί πολύ χρόνο αλλά και πολύ χώρο.

3.3 Τρόποι αναπαράστασης ιεραρχίας

n-δέντρα

Μια ιεραρχική ομαδοποίηση αντιπροσωπεύεται γενικά από ένα διάγραμμα δέντρου. Συγκεκριμένα εάν $\Pi = (x_1, x_2, \ldots, x_n)$ το σύνολο των παρατηρήσεων, τότε ένα n-δέντρο στο Π ορίζεται ώς ένα σύνολο Τ από υποσύνολα του Π, που ικανοποιούν τις παρακάτω ιδιότητες (Bobisud, 1972 McMorris et al., 1983 Gordon, 1996):

- Π ε Τ
- Το κενό σύνολο Φ ε Τ
- $(x_i)\varepsilon T \quad \forall i = 1,2...,n$
- Εάν $A, B\varepsilon T$ τότε $A \cap B$ ε (Φ, A, B)

Στην εικόνα 3.1 βλέπουμε το παράδειγμα ενός n-δέντρου για 5 παρατηρήσεις. Οι εξωτερικοί κόμβοι ή αλλιώς τα φύλλα του δέντρου είναι οι αρχικές παρατηρήσεις του συνόλου

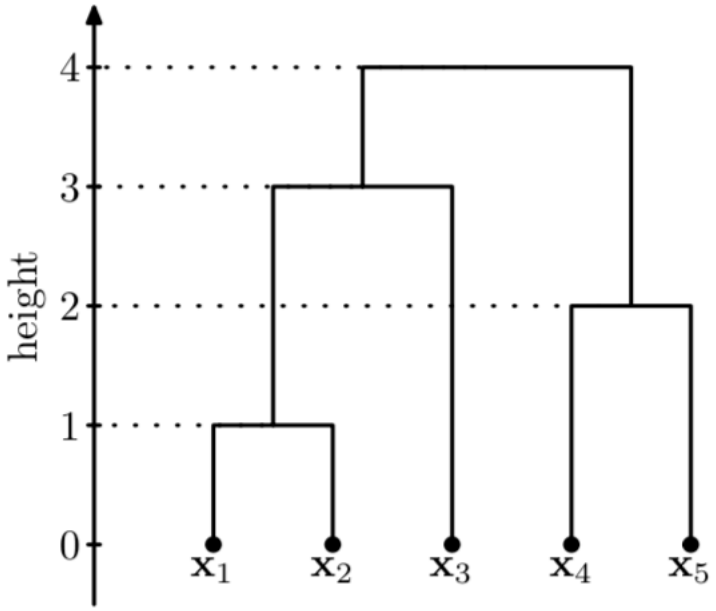
$\Pi = (x_1, x_2, x_3, x_4, x_5)$, ενώ οι εσωτερικοί κόμβοι είναι τα συμπλέγματα που προκύπτουν από συγχωνεύσεις συμπλεγμάτων μικρότερου ύψους. Επιπλέον, παρατηρούμε ότι το σύνολο των εσωτερικών κόμβων είναι n-1, όσες δηλαδή και οι επαναλήψεις που τρέχει ο αλγόριθμος single-link.

Δενδρογράφημα

Ο παραπάνω τρόπος αναπαράστασης δεν προσδίδει ιδιαίτερες πληροφορίες για τους εσωτερικούς κόμβους πέρα από το περιεχόμενό τους. Γι' αυτό το λόγο, γίνεται χρήση των δενδρογραφημάτων τα οποία αποτελούν δέντρα με τιμές (Gordon, 1996). Ένα δενδρογράφημα είναι ένα n-δέντρο, στο οποίο κάθε εσωτερικός κόμβος συσχετίζεται με ένα ύψος που ικανοποιεί την ιδιότητα:

$$h(A) \leq h(B) \Leftrightarrow A \subseteq B$$

για κάθε υποσύνολο των παρατηρήσεων A και B, εάν $A \cap B \neq \Phi$, όπου $h(A)$ και $h(B)$ είναι τα ύψη των A και B αντίστοιχα.



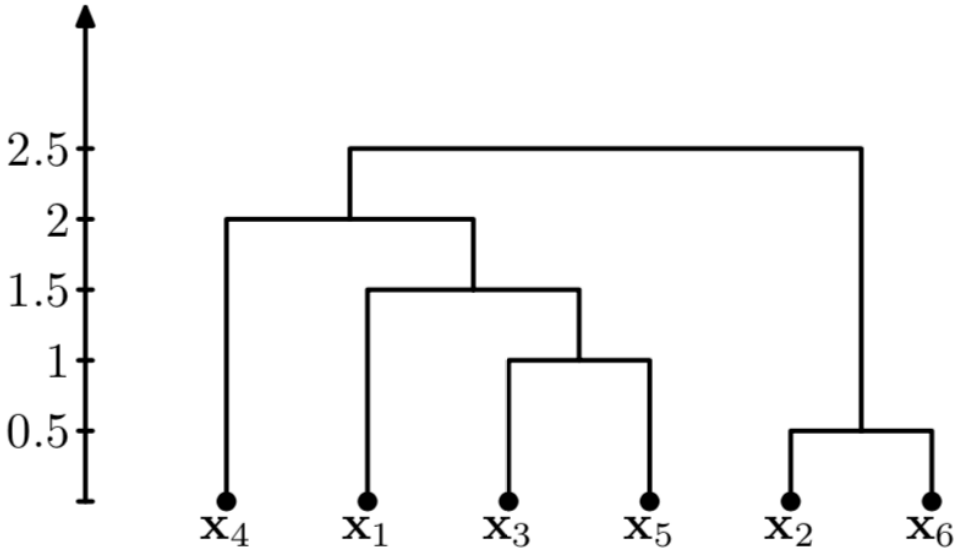
Εικόνα 3.3 Δενδρογράφημα συνόλου 5 παρατηρήσεων

Η εικόνα 3.3 αποτελεί μία αναπαράσταση ενός δενδρογραφήματος με 5 παρατηρήσεις, όπου οι διακεκομμένες γραμμές υποδεικνύουν τα ύψη των εσωτερικών κόμβων. Για κάθε ζεύγος παρατηρήσεων (x_i, x_j), έστω h_{ij} το ύψος του μικρότερου εσωτερικού κόμβου στον οποίο ανήκει η x_i, αλλά και η x_j. Μικρή τιμή της ποσότητας h_{ij} υποδεικνύει ότι οι παρατηρήσεις x_i και x_j βρέθηκαν στο ίδιο σύμπλεγμα σε σχετικά μικρό ύψος, γεγονός που συμβολίζει μεγάλη ομοιότητα μεταξύ των x_i και x_j.

Τα ύψη σε ένα δενδρογράφημα ικανοποιούν την ακόλουθη υπερμετρική συνθήκη (Johnson, 1967):

$$h_{ij} \leq max(h_{ik}, h_{jk}) \qquad \forall i, j, k \varepsilon (1,2,...,n)$$

Στην πραγματικότητα η παραπάνω συνθήκη είναι ικανή και αναγκαία για ένα δενδρογράφημα (Gordon, 1987). Εναλλακτικά, αντί για τα ύψη στον κάθετο άξονα μπορεί να γίνει καταγραφή της τιμής της εγγύτητας, για την οποία πραγματοποιείται κάθε μία από τις n-1 συγχωνεύσεις (εικόνα 3.4).



Εικόνα 3.4 Δενδρογράφημα συνόλου 6 παρατηρήσεων

Στο δενδρογράφημα της εικόνας 3.4 παρατηρούμε την ιδιότητα της μονοτονίας την οποία αναφέραμε νωρίτερα, ότι δηλαδή όσο πιο ψηλά βρισκόμαστε στο δέντρο, τόσο μεγαλύτερη είναι η ανομοιογένεια μεταξύ συμπλεγμάτων που συγχωνεύονται. Παρατηρούμε ότι το παραπάνω γράφημα παρέχει πληροφορίες τόσο για τη σειρά των συγχωνεύσεων, όσο και για την ομοιομορφία μεταξύ των συμπλεγμάτων, όπως αυτή ορίζεται από τη σχέση 2. Ο ορισμός διαφορετικών τιμών στις παραμέτρους στη σχέση του Lance-Williams (σχέση 3) όπως και μικρές διαφοροποιήσεις στα δεδομένα μπορούν να οδηγήσουν σε ένα τελείως διαφορετικό δενδρογράφημα. Επιπλέον η μέθοδος single-link είναι αμετάβλητη υπό μονοτονικούς μετασχηματισμούς (όπως ο λογαριθμικός μετασχηματισμός) των αρχικών δεδομένων (Johnson, 1967), γεγονός που σημαίνει ότι το δενδρογράφημα που απορρέει είναι το ίδιο.

3.4 Αξιολόγηση Ιεραρχίας

Cophenetic Correlation

Ο παραπάνω αλγόριθμος δημιουργεί ιεραρχίες συμπλεγμάτων για κάθε σύνολο δεδομένων που δέχεται ως είσοδο, ακόμα και αν αυτό δεν έχει στη πραγματικότητα δομή ιεραρχίας. Επομένως προκειμένου να γίνει αξιόλογηση του κατά πόσο το παραγόμενο δενδρογράφημα ανταποκρίνεται στην πραγματική δομή των δεδομένων, γίνεται χρήση της cophenetic correlation (Sokal και Rohlf). Έστω T το παραγόμενο δενδρογράφημα ύστερα από τη χρήση του αλγορίθμου single-link σε ένα σύνολο παρατηρήσεων $\Pi = (x_1, x_2, \ldots, x_n)$ και $f(x_i, x_j)$ η συνάρτηση για τον υπολογισμό του πίνακα εγγύτητας (σχέση 1). Επιπλέον, έστω $t(x_i, x_j)$ η εγγύτητα κατά την οποία οι παρατηρήσεις x_i και x_j βρέθηκαν για πρώτη φορά στο ίδιο σύμπλεγμα στο δενδρογράφημα (για παράδειγμα στην εικόνα 3.4 το $t(x_2, x_6) = 0.5$). Εάν $\bar{f}$ και $\bar{t}$ οι μέσες τιμές των $f(x_i, x_j)$ και $t(x_i, x_j)$ αντίστοιχα, τότε η ποσότητα cophenetic correlation δίνεται από τον παρακάτω τύπο:

$$COPH = \frac{\sum_{i<j}(f(x_i, x_j) - \bar{f})(t(x_i, x_j) - \bar{t})}{\sqrt{[\sum_{i<j}(f(x_i, x_j) - \bar{f})^2][\sum_{i<j}(t(x_i, x_j) - \bar{t})^2]}} \qquad (4)$$

Η παραπάνω ποσότητα συνήθως παίρνει τιμές μεταξύ -1 και 1 (βέλτιστο αποτέλεσμα) και είναι ένα μέτρο για το πόσο πιστά ένα δενδρογράφημα διατηρεί τις ζεύξεις μεταξύ των αρχικών μη τροποποιημένων παρατηρήσεων. Οι διαγώνιες τιμές των f( ) ,t( ) αγνοούνται και λόγω της ιδιότητας συμμετρίας, η σύγκριση μπορεί να περιορίζεται στις τιμές κάτω της διαγωνίου. Επιπλέον, η παραπάνω ποσότητα (σχέση 4) χρησιμοποιείται αρκετά συχνά και για τη σύγκριση μεταξύ διαφόρων μεθόδων ιεραρχικής ομαδοποίησης, επιλέγοντας αυτή με τη μεγαλύτερη τιμή.

Δείκτης Delta

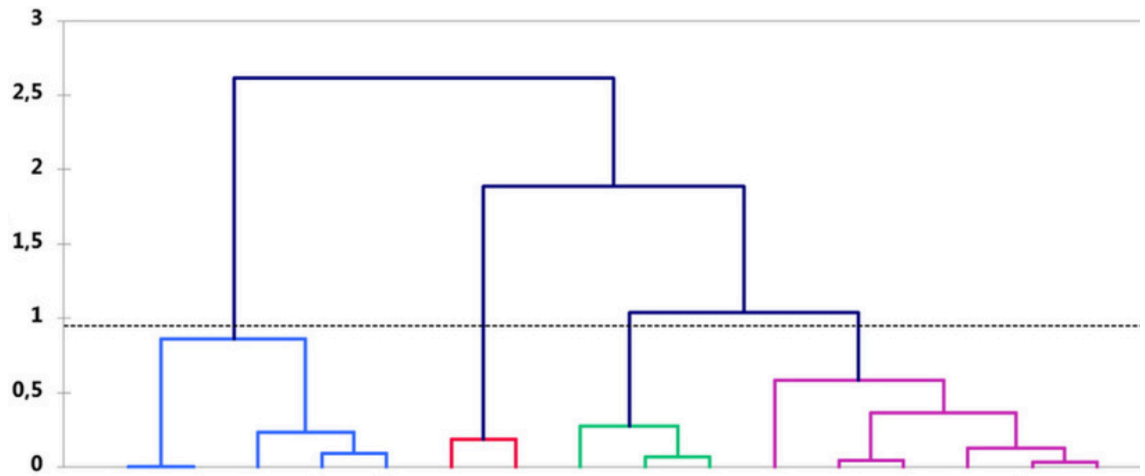
Ένα δεύτερο μέτρο καλής προσαρμογής που ονομάζεται δέλτα περιγράφεται από τον Mather (1976). Αυτά τα στατιστικά στοιχεία μετράνε τον βαθμό της στρέβλωσης και όχι τον βαθμό ομοιότητας (όπως και με τον cophenetic correlation). Οι δύο συντελεστές δέλτα δίνονται από:

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |f(x_j, x_k) - t(x_j, x_k)|^{\frac{1}{A}}}{\sum_{j<k} t(x_j, x_k)^{\frac{1}{A}}} \right]^A \qquad (5),$$

όπου A είναι είτε 0.5 είτε 1 και οι τιμές κοντά στο μηδέν είναι οι επιθυμητές.

3.5 Προσδιορισμός Διαμέρισης

Η πλειονότητα των μεθόδων διαμέρισης απαιτεί εκ των προτέρων καθορισμό του αριθμού των συστάδων (π.χ k means), επομένως αυτές οι μέθοδοι θα πρέπει να χρησιμοποιούνται εφόσον έχουμε αποκτήσει προκαταρκτική εμπειρία μέσω κάποιων άλλων αναλύσεων των δεδομένων. Παρ' όλα αυτά στην περίπτωση της ιεραρχικής συσσώρευσης και του αλγορίθμου single-link, η γνώση του αριθμού των συμπλεγμάτων δεν είναι αναγκαίο να προσδιοριστεί εκ των προτέρων. Δεδομένου του παραγόμενου δενδρογραφήματος η τελική διαμέριση μπορεί να προσδιοριστεί φέρνοντας παράλληλη στον οριζόντιο άξονα σε ύψος που επιλέγουμε εμείς. Τα υποσύνολα που ενώνονται σε απόσταση κάτω από αυτήν την τιμή τοποθετούνται στο ίδιο σύμπλεγμα. Αντίθετα υποσύνολα που ενώνονται σε απόσταση μεγαλύτερη από αυτή την τιμή τοποθετούνται σε διαφορετικά συμπλέγματα. Προφανώς διαφορετικά "κοψίματα" του δέντρου μπορεί να αντιστοιχούν σε διαφορετικές διαμερίσεις του αρχικού συνόλου.



Εικόνα 3.5 Δενδρογράφημα συνόλου 17 παρατηρήσεων

Για παράδειγμα στην εικόνα 3.5 παρατηρούμε το δενδρογράφημα που προέκυψε σε ένα σύνολο 17 παρατηρήσεων και την ευθεία που τέμνει τον κάθετο άξονα σε σημείο πολύ κοντά στο 1. Ο αριθμός των συμπλεγμάτων που προκύπτει από αυτή την επιλογή θα είναι K=4, όπου τα μέλη του κάθε συμπλέγματος θα είναι τα φύλλα των αντίστοιχων υποδέντρων. Αντίστοιχα μία επιλογή κοντά στο 1.5 θα είχε ως αποτέλεσμα 3 συμπλέγματα, ενώ μία επιλογή στο 2 θα έδινε K=2. Η επιλογή του σημείου τομής οπτικά αρκετά συχνά είναι παραπλανητική και δεν οδηγεί σε βέλτιστη διαμέριση, γεγονός που κρίνει αναγκαία τη χρήση διαφόρων μεθόδων για τον προσδιορισμό του. Παρόλο που η σωστή απάντηση σε αυτό το πρόβλημα μπορεί να εξαρτηθεί από πολλούς παράγοντες, μπορούμε να υποθέσουμε με ασφάλεια ότι ο βέλτιστος διαχωρισμός είναι αυτός που παρέχει τις πιο συμπαγείς και μέγιστα διαχωρισμένες ομάδες.

Την παραπάνω λογική υλοποιεί ο δείκτης Calin´ski και Harabasz (1974) επίσης γνωστός και ως κριτήριο αναλογικής διασποράς (VRC). Λαμβάνοντας υπόψη ένα σύνολο n παρατηρήσεων $\Pi = (x_1, x_2, \ldots, x_n)$ και μία διαμέριση με k συμπλέγματα, το VRC υπολογίζεται ως εξής:

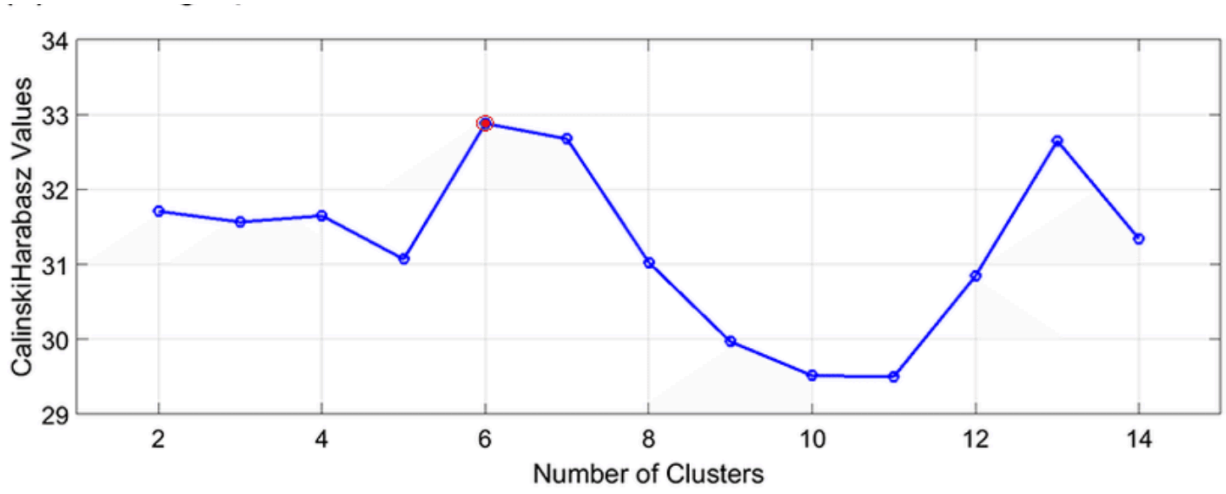
$$VRC_k = \frac{trace(B)}{trace(W)} x \frac{n-k}{k-1} \qquad (6),$$

όπου τα W και B είναι οι πίνακες διασποράς εντός συμπλέγματος και εκτός συμπλέγματος αντίστοιχα :

$$W = \sum_{i=1}^{k} \sum_{l=1}^{n_i} (\overrightarrow{x}_i(l) - \bar{x}_i)(\overrightarrow{x}_i(l) - \bar{x}_i)' \qquad B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

όπου n_i είναι ο αριθμός των παρατηρήσεων στο i-οστό σύμπλεγμα, $\overrightarrow{x_i}(l)$ είναι η l-οστή παρατήρηση σε αυτό το σύμπλεγμα, $\bar{x}_i$ ο δειγματικός μέσος του i-οστού συμπλέγματος και $\bar{x}$ ο δειγματικός μέσος όλων των παρατηρήσεων. Λαμβάνοντας υπόψη τους παραπάνω τύπους, το ίχνος του B είναι το άθροισμα των διακυμάνσεων μεταξύ συμπλεγμάτων, ενώ το ίχνος του W είναι το άθροισμα των διακυμάνσεων εντός συμπλεγμάτων. Μια καλή διαμέριση θα πρέπει να έχει υψηλές τιμές για το B (που είναι μια ένδειξη για καλά διαχωρισμένα συμπλέγματα), χαμηλές τιμές για το W (μια ένδειξη για συμπαγή συμπλέγματα), έτσι ώστε όσο υψηλότερη είναι η ποιότητα του διαχωρισμού, τόσο μεγαλύτερη είναι η τιμή αυτού του λόγου.

Προκειμένου να βρούμε τη βέλτιστη τιμή του k, άρα και το σημείο τομής, υπολογίζουμε το VRC για έναν αριθμό διαφορετικών k (π.χ. από 2 έως 14) και διατηρούμε την τιμή εκείνη που οδηγεί στην υψηλότερη τιμή VRC. Επιπλέον, η μονότονη αύξηση του δείκτη VRC συναρτήσει του k, υποδηλώνει έλλειψη οποιασδήποτε δομής ομάδας, ενώ αντίθετα η μονότονη μείωση υποδηλώνει δομή ιεραρχίας.



Εικόνα 3.6 Διάγραμμα του δείκτη Calin´ski και Harabasz για διάφορες τιμές του k

Ο εντοπισμός του βέλτιστου αριθμού k μπορεί να πραγματοποιηθεί μέσω ενός διαγράμματος VRC συναρτήσει του k. Στην εικόνα 3.6 βλέπουμε το διάγραμμα αυτό για ένα σύνολο παρατηρήσεων όπου ο βέλτιστος αριθμός επιτυγχάνεται για k=6, τέμνοντας επομένως το δέντρο στο αντίστοιχο σημείο ώστε να προκύψουν 6 συμπλέγματα.

3.6 Πλεονεκτήματα και Μειονεκτήματα

Μειονεκτήματα Single-link Hierarchical Clustering

- **Χρονική πολυπλοκότητα:**

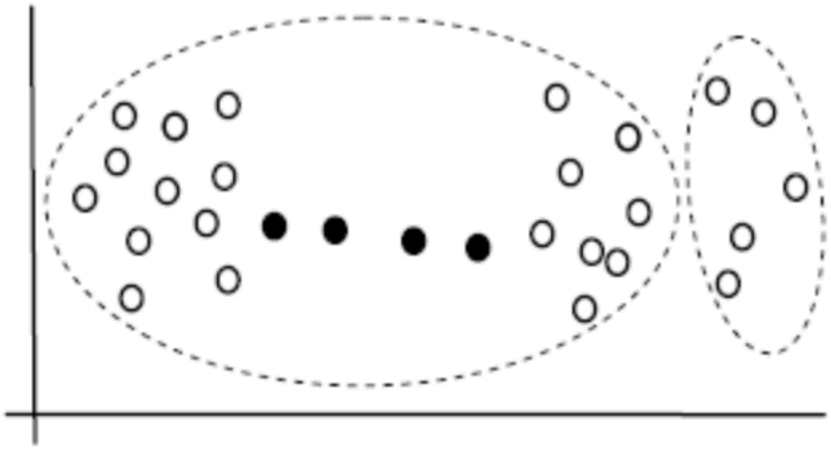
Σύμφωνα με την ανάλυση που έγινε νωρίτερα είδαμε ότι η χρονική πολυπλοκότητα της μεθόδου είναι $O(n^3)$ που με χρήση κατάλληλων δομών μπορεί να μειωθεί σε $O(n^2)$, γεγονός που μπορεί να αποτελέσει απαγορευτικό παράγοντα, όταν το σύνολο των παρατηρήσεων είναι αρκετά μεγάλο.

- **Φαινόμενο της αλυσίδας:**

Η μέθοδος single-link έχει ως μοναδική προϋπόθεση μόνο μία εγγύτητα $f(x_i, x_j)$ (όπου $x_i \varepsilon G$ και $x_j \varepsilon H$) να είναι ελάχιστη, προκειμένου τα 2 συμπλέγματα G, H να θεωρούνται κοντά μεταξύ τους ανεξαρτήτως των υπόλοιπων παρατηρήσεων. Συνεπώς, θα έχει την τάση να συνδυάζει, σε σχετικά χαμηλές τιμές, παρατηρήσεις που συνδέονται με μια σειρά στενών ενδιάμεσων παρατηρήσεων. Τα συμπλέγματα που παράγονται με τη μέθοδο single-link μπορούν να παραβιάσουν την ιδιότητα “συμπαγοποίησης", δηλαδή ότι όλες οι παρατηρήσεις μέσα σε κάθε σύμπλεγμα τείνουν να είναι παρόμοιες μεταξύ τους με βάση τις παρεχόμενες εγγύτητες των παρατηρήσεων $f(x_i, x_j)$. Εάν ορίσουμε τη διάμετρο D_G μιας ομάδας παρατηρήσεων ως τη μεγαλύτερη εγγύτητα μεταξύ των μελών της

$$D_G = max_{i,j \varepsilon G} f(x_i, x_j),$$

τότε η μέθοδος single-link μπορεί να παράγει συμπλέγματα με πολύ μεγάλες διαμέτρους.



Εικόνα 3.7 Διαμέριση παρατηρήσεων με τη μέθοδο single-link και φαινόμενο αλυσίδας

Στην εικόνα 3.7 παρατηρούμε ένα παράδειγμα φαινομένου της αλυσίδας, όπου η διαμέριση που προέκυψε με τη μέθοδο single-link δεν είναι βέλτιστη διασπώντας το ένα από τα 2 φυσικά συμπλέγματα. Γενικά επιμήκη σημειακά σύννεφα αναγνωρίζονται, αλλά δεν είναι δυνατή η ανίχνευση συμπλεγμάτων που συνδέονται με ενδιάμεσες παρατηρήσεις. Η εσωτερική συνοχή των συμπλεγμάτων είναι απολύτως αδιάφορη και ένα μικρό αρχικό σύμπλεγμα μπορεί εύκολα να προσελκύσει τις άλλες παρατηρήσεις μία προς μία στα στάδια ομαδοποίησης. Συνεπώς, είναι ευαίσθητη στο θόρυβο και τις ακραίες παρατηρήσεις.

- **Οι αποφάσεις συγχώνευσης είναι τελικές :**

Όταν αποφασιστεί η συγχώνευση δύο συμπλεγμάτων δεν μπορεί να ανακληθεί αργότερα. Αυτή η προσέγγιση εμποδίζει ένα τοπικό κριτήριο βελτιστοποίησης να γίνει ένα κριτήριο συνολικής βελτιστοποίησης. Υπάρχουν ορισμένες τεχνικές που προσπαθούν να ξεπεράσουν τον περιορισμό του ότι οι συγχωνεύσεις είναι τελικές. Μια προσέγγιση επιχειρεί να διορθώσει την ιεραρχική ομαδοποίηση μετακινώντας κλαδιά του δέντρου, έτσι ώστε να βελτιωθεί η τιμή μιας αντικειμενικής συνάρτησης. Μια άλλη προσέγγιση χρησιμοποιεί τη μέθοδο k-means για τη δημιουργία πολλών μικρών συμπλεγμάτων, και στη συνέχεια εκτελεί τη μέθοδο single-link χρησιμοποιώντας αυτά τα μικρά συμπλέγματα ως σημείο εκκίνησης.

Πλεονεκτήματα Single-link Hierarchical Clustering

- **Απλή υλοποίηση και εφαρμοφή:**

Η μέθοδος μπορεί να υλοποιηθεί αρκετά εύκολα και δεν προϋποθέτει τον προσδιορισμό του αριθμού των συμπλεγμάτων εκ των προτέρων. Επιπλέον, μπορεί να γίνει χρήση οποιασδήποτε συνάρτησης εγγύτητας κατά τον προσδιορισμό του αντίστοιχου πίνακα, ο οποίος αποτελεί και το μοναδικό όρισμα της μεθόδου.

