Τ.Ε.Ι. ΠΑΤΡΑΣ–ΠΑΡΑΡΤΗΜΑ ΑΜΑΛΙΑΔΑΣ

ΤΜΗΜΑ ΕΦΑΡΜΟΓΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

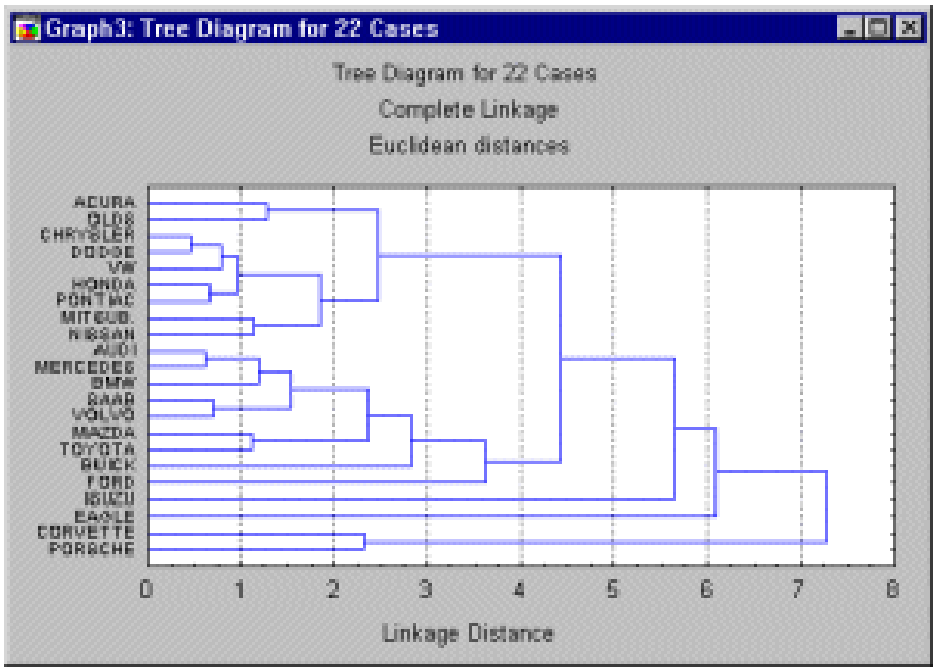
ΟΜΑΔΟΠΟΙΗΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ

CLUSTER ANALYSIS

ΣΠΟΥΔΑΣΤΕΣ:ΧΑΛΒΑΤΖΗΣ ΔΗΜΗΤΡΗΣ

ΓΙΑΝΝΟΠΟΥΛΟΥ ΑΝΑΣΤΑΣΙΑ

ΚΑΘΗΓΗΤΡΙΑ: ΝΙΚΟΛΟΠΟΥΛΟΥ ΕΙΡΗΝΗ



ΑΜΑΛΙΑΔΑ 2011

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΧΑΡΙΣΤΗΡΙΟ

Με την ολοκλήρωση της πτυχιακής μας εργασίας νοιώθουμε την ανάγκη να εκφράσουμε ένα μεγάλο ευχαριστώ σε όλους που από την θέση των καθηγητών μας κατά την διάρκεια της φοίτησης μας στη σχολή Εφαρμογών Πληροφορικής στη Διοίκηση και την Οικονομία μας εφοδίασαν με τις απαραίτητες γνώσεις Θα θέλαμε να ευχαριστήσουμε όλους όσους στήριξαν την προσπάθεια μας να ολοκληρώσουμε την πτυχιακή μας εργασία, με θέμα Cluster Analysis,.Ευχαριστούμε ιδιαίτερα την κα Ειρήνη Νικολοπούλου για την ανάθεση της πτυχιακής εργασίας, την βοήθεια και την μεταξύ μας συνεργασία.

ΠΕΡΙΛΗΨΗ

Η εργασία με τίτλο Cluster Analysis έχει σαν στόχο να αναδείξει το βαθμό που μπορούν οι μέθοδοι της cluster να αξιοποιήσουν τα οποιαδήποτε δεδομένα, είτε αριθμητικά είτε σε οποιοδήποτε περιβάλλον δεδομένα, για να εξάγουν χρήσιμα συμπεράσματα. Στο περιεχόμενο της εργασίας παρουσιάζεται η ομαδοποίηση δεδομένων και μέθοδοι ομαδοποίησης, καθώς και η σημασία και η έννοια των cluster καθώς και η μεθοδολογία της ανάλυσης αυτής. Στο τέλος θα εξαχθούν αποτελέσματα και συμπεράσματα από την μελέτη που έγινε από την τράπεζα Πειραιώς που εξετάζεται ο κίνδυνος χώρας για ένα σύνολο από χώρες στις οποίες οι ελληνικές τράπεζες ίσως ενδιαφέρονταν να εγκρίνουν ή έχουν εγκρίνει δάνεια, σύμφωνα με τη cluster και αποτελέσματα από μια δεύτερη μελέτη για την κατάσταση δυναμικού των ΤΕΕ σε κάθε νομό της Ελλάδας για το έτος 2004-2005.

ABSTRACT

The project of Cluster Analysis has an objective, to promote in a point with which method from cluster to use any data either arithmetical either any other environment in order to conclude with useful information. In the content of the project occurs the grouping of data and the grouping method and the importance and the mean of cluster and the methodology analysis. At the end will show results and useful conclusions from the research that Piraeus Bank made which exams the danger of the country from other countries that maybe Greek banks may interests to approve or they have already approve loans, according to cluster and some results from a second research for the state resources of TEE for every county of Greece for the year 2004-2005.

ΕΙΣΑΓΩΓΗ

Συστάδα θεωρούμε μια ομάδα, δηλαδή ένα σύνολο από στοιχειά τα οποία είναι όμοια μεταξύ τους και έχουν διαφορές από αλλά στοιχειά που ανήκουν σε άλλες συστάδες.

Η ανάλυση κατά συστάδες έχει ως σκοπό τον διαχωρισμό από ένα σύνολο στοιχείων σε υποσύνολα, έτσι ώστε να υπάρχει ομοιογένεια μέσα σε ένα υποσύνολο και ανομοιογένεια μεταξύ των στοιχείων που βρίσκονται σε διαφορετικά υποσύνολα.

Σημαντική έννοια στην ανάλυση αυτή είναι η απόσταση, που μπορεί να χρησιμοποιηθεί για να κατασκευαστεί πινάκας αποστάσεων. Ο πινάκας αυτός θα έχει μηδενικά στοιχειά στη διαγώνιο και την απόσταση μεταξύ δυο συστάδων, για παράδειγμα i,j , στη θέση (i,j). Η απόσταση υπολογίζεται με την Ευκλείδεια απόσταση όταν δεν προσδιοριστεί κάποια συγκεκριμένη μετρική.

ΚΕΦΑΛΑΙΟ 1 : ΟΜΑΔΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

1.1Εισαγωγή

Με την ομαδοποίηση δημιουργούμαι ομάδες από δεδομένα ενός ευρύτερου συνόλου που έχουν κοινά γνωρίσματα και κοινούς σκοπούς.

Στα πλαίσια αυτού του κεφαλαίου επιχειρούμε μια αναφορά στο σύνολο των μεθόδων ομαδοποίησης δεδομένων που καταγράφονται στη βιβλιογραφία εξετάζοντας τα βασικά βήματα μιας διαδικασίας, δηλαδή και το πώς ορίζονται οι σχετικές παράμετροι.

1.2Μέθοδοι μελέτης της ομοιότητας ή της διαφοράς συνόλων

Στην προσπάθεια μας να μελετήσουν την ομοιότητα ή διαφορά ενός σύνολο δεδομένων είναι απαραίτητο να υιοθετήσουμε τη χρήση μετρικών ομοιότητας. Σε αυτή την κατεύθυνση τα δεδομένα αναπαρίστανται ως σημεία ενός πολυδιάστατου χώρου, στον οποίο οι αποστάσεις των σημείων εκφράζουν και τη μεταξύ τους ομοιότητα. Μάλιστα το πλήθος των διαστάσεων καθορίζεται από τον αριθμό των μεταβλητών που περιγράφουν τα δεδομένα. Με στόχο όμως ένα μετρό ομοιότητας να αποτελεί μια καλή μετρική τα ακόλουθα κριτήρια θα πρέπει να ισχύουν:

1. Συμμετρία

Για δυο οντότητες *x* και y , η απόσταση τους να ικανοποιεί τη σχέση: .

$d(x,y)=d(y,x)\geq 0$

2. Ανισότητα τριγώνου

Για τρεις οντότητες *x*, y και z οι μεταξύ τους αποστάσεις πρέπει να ικανοποιούν τη σχέση: $d(x,y)\leq d(z,x)+d(z,y)$. Δηλαδή κάθε πλευρά του τριγώνου που δημιουργείται είναι μικρότερη ή ίση του αθροίσματος των άλλων 2 πλευρών (η γνωστή τριγωνική ανισότητα).

3. Διαφοροποίηση των ανόμοιων σημείων

Για δυο οντότητες x και y, εάν $d(x, y) \neq 0$ τότε $x \neq y$ δηλαδή η ύπαρξη απόστασης καθορίζει δυο διαφορετικά σημεία στο χώρο.

4. Ταυτοποίηση των όμοιων σημείων

Για δυο οντότητες που είναι ίδιες x και x΄, $d(x, x') = 0$, πρόκειται δηλαδή για σημεία που συμπίπτουν στο χώρο.

1.3 Μετρικές ομοιότητας

Έτσι λοιπόν με βάσει τα παραπάνω, το είδος των δεδομένων που μελετάμε, καθώς και το πρόβλημα κατηγοριοποίησης που λύνουμε, πρέπει να επιλέγουμε την κατάλληλη μετρική ομοιότητα. Για αυτό στο σημείο που βρισκόμαστε παρουσιάζουμε ορισμένες από τις πιο συχνά χρησιμοποιούμενες μετρικές ομοιότητες.

