

**ΜΠΣ «ΜΕΘΟΔΟΛΟΓΙΑ ΒΪΟΙΑΤΡΙΚΗΣ ΕΡΕΥΝΑΣ, ΒΙΟΣΤΑΤΙΣΤΙΚΗ
ΚΑΙ ΚΛΙΝΙΚΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»**



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ανάπτυξη λογισμικού σε γλώσσα προγραμματισμού python για ομαδοποίηση παρατηρήσεων με τις μεθόδους της απλής συνένωσης, της πλήρους συνένωσης, των κέντρων βάρους και k-means.»

Χατζηλιάδη Παναγιώτα Ευανθία

ΠΕΡΙΛΗΨΗ

Οι μέθοδοι ομαδοποίησης μπορούν να διαχωριστούν σε δύο διαφορετικές κατηγορίες ανάλογα με τον τρόπο με τον οποίο προχωρούν στη διαμόρφωση των ομάδων: στις ιεραρχικές και στις μη ιεραρχικές μεθόδους.

Σκοπός της παρούσας εργασίας είναι η ανάπτυξη ενός λογισμικού το οποίο θα εκτελεί ομαδοποίηση ενός συνόλου παρατηρήσεων με τη μέθοδο της απλής συνένωσης (Single Linkage Method), τη μέθοδο της πλήρους συνένωσης (Complete Linkage Method), τη μέθοδο των κέντρων βάρους (Centroid Method) και τέλος την k-means μέθοδο. Στο πλαίσιο ανάπτυξης αυτού του λογισμικού χρησιμοποιήθηκε η γλώσσα προγραμματισμού python .

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΕΙΣΑΓΩΓΗ	3
2. ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ	5
2.1 Η Μεθοδος της απλης συνενωσης (SIMPLE LINKAGE METHOD)	
2.2 Η Μεθοδος της πληρους συνενωσης (COMPLETE LINKAGE METHOD).	
2.3. Η Μεθοδος των κέντρων βαρους (CENTROID METHOD)	
3. ΜΗ ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ.....	7
3.1 Η Μέθοδος K-Means.....	
4. ΜΕΘΟΔΟΙ.....	9
5. ΑΝΑΦΟΡΕΣ.....	11

ΠΙΝΑΚΑΣ ΣΧΗΜΑΤΩΝ

Σχήμα 1: Έξοδος προγράμματος 1	9
Σχήμα 2: Έξοδος προγράμματος 2	10

1. ΕΙΣΑΓΩΓΗ

Οι μέθοδοι ομαδοποίησης (Cluster Analysis) είναι τεχνικές της πολυμεταβλητής στατιστικής οι οποίες αποσκοπούν στη δημιουργία ομογενών ομάδων έτσι ώστε τα στοιχεία (παρατηρήσεις) που βρίσκονται στην ίδια ομάδα να παρουσιάζουν παρόμοια συμπεριφορά από άποψη κατανομής ενώ τα στοιχεία διαφορετικών ομάδων να αντιστοιχούν σε ‘απομακρυσμένες’ κατανομές.

Οι μέθοδοι ομαδοποίησης χωρίζονται σε δύο διαφορετικές κατηγορίες: στις ιεραρχικές και τις μη ιεραρχικές μεθόδους. Στις ιεραρχικές μεθόδους ο αριθμός των ομάδων δεν είναι γνωστός εκ των προτέρων. Εν αντιθέσει, στις μη ιεραρχικές μεθόδους θεωρείται ότι ο αριθμός των ομάδων είναι γνωστός απο πρίν. Με έναν επαναληπτικό αλγόριθμο τοποθετούνται οι παρατηρήσεις στις ομάδες ανάλογα με το ποιά ομάδα είναι πιο κοντά στην εκάστοτε παρατήρηση. Η πιο γνωστή μη ιεραρχική μέθοδος ομαδοποίησης είναι η μέθοδος Mac Queen ή k- means method.