- **Έλλειψη Τυχαιότητας**:

Διαδοχικές εφαρμογές της μεθόδου χρησιμοποιώντας το ίδιο σύνολο παρατηρήσεων έχουν το ίδιο αποτέλεσμα. Τα βήματα που ακολουθεί είναι συγκεκριμένα και δεν επηρεάζονται από κάποιον τυχαίο παράγοντα.

- **Δημιουργία εμφωλευμένων συσχετίσεων:**

Εκτός της τελικής διαμέρισης, εξετάζει και τις επιμέρους συσχετίσεις μεταξύ των παρατηρήσεων, γεγονός ιδιαίτερα σημαντικό στη μελέτη παρατηρήσεων που σχετίζονται με τομείς που βρίσκονται υπό εξερεύνηση.

- **Καλή εφαρμογή στα μη ελλειπτικά συμπλέγματα**:

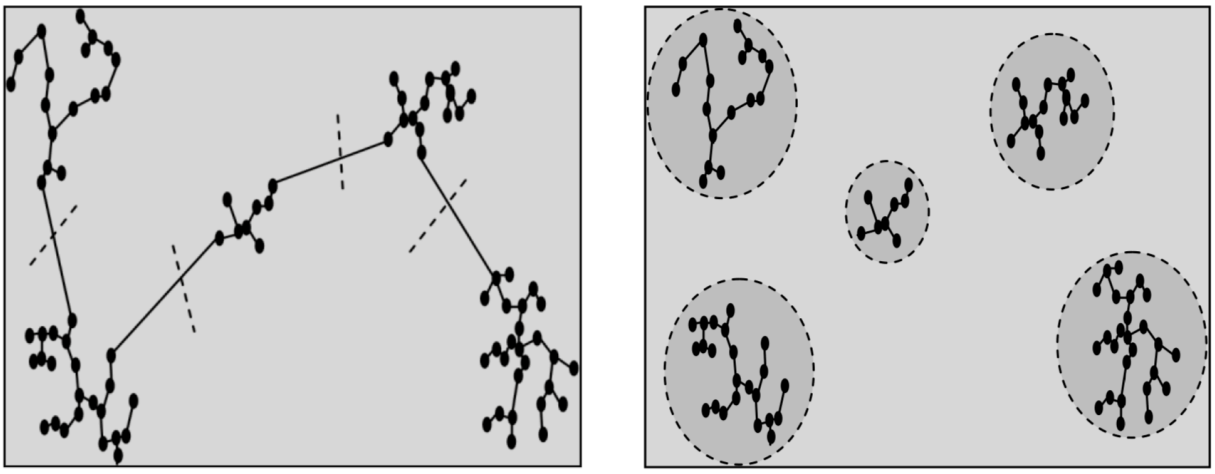
Ο αλγόριθμος single-link είναι ιδιαίτερα κατάλληλος για την αναγνώριση μη ελλειπτικών δομών σε ένα σύνολο παρατηρήσεων, δεδομένου ότι δεν υπάρχει έντονος θόρυβος έτσι ώστε να επηρεαστεί από το φαινόμενο της αλυσίδας.

3.7 Single-link και ελάχιστα διασυνδετικό δέντρο

Για να παρουσιάσουμε τις λεπτομέρειες ξεκινάμε με μια σύντομη εισαγωγή των MSTs και με ορισμένους αποτελεσματικούς αλγορίθμους για την εύρεση ενός MST. Το δέντρο είναι μια έννοια στη θεωρία των γραφημάτων. Ένα δέντρο είναι ένα συνδεδεμένο γράφημα χωρίς κύκλους (Jain και Dubes, 1988). ST (spanning tree) είναι ένα δέντρο που περιέχει όλες τις κορυφές του γραφήματος. Όταν κάθε ακμή σε ένα γράφημα ζυγίζεται από την ανομοιότητα μεταξύ των δύο κορυφών που συνδέει την άκρη, το βάρος ενός δέντρου είναι το άθροισμα βάρους των ακμών του δέντρου. Ένα MST ενός γραφήματος G είναι ένα δέντρο που έχει ελάχιστο βάρος μεταξύ όλων των άλλων δέντρων του G.

Έχουν αναπτυχθεί πολλοί αλγόριθμοι για την εύρεση ενός MST. Δύο δημοφιλείς αλγόριθμοι για την εύρεση ενός MST έχουν αναπτυχθεί από τον Kruskal (1956) με πολυπλοκότητα $O(ElogE)$, και τον Prim (1957) με πολυπλοκότητα $O(VlogV + E)$. Στους αλγορίθμους αυτούς, οι ακμές ανήκουν σε ένα από τα δύο σύνολα A και B σε οποιοδήποτε στάδιο, όπου A είναι το σύνολο που περιέχει τις ακμές που έχουν αντιστοιχιθεί στο MST, και το B είναι το σύνολο ακμών που δεν έχουν εκχωρηθεί. Ο Prim πρότεινε έναν επαναληπτικό αλγόριθμο που ξεκινά με οποιαδήποτε από τις δοσμένες κορυφές και αρχικά αποδίδει στο A τη μικρότερο ακμή, ξεκινώντας από αυτή την κορυφή. Στη συνέχεια, ο αλγόριθμος συνεχίζει να εκχωρεί στο A τη μικρότερη ακμή από το B που συνδέει τουλάχιστον ένα τμήμα από το A, χωρίς να σχηματίζει κλειστό βρόχο μεταξύ των τμημάτων που βρίσκονται ήδη στον A. Ο αλγόριθμος θα σταματήσει όταν υπάρχουν n - 1 ακμές στο A. Το ελάχιστο διασυνδετικό δέντρο που δημιουργείται από αυτούς τους αλγορίθμους μπορεί να μην είναι μοναδικό, εάν υπάρχουν ίσες ακμές ελάχιστου μήκους.

Επομένως με βάση τον πίνακα εγγύτητας μπορούμε να κατασκευάσουμε ένα γράφημα με n κόμβους και $\frac{n(n-1)}{2}$ ακμές, όπου κάθε ακμή θα έχει ως βάρος την αντίστοιχη ομοιότητα $f(x_i, x_j)$, όπου x_i και x_j είναι οι δύο κόμβοι-παρατηρήσεις που συνδέει. Στη συνέχεια, (κάνοντας χρήση τη λίστα γειτνίασης για την αποθήκευση του γραφήματος και το σωρό Fibonacci για την ουρά προτεραιότητας) εφαρμόζουμε τον αλγόριθμου του Prime και βρίσκουμε το ελάχιστο διασυνδετικό δένδρο με χρονική πολυπλοκότητα $O(n^2)$. Επομένως, η επιθυμητή διαμέριση προκύπτει διαγράφοντας όλες τις ακμές που έχουν βάρος μεγαλύτερο από την τιμή που έχουμε επιλέξει να κόψουμε το δέντρο (εικόνα 3.8). Οι συνεκτικές συνιστώσες του παραγόμενου γραφήματος αποτελούν τα συμπλέγματα της διαμέρισης των αρχικών παρατηρήσεων.

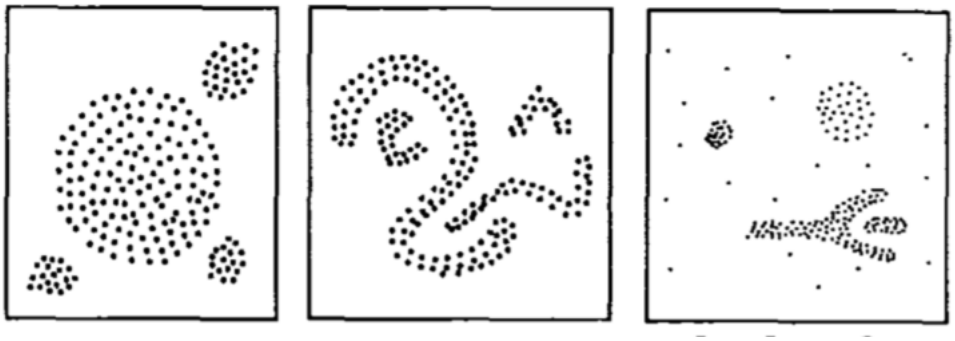


Εικόνα 3.8 α) Ελάχιστο διασυνδετικό δέντρο των παρατηρήσεων β) Παραγόμενη διαμέριση

Κεφάλαιο 4

4.1 Συμπλέγματα με βάση την πυκνότητα

Η αδυναμία πολλών αλγορίθμων ομαδοποίησης να διαχειριστούν ακραίες παρατηρήσεις καθώς και συμπλέγματα διαφορετικού σχήματος, οδήγησε στη δημιουργία εναλλακτικών τεχνικών που βασίζονται στην έννοια της πυκνότητας.



Εικόνα 4.1: 3 διαφορετικά σύνολα παρατηρήσεων

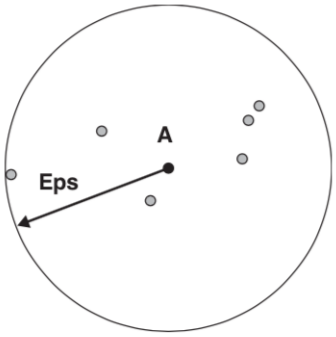
Στην εικόνα 4.1 παρατηρούμε συμπλέγματα διαφορετικής πυκνότητας, σχήματος αλλά και παρατηρήσεις που δεν ανήκουν σε κάποιο σύμπλεγμα. Ο κύριος λόγος που μπορούμε να διακρίνουμε τα συμπλέγματα και στις 3 περιπτώσεις είναι ότι εντός κάθε συμπλέγματος η πυκνότητα των παρατηρήσεων είναι μεγαλύτερη σε σχέση με αυτή που υπάρχει στα σημεία με τις ακραίες παρατηρήσεις. Παρόλο που δεν υπάρχουν τόσες προσεγγίσεις για τον ορισμό της πυκνότητας όπως υπάρχουν για τον ορισμό της ομοιότητας, υπάρχουν αρκετές διαφορετικές μέθοδοι.

Στη συνέχεια, προσπαθούμε να επισημοποιήσουμε αυτή τη διαισθητική έννοια των "συμπλεγμάτων" και “θορύβου” με τη βοήθεια μιας κεντρικής προσέγγισης της πυκνότητας, για ένα σύνολο παρατηρήσεων $\Pi = \{x_1, x_2, \ldots, x_n\}$ κάποιου m-διαστάσεων χώρου S. Μια σημαντική έννοια στους αλγόριθμους με βάση την πυκνότητα είναι η Eps-γειτονιά μιας παρατήρησης **x**. Στη συνέχεια, η Eps-γειτονιά της **x** συμβολίζεται με $N_{Eps}(x)$ και ορίζεται ως εξής.

Ορισμός 4.1: (Eps-γειτονιά της παρατήρησης x) Η Eps-γειτονιά της παρατήρησης x αποτελεί το σύνολο:

$$N_{Eps}(x) = \{y \epsilon \Pi : d(x, y) \leq Eps\}$$

όπου Π το σύνολο των παρατηρήσεων και $d()$ οποιαδήποτε μετρική συνάρτηση.



Εικόνα 4.2: Eps-γειτονιά για ευκλείδεια μετρική στον $\mathbb{R}^2$

Το σχήμα μιας γειτονιάς καθορίζεται από την επιλογή της μετρικής. Για παράδειγμα, όταν χρησιμοποιούμε τη Manhattan μετρική στο $\mathbb{R}^2$, το σχήμα της γειτονίας είναι ορθογώνιο, ενώ κάνοντας χρήση της Ευκλείδειας μετρικής θα έχει τη μορφή κύκλου στο $\mathbb{R}^2$(εικόνα 4.2). Με βάση τον παραπάνω ορισμό, μια αφελής προσέγγιση για κάθε παρατήρηση εντός κάποιου συμπλέγματος θα ήταν να θεωρήσουμε ότι η Eps-γειτονία τους περιέχει κάποιο ελάχιστο αριθμό παρατηρήσεων. Παρόλα αυτά, η παραπάνω θεώρηση αποτυγχάνει γιατί υπάρχουν δύο είδη παρατηρήσεων σε ένα σύμπλεγμα: παρατηρήσεις εντός του συμπλέγματος (κεντρικές παρατηρήσεις) και παρατηρήσεις στα σύνορα του συμπλέγματος (συνοριακές παρατηρήσεις), όπως θα δούμε και στη συνέχεια. Σε γενικές γραμμές, η Eps-γειτονία μιας συνοριακής παρατήρησης x περιέχει σημαντικά λιγότερες παρατηρήσεις σε σχέση με τη γειτονία μιας κεντρικής παρατήρησης. Για την αντιμετώπιση του παραπάνω ζητήματος πρέπει να ορίσουμε τους κύριους χαρακτηρισμούς των παρατηρήσεων.

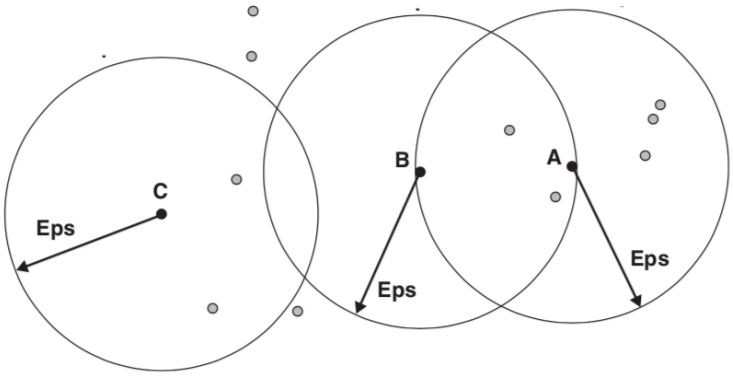
Κεντρικές παρατηρήσεις: Αυτές οι παρατηρήσεις βρίσκονται στο εσωτερικό ενός συμπλέγματος που ορίζεται με βάση την πυκνότητα. Μια παρατήρηση θεωρείται κεντρική παρατήρηση εάν ο αριθμός των παρατηρήσεων εντός μιας δεδομένης γειτονιάς γύρω από τη παρατήρηση, όπως προσδιορίζεται από τη συνάρτηση απόστασης και τη παράμετρο αποστάσεως Eps, υπερβαίνει ένα ορισμένο κατώτατο όριο N_{min} (εικόνα 4.3 παρατήρηση A για $N_{min} = 7$) . Ισοδύναμα :

$$|N_{Eps}(x)| \geq N_{min} \Leftrightarrow x \text{ κεντρική παρατήρηση,}$$

όπου $|N_{Eps}(x)|$ ο αριθμός των παρατηρήσεων στο $N_{Eps}(x)$.

Συνοριακές παρατηρήσεις: Μια συνοριακή παρατήρηση δεν είναι κεντρική παρατήρηση αλλά βρίσκεται στη γειτονιά κάποιας κεντρικής παρατήρησης (εικόνα 4.3 παρατήρηση B για $N_{min} = 7$). Επιπλέον, μπορεί να βρίσκεται στη γειτονιά παραπάνω από μία κεντρικών παρατηρήσεων.

Ακραίες παρατηρήσεις-θόρυβος: Ακραίες θα είναι οι παρατηρήσεις που δεν είναι ούτε συνοριακές αλλά ούτε κεντρικές (εικόνα 4.3 παρατήρηση C για $N_{min} = 7$).



Εικόνα 4.3: C: ακραία , B: συνοριακή , A: κεντρική παρατήρηση

Ο χαρακτηρισμός των σημείων είναι άμεσα συνδεδεμένος με την επιλογή των τιμών για τις παραμέτρους Eps και N_{min} , οι οποίες είναι και οι βασικές παράμετροι της μεθόδου DBSCAN, όπως θα δούμε και στη συνέχεια. Μεγάλες τιμές της παραμέτρου Eps αλλά και μικρές τιμές της N_{min} εν γένει θα έχουν ως αποτέλεσμα πολλές παρατηρήσεις να θεωρούνται κεντρικές, ενώ αντίθετα μικρές τιμές της Eps και μεγάλες της N_{min} οδηγούν στο χαρακτηρισμό πολλών παρατηρήσεων ως θόρυβο. Επιπλέον, προκειμένου να ορίσουμε την έννοια του συμπλέγματος θα έχουμε τους παρακάτω ορισμούς:

Όρισμός 4.2: (Άμεσα πυκνά-προσβάσιμη) Μία παρατήρηση x θα είναι άμεσα πυκνά-προσβάσιμη από μία παρατήρηση y (με βάση κάποιο Eps και N_{min}) εάν:

1. $x \varepsilon N_{Eps}(y)$

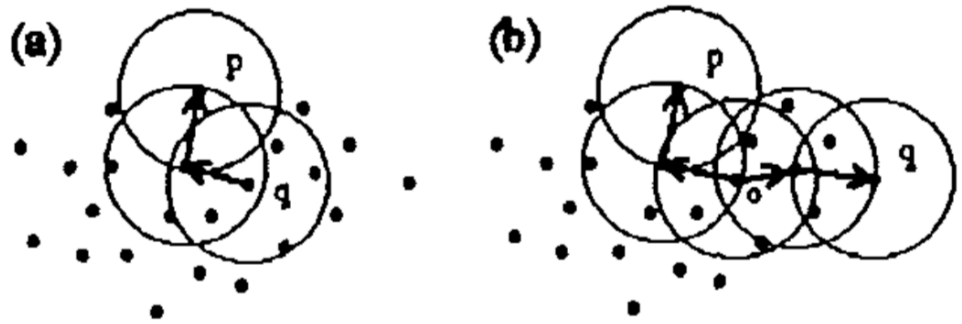
2. $|N_{Eps}(y)| \geq N_{min}$ (συνθήκη κεντρικής παρατήρησης)

Προφανώς η παραπάνω έννοια είναι συμμετρική για ζεύγη κεντρικών παρατηρήσεων, κάτι το οποίο δεν συμβαίνει όμως όταν μία απο τις 2 παρατηρήσεις δεν είναι κεντρική (εικόνα 4.3 παρατηρήσεις A και B).

Όρισμός 4.3: (Πυκνά-προσβάσιμη) Μία παρατήρηση x θα είναι πυκνά-προσβάσιμη από μία παρατήρηση y, εάν υπάρχει ακολουθία παρατηρήσεων $x = x_1, x_2, \ldots, x_i = y$ τέτοια ώστε κάθε x_l να είναι άμεσα πυκνά-προσβάσιμη από τη x_{l+1} για κάθε $l = 1,2,...,i-1$

Δεδομένου ότι ο ορισμός 4.3 αποτελεί μία επέκταση του προηγούμενου, δεν θα είναι συμμετρική έννοια με μοναδική εξαίρεση όταν κάθε x_l είναι κεντρική παρατήρηση. Παρ' όλα αυτά θα ισχύει η μεταβατικότητα και για τις 2 παραπάνω έννοιες.

Όρισμός 4.4: (Πυκνά-συνδεδεμένη) Δύο παρατηρήσεις x και y λέγεται ότι είναι πυκνά-συνδεδεμένες σε σχέση με Eps και Nmin, εάν υπάρχει μία παρατήρηση z τέτοια, ώστε τόσο η x όσο και η y είναι πυκνά-προσβάσιμες από τη z σε σχέση με Eps και Nmin.



Εικόνα 4.4: (α) p πυκνά-προσβάσιμη απο q , **(β)** p,q πυκνά συνδεδεμένες

Με βάση τα παραπάνω, μπορούμε να καθορίσουμε την έννοια του συμπλέγματος μέσω της έννοιας της πυκνότητας, ως ένα σύνολο παρατηρήσεων που είναι πυκνά συνδεδεμένο, το οποίο είναι μέγιστο ως προς την πυκνά-προσβασιμότητά του. Αντίστοιχα, θόρυβος είναι το σύνολο των παρατηρήσεων που δεν ανήκουν σε κάποιο σύμπλεγμα. Μαθηματικά οι έννοιες αυτές ορίζονται ως εξής:

Όρισμός 4.5: (Σύμπλεγμα) Εάν $\Pi = \{x_1, x_2, \ldots, x_n\}$ το σύνολο των παρατηρήσεων, ένα μη κενό υποσύνολο C του Π, θα καλείται σύμπλεγμα αν ικανοποιεί τις ακόλουθες προϋποθέσεις:

1. $\forall x, y \varepsilon \Pi$, εάν $x \varepsilon C$ και y είναι πυκνά-προσβάσιμη από τη x δεδομένου των Eps και N_{min}, τότε $y \varepsilon C$

2. $\forall x, y \varepsilon C$, x και y είναι πυκνά-συνδεδεμένες δεδομένου των Eps και N_{min},

Όρισμός 4.6:(Θόρυβος) Εάν $C_1, C_2, \ldots, C_k$ είναι τα συμπλέγματα του συνόλου $\Pi = \{x_1, x_2, \ldots, x_n\}$ των παρατηρήσεων δεδομένου των Eps και N_{min}, θα ορίζεται ως θόρυβος το υποσύνολο $N = \{x \varepsilon \Pi \mid \forall i : x \notin C_i\}$

Άμεση συνέπεια των ορισμών αυτών θα είναι η παρακάτω παρατήρηση:

Παρατήρηση: Κάθε σύμπλεγμα θα πρέπει να περιέχει τουλάχιστον N_{min} παρατηρήσεις

Απόδειξη

Εάν x μία παρατήρηση ενός συμπλέγματος C, τότε θα πρέπει να είναι πυκνά-συνδεδεμένη με τον εαυτό της μέσω κάποιας παρατήρησης y (μπορεί y=x). Επομένως για να ισχύει αυτό θα πρέπει η y να είναι κεντρική παρατήρηση και η γειτονιά της θα πρέπει να έχει τουλάχιστον N_{min} παρατηρήσεις οι οποίες θα ανήκουν στο C.

□

4.2 DBSCAN

Το 1996 οι Martin Ester, Hans-Peter Kriegel, Jörg Sander και Xiaowei Xu πρότειναν έναν αλγόριθμο ομαδοποίησης με βάση την πυκνότητα που ονομάζεται DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Ο αλγόριθμος χρησιμοποιεί μια βοηθητική συνάρτηση με όνομα ExpandCluster. Το βασικό τμήμα του αλγορίθμου παίρνει τρεις παραμέτρους εισόδου: το σύνολο των παρατηρήσεων $\Pi = \{x_1, x_2, \ldots, x_n\}$, την ακτίνα της γειτονιάς Eps και τον ελάχιστο αριθμό παρατηρήσεων εντός της κάθε γειτονιάς N_{min}. Κάθε παρατήρηση έχει ένα χαρακτηριστικό που ονομάζεται ClusterId, το οποίο αποθηκεύει το αναγνωριστικό του συμπλέγματος και αρχικά είναι ίσο με το ΑΤΑΞΙΝΟΜΗΤΟΣ. Πρώτον, ο αλγόριθμος δημιουργεί μια ετικέτα για το πρώτο σύμπλεγμα που θα βρεθεί. Στη συνέχεια, οι παρατηρήσεις στο Π διαβάζονται. Η τιμή του χαρακτηριστικού ClusterId της πρώτης αναγνωσμένης παρατήρησης είναι ίση με ΑΤΑΞΙΝΟΜΗΤΟΣ. Ενώ ο αλγόριθμος αναλύει κάθε παρατήρηση τη μία μετά την άλλη, είναι πιθανό τα χαρακτηριστικά ClusterId ορισμένων παρατηρήσεων να αλλάξουν πριν αναλυθούν αυτές οι παρατηρήσεις. Μια τέτοια περίπτωση μπορεί να συμβεί όταν μια παρατήρηση είναι πυκνά-προσβάσιμη από μια κεντρική παρατήρηση που εξετάστηκε νωρίτερα. Τέτοιες παρατηρήσεις που είναι πυκνά προσβάσιμες θα ανατεθούν στο σύμπλεγμα μιας κεντρικής παρατήρησης και δεν θα αναλυθούν αργότερα. Εάν μία παρατήρηση p που αναλύεται επί του παρόντος δεν έχει ακόμη ταξινομηθεί (η τιμή του χαρακτηριστικού του ClusterId είναι ίση με το ΑΤΑΞΙΝΟΜΗΤΟΣ), τότε θα καλείται η λειτουργία ExpandCluster για αυτή την παρατήρηση. Εάν το p είναι κεντρική παρατήρηση, τότε όλες οι παρατηρήσεις στο $N_{Eps}(p)$ εκχωρούνται από τη συνάρτηση ExpandCluster στο σύμπλεγμα με μια ετικέτα ίση με την τρέχουσα ετικέτα συμπλέγματος. Στη συνέχεια, δημιουργείται μια νέα ετικέτα συμπλέγματος από το DBSCAN. Διαφορετικά, εάν η p δεν είναι κεντρική παρατήρηση, το χαρακτηριστικό ClusterId της παρατήρησης p ορίζεται σε ΘΟΡΥΒΟΣ, πράγμα που σημαίνει ότι θα αντιμετωπιστεί πειραματικά ως θόρυβος. Μετά την ανάλυση όλων των παρατηρήσεων στο Π, το χαρακτηριστικό ClusterId κάθε παρατήρησης αποθηκεύει μια αντίστοιχη ετικέτα συμπλέγματος ή η τιμή της είναι ίση με ΘΟΡΥΒΟΣ. Με άλλα λόγια, το Π περιέχει μόνο σημεία τα οποία έχουν εκχωρηθεί σε συγκεκριμένα συμπλέγματα ή είναι θόρυβος.

Η συνάρτηση ExpandCluster παίρνει πέντε παραμέτρους: Π το σύνολο των παρατηρήσεων, την παρατήρηση p που βρίσκεται υπό επεξεργασία, ClId την τρέχουσα τιμή του ClusterId, την ακτίνα της γειτονιάς Eps και MinPts, τον ελάχιστο αριθμό παρατηρήσεων στην γειτονιά που απαιτείται για να αποτελούν ένα σύμπλεγμα. Η λειτουργία ξεκινά με τον υπολογισμό της γειτονιάς της παρατήρησης p. Αν το πλήθος των παρατηρήσεων της γειτονιάς $N_{Eps}(p)$ της p είναι μικρότερο από MinPts, τότε δεν είναι μια κεντρική παρατήρηση. Επιπλέον, η τιμή του χαρακτηριστικού του ClusterId ρυθμίζεται προσωρινά σε NOISE και το ExpandCluster αναφέρει την αποτυχία δημιουργίας του συμπλέγματος. Διαφορετικά, εάν ο αριθμός των παρατηρήσεων στην περιοχή $N_{Eps}(p)$ του p είναι επαρκής, η p αναγνωρίζεται ως κεντρική παρατήρηση. Επομένως, όλες οι παρατηρήσεις που είναι πυκνά προσβάσιμες από τη p θα αποτελούν ένα σύμπλεγμα. Αφού προσδιορίσαμε τη γειτονιά της p, όλες οι παρατηρήσεις αυτής της γειτονιάς γίνονται μέλη του συμπλέγματος που είναι ήδη υπο δημιουργία (μια ClId ετικέτα αποδίδεται στα πεδία ClusterId αυτών των παρατηρήσεων). Η γειτονιά της παρατήρησης p (εκτός από τη p), αποθηκεύεται σε μία ουρά προτεραιότητας προκειμένου να εξεταστεί ποιές απο αυτές αποτελούν οριακές και ποιές κεντρικές παρατηρήσεις, επεκτείνοντας έτσι το σύμπλεγμα. Η γειτονιά κάθε παρατήρησης της ουράς που είναι κεντρική παρατήρηση θα προστεθεί εκ νέου στην ουρά, κάτι το οποίο δεν θα συμβεί με τη γειτονιά των οριακών παρατηρήσεων. Οι παρατηρήσεις που ανήκουν στις γειτονιές των κεντρικών παρατηρήσεων της ουράς και έχουν ταξινομηθεί νωρίτερα ως θόρυβος, τώρα αντιστοιχίζονται στο τρέχον σύμπλεγμα.