5. Συνιστώσες Συσχέτισης

Οι συνιστώσες συσχέτισης, γνωστές και ως γωνιώδεις συνιστώσες, λόγω της γεωμετρικής αναπαράστασής τους, αποτελούν την πιο συχνή μετρική ομοιότητα που χρησιμοποιείτε στα προβλήματα κατηγοριοποίησης των κοινωνικών επιστημών. Η πιο δημοφιλής είναι αυτή του γινομένου συσχέτισης (product-moment correlation coefficient) που παρουσιάστηκε από τον Karl Pearson η οποία παρουσιάζεται στον ακόλουθο τύπο,

$$r_{jk} = \frac{\sum (x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})}{\sqrt{\sum (x_{ij} - \bar{x_j})^2 \ (x_{ik} - \bar{x_k})^2}}$$

Όπου x_{ij} η τιμή της μεταβλητής i για τη j οντότητα, $\bar{x_j}$ η μέση τιμή όλων των μεταβλητών για τη j οντότητα και το πλήθος των μεταβλητών N. Εδώ πρέπει να επισημάνουμε ότι η παραπάνω μετρική χρησιμοποιείται κυρίως για διακριτές μεταβλητές και στην περίπτωση δυαδικών δεδομένων μεταφράζεται στη γνωστή μετρική του συνημίτονου (Phi coefficient). Οι τιμές αυτής της μετρικής ανήκουν στο

διάστημα [-1,+1], ενώ η τιμή 0 ισοδυναμεί σε μη συσχέτιση των οντοτήτων. Οι συνιστώσες συσχετίσεις θεωρούνται σχηματικές μετρικές, αφού στοχεύουν στην αναπαράσταση κάθε οντότητας ως διανύσματος στο χώρο διαστάσεων και τη μεταξύ τους σύγκριση. Σε αυτό το σημείο εντοπίζεται και το βασικό μειονέκτημα της συνιστώσας συσχέτισης ως μετρικής ομοιότητας, αφού δυο διανύσματα μπορεί να έχουν γινόμενο συσχέτισης +1 (που μεταφράζεται σε μεταξύ τους ομοιότητα), αλλά η τοποθέτηση τους στο χώρο να διαφέρει. Ταυτόχρονα η μετρική αυτή δεν ικανοποιεί το κριτήριο της ανισότητας τριγώνου και από στατιστική άποψη ο υπολογισμός της συσχέτισης μεταξύ οντοτήτων λαμβάνοντας υπόψη τη μέση τιμή των τιμών τον διαφορετικών μεταβλητών που δεν είναι πάντα αξιόπιστος.(Everitt B.S. 2002)

6. Μετρικές Απόστασης

Οι μετρικές απόστασης μπορούν να χαρακτηριστούν και ως μετρικές διαφορών (αφού κυρίως μετρούν την απόσταση ανάμεσα σε δυο οντότητες x_i και x_j) . Για δυο σχετικές οντότητες η μεταξύ τους απόσταση είναι μηδενική. Η πιο γνωστή μετρική απόστασης είναι η Ευκλείδεια νόρμα:

$$d_{ij} = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

όπου d_{ij} η απόσταση ανάμεσα στην i και j οντότητα, και x_{ik} η τιμή της μεταβλητής για τη k οντότητα. Άλλες μετρικές απόστασης που συχνά χρησιμοποιούνται είναι, η απόσταση Manhattan που ορίζεται ως

$$d_{ij} = \sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|,$$

οι μετρικές Minkowski που δίνονται από τον τύπο $d_{ij} = (\sum_{k=1}^{d} |x_{ik} - x_{jk}|)^{1/r}$ και η μετρική Mahalanobis: $d_{ij} = (X_i - X_j)' S^{-1} (X_i - X_j)$ όπου $S = \frac{1}{N-1} \sum_{k=1}^{N} (x_i - \bar{x})(x_j - \bar{x})$ και $\bar{x} = \frac{1}{N} \sum_{k=1}^{N} x_i$.

Από τους παραπάνω τύπους μπορούμε να διαπιστώσουμε ότι η μετρική Manhattan και η Ευκλείδεια απόσταση, προκύπτουν από την τυπολογία των Minkowski μετρικών για *r=1* και *r=2* αντίστοιχα.

Σημαντικό μειονέκτημα των μετρικών είναι η επίδραση που έχει η αλλαγή στην κλίμακα μέτρησης των μεταβλητών. Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί με κανονικοποίηση των μετρήσιμων τιμών ως προς μια τιμή αναφοράς. Τυπικό παράδειγμα αποτελεί η κανονικοποίηση των μετρήσιμων τιμών χρησιμοποιώντας ως τιμές αναφοράς τη μέση τιμή και η απόκλιση κάθε μεταβλητή:

$$x_{ik}^{standard} = \frac{x_{ik} - x_k}{s_k}$$

7. Συνιστώσες σχέσης

Οι συνιστώσες σχέσης χρησιμοποιούνται για τη μελέτη της ομοιότητας μεταξύ οντοτήτων που χαρακτηρίζονται από δυαδικές μεταβλητές στις οποίες η παρουσία ενός χαρακτηριστικού δηλώνεται με 1 και η απουσία του με 0. Αν και στη σχετική βιβλιογραφία έχουν προταθεί αρκετές συνιστώσες σχέσης ως μετρικές ομοιότητας ,οι πιο συχνά χρησιμοποιούμενες είναι οι ακόλουθες:

α) απλή μετρική ταιριάσματος,

β) συνιστώσα του Jaccard και

γ) συνιστώσα του Gower.

8. Πιθανοτικές συνιστώσες ομοιότητας

Οι πιθανοτικές συνιστώσες ομοιότητας διαφέρουν σε σχέση με τις προηγούμενες αφού δε στοχεύουν στον υπολογισμό κάποιας μετρικής ομοιότητας, αλλά μελετώντας τα δεδομένα προσπαθούν να εξάγουν χρήσιμες πληροφορίες για την κατηγοριοποίησή τους. Βασικός περιορισμός στη χρήση των πιθανοτικών συνιστωσών ομοιότητας είναι η ύπαρξη δυαδικών μεταβλητών, αφού δεν έχουν αναπτυχθεί παρόμοιες τεχνικές για ποσοτικές ή ποιοτικές μεταβλητές.

1.4 Χαρακτηριστικά ομάδων

Κάθε ομάδα ομοειδών στοιχείων (συστάδα) που προκύπτει από μια διαδικασία ανάλυσης συστάδων έχει ορισμένα χαρακτηριστικά όπως: πυκνότητα, διακύμανση, διάσταση, σχήμα και διαχωρισμό.

1.Η πυκνότητα ορίζεται από το πλήθος των ομοειδών στοιχείων που τοποθετούνται στο χώρο.

2.Η διακύμανση αναφέρεται στις αποστάσεις που εμφανίζουν τα σημεία μιας συστάδας από το κέντρο της.

3.Η διακύμανση δε θα πρέπει να θεωρηθεί ως ένα στατιστικό μέτρο αλλά ως τη σχετική τοποθέτηση των σημείων μιας συστάδας ως προς το χώρο. Οι συστάδες χαρακτηρίζονται ως «συμπαγείς» όταν τα σημεία τοποθετούνται κοντά στο κέντρο βάρους διαφορετικά «*χαλαρές*».

4.Η διάσταση μιας συστάδας εκφράζει την ακτίνα της σχηματιζόμενης έλλειψης, η οποία ουσιαστικά εκφράζει το σχήμα μιας συστάδας στοιχείων. Οι συχνά εμφανιζόμενες αναπαραστάσεις συστάδων για δεδομένα πολλών μεταβλητών είναι κουκίδες ή σφαίρες.

5.Το χαρακτηριστικό του διαχωρισμού αναφέρεται στη δυνατότητα ,οι συστάδες να αλληλοκαλύπτονται ή όχι. Η επιλογή τεχνικής θα πρέπει να γίνεται λαμβάνοντας υπόψη το προβλήματα κατηγοριοποίησης, το είδος των δεδομένων, τις μεταβλητές που χαρακτηρίζουν τα δεδομένα και τη μετρική ομοιότητας. Μια σχηματική ταξινόμηση των τεχνικών κατηγοριοποίησης φαίνεται στον ακόλουθο πίνακα:

Ταξινόμηση Μεθόδων Κατηγοριοποίησης	
Hierarchical Methods	Ιεραρχικοί μέθοδοι ανάλυσης συστάδων
Agglomerative	Συσσωρευτικοί μέθοδοι
Divisive	Διαιρετικοί μέθοδοι
Partitioning methods	Μέθοδοι κατάτμησης
Relocation Algorithms	Αλγόριθμοι μεταφοράς
Propablistic Clustering	Πιθανοτίκες μέθοδοι
K-medoids Methods	Μέθοδος κεντρικών σημείων
K-means Methods	Μη ιεραρχική μέθοδος συστάδων
Density-based Methods	Πυκνότητας Μέθοδοι
i.Density-based Connectivity Clustering	Πυκνότητας-βασισμένοι στην γειτνίαση
ii.Density Functions Clustering	Συναρτήσεις πυκνότητας
Graph-based Methods	Γραφικοί μέθοδοι
Methods Based on Co-occurrence of Categorical Data	Μέθοδοι βασισμένοι στην ταυτόχρονη εμφάνιση κατηγορικών δεδομένων
Constraint-based Clustering	

Clustering Algorithms used in Machine Learning	
Gradient Descent and Artificial Neural Networks	Η κλίση της εφαπτομένης και τεχνητά νευρωνικά δίκτυα
Evolutionary methods	Γενετικοί μέθοδοι
Scalable Clustering Algorithms	
Algorithms for High Dimensional Data	
Subspace Clustering	Ομαδοποίηση υποσυνόλων
Projection Techniques	Τεχνικές πρόβλεψης
Co-Clustering Techniques	Συνδυαστικές τεχνικές

Πίνακας : Ιεραρχική ταξινόμηση των μεθόδων κατηγοριοποίησης

Γενικά οι τεχνικές ομαδοποίησης κείμενων που χρησιμοποιούνται συχνότερα μπορούν να χωριστούν σε δυο μεγάλες κατηγορίες: α) τεχνικές ιεραρχικής ομαδοποίησης (Hierarchical Clustering), οι οποίες παρέχουν μια ιεραρχία από παραγόμενες συστάδες και β) τεχνικές διαμερίσεις (Partitioning), οι οποίες παρέχουν ένα σύνολο από ομάδες του αρχικού συνόλου κείμενων.

Παρακάτω αναφέρουμε μερικές μεθόδους ομαδοποίησης από τις παραπάνω με μεγαλύτερη λεπτομέρεια.

1.5 Πιθανοτίκες Μέθοδοι Ομαδοποίησης

Στις πιθανοτίκες μεθόδους κατηγοριοποίησης θεωρούμε τα δεδομένα ως ένα δείγμα ανεξάρτητα επιλεγμένο από ένα μεικτό μοντέλο διαφορετικών πιθανοτικών κατανομών. Ο βασικός άξονας αυτής της μεθόδου στηρίζεται στο ότι τα δεδομένα παράγονται ως εξής:

1) επιλέγοντας τυχαία ένα μοντέλο *j* με πιθανότητα τ_j, *j=1:k.* και

2) σημειώνοντας ένα στοιχείο *x* από την αντίστοιχη κατανομή. Γνωρίζουμε ότι η περιοχή που ορίζεται γύρω από το μέσο όρο κάθε κατανομής αποτελεί μια φυσική συστάδα, με γνωστή μέση τιμή και διασπορά. Επομένως κάθε στοιχείο εισόδου μεταφέρει και την ταυτότητα της συστάδας που ανήκει. Θεωρώντας ότι κάθε στοιχείο ανήκει σε μια και μόνο συστάδα, μπορούμε να υπολογίσουμε την πιθανότητα αντιστοίχησης του στοιχείου *x* στο *j* μοντέλο: Pr(Cj/x).