Οι ιεραρχικές μέθοδοι χωρίζονται σε **συσσωρευτικές μεθόδους (agglomerative methods)** οι οποίες ακολουθούν μία σειρά δαδοχικών συγχωνεύσεων η παρατηρήσεων σε ομάδες και σε **διαιρετικές μεθόδους (divisive methods)** οι οποίες χωρίζουν ένα σύνολο η παρατηρήσεων διαδοχικά σε μικρότερες ομάδες. Ο βασικός αλγόριθμος όλων των συσσωρευτικών μεθόδων είναι περίπου ο ίδιος. Όλες οι μέθοδοι χρησιμοποιούν κάποιο συντελεστή ομοιότητας ή μία απόσταση που υπολογίζεται για όλους τους συνδυασμούς ανά δύο των υπό εξέταση παρατηρήσεων και έτσι διαμορφώνεται ο πίνακας αποστάσεων. Ο αλγόριθμος επιδρά στον πίνακα αποστάσεων και δημιουργεί ένα δενδρόγραμμα το οποίο απεικονίζει τις διαδοχικές συγχωνεύσεις των παρατηρήσεων μέχρι το επίπεδο που σχηματίζεται μία μόνο

ομάδα. Οι συσσωρευτικές μέθοδοι διαφέρουν μεταξύ τους ως προς τον ορισμό της ομοιότητας των παρατηρήσεων κάθε ομάδας. Οι συνηθέστερες μέθοδοι είναι:

- Η μέθοδος της απλής συνένωσης (Simple Linkage Method).
- Η μέθοδος της πλήρους συνένωσης (Complete Linkage Method).
- Η μέθοδος των σταθμισμένων μέσων (Weighted Average Linkage Method).
- Η μέθοδος των κέντρων βάρους (Centroid Method).
- Η μέθοδος του Ward (Ward's Method).
- Η μέθοδος του Gower (Gower's method).
- Η μέθοδος της διαμέσου (Median Method).
- Η μέθοδος του μέσου όρου των ομάδων (Group Average Method).
- Η μέθοδος των Lance & Williams.

Στην παρούσα εργασία θα περιγραφούν οι μέθοδοι της απλής συνένωσης (**Single Linkage Method**), της πλήρους συνένωσης (**Complete Linkage Method**), η μέθοδος των κέντρων βάρους (**Centroid Method**) καθώς και η **k-means μέθοδος**..

2. ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ

2.1 Η ΜΕΘΟΔΟΣ ΤΗΣ ΑΠΛΗΣ ΣΥΝΕΝΩΣΗΣ (SIMPLE LINKAGE METHOD)

Η μέθοδος της απλής συνένωσης (Simple Linkage Method) είναι η παλαιότερη και απλούστερη όλων των ιεραρχικών μεθόδων ομαδοποίησης. Στη μέθοδο αυτή η απόσταση μεταξύ δύο ομάδων ορίζεται ως η μικρότερη απόσταση μεταξύ ενός στοιχείου της μιας ομάδας και ενός στοιχείου της άλλης ομάδας. Γι'αυτό η μέθοδος αυτή ονομάζεται και «**μέθοδος του κοντινότερου γείτονα**» .

Το πιο βασικό μειονέκτημα αυτής της μεθόδου ομαδοποίησης είναι ότι αντί να δημιουργεί καινούριες ομάδες, έχει την τάση να συνδέει απομονωμένα σημεία με ήδη υπάρχουσες ομάδες. Έτσι, δύο ομάδες που είναι εμφανώς διαφορετικές θα συγχωνευθούν εάν υπάρχει κάποιο σημείο ή ένα σύνολο σημείων που να τις συνδέει. Αυτό έχει σαν αποτέλεσμα οι ομάδες οι οποίες προκύπτουν με τη μέθοδο της απλής συνένωσης να είναι κακώς διαμορφωμένες , με δύο μέλη που ανήκουν στη ίδια ομάδα συνδεδεμένα με μία αλυσίδα ενδιάμεσων σημείων. Αυτό το φαινόμενο ονομάζεται φαινόμενο της αλυσίδας (chaining effect). Το σημαντικό πλεονέκτημα αυτής της μεθόδου είναι ότι δεν επηρεάζεται από ακραίες τιμές.

2.2. Η ΜΕΘΟΔΟΣ ΤΗΣ ΠΛΗΡΟΥΣ ΣΥΝΕΝΩΣΗΣ (COMPLETE LINKAGE METHOD).