Η παραπάνω διαδικασία με μορφή ψευδοκώδικα θα είναι ως εξής:

Αλγόριθμος 4.1 DBSCAN

Δεδομένα: Σύνολο των παρατηρήσεων Π ,ακτίνα Eps, ελ.αριθμός παρατηρήσεων N_{min}
Αποτέλεσμα: ClusterId η οποία εκχωρεί μια ετικέτα συμπλέγματος σε κάθε παρατήρηση

```
1. ClusterId = ετικέτα για το πρώτο σύμπλεγμα
2. Για κάθε παρατήρηση p στο Π επανάλαβε
3.     Εάν ( p.Clusterid = ΑΤΑΞΙΝΟΜΗΤΟΣ ) τότε
4.         Εάν ExpandCluster(Π,p,ClusterId,Eps,N_min) τότε
5.             ClusterId = NextId(ClusterId)
6.         τέλος
7.     τέλος
8. τέλος
```

Συνάρτηση 4.1 ExpandCluster

Δεδομένα: Σύνολο Π ,ακτίνα Eps, ελ.αριθμός παρατηρήσεων N_{min}, παρατήρηση p, τρέχον σύμπλεγμα ClId
Αποτέλεσμα: Αληθής ή Ψευδής

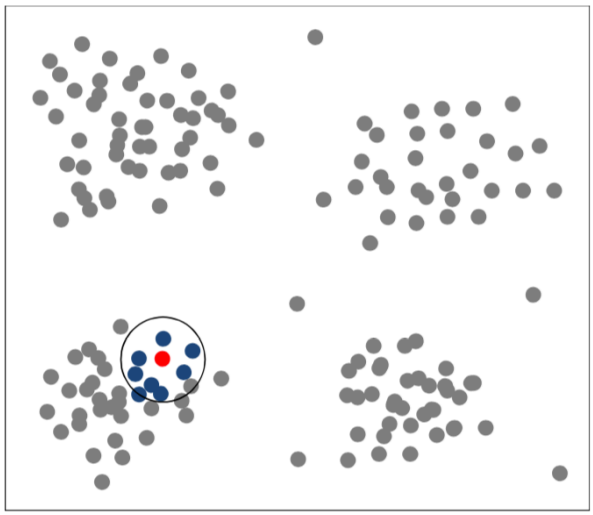
```
1. Σ (ουρά προτεραιότητας) = N_Eps(p)
2. Εάν |Σ| < N_min τότε
3.    p.Clusterid = ΘΟΡΥΒΟΣ
4.    επέστρεψε Ψευδής
5. αλλιώς
6.    για κάθε q στο Σ επανάλαβε      // μαζί με το p
```

```
7.      q.ClusterId= ClId
8.   τέλος
9.   διάγραψε p από το Σ
10. Όσο |Σ| > 0 επανάλαβε
11.      curPoint = πρώτη παρατήρηση στο Σ
12.      Σ' = N_Eps(curPoint)
13.      Εάν |Σ'| >= N_min  τότε
14.         Για κάθε q στο Σ' επανάλαβε
15.            Εάν  ( q.ClusterId = ΑΤΑΞΙΝΟΜΗΤΟΣ ) τότε
16.                  q.ClusterId = ClId
17.                  πρόσθεσε q στο Σ
18.            Αλλίως εάν ( q.ClusterId = ΘΟΡΥΒΟΣ ) τότε
19.                     q.ClusterId = ClId
20.            τέλος
21.         τέλος
22.      τέλος
23.      διάγραψε curPoint από το Σ
24. τέλος
```

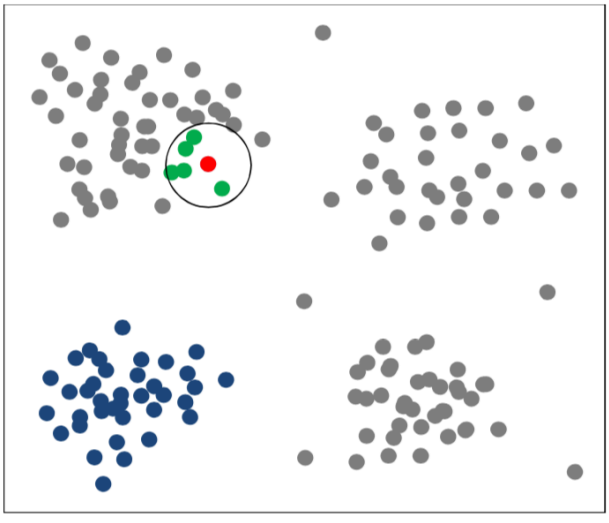
Παρατηρούμε ότι λόγω του γεγονότος ότι η γειτονιά της παρατήρησης p που μεταβιβάζεται ως παράμετρος της συνάρτησης ExpandCluster μπορεί να περιέχει παρατηρήσεις που έχουν ταξινομηθεί νωρίτερα ως θόρυβος, ορισμένοι υπολογισμοί στη συνάρτηση ExpandCluster είναι περιττοί (Kryszkiewicz & Skonieczny, 2005). Αυτές οι παρατηρήσεις θα υποβληθούν εκ νέου σε επεξεργασία. Επιπλέον αξίζει να σημειωθεί ότι οι οριακές παρατηρήσεις μπορεί να ανήκουν σε πολλά συμπλέγματα. Παρόλο που ο αλγόριθμος DBSCAN εκχωρεί αυτές τις παρατηρήσεις μόνο σε ένα σύμπλεγμα, θα ήταν δυνατό να αλλάξει ο αλγόριθμος έτσι ώστε οι οριακές παρατηρήσεις να αντιστοιχίζονται σε όλα τα πιθανά συμπλέγματα. Επιπλέον η συνάρτηση ExpandCluster επεξεργάζεται όλες τις παρατηρήσεις που περιέχονται στην ουρά. Κάθε μία από αυτές τις παρατηρήσεις αφαιρείται από την ουρά, αφού υποβληθεί σε επεξεργασία. Όταν η ουρά είναι άδεια (δηλαδή ελεγχθούν όλες οι παρατηρήσεις που βρέθηκαν), τότε η λειτουργία τελειώνει.

Εικόνες 4.5 - 4.10 Παράδειγμα Εφαμοργής του Αλγορίμου DBSCAN

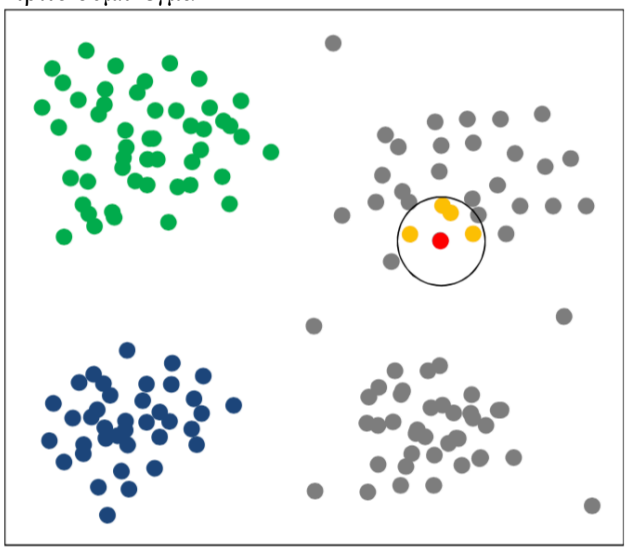
Εικόνα 4.5 Η γειτονιά της πρώτης κεντρικής παρατήρησης εισάγεται στο σύμπλεγμα



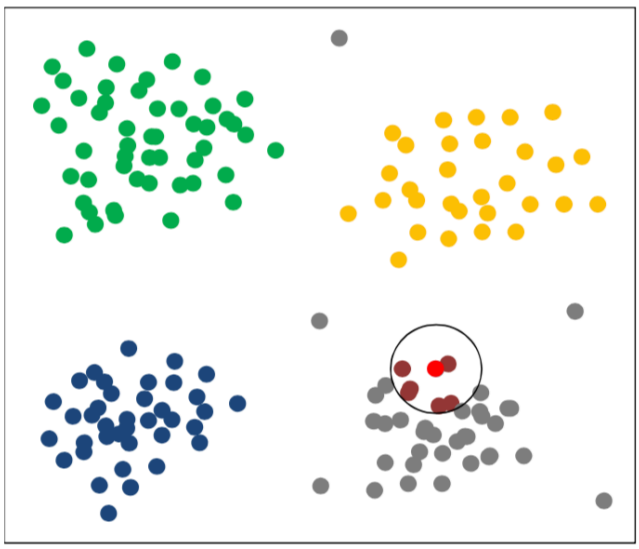
Εικόνα 4.6 Η μεταγενέστερη αντιστοίχιση πυκνά προσβάσιμων παρατηρήσεων αποτελεί την πρώτη συστάδα. Το αρχικό σύνολο Σ καθορίζεται για το δεύτερο σύμπλεγμα



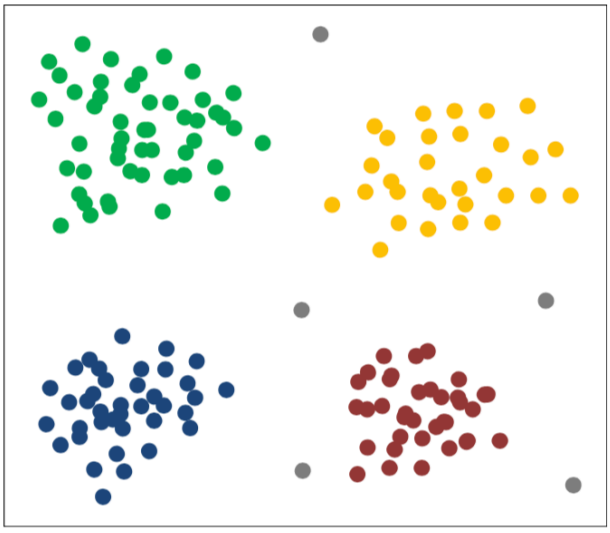
Εικόνα 4.7 Το δεύτερο σύμπλεγμα φτάνει στο μέγιστο του μέγεθος. Το αρχικό σύνολο Σ καθορίζεται για το τρίτο σύμπλεγμα



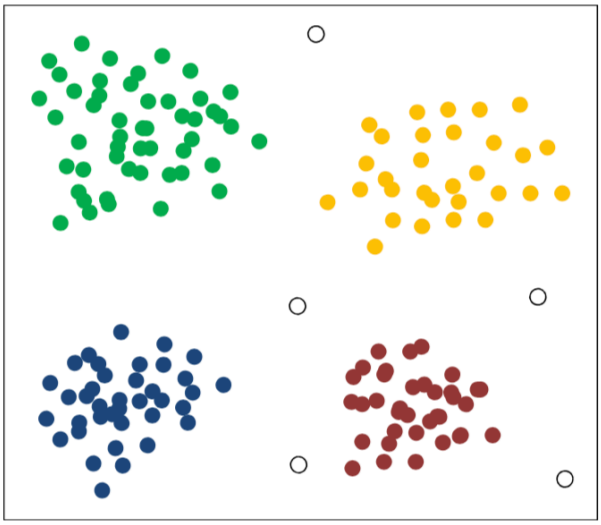
Εικόνα 4.8 Το τρίτο σύμπλεγμα φτάνει στο μέγιστο του μέγεθος. Το αρχικό σύνολο Σ καθορίζεται για το τέταρτο σύμπλεγμα



Εικόνα 4.9 Το τελικό αποτέλεσμα ομαδοποίησης με DBSCAN



Εικόνα 4.10 Θόρυβος (κενές τελείες)



Ανάλυση πολυπλοκότητας χρόνου και χώρου

Η βασική χρονική πολυπλοκότητα του αλγόριθμου DBSCAN είναι O (n × χρόνος για να βρεθούν σημεία στην γειτονιά $N_{Eps}(p)$ κάθε παρατήρησης p), όπου n είναι ο αριθμός των παρατηρήσεων. Στη χειρότερη περίπτωση,αυτή η πολυπλοκότητα είναι $O(n^2)$ χρησιμοποιώντας τον πιο απλό τρόπο (ελέγχοντας απλώς την απόσταση όλων των άλλων n - 1 παρατηρήσεων στο Π) θεωρώντας πάντα ότι οι διαστάσεις των παρατηρήσεων είναι πολύ μικρότερες από τον αριθμό n. Υποστηρίζεται ότι η διαδικασία εύρεσης της γειτονιάς $N_{Eps}(p)$ μπορεί να γίνει με πιο αποτελεσματικό τρόπο με χρήση R^* δέντρων με πολυπλοκότητα $O(logn)$, με τη μέση πολυπλοκότητα του χρόνου λειτουργίας να

μπορεί να μειωθεί στο $O(nlogn)$. Ο παραπάνω ισχυρισμός βασίζεται στο ότι το ύψος του δέντρου βρίσκεται στο O (log n), και μια αναζήτηση της γειτονιάς $N_{Eps}(p)$ πρέπει να διασχίσει μόνο ένα περιορισμένο αριθμό διαδρομών στο δέντρο (ειδικά για μικρά Eps). Ωστόσο, δεν υπάρχει θεωρητική εγγύηση ότι αυτό ισχύει, δεδομένου ότι δεν μπορούμε να είμαστε βέβαιοι ότι κάθε αναζήτηση γειτονιάς θα διασχίσει ένα περιορισμένο αριθμό μονοπατιών στο δέντρο, ειδικά εάν η μεταβλητή Eps είναι μεγάλη.

Η απαίτηση χώρου του DBSCAN, ακόμη και για παρατηρήσεις μεγάλης διαστάσεως είναι O (n), επειδή για κάθε παρατήρηση αποθηκεύεται μόνο μία ετικέτα η οποία θα αναγράφει το σύμπλεγμα ή το ότι αποτελεί θόρυβος, αν δεν τοποθετηθεί σε κάποιο σύμπλεγμα.

Ορθότητα αλγορίθμου

Η ορθότητα του αλγορίθμου DBSCAN αποδεικνύεται εύκολα με βάση τα 2 παρακάτω λήμματα, τα οποία αναφέρουν ότι δεδομένου των παραμέτρων Eps και N_{min}, μπορούμε να ανακαλύψουμε ένα σύμπλεγμα σε μια προσέγγιση δύο βημάτων. Πρώτον επιλέγουμε μία αυθαίρετη παρατήρηση από το σύνολο των παρατηρήσεων Π, που ικανοποιεί την προϋπόθεση της κεντρικής παρατήρησης. Δεύτερον, βρίσκουμε όλα τα σημεία που είναι πυκνά προσβάσιμα από το αρχικό σύνολο Σ.

Λήμμα 4.1: Έστω p μία παρατήρηση στο σύνολο Π με $|N_{Eps}(p)| \geq N_{min}$. Τότε το σύνολο O όλων των πυκνά προσβάσιμων παρατηρήσεων από την p, δεδομένου των παραμέτρων Eps και N_{min}, θα αποτελεί ένα σύμπλεγμα.

Λήμμα 4.2: Έστω C ένα σύμπλεγμα δεδομένου των παραμέτρων Eps και N_{min} και p οποιαδήποτε παρατήρηση στο C με $|N_{Eps}(p)| \geq N_{min}$. Το σύνολο όλων των πυκνά προσβάσιμων παρατηρήσεων από την p, δεδομένου των παραμέτρων Eps και N_{min}, θα είναι ίσο με το C.

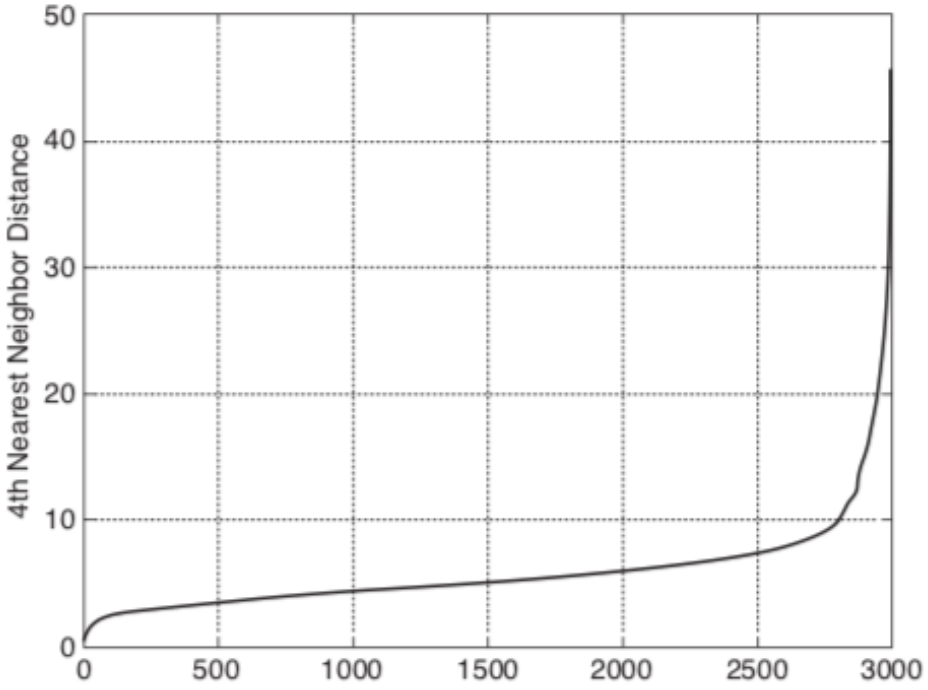
Το λήμμα 4.2 μας εξασφαλίζει ότι οποιαδήποτε από τις κεντρικές παρατηρήσεις εντοπίσουμε πρώτα το σύμπλεγμα που θα προκύψει θα είναι το ίδιο, ενώ αντίστοιχα το λήμμα 4.1 εξασφαλίζει ότι ο εντοπισμός της αρχικής κεντρικής παρατήρησης και η εύρεση όλων των πυκνά προσβάσιμων παρατηρήσεων μέσω αυτής θα οδηγήσει στην εύρεση του συμπλέγματος.

4.3 Προσδιορισμός παραμέτρων

Η επιλογή των παραμέτρων Eps και N_{min} αποτελεί ένα από τα σημαντικότερα ζητήματα και πρέπει να διευθετηθεί πρωτού εφαρμοστεί ο αλγόριθμος DBSCAN, καθώς διαφορετικές τιμές των παραμέτρων μπορεί να οδηγήσουν σε εντελώς διαφορετικά αποτελέσματα. Εν γένει, όπως έχει

αναφερθεί και νωρίτερα, μεγάλες τιμές της παραμέτρου Eps αλλά και μικρές τιμές της N_{min} θα έχουν ως αποτέλεσμα πολλές παρατηρήσεις να θεωρούνται κεντρικές, ενώ αντίθετα μικρές τιμές της Eps και μεγάλες της N_{min} οδηγούν στο χαρακτηρισμό πολλών παρατηρήσεων ως θόρυβο, έχοντας ως αποτέλεσμα την αδυναμία εντοπισμού συμπλεγμάτων τα οποία είναι σχετικά αραιά.

Μια βασική προσέγγιση για τον προσδιορισμό των παραμέτρων Eps και N_{min}, είναι να δούμε τη συμπεριφορά της απόστασης κάθε παρατήρησης από τον k-οστό πλησιέστερο γείτονα της, τον οποίο θα ονομάσουμε k-dist. Για τις παρατηρήσεις που ανήκουν σε κάποιο σύμπλεγμα η τιμή του k-dist θα είναι μικρή, εάν το k δεν είναι μεγαλύτερο από το μέγεθος του συμπλέγματος. Σημειώστε ότι θα υπάρξει κάποια διαφοροποίηση, ανάλογα με την πυκνότητα του συμπλέγματος και την τυχαία κατανομή των παρατηρήσεων, αλλά κατά μέσο όρο το εύρος της μεταβολής δεν θα είναι τεράστιο, εάν οι πυκνότητες των συμπλεγμάτων δεν είναι ριζικά διαφορετικές. Ωστόσο, για παρατηρήσεις που δεν βρίσκονται σε σύμπλεγμα, όπως παρατηρήσεις θορύβου, το k-dist θα είναι σχετικά μεγάλο. Επομένως, αν υπολογίσουμε το k-dist για όλες τις παρατηρήσεις για κάποιο k, τα ταξινομήσουμε με αυξανόμενη η φθίνουσα σειρά και στη συνέχεια σχεδιάσουμε τις ταξινομημένες τιμές, περιμένουμε να δούμε μια απότομη αλλαγή στην τιμή του k-dist, που αντιστοιχεί σε μία τιμή για την παράμετρο Eps. Αν επιλέξουμε αυτή την απόσταση ως την παράμετρο Eps και πάρουμε την τιμή k ως την παράμετρο N_{min}, τότε οι παρατηρήσεις για τις οποίες το k-dist είναι μικρότερο από το Eps θα επισημαίνονται ως κεντρικές, ενώ οι υπόλοιπες θα επισημαίνονται ως θόρυβος ή συνοριακές.



Εικόνες 4.11 Διάγραμμα 4-dist για ένα δυσδιάστατο σύνολο παρατηρήσεων

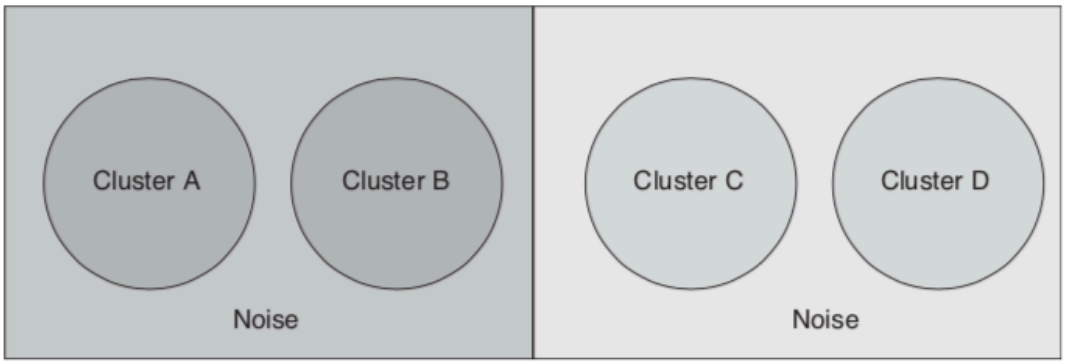
Ο εντοπισμός της μεταβολής μπορεί να γίνει με τη βοήθεια ενός διαγράμματος k-dist συναρτήσει κάθε παρατήρησης, όπως αυτό της εικόνας 4.11. Στην παραπάνω εικόνα παρατηρούμε ότι για $N_{min} = 4$, η κατάλληλη τιμή της παραμέτρου Eps θα είναι γύρω στο 10. Η τιμή του Eps που καθορίζεται με αυτόν τον τρόπο εξαρτάται από το k, αλλά δεν μεταβάλλεται δραματικά καθώς τα k αλλάζουν. Εάν η τιμή του k είναι πολύ μικρή, τότε ακόμη και ένας μικρός αριθμός σημείων που βρίσκονται κοντά στο σημείο που είναι θόρυβος ή ακραίες τιμές, θα χαρακτηριστούν εσφαλμένα ως συστάδες. Εάν η τιμή του k είναι πολύ μεγάλη, τότε μικρές συστάδες (μεγέθους μικρότερες από k) είναι πιθανό να χαρακτηριστούν ως θόρυβος. Ο αρχικός αλγόριθμος DBSCAN χρησιμοποίησε μια τιμή k = 4, η οποία φαίνεται να είναι μια λογική τιμή για τα περισσότερα δυσδιάστατα σύνολα δεδομένων. Για δεδομένα 2 διαστάσεων συνήθως επιλέγεται η τιμή k=4, ενώ για μεγαλύτερες διαστάσεις θέτουμε k = m + 1, όπου m το σύνολο των χαρακτηριστικών κάθε παρατήρησης.

4.4 Πλεονεκτήματα και Μειονεκτήματα

Μειονεκτήματα DBSCAN

- **Συμπλέγματα διαφορετικής πυκνότητας:**

Ο αλγόριθμος DBSCAN μπορεί να έχει προβλήματα με την πυκνότητα, εάν η πυκνότητα των συμπλεγμάτων διαφέρει αρκετά. Για παράδειγμα, στην εικόνα 4.12 έχουμε τέσσερα συμπλέγματα ενσωματωμένα στο θόρυβο τα οποία διαφοροποιούνται όσον αφορά την πυκνότητά τους. Η πυκνότητα των συμπλεγμάτων και των περιοχών θορύβου υποδεικνύεται από το διαφορετικό χρωματισμός τους, θεωρώντας ως πιο πυκνές τις πιο σκούρες περιοχές. Ο θόρυβος γύρω από το ζευγάρι των πυκνότερων συμπλεγμάτων A και B έχει την ίδια πυκνότητα με τα συμπλέγματα C και D. Αν η τιμή της μεταβλητής Eps είναι αρκετά μεγάλη ώστε ο αλγόριθμος DBSCAN να βρει τα C και D ως συμπλέγματα, τότε τα A, B και οι παρατηρήσεις που τους περιβάλλουν θα αποτελούν ένα μόνο σύμπλεγμα.



Εικόνα 4.12: 4 συμπλέγματα διαφορετικής πυκνότητας ενσωματωμένα σε θόρυβο

Αν η τιμή της μεταβλητής Eps είναι αρκετά μικρή ώστε ο αλγόριθμος DBSCAN να βρει τα A και B ως ξεχωριστά συμπλέγματα και τις παρατηρήσεις που τα περιβάλλουν ως θόρυβο, τότε τα C και D και οι παρατηρήσεις που τα περιβάλλουν θα σημειωθούν επίσης ως θόρυβος.

- **Δεν είναι εξ ολοκλήρου ντετερμινιστικός**:

Οι συνοριακές παρατηρήσεις μπορούν να προσεγγιστούν από περισσότερα από ένα συμπλέγματα και μπορούν να αποτελούν μέρος είτε ενός συμπλέγματος είτε άλλου, ανάλογα με τη σειρά που επεξεργάζονται οι παρατηρήσεις. Για τα περισσότερα σύνολα παρατηρήσεων, αυτή η κατάσταση δεν εμφανίζεται συχνά και έχει μικρή επίδραση στο αποτέλεσμα της ομαδοποίησης, καθώς τόσο στις κεντρικές παρατηρήσεις όσο και στις παρατηρήσεις θορύβου, ο αλγόριθμος DBSCAN είναι ντετερμινιστικός.

- **Χρονική πολυπλοκότητα:**

Σύμφωνα με την ανάλυση που έγινε νωρίτερα, είδαμε ότι η χρονική πολυπλοκότητα της μεθόδου είναι $O(n^2)$, γεγονός που μπορεί να αποτελέσει απαγορευτικό παράγοντα όταν το σύνολο των παρατηρήσεων είναι αρκετά μεγάλο.

- **Εξάρτηση από τη μετρική:**

Η ποιότητα του DBSCAN εξαρτάται από τη μετρική που χρησιμοποιείται στην συνάρτηση Expandcluster για την εύρεση της περιοχής $N_{Eps}(p)$ κάθε παρατήρησης p. Η πιο συνηθισμένη μετρική που χρησιμοποιείται είναι η ευκλείδεια απόσταση. Ειδικά για παρατηρήσεις μεγάλης διαστάσεως, αυτή η μετρική μπορεί να καταστεί σχεδόν άχρηστη λόγω της αποκαλούμενης «κατάρας των διαστάσεων», καθιστώντας δύσκολη την εύρεση μιας κατάλληλης τιμής για την παράμετρο Eps. Αυτή η επίδραση, ωστόσο, υπάρχει και σε οποιοδήποτε άλλο αλγόριθμο που βασίζεται στην ευκλείδεια απόσταση.