Τα πλεονεκτήματα των πιθανοτικών μεθόδων είναι τα εξής: α) μπορούν να χειριστούν εγγραφές δεδομένων πολύπλοκης δομής, β) μπορούν εύκολα να υλοποιηθούν και επομένως να ενσωματωθούν σε ένα περιβάλλον κατηγοριοποίησης.

1.6 Μέθοδοι ομαδοποίησης βάση πυκνότητας

Οι μέθοδοι αυτοί διερευνούν την τοποθέτηση των δεδομένων στο χώρο προσπαθώντας να ανιχνεύσουν περιοχές υψηλής πυκνότητας που μπορούν να θεωρηθούν φυσικές κατηγορίες- συστάδες ομοειδών δεδομένων. Ένα βασικό μειονέκτημα αυτών των μεθόδων έγκειται στο γεγονός ότι δύναται να χρησιμοποιηθούν μόνο για συστάδες σφαιρικού σχήματος στο χώρο.

Πιο συγκεκριμένα σε αυτή την μέθοδο ομαδοποίησης η έννοια της πυκνότητας και των ορίων διαχωρισμού των συστάδων παίζει κεντρικό ρόλο. Για περισσότερη ευκολία στους υπολογισμούς η χρήση αποδοτικών δομών δεικτοδότησης (R*-trees), κρίνεται απαραίτητη. Οι βασικοί παράμετροι των μεθόδων πυκνότητας περιλαμβάνουν:

α) ένα στοιχείο *χ να γειτονεύει:* ε-*neighborhood*, $(\mathrm{N}_e(x) = \{y \in X / d(x,y) \le e\}$, δηλαδή το σύνολο των στοιχείων με απόσταση μικρότερη ενός δοσμένου αριθμού ε,

β) να υπάρχει ένα στοιχείο πυρήνα (core object), δηλαδή ένα στοιχείο με βαθμό γειτνίασης- πλήθος γειτονικών στοιχείων,

γ) η απόσταση ενός στοιχείου *y* από ένα στοιχείο πυρήνα και

δ) τη συνεκτικότητα δυο στοιχείων *y, z,* τα οποία μπορούν να προσεγγιστούν από ένα κοινό στοιχείο πυρήνα.

Έχοντας καθορίσει τη συνεκτικότητα όλων των στοιχείων (δεδομένων εισόδου), κατηγοριοποιούμε τα δεδομένα σε συστάδες που χαρακτηρίζονται από το αντίστοιχο στοιχείο πυρήνα. Φυσικά υπάρχουν διαφορετικές προσεγγίσεις που βασίζονται στον υπολογισμό συναρτήσεων πυκνότητας των δεδομένων και την κατηγοριοποίησή τους.

1.7 Συγκεντρωτικές μέθοδοι ομαδοποίησης

Οι μέθοδοι αυτοί , χρησιμοποιούνται για την κατηγοριοποίηση δεδομένων στις οποίες επιτρέπουμε τις επικαλύψεις των συστάδων που προκύπτουν. Πιο συγκεκριμένα αυτή η κατηγορία μεθόδων βρίσκει ευρεία εφαρμογή σε προβλήματα γλωσσικής έρευνας, αφού η επικάλυψη των συστάδων, ουσιαστικά μεταφράζεται στη δυνατότητα αναπαράστασης των λέξεων με παραπάνω από μια ερμηνείες. Η γενική φιλοσοφία των μεθόδων στηρίζεται στη δημιουργία συστάδων με σκοπό τη μεγιστοποίηση μιας αντικειμενικής συνάρτησης σε πολλαπλά βήματα.

1.8 Γραφοθεωρητικές Μέθοδοι Ομαδοποίησης

Οι μέθοδοι της γραφοθεωρητικής κατηγοριοποίησης δεδομένων βασίζονται στη θεωρία της Ανάλυσης Γράφων. Η ιδέα σε αυτές τις μεθόδους είναι αρκετά απλή. Αρχικά υπολογίζουμε ένα γράφημα εγγύτητας (proximity graph) για τα δεδομένα εισόδου. Στη συνέχεια το δεύτερο βήμα είναι να διαγράφουμε οποιαδήποτε ακμή του γραφήματος που είναι μεγαλύτερη από τις γειτονικές της (σύμφωνα με κάποιο κριτήριο σύγκρισης). Στο τρίτο βήμα το δάσος (σύνολο δέντρων) που τελικά αποτελεί το σύνολο των διαμορφούμενων συστάδων (κάθε δέντρο αποτελεί και μια ομάδα ομοειδών στοιχείων).

1.9 Εξελικτικοί Μέθοδοι Ομαδοποίησης

Οι γενετικοί αλγόριθμοι (Genetic Algorithms), βρίσκουν συχνά εφαρμογή σε μεθόδους κατηγοριοποίησης. Πιο συγκεκριμένα κωδικοποιώντας ένα σύνολο τυχαίων πιθανών λύσεων ενός προβλήματος κατηγοριοποίησης ως ένα σύνολο χρωμοσωμάτων (αρχικός πληθυσμός) και χρησιμοποιώντας ένα σύνολο λειτουργιών όπως: επιλογή, μετάλλαξη, διασταύρωση κ.α., μετασχηματίζουμε τα αρχικά χρωμοσώματα (οπότε και προκύπτει ένας νέος πληθυσμός). Μια αντικειμενική συνάρτηση αξιολογεί κάθε χρωμόσωμα και τη δυνατότητα μεταφοράς του στην επόμενη γενιά. Το χρωμόσωμα που τελικά επιζεί κωδικοποιεί και τη λύση του προβλήματος κατηγοριοποίησης. Οι γενετικοί αλγόριθμοι προσφέρουν μια σφαιρική μελέτη του χώρου των πιθανών λύσεων (διαμορφούμενων συστάδων) μιας ομαδοποίησης, αφού ξεκινώντας από ένα σύνολο τυχαίων πιθανών λύσεων οι τελεστές της επιλογής ή μετάλλαξης εγγυούνται τη συνεχή δημιουργία νέων λύσεων, με αποτέλεσμα την εξαντλητική διερεύνηση των συστάδων που μπορούν να διαμορφωθούν (ανεξάρτητα από τον αρχικά επιλεγμένο πληθυσμό). Αντίθετα οι επαναληπτικές διαιρετικές μέθοδοι κατηγοριοποίησης προσφέρουν μια περιορισμένη διερεύνηση του χώρου των πιθανών λύσεων που εξαρτάται από τον αρχικό διαμοιρασμό των δεδομένων σε συστάδες. Προβλήματα εμφανίζονται στην κωδικοποίηση των δεδομένων μέσω χρωμοσωμάτων και στον καθορισμό του μεγέθους του αρχικού πληθυσμού.

ΚΕΦΑΛΑΙΟ 2: ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ (CLUSTER ANALYSIS)

2.1 Εισαγωγή

Η μέθοδος της ανάλυσης συστάδων αποτελεί μια στατιστική διαδικασία πολλών μεταβλητών, η οποία ξεκινώντας από ένα σύνολο δεδομένων, επιχειρεί να το οργανώσει σε ομάδες ομοειδών στοιχείων που ονομάζουμε συστάδες (clusters). Οι ομάδες αυτές δεν είναι εκ των προτέρων γνωστές αλλά προκύπτουν δυναμικά. Αντίθετα σε μια διαδικασία ταξινόμησης ή της επιβλεπόμενης μάθησης (supervised learning), οι κλάσεις/ κατηγορίες στις οποίες αντιστοιχίζονται τα δεδομένα, είναι εκ των προτέρων γνωστές και αποτελούν είσοδο στην αντίστοιχη μέθοδο. Οι περισσότερες εφαρμογές ομαδοποίησης δεδομένων αφορούν:

1. την ανάπτυξη μιας ταξινόμησης,
2. τη διερεύνηση σχημάτων για την ομαδοποίηση οντοτήτων,
3. την παραγωγή υποθέσεων από ανάλυση των δεδομένων και την αναπαράσταση τους,
4. την επαλήθευση υποθέσεων σε ένα σύνολο δεδομένων.

Η πιο συνηθισμένη μέθοδος των ομάδων είναι η ιεραρχική ανάλυση όπου χρησιμοποιεί δύο τεχνικές:

1.Συσσωρευτική ανάλυση

2.Επιμεριστική ανάλυση

Στη συσσωρευτική ανάλυση συγκεντρώνονται οι παρατηρήσεις σε μια και μόνο ομάδα ,ενώ στην επιμεριστική γίνεται το αντίθετο, δηλαδή από μια ομάδα επιμερίζονται σε τόσες ομάδες όσες είναι και οι παρατηρήσεις.

Τα κριτήρια για την ομαδοποίηση των παρατηρήσεων:

1.Κριτήριο εγγύτερου γείτονα :Παίρνουμε τις πιο κοντινές παρατηρήσεις.

2.Κριτήριο απώτερου γείτονα: Παίρνουμε τις πιο μακρινές παρατηρήσεις.

3.Κριτήριο μέσου δεσμού: Η απόσταση που θα έχουν δύο ομάδες μεταξύ τους θα είναι και η απόσταση που θα έχουν και οι υπόλοιπες ομάδες.

Τα αποτελέσματα φαίνονται στο συσσωρευτικό σχέδιο και στο δενδρόγραμμα: Στο συσσωρευτικό σχέδιο η πρώτη σειρά αφορά το πρώτο στάδιο, η δεύτερη το δεύτερο στάδιο κτλ. Στο δενδρόγραμμα βλέπουμε τα αποτελέσματα από το συσσωρευτικό σχέδιο. Οι κάθετες γραμμές είναι οι συνδυασμοί των ομάδων και το μήκος η απόσταση.

.