Η μέθοδος της πλήρους συνένωσης (Complete Linkage Method) σε αντίθεση με τη μέθοδο της απλής συνένωσης, χρησιμοποιεί ως απόσταση μεταξύ των ομάδων την απόσταση των πιο απομακρυσμένων ζευγών σημείων. Το ένα από αυτά τα σημεία ανήκει στην μία ομάδα και το άλλο στην άλλη. Εναλλακτική ονομασία αυτής της μεθόδου είναι « **μέθοδος του μακρινότερου γείτονα**».

Οι ομάδες που δημιουργούνται με τη μέθοδο αυτή είναι συνήθως μεγάλες και συμπαγείς. Υπάρχει όμως ο κίνδυνος ομάδες που φαίνονται όμοιες να μη μπορούν να συγχωνευτούν όταν υπάρχει κάποιο ζεύγος σημείων που απέχουν αρκετά μεταξύ τους. Επίσης, σε αντίθεση με τη μέθοδο της απλής συνένωσης, η συγκεκριμένη μέθοδος επηρεάζεται από την ύπαρξη ακραίων τιμών.

2.3. Η ΜΕΘΟΔΟΣ ΤΩΝ ΚΕΝΤΡΩΝ ΒΑΡΟΥΣ (CENTROID METHOD)

Η μέθοδος των **Κέντρων Βάρους (Centroid Method)** λαμβάνει ως κριτήριο συνένωσης, την ελάχιστη απόσταση μεταξύ των κέντρων βάρους των ομάδων. Σημαντικό μειονέκτημα της συγκεκριμένης μεθόδου είναι ότι μπορεί να εφαρμοστεί μόνο σε ποσοτικά δεδομένα. Η μέθοδος των κέντρων βάρους συχνά παράγει συμπαγείς και ελλειπτικές ομάδες.

3. ΜΗ ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ

3.1 K MEANS METHOD

Η K MEANS μέθοδος χρησιμοποιεί την έννοια του κέντρου της ομάδας και εν συνεχεία κατατάσσει τα στοιχεία ανάλογα με την απόστασή τους από τα κέντρα όλων των ομάδων. Το κέντρο της κάθε ομάδας είναι η μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας. Ο αλγόριθμος αυτός έχει καλύτερη απόδοση για μεγάλα σύνολα δεδομένων καθώς είναι πολύ πιο γρήγορος σχετικά με την ιεραρχική ομαδοποίηση.

Ιδιαίτερο χαρακτηριστικό της k-means μεθόδου ομαδοποίησης είναι ότι εντός της κάθε ομάδας τα στοιχεία έχουν όσο το δυνατόν μικρότερη απόσταση από το κέντρο βάρους της ομάδας, ενώ μεταξύ των ομάδων τα στοιχεία της μιας ομάδας απέχουν όσο το δυνατόν περισσότερο από το κέντρο βάρους της άλλης ομάδας.