<u>Πλεονεκτήματα DBSCAN</u>

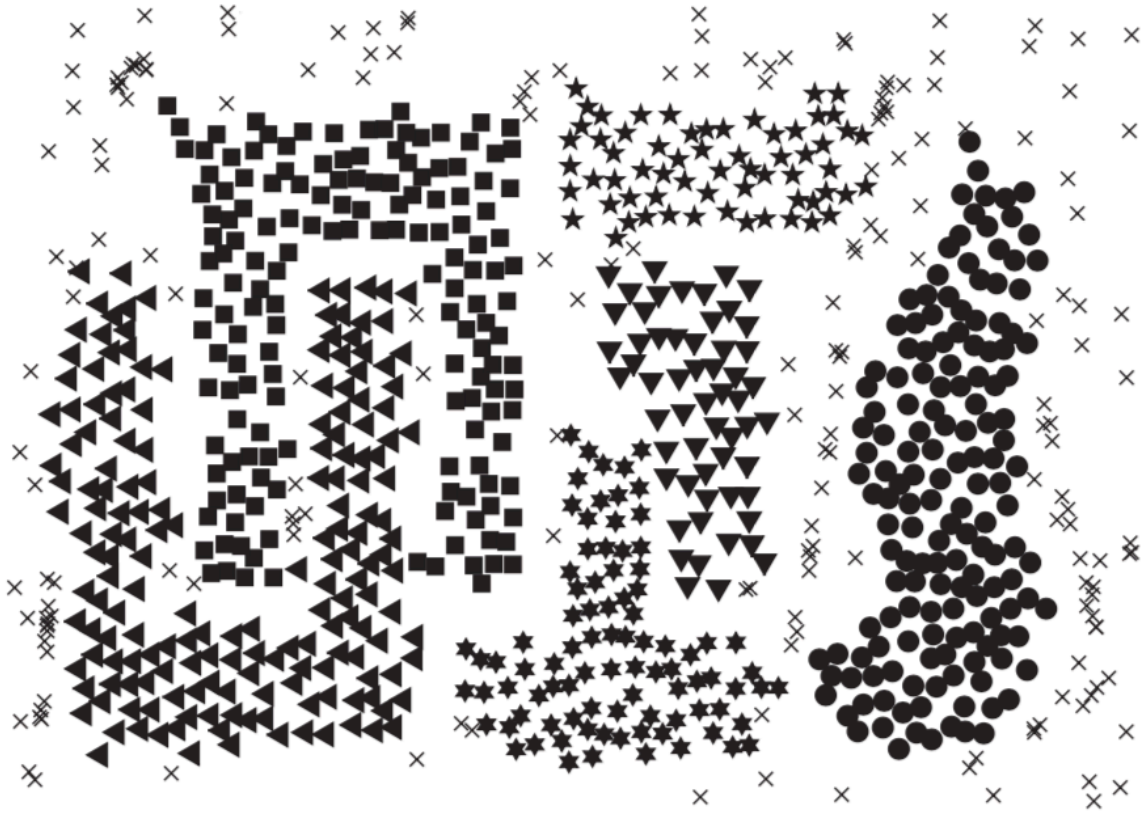
- **Εύρεση συμπλεγμάτων με αυθαίρετα σχήματα και μεγέθη:**

Ο αλγόριθμος DBSCAN μπορεί να εντοπίσει συμπλέγματα διαφόρων σχημάτων, τα οποία δεν θα μπορούσε να διακρίνει π.χ. ο αλγόριθμος K means. Επιπλέον έχει τη δυνατότητα εντοπισμού συμπλεγμάτων που περιβάλονται από άλλα συμπλέγματα χωρίς όμως να ακουμπάνε, αλλά και συμπλέγματα ενσωματωμένα μέσα σε θόρυβο (εικόνα 4.13). Συνεπώς, δεν τοποθετεί όλες τις παρατηρήσεις απαραίτητα σε κάποιο σύμπλεγμα, χαρακτηρίζοντας ως θόρυβο όσες δεν ανήκουν σε

κάποιο σύμπλεγμα. Η παραπάνω έννοια ενδεχομένως να βρίσκεται πιο κοντά στην πραγματικότητα για συγκεκριμένες εφαρμογές, όπου απαιτείται ξεχωριστή μελέτη της συμπεριφοράς τέτοιων παρατηρήσεων. Σε αυτή την κατηγορία συνήθως ανήκουν και ακραίες παρατηρήσεις, οι οποίες δεν επηρεάζουν τη διαμόρφωση των υπολοίπων συμπλεγμάτων.

- **Απλότητα στην εφαρμογή:**

Η εφαρμογή του αλγορίθμου δεν προαπαιτεί τη γνώση του αριθμού των συμπλεγμάτων παρά μόνο 2 μεταβλητών (Eps , N_{min}). Επιπλέον, η διάταξη των παρατηρήσεων δεν επηρεάζει το αποτέλεσμα όσον αφορά τις κεντρικές παρατηρήσεις και το θόρυβο, με μικρές αποκλίσεις σπανίως ως προς τις συνοριακές.



Εικόνα 4.13: Συμπλέγματα που προέκυψαν από εφαρμογή του DBSCAN

Κεφάλαιο 5

5.1 Αξιολόγηση Συμπλεγμάτων

Η αξιολόγηση των παραγόμενων συμπλεγμάτων αποτελεί βασικό κομμάτι της διαδικασίας διερεύνησης ενός συνόλου παρατηρήσεων. Κάθε αλγόριθμος ομαδοποίησης παράγει κάποια διαμέριση, ανεξάρτητα από το αν αυτή ανταποκρίνεται στην πραγματικότητα, δηλαδή στην πραγματική δομή των παρατηρήσεων. Μία από τις εφαρμογές της παραπάνω αξιολόγησης αποτελεί και η σύγκριση διαφορετικών αλγορίθμων ομαδοποίησης, οι οποίοι εφαρμόζονται πάνω στο ίδιο σύνολο παρατηρήσεων. Παρόλο που είναι δυνατόν να αναπτυχθούν διάφορα αριθμητικά μέτρα για την εκτίμηση της εγκυρότητας των συμπλεγμάτων, υπάρχουν ορισμένες προκλήσεις. Πρώτον, ένα μέτρο της εγκυρότητας του συμπλέγματος μπορεί να είναι αρκετά περιορισμένο στο πεδίο εφαρμογής του. Για παράδειγμα, οι περισσότερες εργασίες σχετικά με τα μέτρα αξιολόγησης έχουν γίνει για τα δυσδιάστατα ή τρισδιάστατα χωρικά δεδομένα. Δεύτερον, χρειαζόμαστε ένα πλαίσιο για την ερμηνεία κάθε μέτρου. Αν δηλαδή λάβουμε τιμή 10 για ένα μέτρο που αξιολογεί πόσο καλά οι ετικέτες συμπλέγματος αντιστοιχούν σε ετικέτες κατηγορίας που παρέχονται εξωτερικά, θα πρέπει να είμαστε σε θέση να κρίνουμε αν αποτελεί καλή ή κακή τιμή.

Τα μέτρα αξιολόγησης ή οι δείκτες που εφαρμόζονται για να κρίνουν την εγκυρότητα ενός συμπλέγματος, ταξινομούνται παραδοσιακά στους παρακάτω τρεις τύπους:

- **Εσωτερικοί δείκτες:** Μετράνε την καταλληλότητα μιας δομής ομαδοποίησης, χωρίς να λαμβάνεται υπόψη κάποια εξωτερική πληροφόρηση. Τα μέτρα εγκυρότητας της ομαδοποίησης που ανήκουν σε αυτή τη κατηγορία, χωρίζονται συχνά σε δύο υποκατηγορίες: Μέτρα συνοχής συμπλέγματος (συμπαγές, σφιχτό), τα οποία καθορίζουν πόσο στενά συσχετίζονται οι παρατηρήσεις σε ένα σύμπλεγμα, και μέτρα διαχωρισμού (απομόνωσης) συμπλέγματος, που καθορίζουν πόσο διακριτά ή καλά διαχωρισμένο είναι ένα σύμπλεγμα από άλλα συμπλέγματα.

- **Εξωτερικοί δείκτες:** Μετράνε τον βαθμό κατά τον οποίο η δομή ομαδοποίησης που ανακαλύφθηκε από έναν αλγόριθμο συμπίπτει με κάποια εξωτερική δομή. Ένα παράδειγμα εξωτερικού δείκτη είναι η εντροπία, η οποία μετρά πόσο καλά οι ετικέτες συμπλέγματος αντιστοιχούν σε εξωτερικά εφοδιασμένες ετικέτες κλάσης. Η ονομασία τους προέρχεται από το γεγονός ότι χρησιμοποιούν πληροφορίες που δεν υπάρχουν στο σύνολο των παρατηρήσεων.

- **Σχετικοί δείκτες αξιολόγησης:** Συγκρίνουν διαφορετικές ομαδοποιήσεις ή συμπλέγματα. Ένα σχετικό μέτρο αξιολόγησης είναι ένας δείκτης εσωτερικός ή εξωτερικός, ο οποίος χρησιμοποιείται για σκοπούς σύγκρισης. Επομένως τα σχετικά μέτρα δεν είναι στην πραγματικότητα ένα ξεχωριστό είδος μέτρου αξιολόγησης, αλλά είναι μια συγκεκριμένη χρήση τέτοιων μέτρων. Παραδείγματα τέτοιων εφαρμογών είδαμε στη περίπτωση του αλγορίθμου K means, όπου συγκρίναμε διαφορετικά αποτελέσματα του αλγορίθμου για διάφορες τιμές της παραμέτρου k, διαλέγοντας εκείνη την τιμή με το μικρότερο SSE.

Στη συνέχεια θα εστιάσουμε σε εξωτερικούς δείκτες αξιολόγησης, τους οποίους θα χρησιμοποιήσουμε προκειμένου να συγκρίνουμε τα αποτελέσματα 3 διαφορετικών αλγορίθμων ομαδοποίησης.

5.2 Εξωτερικοί δείκτες

Όταν έχουμε εξωτερικές πληροφορίες σχετικά με τις παρατηρήσεις, είναι συνήθως με την μορφή εξωτερικών ετικετών κλάσης για τις παρατηρήσεις. Στις περιπτώσεις αυτές, η συνήθης διαδικασία είναι να μετρηθεί ο βαθμός αντιστοιχίας μεταξύ των ετικετών συμπλέγματος και των ετικετών κλάσης. Τα κίνητρα μιας τέτοιας ανάλυσης είναι η σύγκριση των διαφόρων αλγορίθμων ομαδοποίησης με την πραγματικότητα ή η αξιολόγηση του βαθμού στον οποίο μια διαδικασία χειρωνακτικής ταξινόμησης μπορεί να παραχθεί αυτόματα μέσω κάποιου αλγορίθμου ομαδοποίησης.

Θεωρούμε δύο διαφορετικά είδη ομάδων. Η πρώτη αποτελείται από μέτρα όπως η εντροπία (entropy), η αγνότητα (purity) και το μέτρο F. Αυτά τα μέτρα αξιολογούν τον βαθμό κατά τον οποίο ένα σύμπλεγμα περιέχει παρατηρήσεις μιας κλάσης. Η δεύτερη ομάδα σχετίζεται με τα μέτρα ομοιότητας για δυαδικά δεδομένα, όπως το μέτρο Jaccard. Τα μέτρα αυτά αξιολογούν τον βαθμό κατά τον οποίο δύο παρατηρήσεις που βρίσκονται στην ίδια κλάση βρίσκονται και στο ίδιο σύμπλεγμα και αντίστροφα . Για λόγους ευκολίας, θα αναφερθούμε σε αυτές τις δύο ομάδες μέτρων ως προσανατολισμένη στην ταξινόμηση και προσανατολισμένη στην ομοιότητα αντίστοιχα.

5.3 Προσανατολισμένοι στην ταξινόμηση

Στην κατηγορία αυτή ανήκουν δείκτες όπως η εντροπία (entropy), η αγνότητα (purity), η ακρίβεια (precision), η ανάκληση (recall) και το μέτρο F, τα οποία χρησιμοποιούνται συνήθως για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης. Στην περίπτωση της ταξινόμησης, μετράμε τον βαθμό στον οποίο οι προβλεπόμενες ετικέτες κλάσης αντιστοιχούν σε πραγματικές

ετικέτες κλάσης. Για τα προαναφερθέντα μέτρα, τίποτα θεμελιώδες δεν αλλάζει χρησιμοποιώντας ετικέτες συμπλέγματος αντί για προβλεπόμενες ετικέτες κλάσης.

Entropy: Ο βαθμός στον οποίο κάθε σύμπλεγμα αποτελείται από παρατηρήσεις μιας κλάσης. Για κάθε σύμπλεγμα υπολογίζεται αρχικά η κατανομή των κλάσεων. Συγκεκριμένα για ένα σύμπλεγμα i, υπολογίζουμε την πιθανότητα ότι ένα μέλος του συμπλέγματος i ανήκει στην κλάση j ως εξής : $p_{ij} = \frac{n_{ij}}{n_i}$, όπου n_i είναι ο αριθμός των παρατηρήσεων μέσα στο σύμπλεγμα i, και n_{ij} είναι ο αριθμός των παρατηρήσεων της κλάσης j στο σύμπλεγμα i. Χρησιμοποιώντας την παραπάνω κατανομή, η εντροπία κάθε συμπλέγματος i υπολογίζεται ως εξής:

$$e_i = -\sum_{j=1}^{L} p_{ij} log_2 p_{ij}$$

όπου L ο αριθμός των κλάσεων. Η συνολική εντροπία για ένα σύνολο συμπλεγμάτων υπολογίζεται ως το άθροισμα από τις εντροπίες κάθε συμπλέγματος, που σταθμίζονται από το μέγεθος κάθε συμπλέγματος. Συγκεκριμένα η συνολική εντροπία θα υπολογίζεται από τον παρακάτω τύπο:

$$e = \sum_{i=1}^{K} \frac{n_i}{n} e_i$$

όπου K ο αριθμός των συμπλεγμάτων και n ο συνολικός αριθμός των παρατηρήσεων.

Purity: Ένας άλλος δείκτης μέτρησης του βαθμού κατά τον οποίο ένα σύμπλεγμα περιέχει παρατηρήσεις μόνο από μία κλάση**.** Χρησιμοποιώντας την προηγούμενη ορολογία, η αγνότητα (purity) του συμπλέγματος i θα υπολογίζεται ως εξής:

$$p_i = max(p_{ij}) \quad \text{για} \quad j = 1,2,...,L$$

ενώ η συνολική τιμή του δείκτη purity μιας διαμέρισης ως εξής :

$$purity = \sum_{i=1}^{K} \frac{n_i}{n} p_i$$

Precision: Το μέγεθος του τμήματος ενός συμπλέγματος που αποτελείται από αντικείμενα συγκεκριμένης κλάσης. Η ακρίβεια του συμπλέγματος i σε σχέση με την κλάση j είναι:

$$precision(i, j) = p_{ij}$$

Recall: Ο βαθμός στον οποίο ένα σύμπλεγμα περιέχει όλες τις παρατηρήσεις συγκεκριμένης κλάσης. Η ανάκληση (recall) του συμπλέγματος i σε σχέση με την κλάση j είναι:

$$recall(i, j) = \frac{n_{ij}}{n_j}$$

όπου n_j είναι ο αριθμός των παρατηρήσεων στην κλάση j.

Δείκτης F: Ένας συνδυασμός τόσο του precision, όσο και του δείκτη recall που μετρά τον βαθμό στον οποίο ένα σύμπλεγμα περιέχει μόνο παρατηρήσεις συγκεκριμένης κλάσης και όλες τις παρατηρήσεις αυτής της κλάσης. Το μέτρο F του συμπλέγματος i σε σχέση με την κλάση j είναι:

$$F(i, j) = \frac{2 \times precision(i, j) \times recall(i, j)}{precision(i, j) + recall(i, j)}$$

5.4 Προσανατολισμένοι στην ομοιότητα

Οι δείκτες σε αυτή την κατηγορία βασίζονται στην αρχή ότι δύο παρατηρήσεις που βρίσκονται στο ίδιο σύμπλεγμα πρέπει να είναι στην ίδια κλάση και αντίστροφα. Προκειμένου να αναφερθούμε στους δείκτες αυτής της κατηγορίας ορίζουμε τους παρακάτω συμβολισμούς (Πίνακας 5.1) :

f_{00} = αριθμός ζευγών παρατηρήσεων που έχουν διαφορετική κλάση και διαφορετικό σύμπλεγμα
f_{01} = αριθμός ζευγών παρατηρήσεων που έχουν διαφορετική κλάση και το ίδιο σύμπλεγμα
f_{10} = αριθμός ζευγών παρατηρήσεων που έχουν την ίδια κλάση και διαφορετικό σύμπλεγμα
f_{11} = αριθμός ζευγών παρατηρήσεων που έχουν την ίδια κλάση και το ίδιο σύμπλεγμα

Πίνακας 5.1: Πίνακας που προσδιορίζει εάν τα ζεύγη παρατηρήσεων είναι της ίδιας κλάσης και του ίδιου συμπλέγματος

	Same cluster	Different cluster
Same class	f_{11}	f_{10}
Different class	f_{01}	f_{00}

Με βάση τους παραπάνω ορισμούς έχουμε τους εξής δείκτες:

Δείκτης Rand: Ο δείκτης Rand υπολογίζει πόσο όμοια είναι η ομαδοποίηση που προέκυψε απο τον αλγόριθμο με αυτή που μας έχει δοθεί. Ο παραπάνω δείκτης θα μπορούσε να θεωρηθεί και ως το ποσοστό των ορθών αποφάσεων που έχουν παραχθεί από τον αλγόριθμο. Υπολογίζεται ως εξής:

$$Rand = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Ο παραπάνω δείκτης παίρνει τιμές στο διάστημα [0,1], με τις τιμές κοντά στο 1 να συμβολίζουν μεγάλη ομοιότητα μεταξύ της διαμέρισης που προέκυψε από τον αλγόριθμο και της πραγματικής ομαδοποίησης των παρατηρήσεων. Ένα θέμα του παραπάνω δείκτη είναι ότι τα f_{00} και f_{11} έχουν το ίδιο βάρος. Το παραπάνω μπορεί να αποτελέσει ένα μη επιθυμητό χαρακτηριστικό σε μερικές εφαρμογές ομαδοποίησης.

Δείκτης Jacard: Ο δείκτης Jaccard χρησιμοποιείται για να ποσοτικοποιήσει την ομοιότητα μεταξύ δύο συνόλων, στην προκειμένη περίπτωση μεταξύ 2 ομαδοποιήσεων. Ο παραπάνω δείκτης παίρνει τιμές στο διάστημα [0,1], με τις τιμές κοντά στο 1 να συμβολίζουν μεγάλη ομοιότητα μεταξύ της διαμέρισης που προέκυψε από τον αλγόριθμο και της πραγματικής ομαδοποίησης των παρατηρήσεων. Ο δείκτης Jaccard ορίζεται ως εξής:

$$Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Παρατηρούμε ότι στον υπολογισμό του παραπάνω δείκτη δεν λαμβάνεται καθόλου υπόψη η ποσότητα f_{00}.

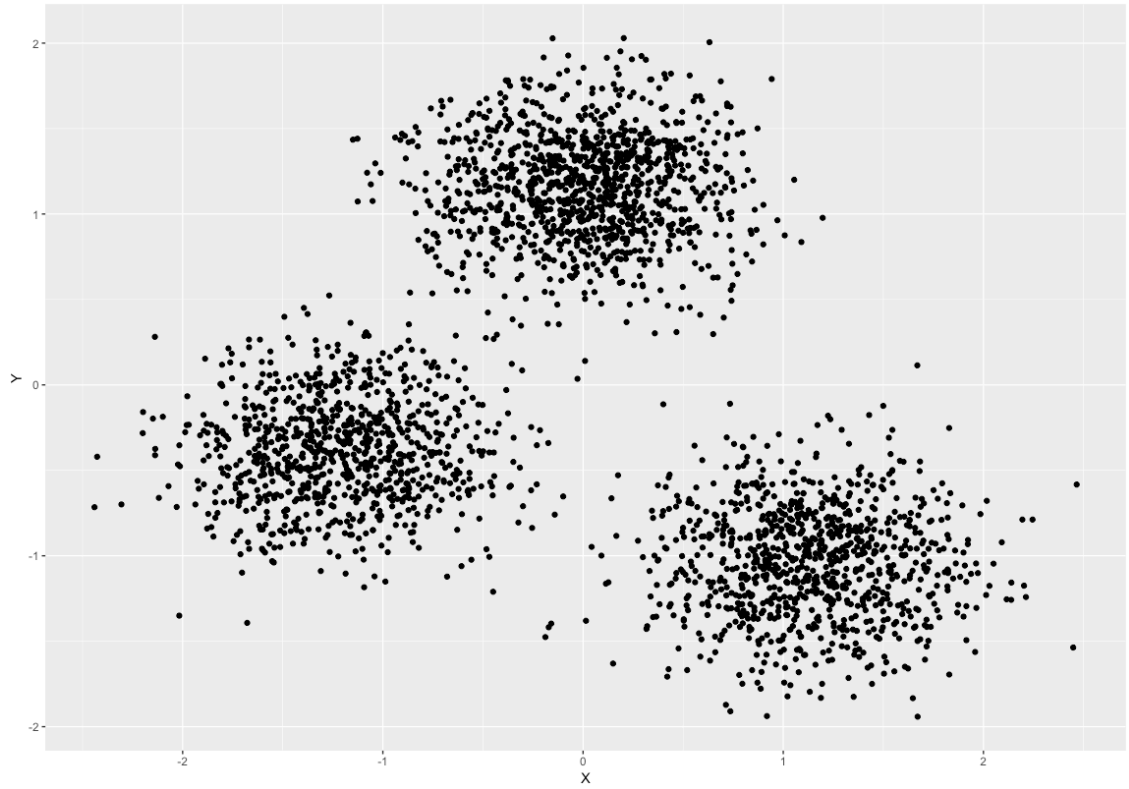
Δείκτης Fowlkes-Mallows: Ο δείκτης Fowlkes-Mallows υπολογίζει την ομοιότητα μεταξύ των συμπλεγμάτων που επιστρέφονται από τον αλγόριθμο ομαδοποίησης και της προκαθορισμένης ομαδοποίησης. Παίρνει τιμές στο διάστημα [0,1], και όσο μεγαλύτερη είναι η τιμή του δείκτη Fowlkes-Mallows, τόσο πιο όμοια είναι τα παραπάνω. Μπορεί να υπολογιστεί χρησιμοποιώντας τον ακόλουθο τύπο:

$$FM = \sqrt{\frac{f_{11}}{f_{11} + f_{10}} \times \frac{f_{11}}{f_{11} + f_{01}}}$$

Κεφάλαιο 6

6.1 Παρουσίαση Δεδομένων

Στην παρούσα ενότητα θα εφαρμόσουμε τους 3 αλγορίθμους ομαδοποίησης σε ένα σύνολο δεδομένων, το οποίο αποτελείται από 2 μεταβλητές (X και Y) και 3000 παρατηρήσεις. Με δεδομένο ότι περιοριζόμαστε στις 2 διαστάσεις, μπορούμε να αναπαραστήσουμε γραφικά το σύνολο δεδομένων μας (εικόνα 6.1).

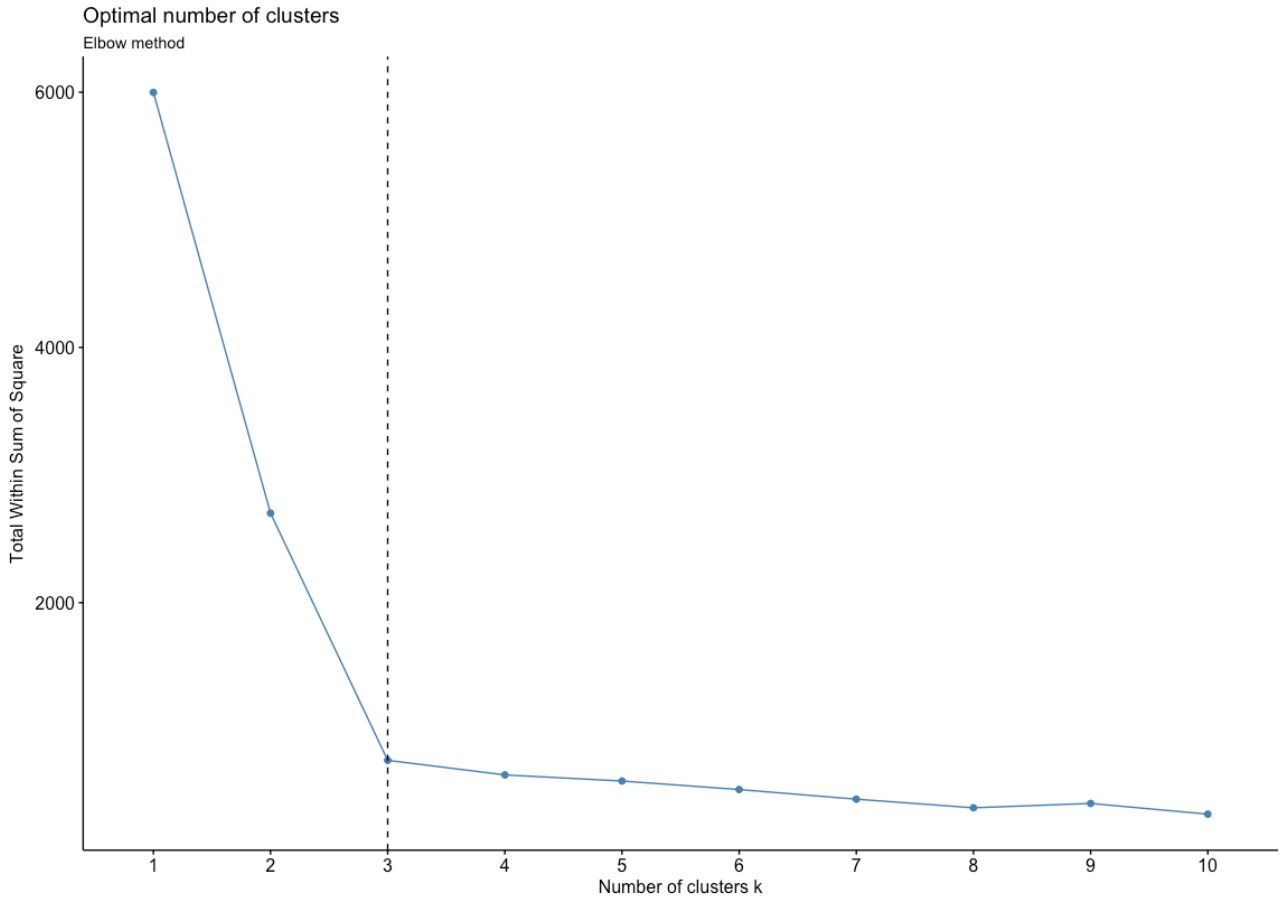


Εικόνα 6.1: Γραφική αναπαράσταση δεδομένων σε 2 διαστάσεις

Από το παραπάνω γράφημα καταλήγουμε στο ότι οι παρατηρήσεις μπορούν να χωριστούν σε 3 συμπλέγματα κυκλικού σχήματος. Η διαπίστωση αυτή βοηθάει αρκετά στην επιλογή του αριθμού των συμπλεγμάτων K, και αναμένουμε οι διάφορες τεχνικές που έχουν αναφερθεί σε προηγούμενα κεφάλαια να μας οδηγήσουν σε παρόμοια συμπεράσματα. Μια τέτοια παρατήρηση δεν θα ήταν εφικτή αν το σύνολο των δεδομένων είχε παραπάνω από 2 μεταβλητές, και αυτό γιατί δεν θα επέτρεπε την απεικόνισή του στο επίπεδο.