Cluster

Complete Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage	Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2			Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	94	97	0	0	0	175	100	93	95	52,14	2	0	175
2	197	198	5,11	0	0	60	101	103	104	53,62	2	51	124
3	128	129	8,25	0	0	85	102	26	45	54,63	2	0	138
4	145	147	9,24	0	0	60	103	137	139	55,02	74	54	185
5	90	92	10,29	0	0	104	104	109	110	58,4	55	0	137
6	36	37	11,24	0	0	147	105	2	16	59,71	74	0	134
7	47	48	12,24	1	1	124	106	1	11	59,94		11	165
8	144	146	14, 8	1	1	75	107	52	57	61,94	79	72	139
9	44	46	14,94	1	1	143	108	5	13	61,94	85	46	132
10	161	170	15,38	1	1	171	109	89	91	62,59	87	5	184
11	85	89	15,48	1	1	41	110	146	151	63,13	6	49	131
12	87	88	15,56	1	1	117	111	184	188	64,48	94	74	141
13	138	141	16,29	0	0	84	112	43	67	66,84	7	90	142
14	114	115	17,74	0	0	139	113	118	135	67,29	7	56	157
15	127	134	17,79	0	0	97	114	170	197	68,73	7	70	155
16	37	54	17,71	0	0	70	115	166	167	70	7	0	132
17	167	172	17,53	0	0	49	116	20	42	70,76	7	71	170
18	70	80	18,27	0	0	67	117	96	98	70,79	77	0	179
19	6	9	18,36	0	0	62	118	69	90	71,11	2	55	150
20	100	109	18,51	0	0	40	119	30	60	71,4	2	0	159
21	18	20	18,80	0	0	62	120	146	105	71,7	110	0	144
22	62	63	18,81	0	0	67	121	2	31	72,69	2	53	150
23	101	102	19,27	0	0	64	122	5	60	75,11	102	0	127
24	104	106	19,7	0	0	111	123	22	30	76,04	60	57	135
25	74	76	19,71	0	0	93	124	103	106	77,37	101	26	177
26	105	107	20,76	0	0	124	125	135	136	78,66	1	0	161
27	13	15	21,06	0	0	45	126	78	87	79,66	67	0	163
28	109	160	21,25	0	0	60	127	49	55	80,49	65	55	142
29	126	127	23,24	0	0	73	128	69	66	81,57	90	0	184
30	50	51	23,14	0	0	112	129	47	67	82,18	7	40	140
31	104	105	23,41	0	0	101	130	108	118	82,53	95	113	151
32	39	40	23,53	0	0	78	131	150	155	83,42	90	0	165
33	31	32	23,53	0	0	121	132	155	177	85,51	115	0	147
34	185	187	25,26	0	0	117	133	114	125	85,59	14	79	165
35	95	96	25,4	0	0	127	134	2	3	87,84	105	52	145
36	116	120	25,72	0	0	93	135	34	41	88,22	1	0	164
37	38	39	25, 6	1	1	124	136	32	36	88,61	121	76	164
38	72	74	25,13	1	34	94	137	21	118	1 1,24	84	1 2	143
39	167	168	25,4n	1	1	47	138	12n	134	1 1,38	65	1 3	164
40	67	68	25,44	1	11	124	139	44	55	1 1,48	84	1 4	162
41	4	24	25,53	1	1	143	140	133	42	1 1,1n	114	128	146
42	98	111	25,64	1	1	77	141	146	154	1 1,4n	131	132	142
43	166	168	26,58	0	37	69	142	43	48	1 1,68	113	127	169
44	111	112	26,46	0	0	85	143	70	84	1 1,84	41	0	163
45	13	14	26,79	27	0	103	144	26	33	1 1,93	7	53	167
46	100	101	27,2	0	0	157	145	2	4	1 2,19	134	31	155
47	157	164	27,29	16	0	103	146	11	139	1 2,71	87	76	165
48	180	182	27,57	0	0	89	147	10	38	1 2,80	93	0	165
49	190	197	27,51	17	0	110	148	160	170	1 2,81	94	97	167
50	190	197	30,21	30	2	152	149	142	151	1 3,03	92	0	170
51	64	65	28,29	0	0	92	150	140	150	1 3,29	72	47	167
52	95	91	29 6	0	0	159	151	100	100	1 3,72	100	46	152
53	108	119	30 1	0	0	164	152	150	131	122,20	157	0	180
54	2	7	30,27	0	0	74	153	150	130	124,01	107	38	160
55	8	12	30,38	0	0	69	154	157	174	124,87	72	36	160
56	120	121	31,22	0	0	145	155	1	2	130,03	103	145	157
57	122	123	31,32	0	19	73	156	180	185	130,77	114	59	168
58	178	179	31,54	0	0	114	157	163	166	141,06	143	59	174
59	183	188	31,72	0	0	163	158	21	95	141,55	121	118	162
60	149	140	32,27	4	0	110	159	30	34	144,11	110	135	173
61	113	117	32,38	0	0	83	160	161	175	146,66	154	0	174
62	7	8	32,54	0	19	124	161	155	145	154,21	125	150	172
63	191	194	34,26	0	0	94	162	27	44	156,24	155	129	167
64	139	140	34,26	12	0	143	163	78	81	156,48	123	27	187
65	25	27	35,15	0	0	125	164	22	25	157,6	139	138	187
66	44	45	35,28	0	9	124	165	111	114	176,65	144	153	177
67	47	50	35,46	19	1	124	166	180	194	142,13	137	11	187
68	43	44	36, 9	1	1	127	167	27	38	144,94	162	142	176
69	4	14	36,1	54	1	114	168	111	131	145,23	156	156	174
70	161	167	36,53	71	1	164	169	43	41	147,53	142	73	176
71	164	169	36,77	1	1	94	170	111	184	148,24	164	171	181
72	137	138	37,16	1	1	113	171	1	6	2,98 4	165	129	187
73	71	73	38,13	0	0	169	172	135	148	198,87	161	141	178
74	3	62	38,28	54	0	105	173	150	43	202,49	168	139	185
75	142	144	40,71	0	0	120	174	16	167	251,02	167	177	181
76	30	79	40,79	0	30	175	175	93	94	266,6	107	1	184
77	86	90	41,21	0	72	117	176	2	30	271,63	169	140	182
78	170	176	41,30	0	0	95	177	100	111	276,21	124	159	183
79	122	120	41,70	57	38	100	178	106	112	286,17	173	149	182
80	160	160	42,05	0	0	102	179	10	96	291,60	1	177	185
81	12	82	42,28	0	0	115	180	10	100	290,17	170	122	189
82	16	19	42,56	0	21	147	181	22	70	300,16	164	159	183
83	22	23	43,20	0	16	120	182	106	108	301,70	172	30	188
84	101	102	44,27	22	0	127	183	22	70	346,40	183	140	187
85	126	182	46,21	9	0	113	184	69	86	330,23	125	129	191
86	133	134	46,56	0	0	163	185	10	137	120,89	175	178	189
87	81	83	46,57	0	0	163	186	18	30	430,42	147	179	190
88	111	113	46,89	44	61	145	187	22	190	433,03	188	136	190
89	188	187	46,95	43	0	157	188	103	135	431,82	172	182	193
90	142	148	48,34	0	0	143	189	10	101	517,05	185	150	195
91	4	6	48,4	0	0	145	190	18	22	534,40	188	157	197
92	65	64	48,56	0	51	125	191	1	65	574,22	171	154	196
93	75	76	49,28	36	0	144	192	21	196	649,42	174	80	194
94	162	164	49,21	0	71	148	193	103	161	774,43	185	172	196
95	60	61	49,28	0	0	119	194	21	95	795 2	190	176	196
96	174	176	50,26	0	48	164	195	10	105	843	189	169	198
97	141	171	51 n	1	1	148	196	1	21	1062 4	193	194	197
98	116	118	51,28	1	94	121	197	1	18	1365,71	194	191	199
99	191	193	51,29	48	65	127	198	1	11	1 4	197	195	1

2.2 Ιστορικό

Ο όρος ανάλυση συστάδων (cluster analysis) χρησιμοποιήθηκε για πρώτη φορά από τον Tyron το 1939 και περιλαμβάνει ένα πλήθος διαφορετικών αλγορίθμων για την ομαδοποίηση αντικειμένων ίδιου είδους σε ξεχωριστές κατηγορίες. Μεγάλο μέρος της ιστορίας της ανάλυσης συστάδων ασχολείται με την ανάπτυξη αλγορίθμων οι οποίοι δεν ήταν τόσο πολύ απαιτητικοί υπολογιστικά καθώς οι πρώιμοι, αρχικοί υπολογιστές δεν ήταν τόσο ισχυροί όσο οι σημερινοί. Μάλιστα οι υπολογιστικές παρακάμψεις έχουν κατά παράδοση χρησιμοποιηθεί σε πολλούς αλγορίθμους της ανάλυσης συστάδων. Οι αλγόριθμοι αυτοί έχουν αποδειχθεί πολύ χρήσιμοι ενώ μπορούν να βρεθούν στα περισσότερα υπολογιστικά προγράμματα.

2.3 Λόγος εμφάνισης

Στη πραγματικότητα η μέθοδος της ανάλυσης συστάδων είναι κατά κάποιον τρόπο ένα αναπόσπαστο κομμάτι της ζωής μας. Για παράδειγμα μια ομάδα γευμάτων σε ένα εστιατόριο το οποίο είναι κοινό σε διάφορα άτομα μπορεί να θεωρηθεί μια συστάδα ατόμων. Σε μαγαζιά λιανικής πώλησης τροφίμων ιδίου τύπου οι διαφορετικοί τύποι κρέατος ή λαχανικών που παρουσιάζονται στις ίδιες ή κοντινές περιοχές. Υπάρχει ένας αναρίθμητο πλήθος παραδειγμάτων στο οποίο η ανάλυση συστάδων παίζει σημαντικό ρόλο. Επιπλέον ένα άλλο παράδειγμα είναι η κατηγοριοποίηση που πρέπει να κάνουν οι βιολόγοι σε διαφορετικά είδη ζώων με όσο το δυνατόν ρεαλιστική περιγραφή μεταξύ των ζώων. Σύμφωνα με το μοντέρνο σύστημα ταξινόμησης που εφαρμόζεται στη βιολογία ο άνθρωπος ανήκει σε διάφορες κατηγορίες όπως τα θηλαστικά, τα ζώα κ.α. Πρέπει να προσέξουμε εδώ ότι σε αυτή την κατηγοριοποίηση όσο υψηλότερο είναι το επίπεδο aggregation τόσο λιγότερο όμοια είναι τα μέλη στην αντίστοιχη κατηγορία. Ο άνθρωπος έχει περισσότερα κοινά με τους πιθήκους για παράδειγμα, παρά με τα πιο μακρινά μέλη όπως τα θηλαστικά (σκύλοι για παράδειγμα).