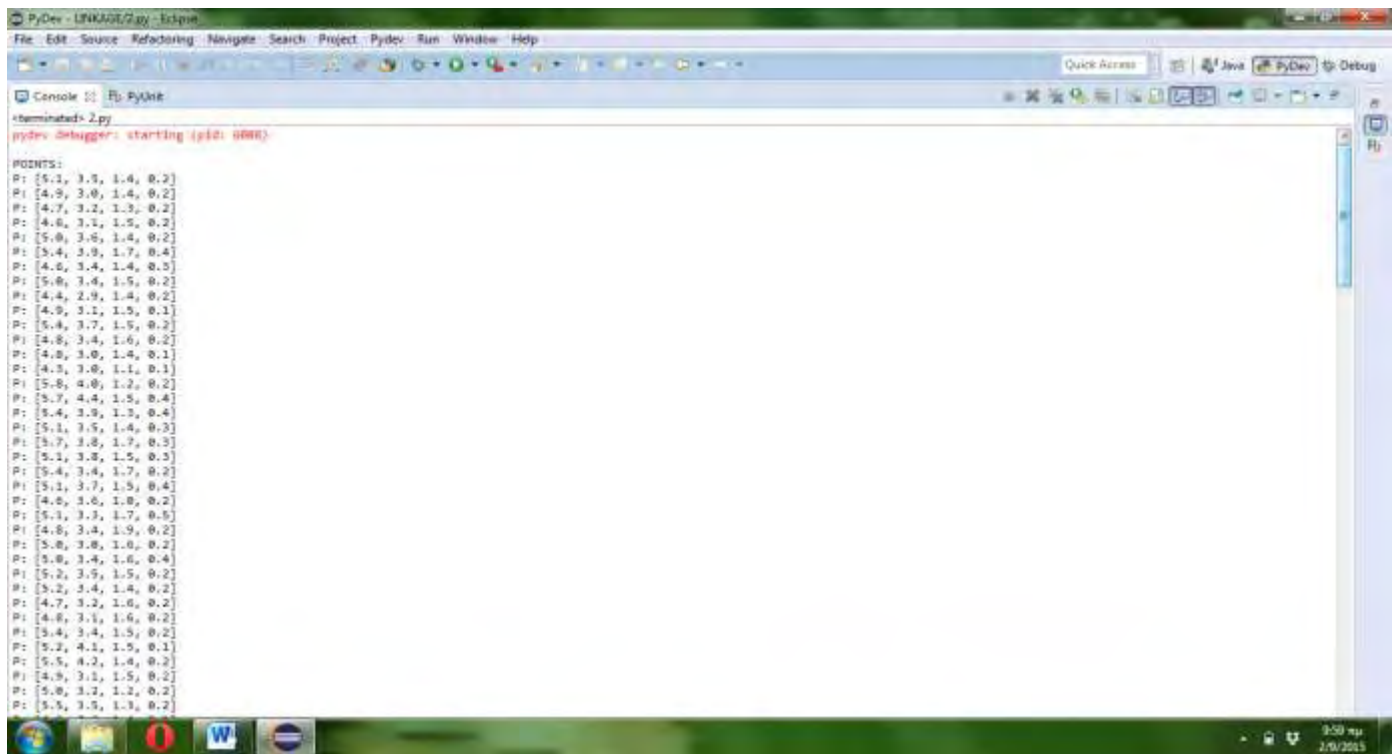
Ο αλγόριθμος από τις πρώτες επαναλήψεις πλησιάζει πολύ στην τελική λύση του ενώ στις επόμενες επαναλήψεις ο,τι διαφορετικό προκύπτει οφείλεται σε μετακινήσεις ενός μικρού αριθμού παρατηρήσεων που κατά πάσα πιθανότητα βρίσκονται στα σύνορα κάποιων ομάδων. Αυτό έχει σαν αποτέλεσμα να μην απαιτούνται πολλές επαναλήψεις ενώ οι τελικές ομάδες που δημιουργούνται από τον αλγόριθμο έχουν συνήθως τον ίδιο αριθμό παρατηρήσεων.

Το μεγαλύτερο μειονέκτημα του αλγορίθμου εμφανίζεται εάν δε γίνει σωστή επιλογή των αρχικών κεντρών. Σε αυτή την περίπτωση οι ομάδες που θα προκύψουν θα διαφέρουν σημαντικά από τη φυσική ομαδοποίηση που θα υπάρχει στα δεδομένα. Ένας τρόπος για να αντιμετωπιστεί αυτό το πρόβλημα είναι να τρέχουμε τη μέθοδο με διαφορετικές επιλογές έτσι ώστε να είμαστε σίγουροι ότι δε θα παγιδεύεται ο αλγόριθμος σε κάποια λύση που δε θα είναι η βέλτιστη.

Επίσης ,σημαντικό μειονέκτημα του αλγορίθμου είναι ότι στην περίπτωση που υπάρχουν ακραίες παρατηρήσεις υπάρχει πιθανότητα να δημιουργηθούν ομάδες με πολύ διασπαρμένα στοιχεία. Αυτό σημαίνει ότι η απόσταση των στοιχείων κάθε ομάδας από τι κέντρο βάρους της θα είναι μεγάλη και αυτό είναι μία ένδειξη ότι η ομαδοποίηση δεν είναι ιδανική. Ακόμη, εάν είναι γνωστό ότι ο πληθυσμός μας αποτελείται απο k ομάδες και εν τέλει το δείγμα μας δεν αντιπροσωπεύεται από κάποια απο αυτές (την πιο σπάνια συνήθως),τότε με το διαχωρισμό σε k ομάδες θα προκύψουν παραπλανητικές ομαδοποιήσεις. Μία λύση αυτού του προβλήματος θα αποτελούσε η εφαρμογή του αλγορίθμου για διάφορες τιμές του k και η σύγκριση εν συνεχεία των αποτελεσμάτων με στόχο την καλύτερη δυνατή ομαδοποίηση.