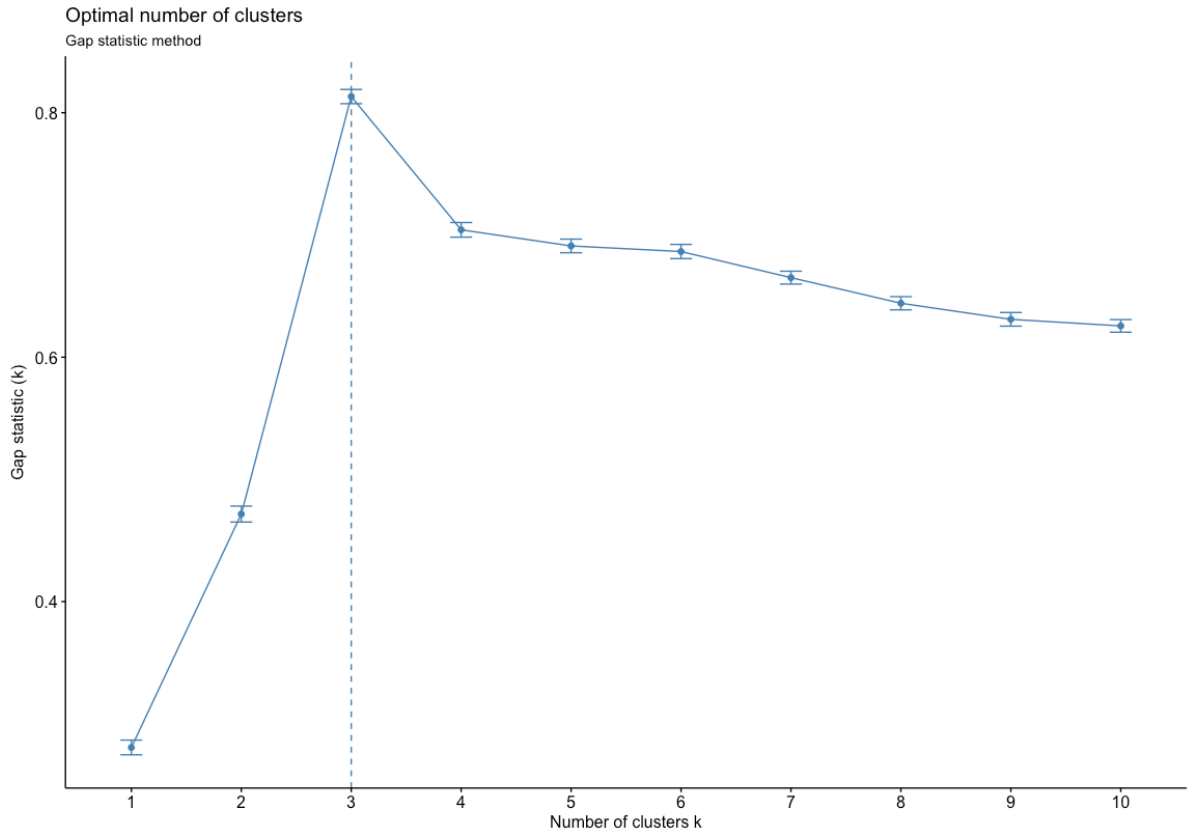
6.2 Εφαρμογή αλγορίθμου K means

Βασική προϋπόθεση για την εφαρμογή του αλγορίθμου K means είναι η επιλογή του αριθμού K. Εφαρμόζοντας τις μεθόδους Elbow, Gap Statistic και Silhouette (εικόνες 6.2, 6.3, 6.4), παρατηρούμε ότι και οι 3 μέθοδοι προτείνουν ως βέλτιστο αριθμό το K = 3, κάτι το οποίο επιβεβαιώνει την αρχική μας παρατήρηση.

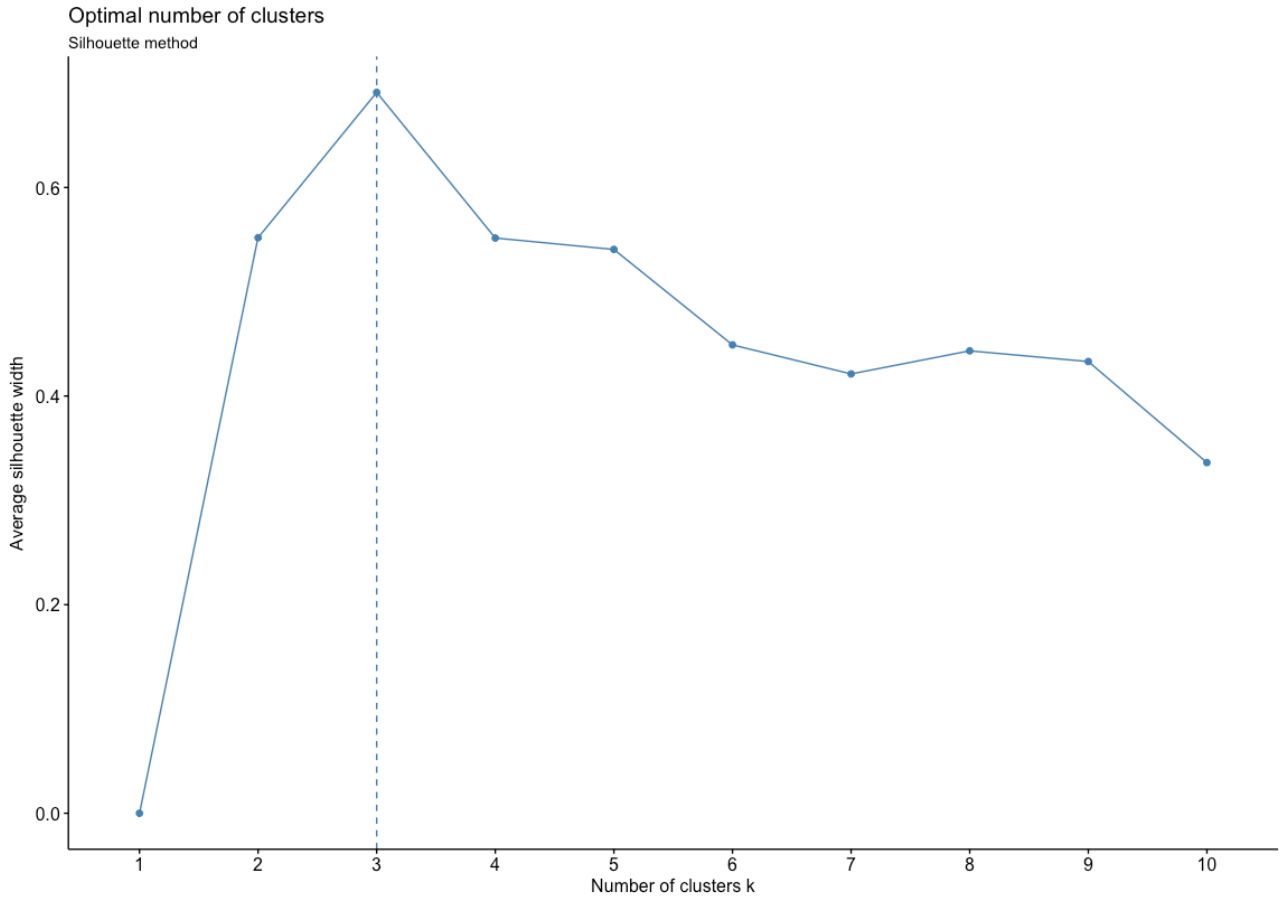


Εικόνα 6.2: Η μέθοδος του Ακγώνα

Έχοντας σαφή εικόνα από την γραφική αναπαράσταση των δεδομένων, παρατηρούμε ότι οι 3 μέθοδοι Elbow, Gap Statistic και Silhouette έδωσαν ικανοποιητικά αποτελέσματα ανταποκρινόμενα στην πραγματικότητα, πράγμα που όμως δεν συμβαίνει απαραίτητα πάντα.



Εικόνα 6.3: Η μέθοδος Gap Statistic



Εικόνα 6.4: Η μέθοδος Silhouette

Έχοντας επιλέξει τον αριθμό K μπορούμε πλέον να κάνουμε εφαρμογή του αλγορίθμου K means (Εικόνα 6.5). Βασιζόμενοι στο κυκλικό σχήμα των συμπλεγμάτων (εικόνα 6.1) επιλέγουμε την ευκλείδεια απόσταση. Επιπλέον, προκειμένου να αντιμετωπιστεί η τυχαιότητα της αρχικοποίησης, η οποία μπορεί να οδηγήσει σε τοπικό ακρότατο και όχι ολικό, πραγματοποιούμε 20 αρχικοποιήσεις των αρχικών μέσων, επιλέγοντας εκείνους με την μικρότερη τιμή της αντικειμενικής συνάρτησης.

Εικόνα 6.5: Ομαδοποίηση με τη μέθοδο K means

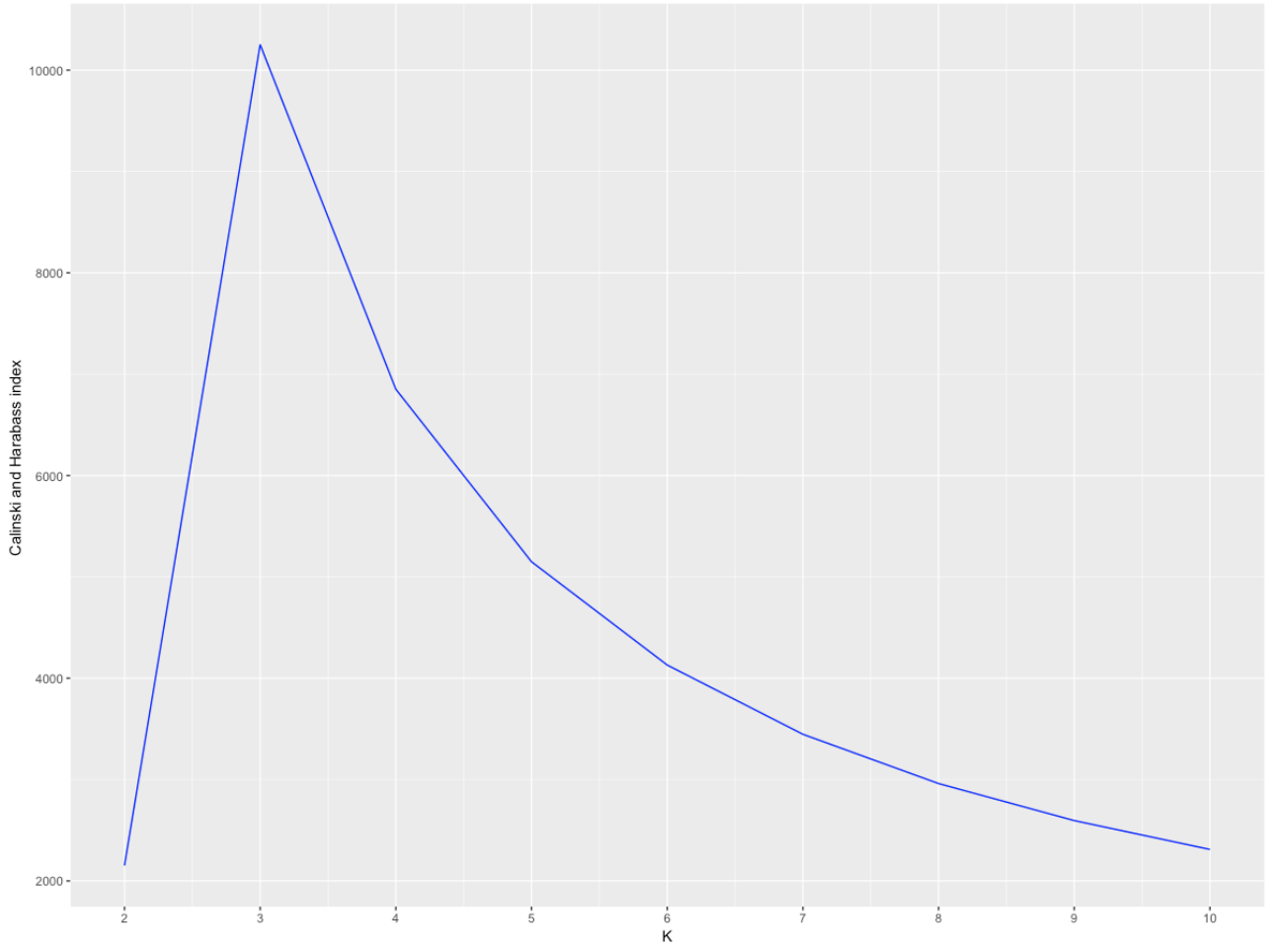
Πίνακας 6.1: Αριθμητικοί μέσοι παραγόμενων συμπλεγμάτων μέσω του αλγορίθμου K means

Σύμπλεγμα	X	Y
1	-1.21	-0.38
2	1.13	-1.04
3	0	1.16

Στον πίνακα 6.1 παρατηρούμε τους μέσους για τους οποίους η αντικειμενική συνάρτηση ελαχιστοποιείται. Με βάση τα αποτελέσματα της διαμέρισης στην εικόνα 6.5, παρατηρούμε ότι μας δίνουν ικανοποιητικά αποτελέσματα. Ο αλγόριθμος K means ήταν σε θέση να εντοπίσει αποτελεσματικά την ορθή διαμέριση του συνόλου των δεδομένων, κάτι το οποίο οφείλεται σε μεγάλο βαθμό στην κυκλική μορφή των συμπλεγμάτων, αλλά και στην απουσία θορύβου.

6.3 Εφαρμογή αλγορίθμου Single - Link

Με δεδομένο ότι ο αλγόριθμος Single link δημιουργεί μία ιεραρχία και όχι άμεσα ομαδοποίηση των δεδομένων, θα πρέπει αρχικά να υπολογίσουμε τον δείκτη Calin΄ski και Harabas για κάθε τιμή K, επιλέγοντας εκείνο το K για το οποίο έχουμε μεγιστοποίηση του δείκτη. Στην εικόνα 6.6 παρατηρούμε ότι η προτεινόμενη τιμή είναι η K = 3, κάτι το οποίο ταυτίζεται τόσο με τα αποτελέσματα των προηγούμενων μεθόδων (Elbow, Gap Statistic, Silhouette), όσο και με την πραγματικότητα.



Εικόνα 6.6: Διάγραμμα του δείκτη Calin΄ski και Harabasz για διάφορες τιμές του k

Η ομαδοποίηση που προκύπτει με τη βοήθεια της μεθόδου Single link κάνοντας χρήση την ευκλείδεια απόσταση και την κατάλληλη τομή στο παραγόμενο ιεραρχικό δέντρο, είναι ίδια με αυτή της μεθόδου K means (εικόνα 6.7 , πίνακας 6.2).

Εικόνα 6.7: Ομαδοποίηση με τη μέθοδο Single - link

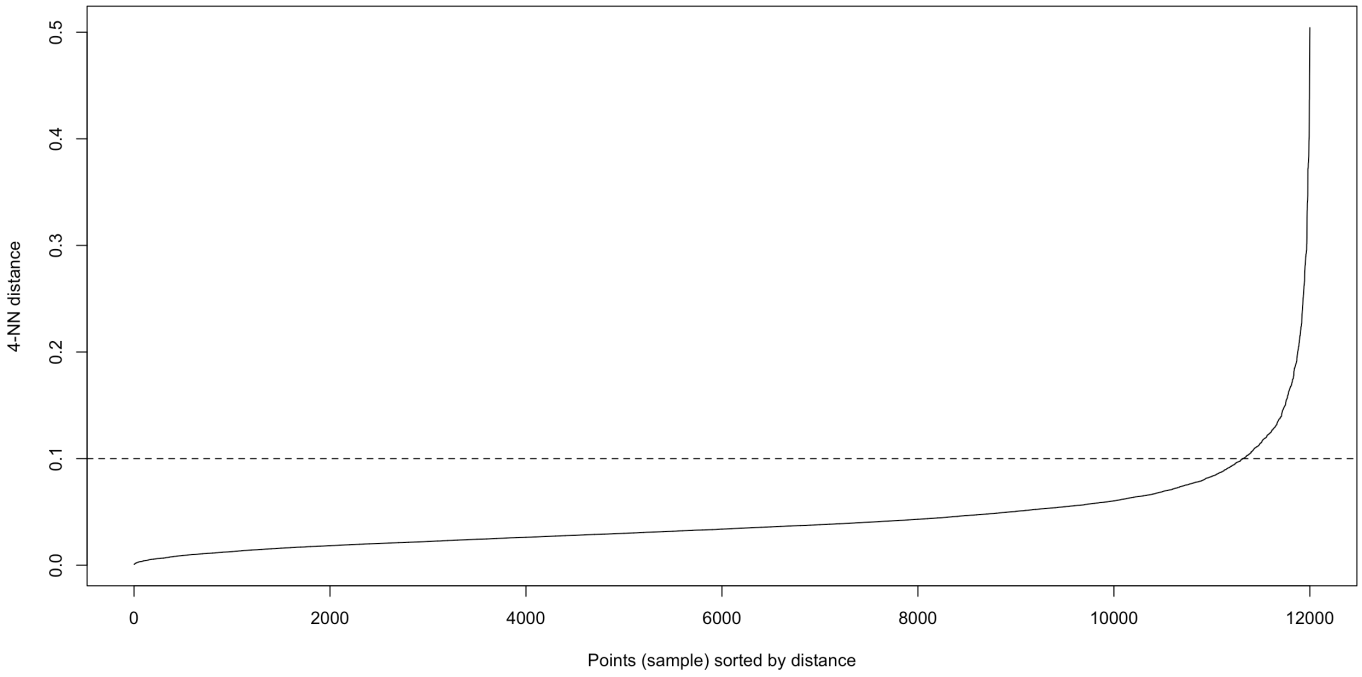
Πίνακας 6.2: Κατανομή σημείων στις 3 κλάσεις με τις μεθόδους K means και Single link

Κλάση	K means	Single link
1	897	897
2	952	1151
3	1151	952

Ο πίνακας 6.2 απεικονίζει πως κατανέμονται τα 3000 σημεία στις 3 κλάσεις με τις 2 μεθόδους (K means, Single link), με μόνη διαφορά την διαφορετική αρίθμηση. Δεδομένου ότι οι κατανομές είναι ίδιες, το ίδιο θα ισχύει και για τους αντίστοιχους αριθμητικούς μέσους (πίνακας 6.1).

6.4 Εφαρμογή αλγορίθμου DBSCAN

Βασική προϋπόθεση για την εφαρμογή του αλγορίθμου DBSCAN είναι ο προσδιορισμός των παραμέτρων Eps και N_{min}. Δεδομένου ότι βρισκόμαστε στις 2 διαστάσεις θεωρούμε $N_{min} = 4$ και σχεδιάζουμε το διάγραμμα 4-NN (εικόνα 6.8).



Εικόνα 6.8: Διάγραμμα 4-NN σε αύξουσα σειρά

Με βάση την εικόνα 6.8 παρατηρούμε την ενδεικτική τιμή για την παράμετρο $Eps = 0.1$.
Μετά από την τιμή 0.1 οι αντίστοιχες αποστάσεις αρχίζουν να αυξάνονται αισθητά. Μέσω επαναλαμβανόμενων δοκιμών, παρατηρήθηκε ότι τα βέλτιστα αποτελέσματα επιτεύχθηκαν για τις τιμές των παραμέτρων $N_{min} = 4$ και $Eps = 0.17$ (εικόνα 6.9). Ο αλγόριθμος κατάφερε να ομαδοποιήσει σωστά την πλειοψηφία των παρατηρήσεων, ορίζοντας ως θόρυβο τις παρατηρήσεις που είναι σχετικά πιο απομακρυσμένες από το αντίστοιχο σύμπλεγμα.

Πίνακας 6.3: Ομαδοποίηση σημείων με τη μέθοδο DBSCAN

0	1	2	3	4
31	883	1146	5	935

Εικόνα 6.9: Ομαδοποίηση με τη μέθοδο DBSCAN

Με βάση την κατανομή των παρατηρήσεων (πίνακας 6.3) καθώς και την εικόνα 6.9, παρατηρούμε ότι το σύμπλεγμα στο οποίο εμφανίστηκαν τα περισσότερα προβλήματα είναι αυτό με τον αριθμό 4, όπου ο αλγόριθμος δεν κατάφερε να εντοπίσει 17 παρατηρήσεις σε σχέση με τους αλγορίθμους Single link , K means. Αντίστοιχα, στο σύμπλεγμα 1 δεν κατάφερε να εντοπίσει 14 παρατηρήσεις, ενώ στο σύμπλεγμα 2 δεν εντοπίστηκαν 5 παρατηρήσεις. Παρατηρούμε ότι για το σύμπλεγμα 2, που είναι πιο "πυκνό", ο αλγόριθμος DBSCAN δίνει καλύτερα αποτελέσματα σε σχέση το σύμπλεγμα 4 που είναι πιο "αραιό". Το παραπάνω αποτέλεσμα γνωρίζουμε ότι είναι ένα από τα μειονεκτήματα του αλγορίθμου.

Κεφάλαιο 7

7.1 Παρουσίαση Παρατηρήσεων

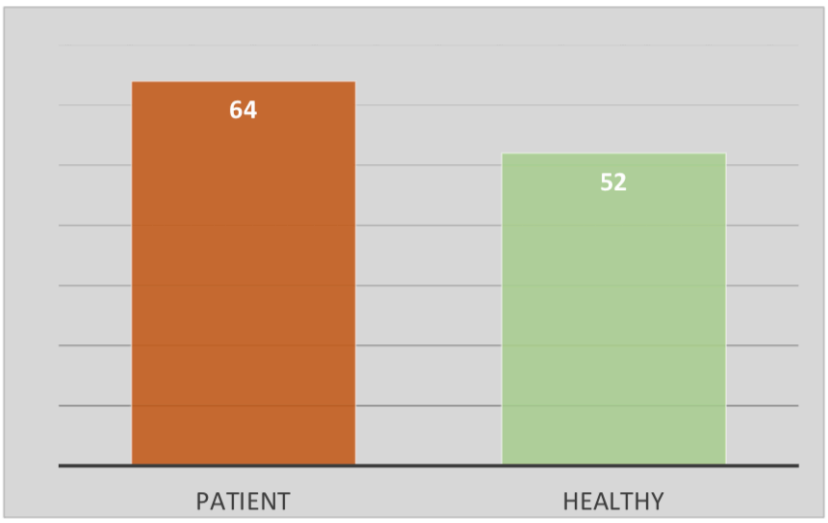
Στην παρούσα ενότητα θα γίνει παρουσίαση των παρατηρήσεων (με τη βοήθεια του στατιστικού πακέτου R) που αφορούν ασθενείς με ή χωρίς καρκίνο του μαστού. Τα δεδομένα αποτελούνται από 10 ποσοτικές ανεξάρτητες μεταβλητές και 1 κατηγορική την οποία θα χρησιμοποιήσουμε για να ελέγξουμε την απόδοση των 3 αλγορίθμων. Οι ανεξάρτητες μεταβλητές αποτελούν ανθρωπομετρικά δεδομένα και παράμετροι που μπορούν να συγκεντρωθούν με μία απλή ανάλυση αίματος. Για καλύτερη κατανόηση των αποτελεσμάτων, θα πραγματοποιήσουμε μελέτη της δομής των δεδομένων με τη βοήθεια διαφόρων γραφημάτων. Τα δεδομένα που θα χρησιμοποιήσουμε θα είναι της μορφής:

Πίνακας 7.1: Υποσύνολο ιατρικών δεδομένων

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Class
48.0	23.5	70.0	2.71	0.467	8.81	9.70	8.00	417	Healthy
83.0	20.7	92.0	3.12	0.707	8.84	5.43	4.06	469	Healthy
82.0	23.1	91.0	4.50	1.01	17.9	22.4	9.28	555	Healthy
68.0	21.4	77.0	3.23	0.613	9.88	7.17	12.8	928	Healthy
86.0	21.1	92.0	3.55	0.805	6.70	4.82	10.6	774	Healthy
49.0	22.9	92.0	3.23	0,732	6.83	13.7	10.3	530	Healthy

Συνολικά έχουμε στη διάθεσή μας 116 παρατηρήσεις, από τις οποίες οι 64 είναι ασθενής με καρκίνο του μαστού και 52 που αφορούν υγιείς περιπτώσεις, κάτι το οποίο μπορούμε να δούμε και από το επόμενο γράφημα (εικόνα 7.1).

Εικόνα 7.1: Κατανομή παρατηρήσεων στις 2 φυσικές κλάσεις



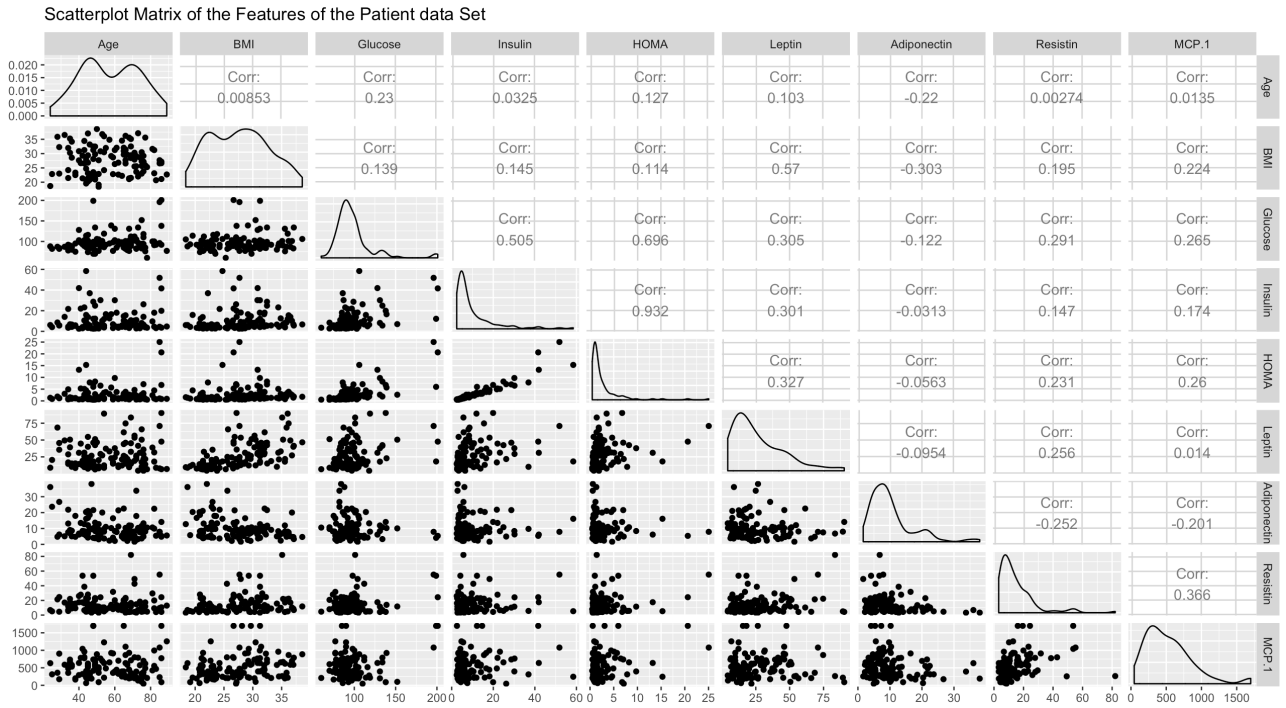
Τα μεγέθη των 2 φυσικών συμπλεγμάτων είναι σχεδόν ίδια γεγονός αρκετά θετικό, καθώς απ ότι είδαμε στην ανάλυση του αλγορίθμου K means η απόδοση της μεθόδου επηρεάζεται όταν τα φυσικά συμπλέγματα διαφέρουν σημαντικά ως προς το μέγεθός τους. Επιπλέον οι κλίμακες των ποσοτικών μεταβλητών είναι διαφορετικές γεγονός που καλούμαστε να επιλύσουμε. Στη συνέχεια παρουσιάζουμε μερικά μέτρα θέσης για τις ποσοτικές μας μεταβλητές:

Age	BMI	Glucose	Insulin	HOMA
Min. : 24.0	Min. : 18.37	Min. : 60.00	Min. : 2.432	Min. : 0.4674
1st Qu. : 45.0	1st Qu. : 22.97	1st Qu. : 85.75	1st Qu. : 4.359	1st Qu. : 0.9180
Median : 56.0	Median : 27.66	Median : 92.00	Median : 5.925	Median : 1.3809
Mean : 57.3	Mean : 27.58	Mean : 97.79	Mean : 10.012	Mean : 2.6950
3rd Qu. : 71.0	3rd Qu. : 31.24	3rd Qu. : 102.00	3rd Qu. : 11.189	3rd Qu. : 2.8578
Max. : 89.0	Max. : 38.58	Max. : 201.00	Max. : 58.460	Max. : 25.0503