Ένας λόγος ύπαρξης της ανάλυσης συστάδων είναι ότι απαντάει σε ένα γενικό πρόβλημα που απασχολεί τους ερευνητές πολλές φορές σε διάφορα ερευνητικά θέματα, όπως να οργανώσουν τα παρατηρούμενα δεδομένα σε δομές, κατηγορίες που έχουν νόημα. Η ανάλυση συστάδων είναι ένα διερευνητικό εργαλείο ανάλυσης δεδομένων το οποίο στοχεύει να κατανείμει διαφορετικά αντικείμενα σε ομάδες με τέτοιο τρόπο ώστε ο βαθμός συσχέτισης μεταξύ δύο στοιχείων της ίδιας ομάδας να είναι όσο το δυνατόν μεγαλύτερος και μικρότερος στη περίπτωση που ανήκουν σε διαφορετικές. Με βάσει τα παραπάνω η ανάλυση συστάδων μπορεί να χρησιμοποιηθεί για να ανακαλύψει δομές στα δεδομένα χωρίς να παρέχει κάποια ερμηνεία, εξήγηση για αυτό. Με άλλα λόγια η ανάλυση συστάδων ανακαλύπτει δομές στα δεδομένα χωρίς να εξηγεί γιατί αυτές υπάρχουν.

2.4 Περιοχή Εφαρμογής

Η τεχνική της ανάλυσης συστάδων εφαρμόζεται σε μια ευρεία ποικιλία προβλημάτων. Ο Hartigan (1975) παρέχει μια εξαιρετική σύνοψη των αποτελεσμάτων της ανάλυσης συστάδων. Για παράδειγμα στο τομέα της ιατρικής, η ταξινόμηση ασθενειών, θεραπειών για ασθένειες ή συμπτωμάτων από ασθένειες μπορεί να οδηγήσουν σε πολύ χρήσιμες κατηγοριοποιήσεις. Στο τομέα της ψυχιατρικής η σωστή διάγνωση της κατηγορίας των συμπτωμάτων όπως η παράνοια, η σχιζοφρένεια, είναι απαραίτητα σε μια επιτυχή θεραπεία. Στην αρχαιολογία οι ερευνητές έχουν προσπαθήσει να δημιουργήσουν κατηγορίες για λίθινα εργαλεία, αντικείμενα τελετουργιών με το να εφαρμόσουν τεχνικές της ανάλυσης συστάδων. Γενικά οποτεδήποτε υπάρχει η ανάγκη να ταξινομήσουμε ένα βουνό πληροφοριών σε διαχειρίσιμες και λογικές κατηγορίες η ανάλυση συστάδων είναι ένα χρήσιμο εργαλείο.

2.5 Πλεονεκτήματα της ανάλυσης συστάδων

Ένα μεγάλο πλεονέκτημα της ανάλυσης συστάδων είναι η γρήγορη σύνοψη των δεδομένων και ειδικότερα όταν τα αντικείμενα ταξινομούνται σε πολλές ομάδες. Η μέθοδος αυτή προσφέρει ένα απλό προφίλ των ατόμων. Δεδομένου του πλήθους των μονάδων της ανάλυσης, για παράδειγμα του μεγέθους των σχολείων, της εθνικότητας των μαθητών, της περιοχής, του μεγέθους της πολιτικής αρμοδιότητας και του οικονομικού επιπέδου καθένα εκ των οποίων περιγράφεται από ένα σύνολο χαρακτηριστικών. Η μέθοδος αυτή προτείνει πως οι ομάδες των στοιχείων καθορίζονται έτσι ώστε οι μονάδες εντός των ομάδων να είναι παρόμοια κατά κάποιο τρόπο και διαφορετικά μεταξύ άλλων ομάδων.

2.6 Μειονεκτήματα της ανάλυσης συστάδων

Από τα μειονεκτήματα της ανάλυσης των συστάδων μπορούμε να αναφέρουμε ότι ένα αντικείμενο μπορεί να τοποθετηθεί σε μια μόνο ομάδα. Τα δεδομένα μέσω της ανάλυσης συστάδων ενδεχομένως να μην ανταποκρίνονται στη πραγματικότητα διότι τα δεδομένα μπορούν να κατηγοριοποιηθούν σε μια μόνο ομάδα και όχι σε πάνω από μια.

2.7 Έλεγχος στατιστικής σημαντικότητας

Η ανάλυση συστάδων ως μέθοδος δεν έχει καμία σχέση με τον έλεγχο στατιστικής σημαντικότητας. Στη πραγματικότητα η μέθοδος αυτή δεν έχει καμία σχέση με τα τυπικά στατιστικά τεστ αλλά είναι μια συλλογή διαφορετικών αλγορίθμων οι οποίοι τοποθετούν τα αντικείμενα σε ομάδες σύμφωνα με καλά ορισμένους κανόνες ομοιότητας. Το σημαντικό εδώ είναι ότι σε αντίθεση με πολλές άλλες στατιστικές διαδικασίες, η ανάλυση συστάδων χρησιμοποιείται συνήθως όταν δεν έχουμε κάποια αρχική υπόθεση αλλά είμαστε ακόμα στη φάση της εξερεύνησης. Κατά μια έννοια η μέθοδος αυτή βρίσκει την πιο σημαντική λύση που ενδεχομένως υπάρχει. Επομένως ο έλεγχος στατιστικής σημαντικότητας δεν είναι κατάλληλος εδώ, ακόμα και σε περιπτώσεις που αναφέρονται χαμηλά επίπεδα για την τιμή σημαντικότητας.

2.8 Μέθοδοι ανάλυσης συστάδων

Στην ανάλυση συστάδων υπάρχουν αρκετοί διαφορετικοί μέθοδοι, αλγόριθμοι που μπορούν να χρησιμοποιηθούν. Αυτές είναι: η ιεραρχική ανάλυση ομάδων (tree clustering), η σύνδεση δύο τρόπων (block clustering) και η μη ιεραρχική ανάλυση ομάδων (k-means clustering).

2.8.1 Ιεραρχική μέθοδος

Βασικό χαρακτηριστικό της ιεραρχικής μεθόδου και η σημαντικότερη διαφορά από τη μέθοδο k-means είναι ότι ο αριθμός των ομάδων δεν είναι γνωστός από την αρχή κάθε μελέτης. Κάθε παρατήρηση αποτελεί μια ομάδα αρχικά και σε κάθε βήμα ο ερευνητής ενώνει σε ομάδες τις πιο κοντινές παρατηρήσεις .(Tyron 1939) Υπάρχουν δυο τύποι μεθόδων με τις οποίες δουλεύονται τα δεδομένα και είναι οι εξής :

1. Η agglomerative μέθοδος που είναι η πιο διαδεδομένη. Σε αυτή την περίπτωση χρησιμοποιείται ο αλγόριθμος που ξεκινά με κάθε παρατήρηση σαν μια ομάδα και ενώνουμε τις πιο κοντινές παρατηρήσεις σε ομάδες.

2. Η divisive μέθοδος που είναι αντίστροφη από την παραπάνω μέθοδο που στην οποία ο μελετητής ξεκινά αντίστροφα, δηλαδή έχει όλες τις παρατηρήσεις σε μια ομάδα και η παρατήρηση που βρίσκετε πιο μακριά από τις υπόλοιπες αποχωρίζεται την ομάδα και σχηματίζει μια νέα ομάδα μονή της. Έπειτα παίρνει τη δεύτερη πιο μακρινή παρατήρηση και τη κατατάσσει είτε σε μια νέα ομάδα είτε στη πρώτη ομάδα και έτσι συνεχίζει μέχρι να μετακινηθούν όλες οι παρατηρήσεις της μελέτης.

Είναι κατανοητό λοιπόν από την περιγραφή που δόθηκε ότι η μέθοδος είναι ιδιαίτερα χρονοβόρα ειδικά όταν υπάρχουν μεγάλα σύνολα δεδομένων.

Τα βήματα που ακόλουθη η μέθοδος είναι τα εξής:

Βήμα 1: Κατασκευή του πινάκα αποστάσεων για όλες τις ομάδες

Βήμα 2: Βρίσκει τις μικρότερες αποστάσεις μεταξύ των παρατηρήσεων και τις ενώνει σε ομάδες

Βήμα 3: Στη περίπτωση που οι παρατηρήσεις δεν έχουν γίνει μια ομάδα επαναλαμβάνετε το βήμα 1

2.8.2. Ιεραρχικοί αλγόριθμοι

Υπάρχουν πολλές επιλογές μεθόδων για να υπολογιστούν οι αποστάσεις των ομάδων. Εμείς θα αναφέρουμε παρακάτω μερικές από αυτές. Οι μέθοδοι αυτοί κατηγοριοποίησης στηρίζονται στην επαναληπτική συνένωση ή διάσπαση επιμέρους συστάδων. Η συνένωση (ή διάσπαση) βασίζεται σε κριτήρια ομοιότητας (και αντίστοιχα διαφοράς) όχι ανάμεσα σε σημεία αλλά σε επιμέρους ομάδες στοιχείων που ονομάζονται κριτήρια σύνδεσης (linkage rules). Η επιλογή του κριτηρίου σύνδεσης επηρεάζει τα αποτελέσματα της διαδικασίας ομαδοποίησης αφού χαρακτηρίζει τα κριτήρια ομοιότητας ανάμεσα στις διαμορφούμενες συστάδες. Τα κριτήρια σύνδεσης που συχνά χρησιμοποιούνται είναι:

- Η μέθοδος του κοντινότερου γείτονα (single linkage) αναφέρεται στη μικρότερη απόσταση ανάμεσα σε δυο ζεύγη συστάδων,
- Η μέθοδος με βάση το μέσο ορό (average linkage) βασίζεται στη μέση απόσταση ανάμεσα σε δυο ζεύγη συστάδων και

- Η μέθοδος του μακρύτερου γείτονα (complete linkage) που αναφέρεται στη μεγαλύτερη απόσταση ανάμεσα σε δυο ζεύγη συστάδων.
- Η μέθοδος του κεντροεϊδούς υπολογίζει την απόσταση των κέντρων των ομάδων. Σε αυτή τη περίπτωση δημιουργούνται ελλειπτικές και συμπαγείς ομάδες.
- Η μέθοδος του Ward ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες, εξασφαλίζει ομάδες με παρόμοιο αριθμό παρατηρήσεων και παράγει ορθά αποτελέσματα γι αυτό χρησιμοποιείται πολύ έντονα στη πράξη (Ward 1963).