4. ΜΕΘΟΔΟΙ

Η ανάπτυξη του λογισμικού έγινε με τη γλώσσα προγραμματισμού python. Το πρόγραμμα δέχεται σαν είσοδο ένα αρχείο txt το οποίο θα πρέπει να περιλαμβάνει τα σημεία που θέλουμε να ομαδοποιήσουμε. Τα στοιχεία είναι πραγματικοί αριθμοί και μπορούν να έχουν όσες διαστάσεις επιθυμούμε. Εν συνεχεία επιλέγουμε τη μέθοδο ιεραρχικής ομαδοποίησης με την οποία θέλουμε να κάνουμε την ομαδοποίηση. Το πρόγραμμα δημιουργεί και εκτυπώνει τα clusters. Παράλληλα, εκτελεί ομαδοποίηση και με τη μέθοδο k means.

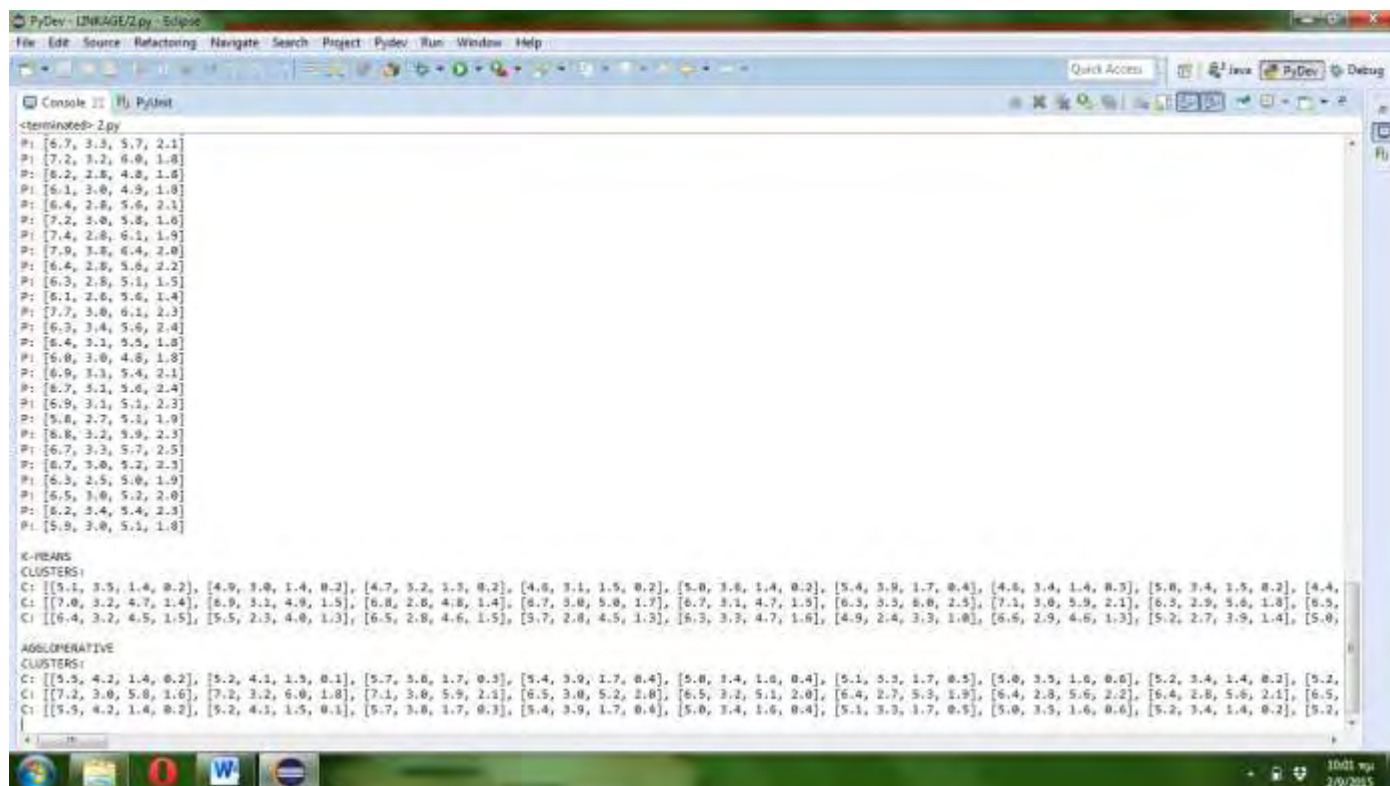


```
PyDev - [F:\KAGGLE\py - Ichneumon]
File Edit Source Refactoring Navigate Search Project PyDev Run Window Help

Console: F:\PyUnit
<terminated> 2.py
pydev debugger: starting (pid: 8888)

POINTS:
P1: [5.1, 3.5, 1.4, 0.2]
P2: [4.9, 3.0, 1.4, 0.2]
P3: [4.7, 3.2, 1.3, 0.2]
P4: [4.6, 3.1, 1.5, 0.2]
P5: [5.0, 3.6, 1.4, 0.2]
P6: [5.4, 3.9, 1.7, 0.4]
P7: [4.0, 3.4, 1.4, 0.3]
P8: [5.0, 3.4, 1.5, 0.2]
P9: [4.4, 2.9, 1.4, 0.2]
P10: [4.9, 3.1, 1.5, 0.1]
P11: [5.4, 3.7, 1.5, 0.2]
P12: [4.8, 3.4, 1.6, 0.2]
P13: [4.0, 3.0, 1.4, 0.1]