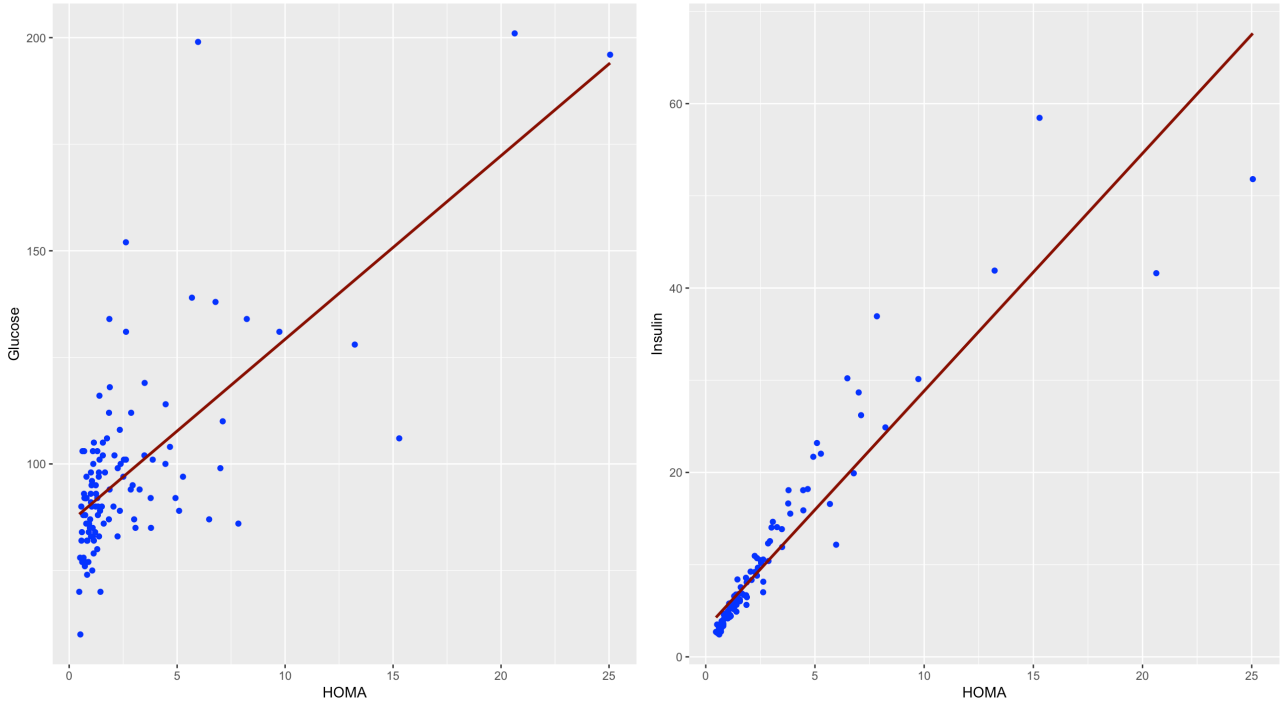
Leptin	Adiponectin	Resistin	MCP.1
Min. : 4.311	Min. : 1.656	Min. : 3.210	Min. : 45.84
1st Qu. : 12.314	1st Qu. : 5.474	1st Qu. : 6.882	1st Qu. : 269.98
Median : 20.271	Median : 8.353	Median : 10.828	Median : 471.32
Mean : 26.615	Mean : 10.181	Mean : 14.726	Mean : 534.65
3rd Qu. : 37.378	3rd Qu. : 11.816	3rd Qu. : 17.755	3rd Qu. : 700.09
Max. : 90.280	Max. : 38.040	Max. : 82.100	Max. : 1698.44

Παρατηρούμε ότι το εύρος των μεταβλητών Insulin,HOMA,Resistin,Adiponection και MCP.1 είναι μεγάλο με τη μέγιστη τιμή να υπερβαίνει αρκετά την αντίστοιχη διάμεσο γεγονός που αποτελεί μία αρχική ένδειξη ύπαρξης ακραίων τιμών για αυτές τις μεταβλητές. Στη συνέχεια στο παρακάτω διάγραμμα συνοψίζονται οι μεταξύ τους ανά 2 σχέσεις

Εικόνα 7.2: Εκτίμηση κατανομών, διαγράμματα διασποράς και γραμμική συσχέτιση



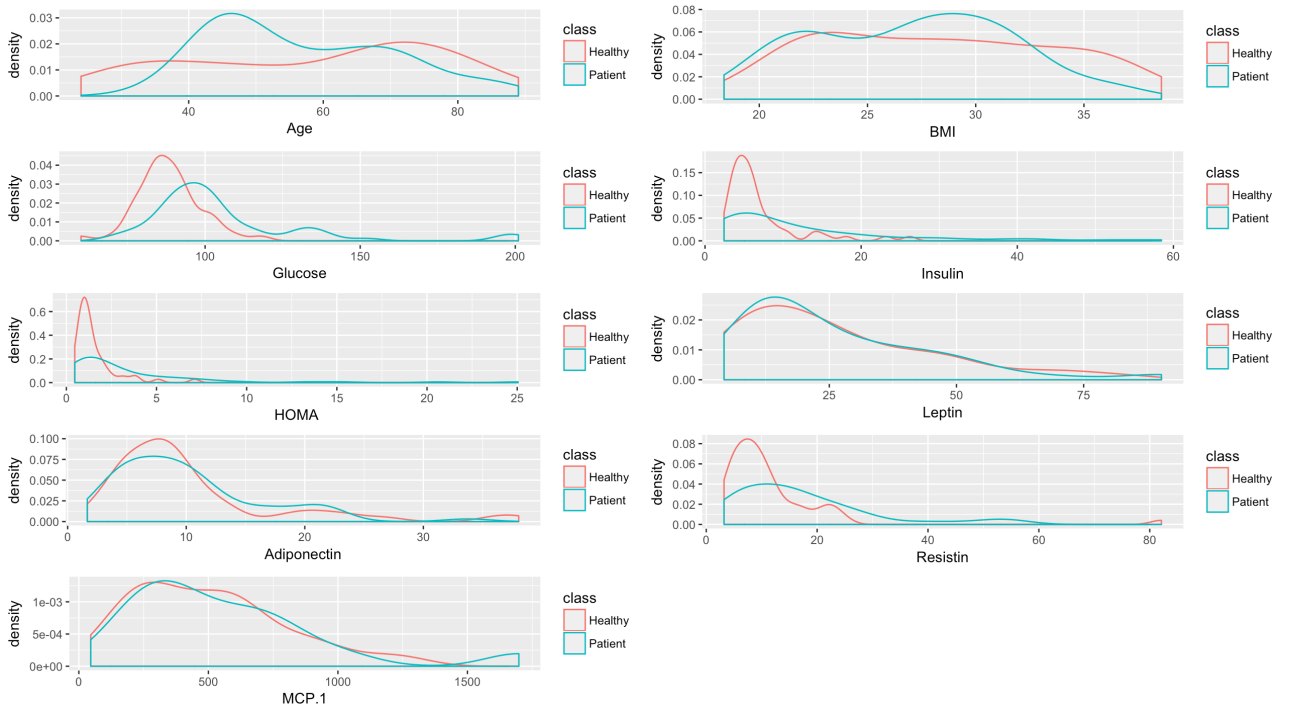
Το παραπάνω διάγραμμα (εικόνα 7.2) μας βοηθάει στο να εστιάσουμε τη προσοχή μας σε συγκεκριμένες μεταβλητές οι οποίες χρήζουν περαιτέρω διερεύνησης. Συγκεκριμένα παρατηρούμε ότι οι μεταβλητές Insulin και HOMA έχουν ισχυρά θετική γραμμική συσχέτιση, με τον αντίστοιχο συντελεστή γραμμικής συσχέτισης να παίρνει την τιμή 0.932, καθώς επίσης και η μεταβλητή HOMA με τη Glucose με συντελεστή 0.696 . Με βάση τις παραπάνω παρατηρήσεις πάμε να εξετάσουμε γραφικά τη γραμμική συσχέτιση μεταξύ των μεταβλητών αυτων.



Εικόνα 7.3: Ευθεία ελαχίστων τετραγώνων για Glucose-HOMA και Insuling-HOMA

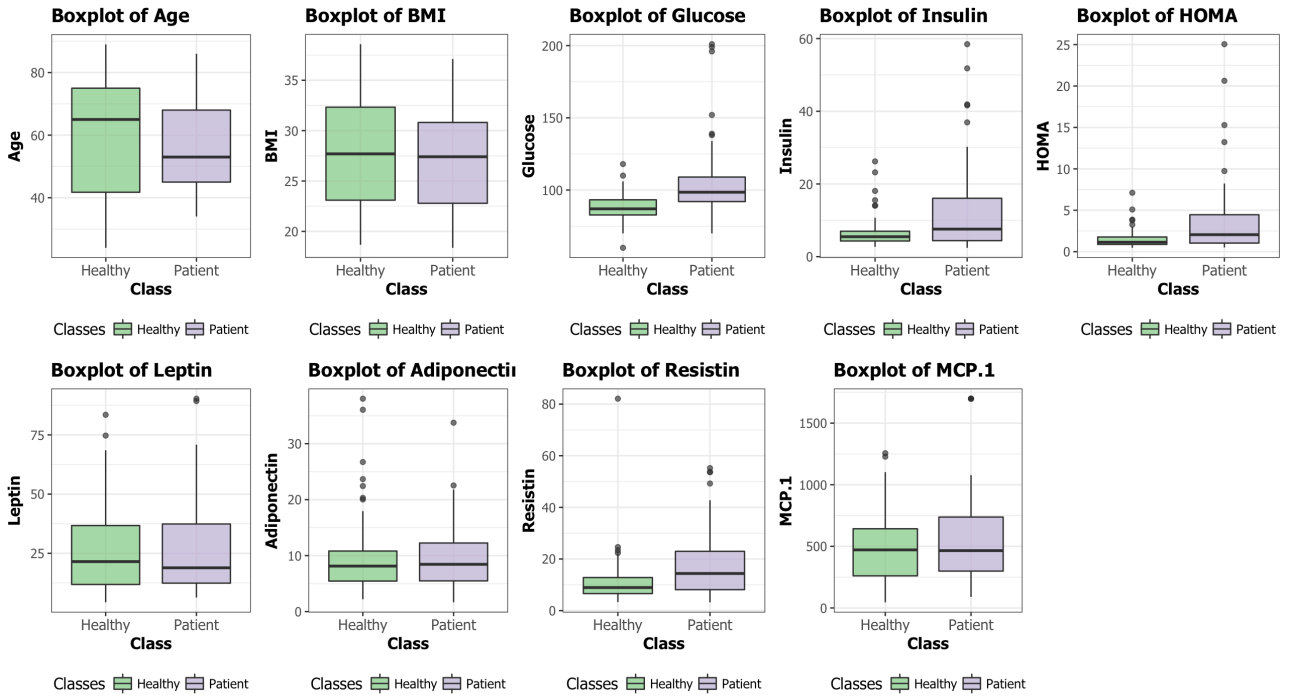
Κάνοντας προσαρμογή ευθείας με τη μέθοδο ελαχίστων τετραγώνων (εικόνα 7.3) παρατηρούμε ότι έχουμε καλύτερη προσαρμογή στο γράφημα HOMA-Insulin το οποίο επαληθεύει την ισχυρή γραμμική εξάρτηση μεταξύ των 2 μεταβλητών. Η πληροφορία αυτή θα μας χρειαστεί στην τελική επιλογή των μεταβλητών που θα χρησιμοποιήσουμε στους 3 αλγορίθμους ομαδοποίησης, καθώς μας ενδιαφέρει να έχουμε όσο το δυνατό λιγότερες διαστάσεις χωρίς να χάνουμε πολύ σε πληροφορία.

Προκειμένου να εντοπίσουμε ποιές μεταβλητές είναι αυτές που διαφοροποιούν περισσότερο τις 2 κλάσεις , πραγματοποιούμε εκτίμηση των κατανομών των μεταβλητών για κάθε κλάση ξεχωριστά, με τη μέθοδο των πυρήνων χρησιμοποιώντας ως συνάρτηση πυρήνα τη γκαουσιανή κατανομή. Τα αποτελέσματα της παραπάνω εκτίμησης φαίνονται στο παρακάτω διάγραμμα (εικόνα 7.4). Παρατηρούμε ότι όσον αφορά τις μεταβλητές MCP.1, Leptin, Adiponectin δεν παρουσιάζεται κάποια έντονη διαφοροποίηση μεταξύ των 2 κλάσεων. Αντίθετα έντονες διαφοροποιήσεις υπάρχουν στις μεταβλητές Age, Glucose, Insulin, HOMA, Resistin και BMI.



Εικόνα 7.4: Εκτίμηση κατανομών κάθε μεταβλητής σε κάθε κλάση

Για την ενίσχυση των παραπάνω συμπερασμάτων αλλά και τον εντοπισμό των μεταβλητών με ακραίες παρατηρήσεις κατασκευάζουμε το θυκόγραμμα κάθε μεταβλητής για κάθε κλάση χωριστά (εικόνα 7.5):



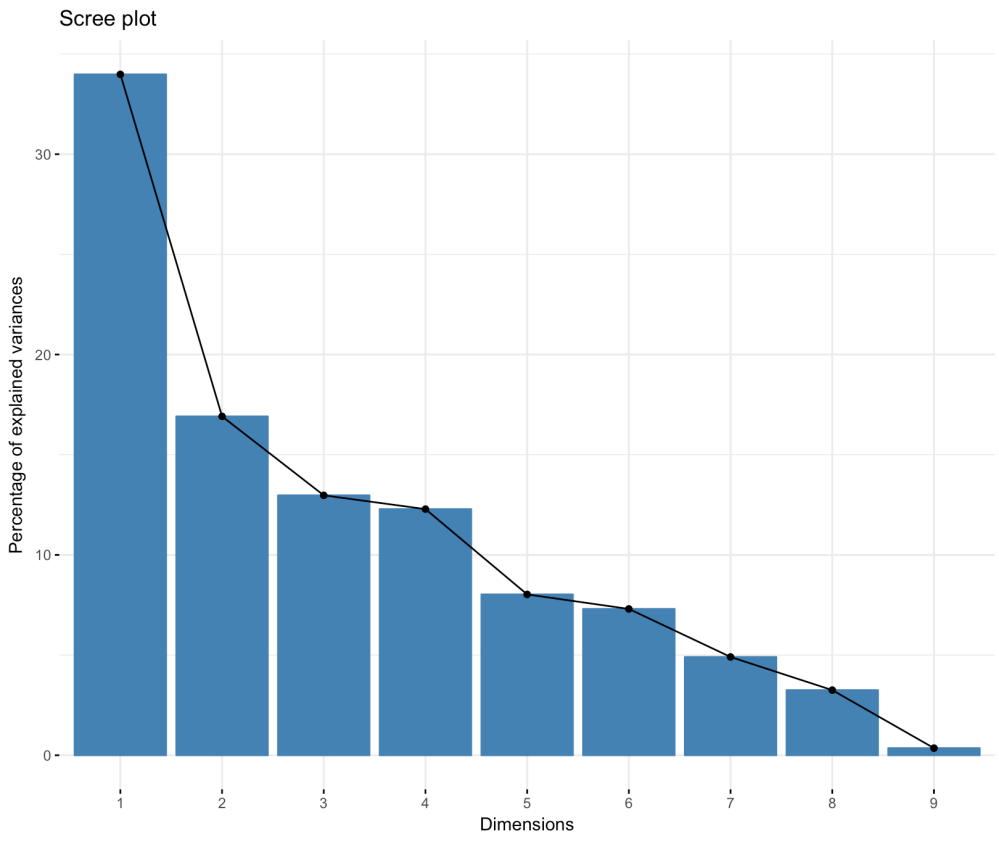
Εικόνα 7.5: Θυκόγραμμα κάθε μεταβλητής σε κάθε κλάση

Παρατηρούμε ότι σε όλες τις μεταβλητές εκτός των περιπτώσεων Age και BMI έχουμε ακραίες παρατηρήσεις, με τις μεταβλητές Adiponectin,HOMA,Insulin,Resistin και Glucose να εμφανίζουν τις περισσότερες. Επιπλέον όσον αφορά τις μεταβλητές Insulin και Resistin υπάρχουν παρατηρήσεις οι οποίες θεωρούνται ακραίες για τη κλάση Healthy αλλά βρίσκονται εντός 1ου και 3ου τεταρτημόριου της κλάσης Patient γεγονός που ενδεχομένως μπορεί να λειτουργήσει ως θόρυβος επηρεάζοντας την απόδοση του αλγορίθμου K means αλλά και του Single-link hierrarchical clustering.

7.2 Μείωση Διαστάσεων

Παρατηρούμε ότι μερικές από τις μεταβλητές δεν συμβάλουν αρκετά στο διαχωρισμό μεταξύ των 2 κλάσεων. Γι αυτό το λόγο αρχικά θα κάνουμε χρήση PCA προκειμένου να εντοπίσουμε μετασχηματισμούς των αρχικών μας επεξηγηματικών μεταβλητών και στη συνέχεια θα πραγματοποιήσουμε μείωση των διαστάσεων του αρχικού συνόλου των παρατηρήσεων. Πραγματοποιώντας μείωση των διαστάσεων πετυχαίνουμε να εκφράσουμε μεγάλο ποσοστό της αρχικής μεταβλητότητας χρησιμοποιώντας λιγότερες μεταβλητές, γεγονός αρκετά θετικό όσον αφορά τη χρονική πολυπλοκότητα των αλγορίθμων αλλά και την αποφυγή “δυσλειτουγίας” της Ευκλείδειας μετρικής στις πολλές διαστάσεις (curse of dimensionality). Εφαρμόζοντας PCA βρίσκουμε τις βασικές συνιστώσες (principal components) καθώς επίσης και το ποσοστό της μεταβλητότητας που εκφράζει η κάθε μία, το οποίο μπορούμε να δούμε στο επόμενο διάγραμμα (εικόνα 7.6):

Εικόνα 7.6: Ποσοστό μεταβλητότητας για κάθε βασική συνιστώσα

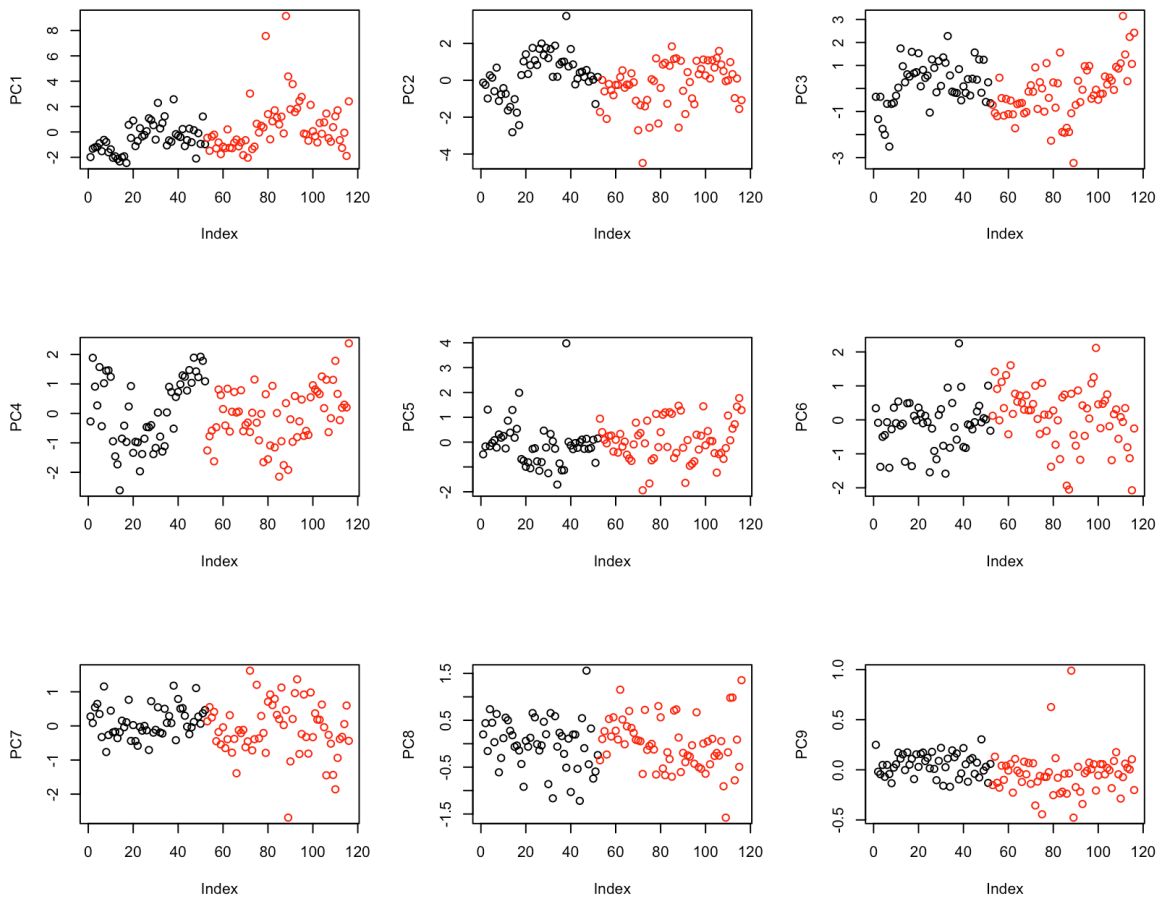


Οι principal components είναι όσες και οι αρχικές μεταβλητές μας, παρατηρούμε όμως ότι η τελευταία εκφράζει πολύ μικρό ποσοστό της μεταβλητότητας (0.4%) ενώ η πρώτη εκφράζει ένα ποσοστό της τάξης 34%.

Πίνακας 7.2: Ποσοστό μεταβλητότητας που εκφράζεται από κάθε κύρια συνιστώσα

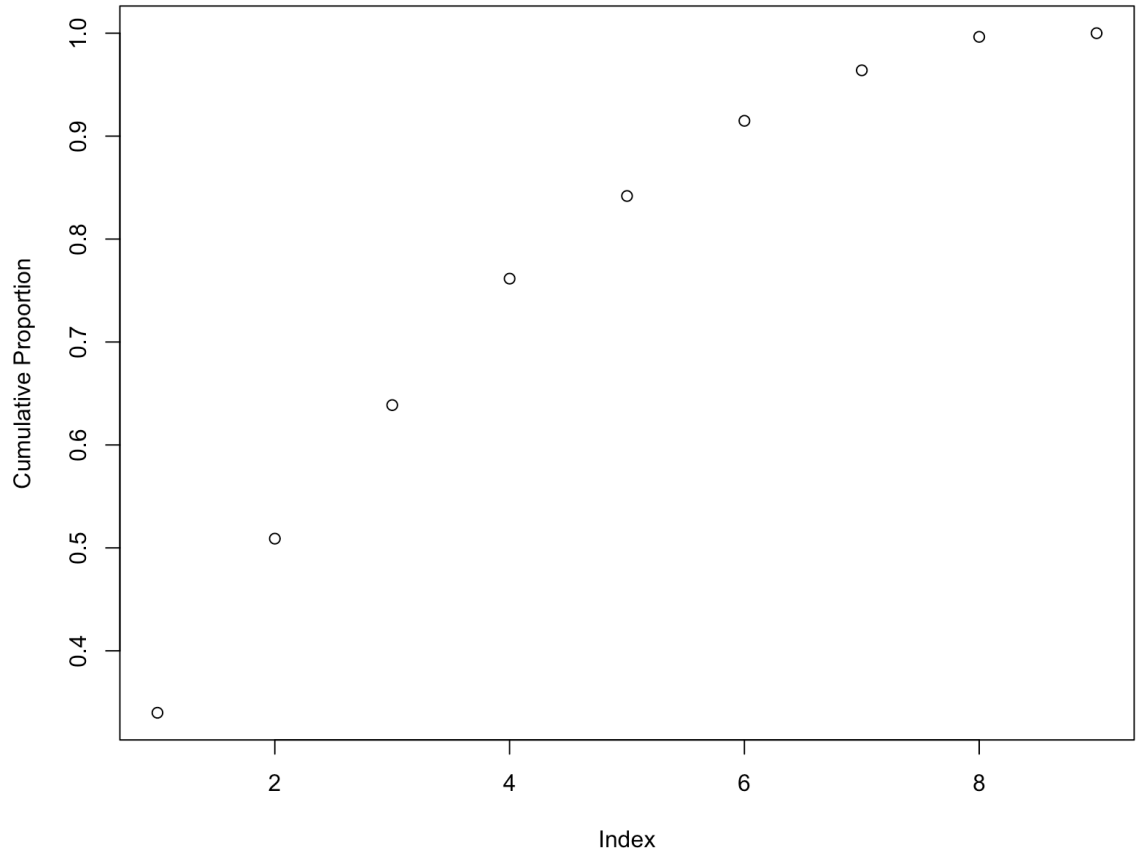
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.7489	1.2338	1.0805	1.0515	0.85002	0.81073	0.66449	0.54095	0.17894
Proportion of Variance	0.3398	0.1691	0.1297	0.1229	0.08028	0.07303	0.04906	0.03251	0.00356
Cumulative Proportion	0.3398	0.5090	0.6387	0.7615	0.84184	0.91487	0.96393	0.99644	1.00000

Παρατηρούμε ότι καμοία από αυτές δεν μπορεί να χρησιμοποιηθεί αποκλειστικά για τη περιγραφή των αρχικών παρατηρήσεων. Επιπλέον στο παρακάτω διάγραμμα (εικόνα 7.7) βλέπουμε τις προβολές των αρχικών μεταβλητών του συνόλου των παρατηρήσεων σε κάθε μία απο τις βασικές συνιστώσες που προέκυψαν μέσω της PCA.



Εικόνα 7.7: Προβολές στις κύριες συνιστώσες (Patient= Κόκκινο , Healthy= Μαύρο)

Δεδομένου ότι δεν υπάρχει κάποια βασική συνιστώσα η οποία να εκφράζει αρκετά μεγάλο ποσοστό της μεταβλητότητας, παρατηρούμε ότι σε κανένα από αυτά δεν έχουμε σαφή διαχωρισμό των 2 φυσικών συμπλεγμάτων. Προκειμένου επομένως να αποφασίσουμε πόσες από αυτές θα επιλέξουμε για τον μετασχηματισμό του αρχικού συνόλου δεδομένων, παρατηρούμε ότι χρησιμοποιώντας τις 6 πρώτες (εικόνα 7.8) θα έχουμε περιγραφή της μεταβλητότητας κατά 92% μειώνοντας τις διαστάσεις κατά 3.

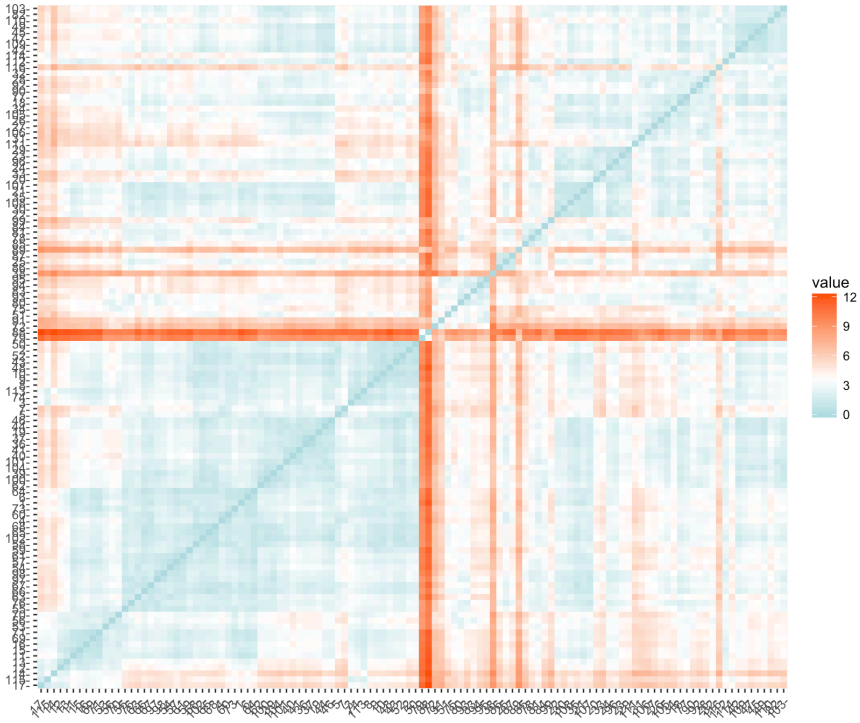


Εικόνα 7.8: Αθροιστικό ποσοστό έκφρασης της μεταβλητότητας

Στη συνέχεια κάνοντας χρήση του μετασχηματισμού $\Upsilon_{116\times6} = X_{116\times9}V_{9\times6}$ όπου V ο πίνακας με τα δεξιά ιδιάζουσα ιδιοδιανύσματα της μεθόδου SVD θα προκύψει ένα σύνολο παρατηρήσεων με λιγότερες επεξηγηματικές μεταβλητές το οποίο θα χρησιμοποιήσουμε για τη περεταίρω ανάλυσή μας.