Πιο συγκεκριμένα όσον αφορά τα μέτρα που χρησιμοποιούνται για τον υπολογισμό των ομοιοτήτων ή των ανομοιοτήτων (αποστάσεων) μεταξύ των αντικειμένων, αυτά αντιστοιχούν σε ένα σύνολο κανόνων που εξυπηρετούν ως κριτήρια ομαδοποίησης ή διαχωρισμού των αντικειμένων. Αυτές οι αποστάσεις, ομοιότητες μπορεί να βασιστούν σε μια μόνο διάσταση ή σε πολλαπλές με κάθε διάσταση να αναπαριστά έναν κανόνα ή μια συνθήκη για την ομαδοποίηση των αντικειμένων. Για παράδειγμα αν θέλαμε να ομαδοποιήσουμε εστιατόρια γρήγορης εστίασης θα πρέπει να λάβουμε υπόψη μας τον αριθμό των θερμίδων που περιέχουν τα γεύματα τους, τις τιμές τους και την υποκειμενική αξιολόγηση ως προς την γεύση. Ο πιο άμεσος υπολογισμός των αποστάσεων μεταξύ των αντικειμένων σε ένα πολλαπλών διαστάσεων χώρο είναι ο υπολογισμός των Ευκλείδειων αποστάσεων. Στη περίπτωση δύο διαστάσεων ή τριών διαστάσεων χώρων το μέτρο αυτό στη πραγματικότητα είναι η γεωμετρική απόσταση των αντικειμένων στο χώρο. Παρόλα αυτά ο αλγόριθμος που χρησιμοποιείται δεν ενδιαφέρεται για το αν οι αποστάσεις αντιστοιχούν σε φυσικές αποστάσεις στο χώρο ή στο επίπεδο.

Η Ευκλείδεια απόσταση είναι συνήθως το πιο χρησιμοποιημένο μέτρο για τον υπολογισμό της απόστασης. Είναι η γεωμετρική απόσταση στο πολλαπλών διαστάσεων χώρο και υπολογίζεται με βάση τον παρακάτω τύπο:

$$\text{Απόσταση}(x,y) = \{ \Sigma_i (x_i - y_i)^2 \}^{½}$$

Να σημειωθεί ότι η Ευκλείδεια (και το τετράγωνο της Ευκλείδειας απόστασης) απόσταση συνήθως υπολογίζεται από τα αρχικά δεδομένα και όχι από τα τυποποιημένα δεδομένα. Το μέτρο αυτό έχει συγκεκριμένα πλεονεκτήματα όπως το γεγονός ότι η απόσταση μεταξύ δύο αντικειμένων δεν επηρεάζεται από την προσθήκη ενός νέου αντικειμένου στα δεδομένα το οποίο μπορεί να είναι μια ακραία τιμή. Παρόλα αυτά οι αποστάσεις μπορεί να επηρεαστούν σημαντικά από τις διαφορές στις κλίμακες που χρησιμοποιούνται για την κάθε διάσταση. Για παράδειγμα εάν μια διάσταση είναι στη κλίμακα των εκατοστών με το να πολλαπλασιάσουμε τις τιμές αυτές με τον αριθμό 10 η ευκλείδεια απόσταση θα επηρεαστεί πολύ και σαν επακόλουθο τα αποτελέσματα της ανάλυσης θα είναι αρκετά διαφορετικά. Γενικά είναι καλό οι κλίμακες των διαστάσεων να μετατρέπονται σε μια κοινή κλίμακα.

Όσον αφορά το τετράγωνο της Ευκλείδειας απόστασης η χρήση του σχετίζεται με το στόχο της προσθήκης μεγαλύτερου βάρους σε αντικείμενα που είναι πιο απομακρυσμένα. Η απόσταση αυτή υπολογίζεται με βάση τον παρακάτω τύπο:

$$\text{Απόσταση}(x,y) = \Sigma_i (x_i - y_i)^2$$

Όσον αφορά την απόσταση με το όνομα ‘City block (Manhattan) το μέτρο αυτό αναφέρεται στη μέση διαφορά κατά μήκος των διαστάσεων. Παρόλα αυτά πρέπει να σημειωθεί ότι η επίδραση μεγάλων διαφορών, ακραίων τιμών έχει μειωμένη ένταση καθώς δεν είναι στο τετράγωνο. Ο τρόπος υπολογισμού του είναι ο παρακάτω:

$$\text{Απόσταση}(x,y) = \Sigma_i |x_i - y_i|$$

Για την απόσταση τώρα που αναφέρεται κάτω από το όνομα του Chebyshev αυτό το μέτρο μπορεί να είναι κατάλληλο σε περιπτώσεις που θέλουμε να ορίσουμε δύο αντικείμενα ως διαφορετικά εάν είναι διαφορετικά έστω και σε μια τουλάχιστον διάσταση. Ο τρόπος υπολογισμού είναι ο παρακάτω:

$$\text{Απόσταση}(x,y) = \text{Maximum}|x_i - y_i|$$

Για την απόσταση βασισμένη στη δύναμη του εκθέτη το μέτρο αυτό χρησιμοποιείται όταν θέλουμε να αυξήσουμε ή να μειώσουμε την δύναμη, το βάρος που υπάρχει στις διαστάσεις όπου τα αντίστοιχα αντικείμενα είναι διαφορετικά. Ο τρόπος υπολογισμού είναι ο παρακάτω:

$$\text{Απόσταση}(x,y) = (\Sigma_i |x_i - y_i|^p)^{1/r}$$

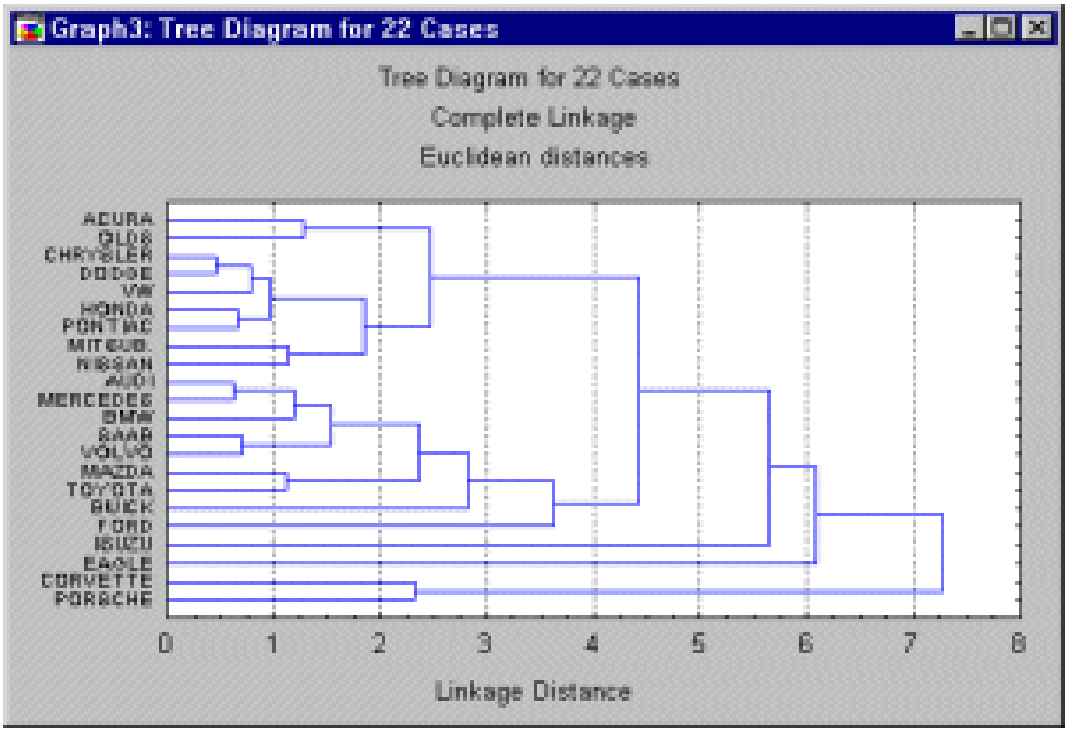
Όπου r και p είναι ορισμένοι παράμετροι. Η παράμετρος p ελέγχει το βαθμιαίο βάρος που τοποθετείται στις διαφορές των διαστάσεων ενώ η παράμετρος r ελέγχει το βαθμιαίο βάρος που τοποθετείται σε μεγάλες διαφορές μεταξύ των αντικειμένων. Στη περίπτωση που οι παράμετροι αυτοί είναι ίση με την τιμή 2 ο καθένας τότε η απόσταση αυτή είναι η Ευκλείδεια απόσταση.

Τέλος υπάρχει και το μέτρο που βασίζεται στο ποσοστό διαφωνίας. Το μέτρο αυτό είναι αρκετά χρήσιμο εάν τα δεδομένα των διαστάσεων που εμπεριέχονται στην ανάλυση είναι κατηγορικά στη φύση. Η απόσταση αυτή υπολογίζεται ως ακολούθως:

$$\text{Απόσταση } (x,y) = (\text{Number of } x_i \neq y_i)/\ i$$

2.8.2.1 Ιεραρχική Ανάλυση Ομάδων (tree clustering)

Ο στόχος της ανάλυσης αυτής είναι να ενώσει τα δεδομένα επιτυχώς σε μεγάλες ομάδες χρησιμοποιώντας ένα μέτρο ομοιότητας ή απόστασης. Το τυπικό αποτέλεσμα αυτού του τύπου ανάλυσης είναι το ιεραρχικό δέντρο. Παρατηρώντας το γράφημα στο αριστερό μέρος του ξεκινάμε με κάθε δεδομένο σε δική του ομάδα.. Αν θέλουμε να το θέσουμε διαφορετικά μειώνουμε το επίπεδο απόφασης μας για το πότε θα δηλώσουμε δύο ή περισσότερα δεδομένα μέλη της ίδιας ομάδας.