P14: [4.3, 3.0, 1.1, 0.1]
P15: [5.8, 4.0, 1.2, 0.2]
P16: [5.7, 4.4, 1.5, 0.4]
P17: [5.4, 3.9, 1.3, 0.4]
P18: [5.1, 3.5, 1.4, 0.3]
P19: [3.7, 3.8, 1.7, 0.3]
P20: [5.1, 3.8, 1.5, 0.3]
P21: [5.4, 3.4, 1.7, 0.2]
P22: [3.1, 3.7, 1.5, 0.4]
P23: [4.0, 3.6, 1.0, 0.2]
P24: [5.1, 3.3, 1.7, 0.5]
P25: [4.8, 3.4, 1.9, 0.2]
P26: [5.0, 3.8, 1.0, 0.2]
P27: [5.0, 3.4, 1.6, 0.4]
P28: [5.2, 3.5, 1.5, 0.2]
P29: [5.2, 3.4, 1.4, 0.2]
P30: [4.7, 3.2, 1.6, 0.2]
P31: [4.6, 3.1, 1.6, 0.2]
P32: [3.4, 3.4, 1.5, 0.2]
P33: [5.2, 4.1, 1.5, 0.1]
P34: [5.5, 4.2, 1.4, 0.2]
P35: [4.9, 3.1, 1.5, 0.2]
P36: [5.0, 3.2, 1.2, 0.2]
P37: [5.5, 3.5, 1.3, 0.2]
```

Σχήμα 1, Έξοδος προγράμματος 1: Τα στοιχεία στα οποία θα γίνει ομαδοποίηση.



Σχήμα 2, Έξοδος προγράμματος 2: Τα αποτελέσματα από την ομαδοποίηση με τη μέθοδο k means και με τη μέθοδο απλής συνένωσης.

ΑΝΑΦΟΡΕΣ

1.) Μέθοδοι Εύρεσης Βέλτιστου Πλήθους Ομάδων για πολυδιάστατα δεδομένα (Διπλωματική Εργασία, Φανή Ζαφειροπούλου, Τμήμα Στατιστικής, Πανεπιστήμιο Πειραιώς).

2) Clustering Algorithms (Instituto Superior Tecnico, Universidade Tecnica de Lisboa).

.