7.3 Εφαρμογή K-means

Βασική προϋπόθεση για την εφαρμογή του αλγορίθμου K means σε ένα σύνολο δεδομένων είναι η γνώση του αριθμού των συμπλεγμάτων k. Συνεπώς προκειμένου να προσδιορίσουμε τη ποσότητα αυτή θα κάνουμε χρήση των 3 μεθόδων που αναφέραμε σε προηγούμενο κεφάλαιο (elbow method, silhouette method, gap statistic method) συγκρίνοντας τα αποτελέσματα.

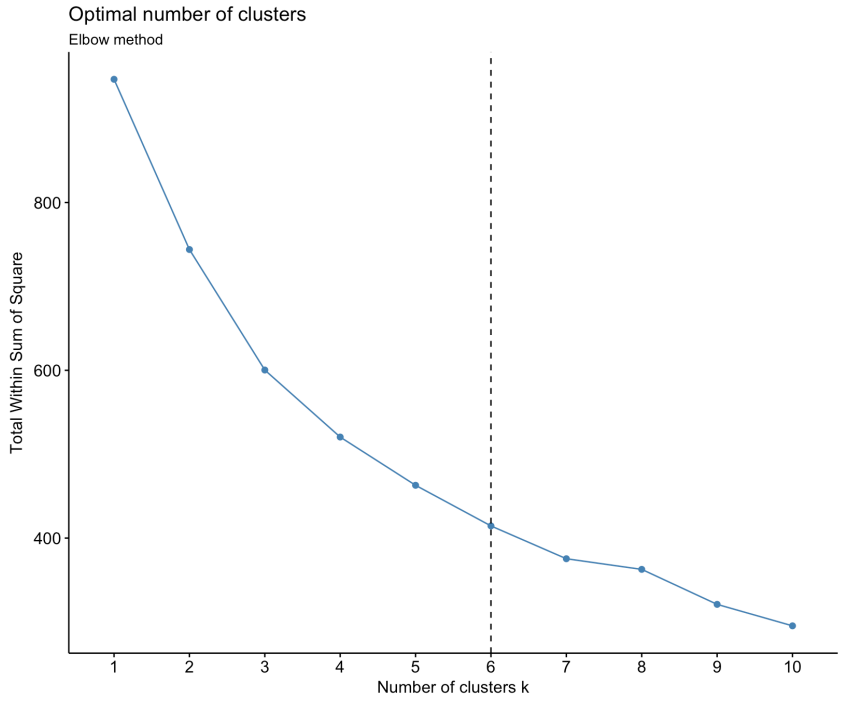


Εικόνα 7.9: Αναπαράσταση ευκλείδειων αποστάσεων

Αρχικά πραγματοποιούμε μία γραφική αναπαράσταση των Ευκλείδειων αποστάσεων (εικόνα 7.9) μεταξύ των παρατηρήσεων, έχοντας πραγματοποιήσει κανονικοποίηση των παρατηρήσεων έτσι ώστε τα αποτελέσματα να μην επηρεάζονται από τις διαφορετικές κλίμακες της κάθε μεταβλητής. Παρατηρούμε ότι 4 υποσύνολα παρατηρήσεων βρίσκονται πιο κοντά μεταξύ τους, με εντονότερο εκείνο που βρίσκεται κάτω αριστερά. Το παραπάνω γράφημα αποτελεί μία πρώτη εικόνα των παρατηρήσεων μας και στη περίπτωση που τα 2 φυσικά συμπλέγματα ήταν καλά διαχωρισμένα θα εμφάνιζε τον πραγματικό αριθμό συμπλεγμάτων. Παρ όλα αυτά το γεγονός ότι 2 παρατηρήσεις βρίσκονται κοντά μεταξύ τους δεν σημαίνει απαραίτητα ότι θα ανήκουν και στο ίδιο σύμπλεγμα, σύμφωνα πάντα με τον αλγόριθμο K means καθώς οι αποστάσεις που μας ενδιαφέρουν στην προκειμένη περίπτωση είναι αποστάσεις από τον αντίστοιχο μέσο.

Elbow method

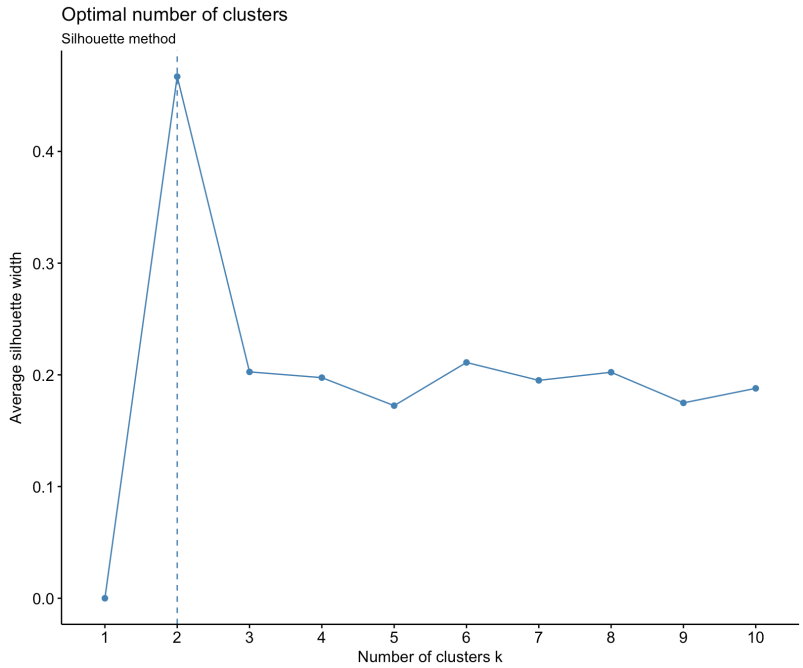
Κάνοντας εφαρμογή του Elbow method προκύπτει το παρακάτω γράφημα (εικόνα 7.10) στο οποίο παρατηρούμε ότι η επιλογή του κατάλληλου σημείου και επομένως του αντίστοιχου αριθμού k δεν είναι μοναδική και ιδιαίτερα εμφανής. Συγκεκριμένα θα μπορούσε να επιλεγεί τόσο το k=6 όσο το k=7, καθώς παρατηρούμε ότι μετά από αυτά τα σημεία η μείωση της τιμής του WSS είναι ελάχιστη. Ενδεικτικά θεωρούμε ως καταλληλότερη την τιμή k=6, καθώς αποτελεί μικρότερο αριθμό συμπλεγμάτων κάτι το οποίο εν γένει είναι προτιμότερο, θυσιάζοντας μόνο μία ελάχιστη αύξηση της τιμής WSS σε σχέση με την περίπτωση k=7. Παρ΄ όλα αυτά παρατηρούμε ότι η τιμή αυτή απέχει από την πραγματικότητα όπου έχουμε μόνο 2 ομάδες.



Εικόνα 7.10: Μέθοδος του Ανκώνα

Average Silhouette Method

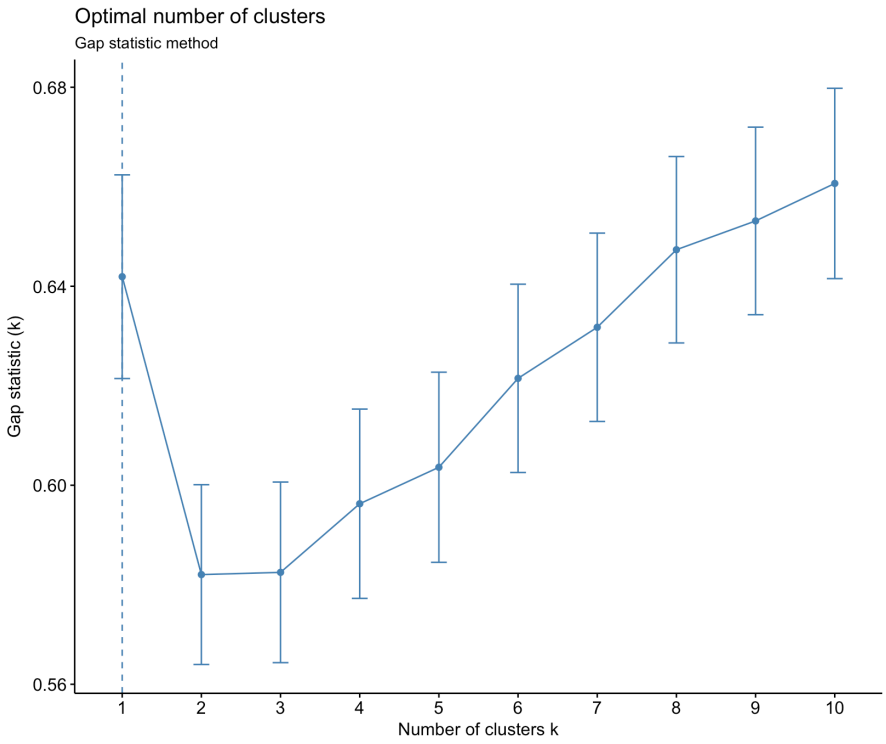
Στη προκειμένη περίπτωση (εικόνα 7.11) παρατηρούμε ότι για k=2 οι παρατηρήσεις είναι τοποθετημένες κατά μέσο όρο καλύτερα σε σχέση με τις διαμερίσεις που προέκυψαν για τις υπόλοιπες τιμές του k. Η μέθοδος αυτή είχε διαφορετικά αποτελέσματα από την Elbow method, τα οποία όμως ανταποκρίνονται στη πραγματική τιμή του αριθμού των συμπλεγμάτων.



Εικόνα 7.11: Average Silhouette μέθοδος

Gap statistic method

Κάνοντας εφαρμογή της μεθόδου Gap Statistic (εικόνα 7.12) για B=500 προέκυψε ως ιδανική τιμή το k=1 εμφανίζοντας μία αύξουσα τάση για τις επόμενες τιμές του k. Η παραπάνω ένδειξη υποδηλώνει την υπάρξη μίας μόνο ομάδας κάτι το οποίο όμως δεν ανταποκρίνεται στην πραγματικότητα αλλά ούτε και ταυτίζεται με τα αποτελέσματα των προηγούμενων μεθόδων.



Εικόνα 7.12: Gap statistic μέθοδος

Συμπερασματικά οι 3 μέθοδοι είχαν διαφοτερικά αποτελέσματα με την Average Silhouette Method να προσδιορίζει επακριβώς την πραγματική τιμή του k. Η εκ των προτέρων γνώση του αριθμού των συμπλεγμάτων βοηθάει στην αντιμετώπιση της διαφορετικότητας των αποτελεσμάτων, που προέκυψαν από τις 3 διαφορετικές μεθόδους. Στη συνέχεια χρησιμοποιώντας 2 συμπλέγματα κάνουμε χρήση του αλγορίθμου K means στα δεδομένα μας. Προκειμένου να αντιμετωπιστεί η τυχαιότητα της αρχικοποίησης η οποία μπορεί να οδηγήσει σε τοπικό ακρότατο και όχι ολικό, πραγματοποιούμε 20 αρχικοποιήσεις των αρχικών μέσων επιλέγοντας εκείνους με την μικρότερη τιμή της αντικειμενικής συνάρτησης. Στους παρακάτω πίνακες βρίσκονται οι μέσοι των συμπλεγμάτων που προέκυψαν από την εφαρμογή του αλγορίθμου καθώς και οι πραγματικές τιμές των μέσων.

Πίνακας 7.3: Προβλεπόμενοι αριθμητικοί μέσοι με τη μέθοδο K - means

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Healthy	56.64	25.19	90.27	6.080	1.377	16.66	11.07	11.28	462.8
Patient	58.42	31.65	110.6	16.69	4.933	43.52	8.664	20.59	656.7

Πίνακας 7.4: Πραγματικές τιμές των αριθμητικών μέσων κάθε συμπλέγματος

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Healthy	58.08	28.32	88.23	6.934	1.552	26.64	10.33	11.62	499.7
Patient	56.67	26.98	105.6	12.51	3.623	26.60	10.06	17.25	563.0

Παρατηρούμε ότι ο αλγόριθμος κατάφερε να διαφοροποιήσει τις κλάσεις όσον αφορά τις μεταβλητές Glucose,Insulin,HOMA,Resistin,MCP.1 χωρίς όμως απαραίτητα να προσεγγίσει τις πραγματικές τιμές των μέσων. Οι μεταβλητές Glucose,Insulin,HOMA είναι αυτές με το μεγαλύτερο βάρος στη δημιουργία της πρώτης βασικής συνιστώσας η οποία εκφράζει το 34% της μεταβλητότητας των παρατηρήσεων. Όσον αφορά όμως τις μεταβλητές Age,BMI,Leptin ο αλγόριθμος απέδωσε μη ικανοποιητικά αντιστρέφοντας τις τιμές των μέσων για τις μεταβλητές Age και BMI οι οποίες έχουν μικρά βάρη στη δημιουργία της πρώτης βασικής συνιστώσας. Στο παρακάτω γράφημα (εικόνα 7.13) βλέπουμε μία γραφική απεικόνιση των συμπλεγμάτων (1= Healthy και 2= Patient) που προέκυψαν μέσω του K means κάνοντας προβολή των παρατηρήσεων στις 2 πρώτες βασικές συνιστώσες οι οποίες εκφράζουν το 50,9% της μεταβλητότητας.

Εικόνα 7.13: Συμπλέγματα μέσω του αλγορίθμου K means

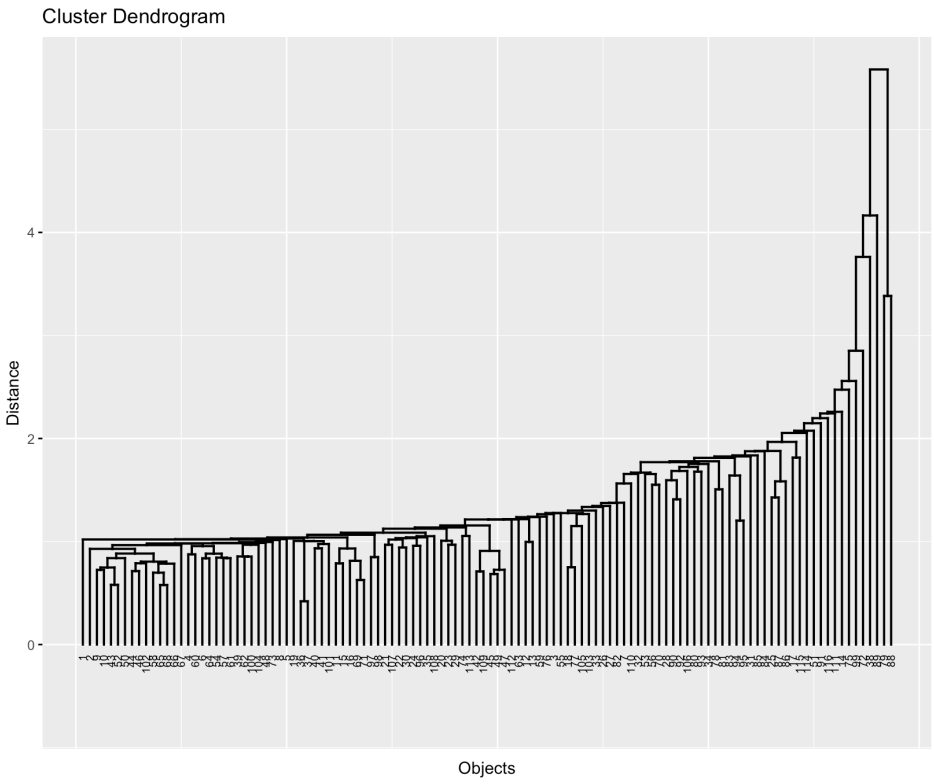
Η αξιολόγηση των παραπάνω αποτελεσμάτων θα πραγματοποιηθεί μέσω του υπολογισμού των εξωτερικών κριτηρίων αξιολόγησης κάνοντας χρήση της πραγματικής ομαδοποίησης των παρατηρήσεων και του παρακάτω πίνακα όπου διακρίνουμε τις τιμές $f_{00}, f_{11}, f_{10}, f_{01}$.

Πίνακας 7.5: Πλήθος ορθών και λανθασμένων ταξινομήσεων με τη μέθοδο K - means

		Actual Values		
Predicted values		Healthy	Patient	
	Healthy	$f_{11} = 37$	$f_{10} = 36$	73
	Patient	$f_{01} = 15$	$f_{00} = 28$	43
		52	64	116

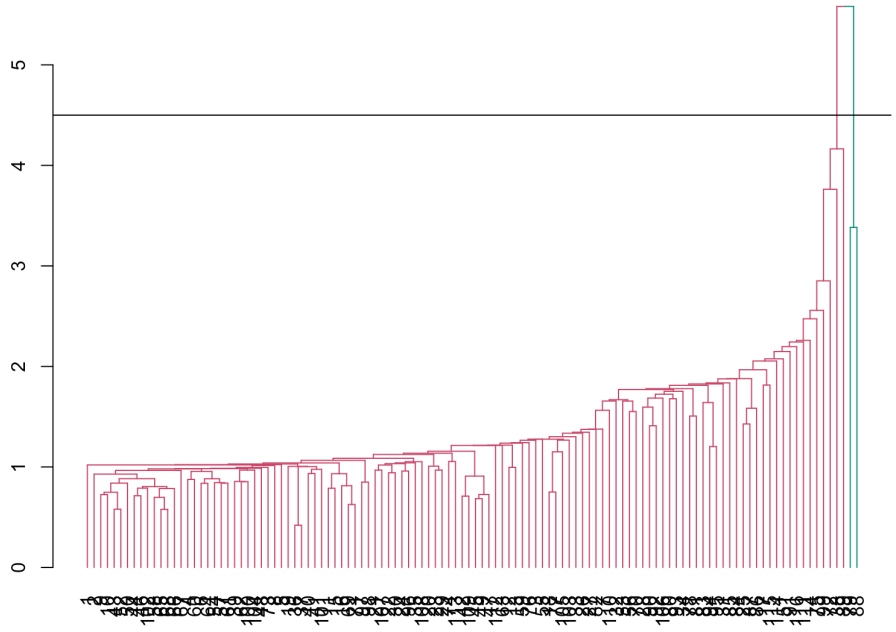
7.4 Εφαρμογή Single link hierarchical Clustering

Η εφαρμοφή του αλγορίθμου single link δεν προϋποθέτει τη γνώση του αριθμού των συμπλεγμάτων εκ των προτέρων. Επομένως κάνοντας χρήση του πίνακα εγγύτητας που έχει υπολογιστεί με χρήση της Ευκλείδειας μετρικής προκύπτει το εξής δενδρογράφημα (εικόνα 7.14):



Εικόνα 7.14: Παραγόμενο δενδρογράφημα του αλγορίθμου Single-link

Με βάση το παραπάνω γράφημα παρατηρούμε ότι το δέντρο δεν είναι ισορροπημένο. Επιπλέον οι παρατηρήσεις 79 και 88 που βρίσκονται τέρμα δεξιά έχουν μεγάλη ανομοιογένεια με τις άλλες παρατηρήσεις. Προκειμένου να γίνει αξιόλογηση κατά πόσο το παραγόμενο δενδρογράφημα ανταποκρίνεται στη πραγματική δομή των δεδομένων γίνεται χρήση της cophenetic correlation η οποία όπως έχουμε δεί εκφράζει τη συσχέτιση μεταξύ της απόστασης κατά την οποία 2 παρατηρήσεις βρέθηκαν στο ίδιο σύμπλεγμα με τη μεταξύ τους απόστασης. Η τιμή του συντελεστή υπολογίστηκε 0.85 τιμή η οποία είναι αρκετά κοντά στο 1 γεγονός που δείχνει υψηλή συσχέτιση. Πρέπει να επιτευχθεί τιμή 0,75 ή μεγαλύτερη, προκειμένου να θεωρηθεί χρήσιμη η ομαδοποίηση. Επιπλέον οι αντίστοιχες τιμές των δεικτών δέλτα είναι $\Delta_1 = 1.05$ και $\Delta_{0.5} = 1.07$ οι οποίες συνήθως χρησιμοποιούνται για σύγκριση ιεραρχικών δομών με βάση διαφορετικούς αλγορίθμους. Παρ΄ όλα αυτά ιδανικά θα θέλαμε να είναι κοντά στο 0 προκειμένου να είχαμε καλή προσαρμογή κάτι το οποίο δεν ισχύει. Δεδομένου ότι το αποτέλεσμα του αλγορίθμου είναι μία ιεραρχία και όχι μία διαμέριση των παρατηρήσεων, θα πρέπει να "κόψουμε" το δέντρο στο κατάλληλο ύψος προκειμένου να προκύψουν 2 συμπλέγματα. Από το παρακάτω δενδρογράφημα (εικόνα 7.14) παρατηρούμε ότι αν φέρουμε κάθετη στην απόσταση 4.5 θα προκύψουν 2 συμπλέγματα, χρησιμοποιώντας κόκκινο και πράσινο χρώμα αντίστοιχα.

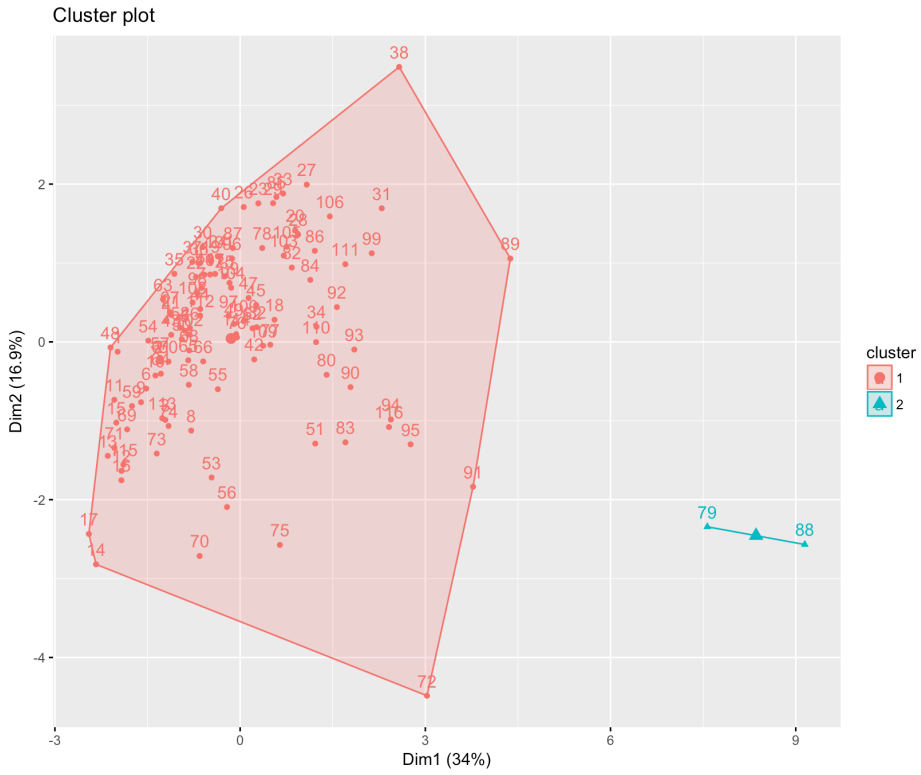


Εικόνα 7.15: Παραγόμενα συμπλέγματα(Κόκκινο = 1 , Πράσινο= 2) με μέθοδο Single link

Παρατηρούμε ότι η διαμέριση που προέκυψε τοποθέτησε σχεδόν όλες τις παρατηρήσεις σε ένα σύμπλεγμα με εξαίρεση τις 2 παρατηρήσεις που αναφέραμε προηγουμένως. Ο αλγόριθμος single - link γνωρίζουμε ότι μπορεί να επηρεαστεί από το φαινόμενο της

αλυσίδας, γεγονός που βλέπουμε να συμβαίνει και εδώ λόγο του ότι οι παρατηρήσεις των 2 συμπλεγμάτων βρίσκονται κοντά μεταξύ τους, εκτός από τις 2 ακριανές παρατηρήσεις.

Δεδομένου ότι ο αλγόριθμος απαιτεί μόνο 2 σημεία να βρίσκονται κοντά μεταξύ τους έτσι ώστε 2 συμπλέγματα να συγχωνευθούν, οδηγούμαστε στην δημιουργία ενός συμπλέγματος με αρκετά μεγάλη διάμετρο που περιέχει την πλειοψηφία των παρατηρήσεων. Προφανώς κάτι τέτοιο δεν είναι επιθυμητό καθώς δεν μας παρέχει καμοία διαφοροποίηση μεταξύ των 2 φυσικών συμπλεγμάτων. Η αφαίρεση αυτών των παρατηρήσεων δεν θα βελτίωνε τα αποτελέσματα καθώς τη θέση τους θα έπαιρναν άλλες παρατηρήσεις. Στο παρακάτω γράφημα (εικόνα 7.16) βλέπουμε μια γραφική απεικόνιση των 2 συμπλεγμάτων έχοντας κάνει προβολή στις 2 πρώτες βασικές συνιστώσες.

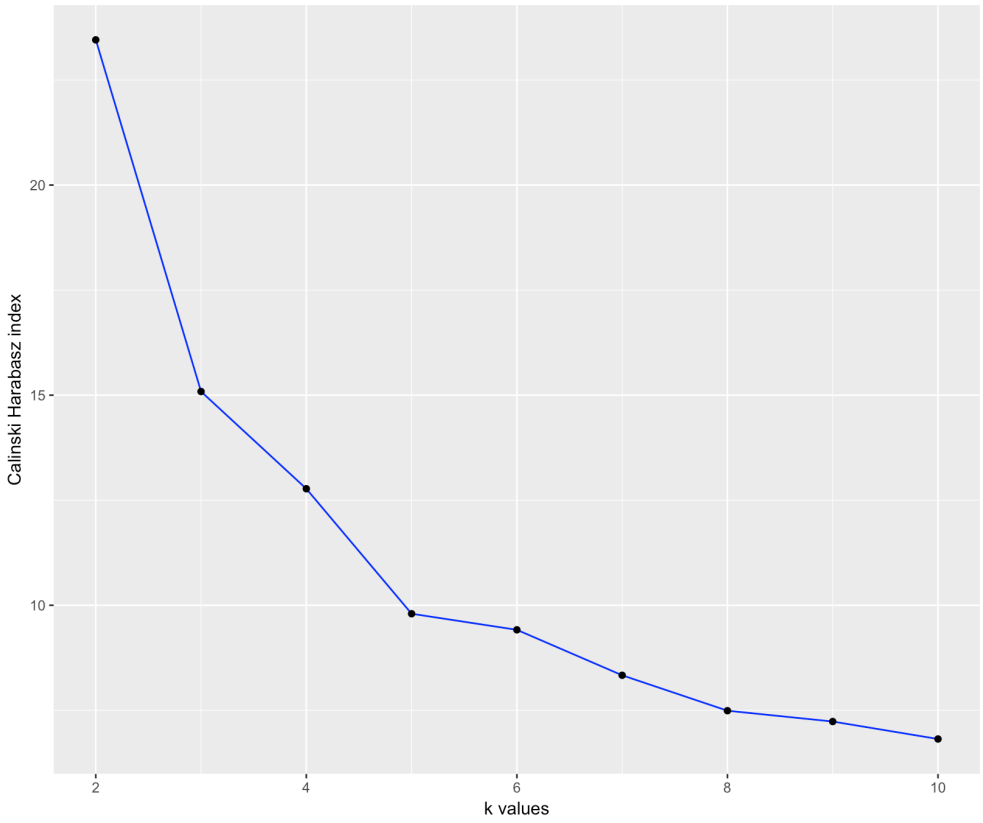


Εικόνα 7.16: Προβολή παραγόμενων συμπλεγμάτων στις 2 πρώτες κύριες συνιστώσες

Τα παραπάνω αποτελέσματα συνοψίζονται και στο πίνακα που ακολουθεί, όπου βλέπουμε πως κατανέμονται οι παρατηρήσεις στα 2 συμπλέγματα καθώς και τις πραγματικές τους κατηγοροποιήσεις.