Το αποτέλεσμα της διαδικασίας αυτής είναι να ενώνουμε όλο και περισσότερα δεδομένα μεταξύ τους και να δημιουργούμε μεγαλύτερες ομάδες με διαφορετικά στοιχεία. Στο τελευταίο στάδιο όλα τα αντικείμενα είναι ενωμένα μεταξύ τους. Σε αυτό το γράφημα ο οριζόντιος άξονας αναπαριστά την απόσταση μεταξύ των δεδομένων. Έτσι για κάθε σύνδεσμο στο γράφημα μπορούμε να διαβάσουμε το κριτήριο της απόστασης για τα στοιχεία, τα δεδομένα είναι ενωμένα σε ένα νέο στοιχείο. Όταν τα δεδομένα είναι χωρισμένα σε ξεκάθαρες ομάδες τότε η μέθοδος

έχει τελειώσει. Το αποτέλεσμα της επιτυχής ανάλυσης με αυτή την μέθοδο είναι να δημιουργηθούν ομάδες.(Herbert 1984)

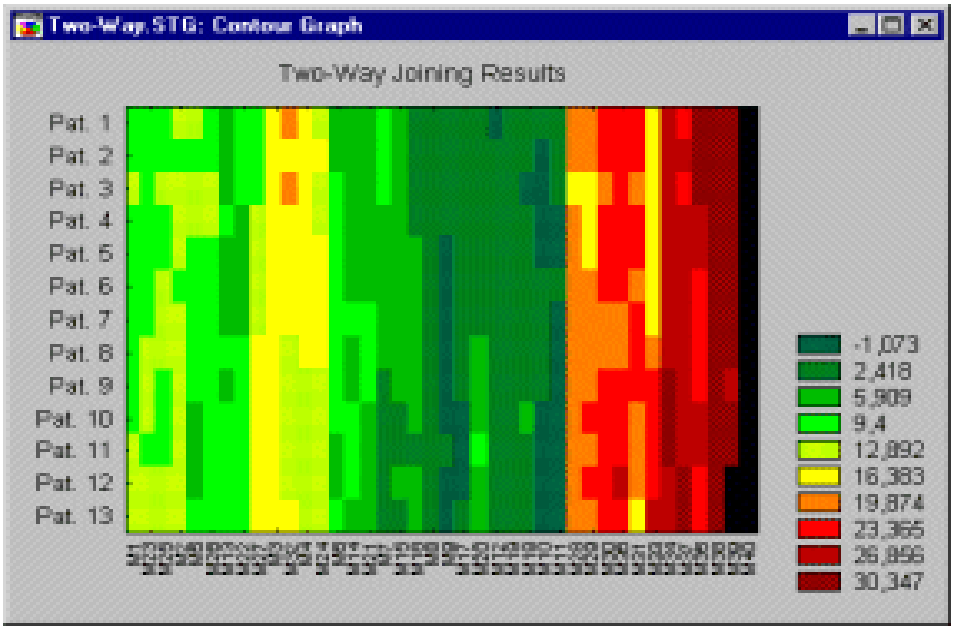
Το μειονέκτημα αυτής της μεθόδου είναι ότι έχει υψηλές απαιτήσεις σε αποθηκευτικό χώρο για τον πίνακα ομοιότητας που χρησιμοποιείται.

Αλγόριθμος Ιεραρχικής ανάλυσης συστάδων (Tan Steinbach Kumar, 1994)

1. Υπολόγισε τον πίνακα (proximity matrix)
2. Κάθε σημείο θεώρησε το ως μια συστάδα
3. Επανέλαβε
4. Ένωσε τις πιο κοντινές συστάδες
5. Ενημέρωσε τον πίνακα (proximity matrix)
6. Μέχρι να παραμείνει μια συστάδα

2.8.2.2 Σύνδεση δυο τρόπων (Block clustering)

Πριν είδαμε ότι η κατηγοριοποίηση γινόταν σε επίπεδο αντικειμένων ενώ σε πολλές άλλες αναλύσεις η κατηγοριοποίηση γίνεται σε επίπεδο περιπτώσεων ή μεταβλητών. Έχει φανεί όμως ότι η κατηγοριοποίηση και με τους δύο τρόπους μαζί προσφέρει σημαντικά αποτελέσματα. Για παράδειγμα σε μια ιατρική έρευνα ένας ερευνητής που έχει μαζέψει δεδομένα σε διαφορετικούς τομείς της φυσικής υγείας σε ένα δείγμα ασθενών με καρδιακά προβλήματα (περιπτώσεις). Ο ερευνητής μπορεί να θέλει να κατηγοριοποιήσει τους ασθενείς έτσι ώστε να εντοπίσει ομάδες ασθενών με παρόμοια συμπτώματα. Ταυτόχρονα όμως ο ερευνητής μπορεί να θέλει κατηγοριοποιήσει τα μέτρα, τους δείκτες της φυσικής υγείας έτσι ώστε να εντοπίσει κατηγορίες δεικτών που εμφανίζουν παρόμοιες φυσικές ικανότητες. Με βάσει τα παραπάνω η μέθοδος αυτού του τρόπου κατηγοριοποίησης είναι πολύ χρήσιμος σε περιπτώσεις που αναμένουμε ότι και οι παρατηρήσεις αλλά και οι μεταβλητές συνεισφέρουν ταυτόχρονα στην ανακάλυψη κατηγοριών που έχουν νόημα. (Hartigan, 1975).



Για παράδειγμα στο θέμα της ιατρική έρευνας που αναφερθήκαμε πριν ο ερευνητής μπορεί να θέλει κατηγορίες ασθενών με παρόμοια χαρακτηριστικά σχετικά με δείκτες φυσικής υγείας. Η δυσκολία με την ερμηνεία των αποτελεσμάτων αυτών είναι ότι οι ομοιότητες μεταξύ διαφορετικών ομάδων μπορεί να προκαλείται από διαφορετικά υποσύνολα των μεταβλητών. Έτσι η δομή που δημιουργείται από τις ομάδες δεν είναι ομοιογενής. Αυτό βέβαια μπορεί να δημιουργεί κάποια σύγχυση στην αρχή συγκρινόμενο και με τις δύο άλλες μεθόδους που υπάρχουν στην ανάλυση συστάδων (την ιεραρχική και την μη ιεραρχική ανάλυση συστάδων). Η μέθοδος αυτή είναι η λιγότερο χρησιμοποιούμενη. Παρόλα αυτά κάποιοι ερευνητές πιστεύουν ότι η μέθοδος αυτή προσφέρει ένα ισχυρό εργαλείο εξερεύνησης (Hartigan, 1975).

2.9 Μη Ιεραρχική Ανάλυση Συστάδων (*k*-Means Clustering)

Η μέθοδος αυτή είναι πολύ διαφορετική σε σχέση με τις δύο προηγούμενες. Ας υποθέσουμε ότι έχουμε ήδη κάποιες υποθέσεις σχετικά με το πλήθος των ομάδων είτε στις παρατηρήσεις είτε στις μεταβλητές. Μπορεί να θέλουμε να θέσουμε στο πρόγραμμα, αλγόριθμο που χρησιμοποιούμε να δημιουργήσει για παράδειγμα 3 ομάδες που είναι όσο το δυνατόν διακριτές μεταξύ τους. Η μέθοδος αυτή θα δημιουργήσει 3 (κ αν θέλαμε κ στο πλήθος ομάδες) με την μεγαλύτερη δυνατή διάκριση μεταξύ τους. Εδώ πρέπει να ειπωθεί ό τι καλύτερος, βέλτιστος αριθμός ομάδων που οδηγεί στο καλύτερο διαχωρισμό δεν είναι γνωστός εξ αρχής και πρέπει να υπολογιστεί από τα δεδομένα.

Να τονίσουμε ότι υπάρχουν δύο βασικές προσεγγίσεις για την επιλογή της αρχικής διαχώρισης των δεδομένων σε συστάδες. Στην πρώτη προσέγγιση χρησιμοποιούμε μια εκτίμηση των κέντρων βάρους των συστάδων και κάθε στοιχείο αντιστοιχίζεται στη συστάδα με το κοντινότερο κέντρο βάρους, ήδη από το πρώτο πέρασμα των δεδομένων. Στη δεύτερη προσέγγιση η επιλογή των αρχικών κέντρων βάρους γίνεται είτε επιλέγοντας το κεντρικό σημείο κάθε συστάδας ***(K-medoids***

methods), είτε χρησιμοποιώντας τα αποτελέσματα μιας άλλης μεθόδου κατηγοριοποίησης. Σε αυτή την κατεύθυνση οι Bradley και Fayyad (1998), πρότειναν την αρχική εφαρμογή της ***k-means*** μεθόδου σε μικρά δείγματα των δεδομένων και την επιλογή των κέντρων βάρους της καλύτερης κατηγοριοποίησης ως αρχικών κέντρων βάρους για την επανάληψη της μεθόδου στο σύνολο των δεδομένων, αν και αυτό συνεπάγεται παραπάνω χρόνο επεξεργασίας.

Ένα παράδειγμα στις φυσικές ικανότητες που είδαμε παραπάνω είναι ότι οι ερευνητές μπορεί να έχουν μια διαίσθηση από κλινικές δοκιμές ότι οι ασθενείς με καρδιακά προβλήματα εμπίπτουν σε τρεις διαφορετικές κατηγορίες σχετικά με τις φυσικές τους δυνατότητες. Μπορεί να αναρωτιόνται εάν αυτή η διαίσθηση μπορεί να ποσοτικοποιηθεί, δηλαδή αν η μη ιεραρχική ανάλυση συστάδων μπορεί να επιβεβαιώσει την διαίσθηση τους. Σε αυτή την περίπτωση ο μέσος όρος των διαφορετικών δεικτών των φυσικών ικανοτήτων για κάθε ομάδα θα αναπαριστούν έναν ποσοτικό τρόπο έκφρασης της διαίσθησης των ερευνητών.

Όσον αφορά το υπολογιστικό μέρος μπορεί κάποιος να σκεφτεί αυτή την μέθοδο ως μια ανάλυση διασποράς αντίστροφα όμως. Ο αλγόριθμος θα ξεκινήσει με κ τυχαίες ομάδες και μετά θα μετακινηθεί σε αντικείμενα μεταξύ των ομάδων με στόχο 1) να ελαχιστοποιήσει την μεταβλητότητα εντός της ομάδας και 2) να μεγιστοποιήσει την μεταβλητότητα μεταξύ των ομάδων. Με άλλα λόγια οι κανόνες ομοιότητας θα εφαρμοστούν πλήρως στα μέλη μια ομάδας και ελάχιστα στα μέλη μεταξύ διαφορετικών ομάδων. Αυτό είναι ανάλογο με τον έλεγχο της ανάλυσης διασποράς που υπολογίζει και συγκρίνει την μεταβλητότητα μεταξύ των ομάδων με την μεταβλητότητα εντός των ομάδων. Στον αλγόριθμο αυτό γίνεται προσπάθεια να μετακινηθούν τα αντικείμενα εντός και εκτός ομάδων με στόχο να επιτευχθούν τα πιο σημαντικά αποτελέσματα της ανάλυσης διασποράς.

Ως προς την ερμηνεία των αποτελεσμάτων εξετάζουμε τους μέσους όρους για κάθε ομάδα για κάθε διάσταση για να εκτιμήσουμε πόσο διακριτοί είναι οι κ ομάδες. Ιδανικά θα έχουμε πολύ διαφορετικούς μέσους όρους για τις περισσότερες αν όχι όλες τις διαστάσεις που χρησιμοποιούνται στην ανάλυση. Το μέγεθος των τιμών της F κατανομής από την ανάλυση της διασποράς που υπολογίζονται για κάθε διάσταση είναι μια άλλη ένδειξη για το πόσο καλά οι διαστάσεις διαχωρίζονται μεταξύ των ομάδων.