Πίνακας 7.6: Πλήθος ορθών και λανθασμένων ταξινομήσεων με τη μέθοδο Single - link

		Actual Values		
Predicted values		Healthy	Patient	
	Healthy	$f_{11} = 52$	$f_{10} = 62$	114
	Patient	$f_{01} = 0$	$f_{00} = 2$	2
		52	64	116



Εικόνα 7.17: Δείκτης Calin΄ski και Harabasz για κάθε k

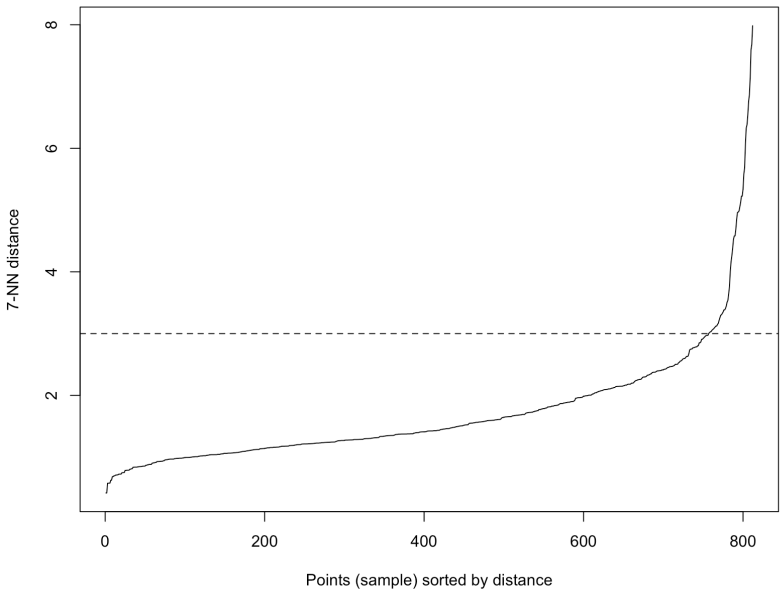
Άμα θέλαμε να εκτιμήσουμε των αριθμό των συμπλεγμάτων πριν φέρουμε την αντίστοιχη κάθετο θα μπορούσαμε και εδώ να κάνουμε χρήση των 3 προηγούμενων μεθόδων (Elbow, Average Silhouette, Gap statistic). Επιπλέον για την εκτίμηση του αριθμού k μπορούμε να κάνουμε και χρήση του δείκτη Calin΄ski και Harabasz παράγοντας το παραπάνω γράφημα (εικόνα 7.17). Στη προκειμένου περίπτωση μεγάλες τιμές του δείκτη υποδηλώνουν καλή διαμέριση. Από το γράφημα βλέπουμε ότι αυτό επιτυγχάνεται για την τιμή k=2 η οποία είναι και η πραγματική τιμή για τον αριθμό των συμπλεγμάτων. Συνεπώς παρ΄ όλο που ο αλγόριθμος single - link δεν προϋποθέτει τη γνώση του αριθμού των συμπλεγμάτων, κάνοντας χρήση του δείκτη Calin΄ski και Harabasz είμασταν σε θέση να το εντοπίσουμε.

7.5 Εφαρμογή DBSCAN

Για την εφαμορφή του αλγορίθμου DBSCAN θα πρέπει να γίνει προσδιορισμός των παραμέτρων Eps και N_{min}. Δεδομένου ότι βρισκόμαστε σε πολυδιάστατα δεδομένα (m=6) θεωρούμε $N_{min} = 7$ και σχεδιάζουμε το γράφημα 7-NN για κάθε παρατήρηση σε αύξουσα σειρά. Από το γράφημα (εικόνα 7.18) παρατηρούμε ότι ενδεικτική τιμή αποτελει το Eps=3 καθώς από εκείνη την τιμή και μετά έχουμε απότομη άνοδο των αποστάσεων. Επομένως κάθε παρατήρηση που η απόσταση από τον 7ο κοντινότερο γείτονα είναι μικρότερη η ίση του 3 θα θεωρηθεί κεντρική παρατήρηση, ενώ αντίστοιχα όλες οι υπόλοιπες θα αποτελούν οριακές παρατηρήσεις ή παρατηρήσεις θορύβου. Η

παραπάνω μεθοδολογία αποτελεί μια οπτική επιλογή της παραμέτρου Eps που μερικές φορές μπορεί να μην είναι και η καταλληλότερη όταν τα 2 φυσικά συμπλέγματα διαφέρουν σημαντικά ως προς την πυκνότητά τους. Προκειμένου επομένως να αξιολογήσουμε αυτή την επιλογή των παραμέτρων εφαρμόζουμε τον αλγόριθμο DBSCAN για $N_{min} = 7$ και $Eps = 3$, τα αποτελέσματα του οποίου μπορούμε να δούμε στο γράφημα που ακολουθεί (εικόνα 7.19). Παρατηρούμε ότι για τις δεδομένες επιλογές ο αλγόριθμος κατάφερε να εντοπίσει 1 σύμπλεγμα με την πλειοψηφία όλων των παρατηρήσεων, τοποθετώντας ένα μικρό πλήθος παρατηρήσεων ως θόρυβο. Το αποτέλεσμα αυτό προφανώς δεν μας ικανοποιεί καθώς δεν έχει πραγματοποιηθεί διαχωρισμός μεταξύ των 2 κλάσεων. Χρησιμοποιώντας την επιπλέον πληροφορία που έχουμε στη διάθεση μας σχετικά με την πραγματική κατάταξη των παρατηρήσεων μπορούμε να διαπιστώσουμε ότι για τις τιμές $N_{min} = 7$ και $Eps = 1.9$ επιτυγχάνετε το μεγαλύτερο ποσοστό σωστής ταξινόμησης.

Εικόνα 7.18: Γράφημα 7-NN κάθε παρατήρησης

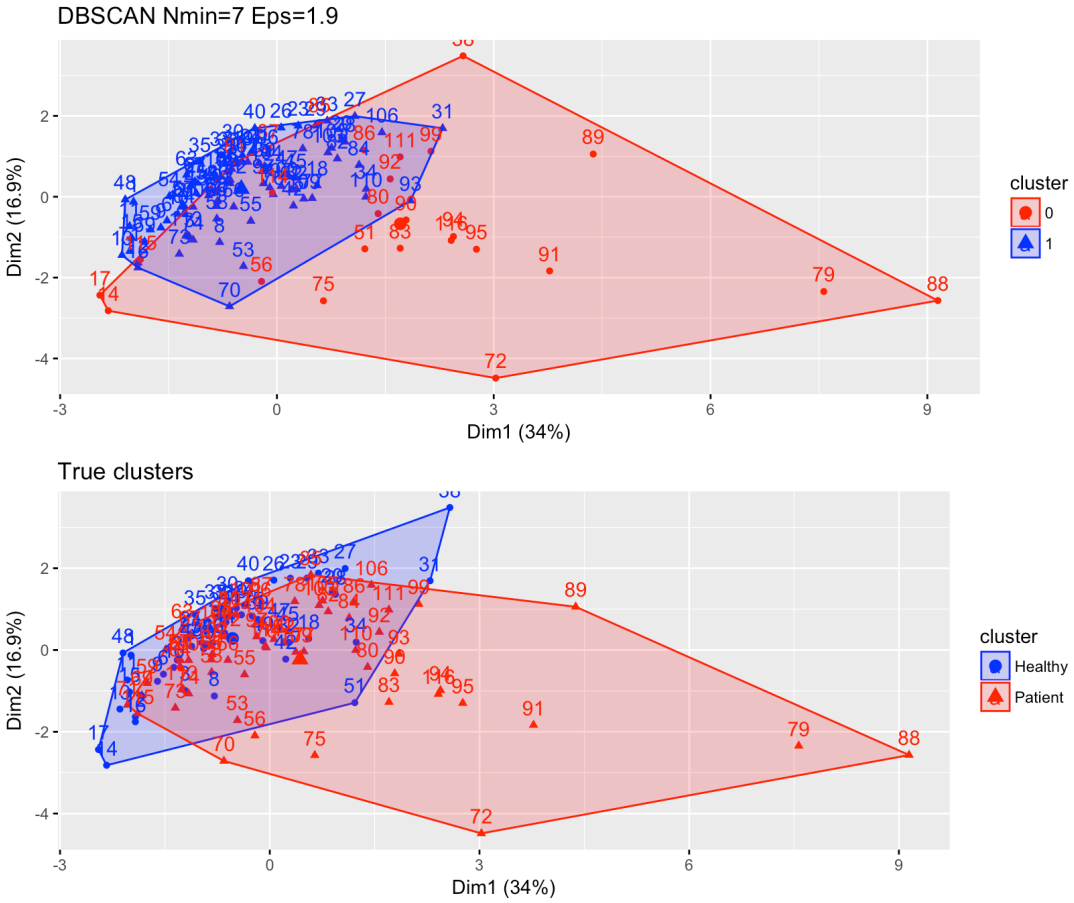


Εικόνα 7.19: Συμπλέγματα DBSCAN για Nmin=7 ,Eps=3

Συγκεκριμένα τα αποτελέσματα του αλγορίθμου για τις τιμές αυτές θα είναι ως εξής:

dbscan	Pts = 116	MinPts = 7	eps=1.9
	0	1	
border	26	17	
seed	0	73	
total	26	90	

Και σε αυτή την περίπτωση ο αλγόριθμος κατάφερε να εντοπίσει ένα σύμπλεγμα με συνολικά 90 παρατηρήσεις από τις οποίες οι 73 αποτελούν κεντρικές και οι 17 συνοριακές. Επιπλέον έχει αξιολογήσει τις υπόλοιπες 26 ως θόρυβο. Προβάλλοντας τα αποτελέσματα της ομαδοποίησης στις 2 πρώτες κύριες διαστάσεις καθώς και την πραγματική ομαδοποίηση των δεδομένων (Εικόνα 7.20) παρατηρούμε ότι η κλάση που έχει προσπαθήσει να εντοπίσει ο αλγόριθμος είναι η Healthy (μπλέ χρώμα) ενώ αντίστοιχα έχει θεωρήσει ως θόρυβο μέρος της κλάσης Patient. Ο βαθμός κατά τον οποίο έχει επιτευχθεί η διαφοροποίηση των 2 κλάσεων θα ελεγχεί στη συνέχεια με τη βοήθεια των μέτρων αξιολόγησης.



Εικόνα 7.20: Συμπλέγματα DBSCAN για Nmin=7 ,Eps=1.9 και φυσικές κλάσεις

Παρόλο που ο αλγόριθμος DBSCAN δεν βασίζεται στη χρήση αριθμητικών μέσων, στους παρακάτω πίνακες βρίσκονται οι μέσοι των συμπλεγμάτων που προέκυψαν από την εφαρμογή του αλγορίθμου καθώς και οι πραγματικές τιμές των μέσων:

Πίνακας 7.7: Προβλεπόμενοι αριθμητικοί μέσοι με τη μέθοδο DBSCAN

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Healthy	57.2	27.35	92.92	7.028	1.6305	23.527	9.209	12.792	484.68
Patient	57.65	28.40	114.7	20.343	6.3799	37.30	13.546	21.421	707.62

Πίνακας 7.8: Πραγματικές τιμές των αριθμητικών μέσων κάθε συμπλέγματος

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Healthy	58.08	28.32	88.23	6.934	1.552	26.64	10.33	11.62	499.7
Patient	56.67	26.98	105.6	12.51	3.623	26.60	10.06	17.25	563.0

Παρατηρούμε ότι ο αλγόριθμος K means πραγματοποίησε καλύτερη προσέγγιση των αντίστοιχων αριθμητικών μέσων, γεγονός αναμενόμενο δεδομένου της διαφορετικής φιλοσοφίας του αλγορίθμου DBSCAN. Επιπλέον στο παρακάτω πίνακα βλέπουμε τη δομή των παραγόμενων συμπλεγμάτων μέσω του αλγορίθμου DBSCAN καθώς και τις αντίστοιχες τιμές $f_{00}, f_{11}, f_{10}, f_{01}$

Πίνακας 7.9: Πλήθος ορθών και λανθασμένων ταξινομήσεων με τη μέθοδο DBSCAN

		Actual Values		
Predicted values		Healthy	Patient	
	Healthy	$f_{11} = 48$	$f_{10} = 42$	90
	Patient	$f_{01} = 4$	$f_{00} = 22$	26
		52	64	116

Ο αλγόριθμος κατάφερε να εντοπίσει 48 από τις 52 παρατηρήσεις της κλάσης Healthy και 22 από τις 64 της κλάσης Patient τις οποίες εντόπισε ως θόρυβο. Επιπλέον όσο περισσότερο αυξάνεται η τιμή της παραμέτρου Eps τόσο περισσότερες παρατηρήσεις εντάσσονται στη κλάση Healthy. Το παραπάνω οφείλεται στο γεγονός ότι ο αλγόριθμος DBSCAN μπορεί να έχει προβλήματα με την πυκνότητα εάν η πυκνότητα των συμπλεγμάτων διαφέρει αρκετά. Στη προκειμένη περίπτωση επομένως θα μπορούσαμε να πούμε ότι η κλάση Healthy είναι πιο πυκνή σε σχέση με τη κλάση Patient. Επιπλέον οι κλάσεις δεν είναι απολύτως διαχωρίσιμες καθώς οι παρατηρήσεις της κλάσης Patient είναι πυκνά προσβάσιμες από κεντρικές παρατηρήσεις της κλάσης Healthy καθώς αυξάνουμε την τιμή της παραμέτρου Eps.

7.6 Σύγκριση Αποτελεσμάτων

Η αξιολόγηση των παραπάνω αποτελεσμάτων θα πραγματοποιηθεί μέσω του υπολογισμού των εξωτερικών κριτηρίων αξιολόγησης κάνοντας χρήση της πραγματικής ομαδοποίησης των παρατηρήσεων και των τιμών $f_{00}, f_{11}, f_{10}, f_{01}$ που προέκυψαν από τους αλγορίθμους K means και DBSCAN. Στη σύγκριση δεν θα συμπεριλάβουμε τα αποτελέσματα του single link αλγορίθμου καθώς δεν κατάφερε να διαφοροποιήσει ικανοποιητικά τις 2 κλάσεις. Αρχικά θα εστιάσουμε την προσοχή μας στους δείκτες οι οποίοι είναι προσανατολισμένοι στην ταξινόμηση:

Δείκτες προσανατολισμένοι στην ταξινόμηση

Πίνακας 7.10: Δείκτες Entropy και Purity για K - means και DBSCAN

	K means		DBSCAN	
	Entropy	Purity	Entropy	Purity
Healthy	0.9999	0.5070	0.9968	0.5333
Patient	0.9332	0.6510	0.6194	0.8462
Συνολικά	0.9751	0.5604	0.9122	0.6034

Η εντροπία αποτελεί μία ένδειξη του βαθμού στον οποίο κάθε σύμπλεγμα αποτελείται από παρατηρήσεις μιας κλάσης και παίρνει τιμές στο διάστημα [0,1] , με τιμές κοντά στο 0 να είναι πιο επιθυμητές. Με βάση τον παραπάνω πίνακα παρατηρούμε ότι ο αλγόριθμος DBSCAN είναι πιο αποτελεσματικός όσον αφορά το συγκεκριμένο δείκτη, ιδιαίτερα όσον αφορά το σύμπλεγμα “Patient” όπου η αντίστοιχη τιμή του είναι αρκετά πιο μικρή από αυτή που προέκυψε από τον αλγόριθμο K means (Πίνακας 7.10).

Παρόμοια συμπεράσματα έχουμε και μέσου του δείκτη Purity, ο οποίος είναι ένας άλλος δείκτης μέτρησης του βαθμού κατά τον οποίο ένα σύμπλεγμα περιέχει παρατηρήσεις μόνο από μία κλάση παίρνοντας τιμές στο διάστημα [0,1] , με τιμές κοντά στο 1 να είναι πιο επιθυμητές. Και σε αυτή τη περίπτωση παρατηρούμε μεγάλη διαφοροποίηση όσον αφορά το δεύτερο σύμπλεγμα όπου η πλειοψηφία των παρατηρήσεων ανήκουν στη κλάση “Patient", έχοντας ως αποτέλεσμα μεγαλύτερες τιμές του δείκτη Purity.

Πίνακας 7.11: Δείκτης Precision για K - means και DBSCAN

	K means		**DBSCAN**	
	Precision		Precision	
	Healthy	Patient	Healthy	Patient
Healthy	0.507	0.493	0.533	0.467
Patient	0.349	0.651	0.154	0.846

Στο παραπάνω πίνακα παρατηρούμε το δείκτη Precision, ο οποίος αποτελεί ένδειξη του μεγέθους του τμήματος ενός συμπλέγματος που αποτελείται από αντικείμενα συγκεκριμένης κλάσης. Παρατηρούμε ότι ο αλγόριθμος DBSCAN είναι ελάχιστα πιο ακριβής όσον αφορά το πρώτο σύμπλεγμα, όπου το ποσοστό των παρατηρήσεων που ανήκουν στη κλάση "Healthy" έχει αυξηθεί ελάχιστα σε 53.3% έναντι 50.7% του αλγορίθμου K means. Αντίστοιχα το δεύτερο σύμπλεγμα μέσω του αλγορίθμου DBSCAN αποτελείται κατά 84.6% από παρατηρήσεις της κλάσης "Patient" , ποσοστό αισθητά καλύτερο από το 65.1% που προέκυψε μέσω του αλγορίθμου K means. Τιμές που προσεγγίζουν το μοναδιαίο πίνακα αποτελούν ιδανικές τιμές του παραπάνω δείκτη γεγονός που αντιστοιχεί σε υψηλή ακρίβεια και των 2 παραγόμενων συμπλεγμάτων.

Πίνακας 7.12: Δείκτης Recall για K - means και DBSCAN

	K means		**DBSCAN**	
	Recall		Recall	
	Healthy	Patient	Healthy	Patient
Healthy	0.7115	0.5625	0.923	0.656
Patient	0.2885	0.4375	0.077	0.344

Ο δείκτης Recall μετράει το βαθμό στον οποίο ένα σύμπλεγμα περιέχει όλες τις παρατηρήσεις συγκεκριμένης κλάσης. Παρατηρούμε ότι παρ όλο που η ακρίβεια του δεύτερου συμπλέγματος που προέκυψε μέσω του αλγορίθου DBSCAN είναι αρκετά μεγάλη ως προς τη κλάση "Patient", περιέχει μόνο το 34.4% των παρατηρήσεων της κλάσης αυτής. Επιπλέον το πρώτο σύμπλεγμα μέσω DBSCAN περιέχει σχεδόν όλες τις παρατηρήσεις της κλάσης "Healthy" (92.3%) και 65.6% της κλάσης "Patient". Και σε αυτή τη περίπτωση επιθυμούμε οι τιμές του δείκτη Recall να προσεγγίζουν το μοναδιαίο πίνακα γεγονός που δεν επιτυγχάνεται για κανένα απο τους 2 αλγορίθμους. Η τιμή του δείκτη Recall του αλγορίθμου DBSCAN παρουσιάζει καλύτερη συμπεριφορά για τις τιμές Recall(1,1)= 0.923 και Recall(2,1)=0.077 αλλά χειρότερη ως προς τι υπόλοιπες. Τα ποσοτά αυτά οφείλονται στο γεγονός ότι ο αλγόριθμος DBSCAN τοποθετεί 90 παρατηρήσεις στο πρώτο σύμπλεγμα σε αντίθεση με τον

αλγόριθμο K means ο οποίος τοποθετεί 73 και στο ότι η ακρίβεια του δεύτερου συμπλέγματος είναι αρκετά πιο μεγάλη από αυτή του αλγορίθμου K means.

Πίνακας 7.13: Δείκτης F για K - means και DBSCAN

	K means		DBSCAN	
	Δείκτης F		Δείκτης F	
	Healthy	Patient	Healthy	Patient
Healthy	0.5921	0.5255	0.6758	0.5454
Patient	0.3159	0.5233	0.1026	0.4891

Στο παραπάνω πίνακα παρατηρούμε το δείκτη F, ο οποίος είναι ένας συνδυασμός τόσο του precision όσο και του δείκτη recall και μετρά τον βαθμό στον οποίο ένα σύμπλεγμα περιέχει μόνο παρατηρήσεις συγκεκριμένης κλάσης και όλες τις παρατηρήσεις αυτής της κλάσης. Τα αποτελέσματα του δείκτη F επιβεβαιώνουν την υπεροχή του αλγορίθμου DBSCAN , κάτι το οποίο ήταν αναμενόμενο δεδομένου ότι αποτελεί συνδιασμός των δεικτών Precision και Recall.

Δείκτες προσανατολισμένοι στην ομοιότητα

Οι δείκτες που ακολουθούν μετράνε την ομοιότητα των αποτελεσμάτων των 2 αλγορίθμων με την πραγματική ομαδοποίηση των παρατηρήσεων.

Πίνακας 7.14: Δείκτες Rand, Jacard και FM για K - means και DBSCAN

	Rand	Jacard	Fowlkes-Mallows
K means	0.5603	0.4205	0.6005
DBSCAN	0.6034	0.5106	0.7017

Με βάση το δείκτη Rand παρατηρούμε ότι το ποσοτό των ορθών αποφάσεων μέσω του αλγορίθμου DBSCAN είναι μεγαλύτερο (60.34 % έναντι 56.03%) με αποτέλεσμα η διαμέριση μέσω αυτού του αλγορίθμου να βρίσκεται πιο κοντά στη πραγματική ομαδοποίηση των παρατηρήσεων. Παρόμοια συμπεράσματα έχουμε και με τους δείκτες Jacard και Fowlkes-Mallows με τις αντίστοιχες τιμές τους για τον αλγόριθμο DBSCAN να είναι μεγαλύτερες και πιο κοντά στη μονάδα.

Συνεπώς με βάση τους παραπάνω δείκτες αξιολόγησης μπορούμε να αποφανθούμε πλέον ότι ο αλγόριθμος DBSCAN είναι αυτός με τα καλύτερα αποτελέσματα. Αντιθέτως, παρόλο που η ακρίβεια της μεθόδου για την κλάση Patient είναι αρκετά μεγάλη (0.846), τα υπόλοιπα αποτελέσματα δεν είναι ιδιαίτερα ικανοποιητικά για την πρόβλεψη του καρκίνου του μαστού.

Κεφάλαιο 8

Συμπεράσματα

Στην παρούσα διπλωματική εργασία διερευνήθηκαν 3 αλγόριθμοι ομαδοποίησης μηχανικής μάθησης (DBSCAN, K means, Single Link), συγκεκριμένα η απόδοσή τους σε ένα σύνολο ιατρικών δεδομένων όπως επίσης και το αν μπορεί να χρησιμοποιηθεί κάποιος από αυτούς για την πρόβλεψη του καρκίνου του μαστού. Η βιβλιογραφική μελέτη που διεξήχθη στη παρούσα εργασία στόχευε στην απόκτηση περαιτέρω γνώσεων σχετικά με τους αλγορίθμους όπως επίσης πλεονεκτημάτων και μειονεκτήματων του καθενός έναντι των υπολοίπων. Κατά τη διάρκεια της μελέτης μας χρησιμοποιήσαμε ετικετοποιημένα δεδομένα, προκειμένου να είμαστε σε θέση να μπορούμε να κρίνουμε για το αν η ομαδοποίηση είναι ακριβής με σκοπό τη σύγκριση των αποτελεσμάτων των 3 μεθόδων.

Τα αποτελέσματα από την εφαρμογή των 3 μεθόδων (DBSCAN, K means, Single Link) ήταν μεικτά. Ο αλγόριθμος K means πέτυχε βέλτιστα αποτελέσματα στα δυσδιάστατα δεδομένα, όπου η επιλογή του αριθμού k που προέκυψε από τις μεθόδους (Elbow, Gap Statistic και Silhouette) ταυτιζόταν με την πραγματικότητα. Παρόμοια αποτελέματα προέκυψαν και από τον αλγόριθμο Single Link, ενώ ο αλγόριθμος DBSCAN δεν κατάφερε να εντοπίσει πλήρως τα 3 φυσικά συμπλέγματα τοποθετώντας μερικές παρατηρήσεις ως θόρυβο. Αντιθέτως, τα αποτελέσματα αντιστρέφονται στα πολυδιάστατα δεδομένα όπου οι αλγόριθμοι Single Link και K means αποδίδουν χειρότερα σε σχέση με τον αλγόριθμο DBSCAN, αδυνατώντας να παράξουν ικανοποιητική ομαδοποίηση των δεδομένων. Σε αυτή την περίπτωση κάνοντας χρήση της πραγματικής ομαδοποίησης τους, πραγματοποιήθηκε σύγκριση των αποτελεσμάτων με τη βοήθεια δεικτών αξιολόγησης. Αξίζει να σημειωθεί πως ιδιαίτερη δυσκολία παρουσιάστηκε επίσης και στην επιλογή των παραμέτρων των μοντέλων, γεγονός στο οποίο βοήθησε η εκ των προτέρων γνώση τη πραγματικής ομαδοποίησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Akume, D. & Weber, G.W. (2002), *Cluster Algorithms: Theory and Methods*

[2] Arbelaitz, O. & Gurrutxaga, I. & Muguerza, J. & Pe´rez, J. M. & Perona, I. (2012), *An extensive comparative study of cluster validity indices*

[3] Clustering Algorithms and Evaluations. Ανακτήθηκε από: https://pdfs.semanticscholar.org/d98c/801b10f60934ef2ff29534071ac6197e1b5e.pdf

[4] Cluster Analysis: Basic Concepts and Algorithms. Ανακτήθηκε από: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf

[5] Craenendonck, T. M. & Blockeel, H. (2015), *Using Internal Validity Measures to Compare Clustering Algorithms*

[6] Ester, M. & Kriegel, H. P. & Sander, J. & Xu X. (1996), *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*

[7] Gan, J. (2017), *High Performance Density-based Clustering on Massive Data (Διδακτορική Διατριβή),* Queensland

[8] Guojun, G. & Chaoqun, M. & Jianhong, W. (2008), *Data Clustering Theory, Algorithms, and Applications*

[9] Hastie, T. & Tibshirani, R. & Friedman, J. (2017), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*

[10] Hierarchical Clustering / Dendrograms. Ανακτήθηκε από: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf

[11] Juhàsz , S. & Legány, C. & Babos, A. (2006), *Cluster Validity Measurement Techniques*

[12] Lasek, P. (2011), *Efficient Density-Based Clustering* (Διδακτορική *Διατριβή*), Warsaw.

[13] Liu, Y. & Li, Z. & Xiong, H. & Gao, X. & Wu, J (2010), *Understanding of Internal Clustering Validation Measures*

[14] Nayeem, R. (2017). *Selection of K in K-Means Algorithm* (Προπτυχιακή Εργασία). Bangladesh University of Engineering and Technology, Dhaka.

[15] Pham, D. T. & Dimov, S. S. & Nguyen C. D. (2004), *Selection of K in K-means clustering*

[16] Stein, B. & Busch, M. (2005), *Density-based Cluster Algorithms in Low-dimensional and High-dimensional Applications*