Ο αλγόριθμος της μη ιεραρχικής ανάλυσης συστάδων είναι ο παρακάτω (Tan Steinbach Kumar, 1994)

Αλγόριθμος

1. Επέλεξε K σημεία ως τα αρχικά κεντρικά σημεία
2. Επανέλαβε
3. Σχημάτισε K συστάδες με το να αντιστοιχηθούν όλα τα σημεία στο κοντινότερο κεντρικό σημείο.
4. Υπολόγισε ξανά τα κεντρικά σημεία της κάθε συστάδας
5. Μέχρι τα κεντρικά σημεία δεν αλλάζουν

Λεπτομέρειες σχετικές με τον αλγόριθμο :

1. Τα αρχικά κεντρικά σημεία συχνά επιλέγονται στην τύχη, οι συστάδες που παράγονται πολλές φορές ποικίλουν μεταξύ τους
2. Τα κεντρικά σημεία είναι συνήθως ο μέσος όρος των σημείων στην συστάδα
3. Το πόσο κοντά είναι μετριέται από την Ευκλείδεια απόσταση, την ομοιότητα συνημίτονου, συσχέτιση, κ.α.

4. Ο αλγόριθμος θα συγκλίνει για τα συνηθισμένα μέτρα ομοιότητας που αναφέρονται παραπάνω
5. Συνήθως η σύγκλιση συμβαίνει στις πρώτες επαναλήψεις, συχνά ο κανόνας για να σταματήσει ο αλγόριθμος αλλάζει αφού λίγα κεντρικά σημεία αλλάζουν συστάδες.
6. Η πολυπλοκότητα είναι O(n*K*I*d), η= ο αριθμός των σημείων, K= ο αριθμό των συστάδων, I= ο αριθμός των επαναλήψεων, d= ο αριθμός των συνεισφορών.

2.10. Ασαφής ανάλυση συστάδων (Fuzzy analysis)

Η μέθοδος ομαδοποίησης fuzzy, επιτρέπει σε κάθε στοιχειό να ανήκει σε περισσότερες από μια συστάδες. Αυτό γίνεται υπολογίζοντας κάποια ποσοστά (memberships) για κάποιο στοιχειό για κάθε συστάδα, ώστε το άθροισμα τους να είναι ίσο με το 1. Με αλλά λόγια στην ασαφής ομαδοποίηση ορίζουμε μια συνάρτηση συμμετοχής. Η συνάρτηση αυτή υποδηλώνει το βαθμό συμμετοχής κάθε στοιχείου στην κάθε ομάδα. Οι τιμές που μπορεί να πάρει ο βαθμός συμμετοχής είναι από μηδέν έως ένα. Όσο πιο κοντά στο ένα είναι ο βαθμός συμμετοχής του i προτύπου στην j ομάδα, τόσο πιο μεγάλη σιγουριά υπάρχει για τη συμμετοχή αυτού στη συγκεκριμένη ομάδα. Αντίθετα, όσο πιο κοντά στο μηδέν είναι ο βαθμός συμμετοχής, μεγαλώνουν οι αμφιβολίες για τη συμμετοχή του. Φυσικά δεν είναι απαραίτητο όλες οι ομάδες να έχουν ασαφή χαρακτήρα. Είναι πιθανό να προκύψουν και απόλυτες ομάδες. Αυτό θα γίνει αν όλα τα στοιχειά μιας ομάδας έχουν βαθμό συμμετέχεις ίσο με ένα.

ΚΕΦΑΛΑΙΟ 3 :Παραδείγματα

Για να κατανοήσουμε το θέμα της εργασίας μας και το τι ακριβώς κάνει η ομαδοποίηση κατά συστάδες (cluster analysis) βρήκαμε κάποια παραδείγματα που χρησιμοποιούν τη συγκεκριμένη ανάλυση. Τα παραδείγματα αυτά τα αναλύουμε παρακάτω.

Παράδειγμα 1

Το παρόν παράδειγμα αναφέρεται σε μια ανάλυση που έχει εκπονηθεί από την τράπεζα Πειραιώς για την ιεράρχηση του βαθμού επικινδυνότητας των οικονομιών. Η μεθοδολογία, που ανέπτυξε η τράπεζα βασίζεται στο στατιστικό αλγόριθμο της ανάλυσης συστάδων που ομαδοποιεί τις χώρες σε ομοιογενείς ομάδες βάσει ενός μεγάλου αριθμού προκαθορισμένων μεταβλητών που αντανακλούν το μακροοικονομικό, χρηματοοικονομικό και πολιτικό κίνδυνο. Με αυτό τον αλγόριθμο κατατάχτηκαν όλες οι χώρες της ανάλυσης σε τρεις κατηγορίες,

α) την κατηγορία υψηλού κινδύνου (R+),

β) την κατηγορία μεσαίου κινδύνου (R),

γ) την κατηγορία χαμηλού κινδύνου (R-).

Όπως ήδη έχουμε αναφέρει η ανάλυση αυτή έγινε με την βοήθεια της στατιστικής μεθοδολογίας της ανάλυσης συστάδων (cluster analysis) όπου οι χώρες υπό εξέταση συγκροτούνται σε ομάδες βάση μιας σειράς μεταβλητών οι οποίες αντανακλούν το βαθμό μακροοικονομικού, χρηματοπιστωτικού και πολιτικού ρίσκου. Πιο αναλυτικά, εάν ομαδοποιούμε τις χώρες βάσει μίας και μόνο μεταβλητής (π.χ. ΑΕΠ) τότε εάν σε μια χρονιά όλες οι χώρες επιδεινωθούν κατά τον ίδιο βαθμό - επειδή ακριβώς οι σχετικές αποστάσεις μεταξύ των χωρών έχουν παραμείνει αμετάβλητες - τότε ο αλγόριθμος δεν θα καταγράψει καμία μεταβολή στην εκτίμηση των χωρών, παρά το γεγονός ότι οι οικονομικές συνθήκες έχουν μεταβληθεί

σημαντικά. Για να ξεπεραστεί αυτό η αναφέρθηκε σε κάθε χώρα/χρονιά σαν να ήταν ξεχωριστή παρατήρηση π.χ. Ρωσία-2007 και η Ρωσία-2008 αποτελούν δύο διαφορετικές παρατηρήσεις και εξετάζουμε εάν ο αλγόριθμος τις κατατάσσει στην ίδια ομάδα. Εάν ναι τότε δεν υπήρξε μεταβολή την αξιολόγηση της χώρας, εάν όχι τότε καταγράφουμε μια αναβάθμιση/υποβάθμιση. Μέσω της ανάλυσης προέκυψαν τρεις ομάδες χωρών για το 2007 και το 2008 που χαρακτηρίσαμε ως προς τον κίνδυνο. Ξεχώρισαν τρεις ομάδες ως προς τον κίνδυνο χώρας, την ομάδα χαμηλού κινδύνου (R-), την ομάδα μεσαίου κινδύνου (R) και την ομάδα υψηλού κινδύνου (R+). Σύμφωνα με την ανάλυση μας, οι μεταβλητές που συνεισφέρουν περισσότερο στην κατάταξη των χωρών σε συγκεκριμένες ομάδες είναι τα συναλλαγματικά διαθέσιμα ως ποσοστό των εισαγωγών, το εξωτερικό χρέος ως ποσοστό των εξαγωγών, το κατά κεφαλή ΑΕΠ, το εξωτερικό εμπόριο (εξαγωγές συν εισαγωγές) ως ποσοστό του ΑΕΠ, το δημοσιονομικό έλλειμμα ως ποσοστό του ΑΕΠ, η κεφαλαιακή επάρκεια του τραπεζικού συστήματος, η πιστοληπτική διαβάθμιση για κυβερνητικό χρέος με ρήτρα ξένου και εγχωρίου νομίσματος μακροχρόνιας και βραχυχρόνιας διάρκειας, η πρόβλεψη της πιστοληπτικής αυτής διαβάθμισης και η διαβάθμιση του κινδύνου για τη μετατροπή και μεταφορά συναλλάγματος εκτός χώρας.

Στον Πίνακα Ι βλέπουμε ότι η μόνη χώρα που υποβαθμίστηκε το 2008 σε σχέση με το προηγούμενο έτος ήταν η Βουλγαρία. Συγκεκριμένα, η διαβάθμιση της Βουλγαρίας ως προς τον κίνδυνο χώρας υποχώρησε δύο θέσεις και η χώρα εντάχθηκε στην ομάδα R+ το 2008 από την ομάδα R- το 2007 εξαιτίας του πολύ υψηλού ελλείμματος στο ισοζύγιο τρεχουσών συναλλαγών (2008: -25,3% του ΑΕΠ) που λόγω της διεθνούς οικονομικής κρίσης επέτεινε την ανάγκη για εξωτερική χρηματοδότηση. Παράλληλα, η εφαρμογή πολιτικής σταθερής διμερούς συναλλαγματικής ισοτιμίας μεταξύ του εγχωρίου νομίσματος και του ευρώ περιόρισε τη δυνατότητα διόρθωσης της εξωτερικής ανισορροπίας. Η αποχώρηση των ξένων επενδυτών από τις διεθνείς αγορές λόγω της κρίσης επέτεινε την πίεση στο εγχώριο νόμισμα για την άρση εφαρμογής της πολιτικής σταθερής συναλλαγματικής ισοτιμίας με το ευρώ.

Πίνακας Ι: Προφίλ Ομάδων ανά έτος

Προφίλ Ομάδων: 2007			Προφίλ Ομάδων: 2008		
R-	R	R+	R-	R	R+
Βουλγαρία	Ρωσσία	Αίγυπτος	Τυνησία	Ρωσσία	Βουλγαρία
Τυνησία	Καζακστάν	ΠΓΔΜ	Κύπρος	Καζακστάν	Αίγυπτος
Κύπρος	Μαρόκο	Ιορδανία	Κροατία	Μαρόκο	ΠΓΔΜ
Κροατία		Σερβία			Ιορδανία
		Τουρκία			Σερβία
		Ρουμανία			Τουρκία
		Ουκρανία			Ρουμανία
					Ουκρανία

Πηγή: Piraeus Bank Research

Μπορεί η συμμετοχή των χωρών σε μια ομάδα (εκτός της Βουλγαρίας) να μην άλλαξε την περίοδο 2007-08, αλλά άλλαξαν τα ποιοτικά χαρακτηριστικά της κάθε ομάδας. Από τον Πίνακα ΙΙ για το 2007 βλέπουμε ότι:

Πίνακας ΙΙ: Χαρακτηριστικά Ομάδων, 2007

	R-	R	R+
Συναλλαγματικά Διαθέσιμα	L	H	M
Εξωτερικό Χρέος	L	H	M
ΑΕΠ	H	M	L
Ισοζύγιο Τρεχουσών Συναλλαγών	L	H	M
Δημοσιονομικό Ισοζύγιο	M	H	L
Κεφαλαιακή Επάρκεια	M	L	H
ICRG	H	M	L
LCLT	H	M	L
LCO	M	H	L
LCST	H	M	L
FCLT	H	M	L
FCO	M	H	L
FCST	H	M	L
TCR	H	M	L

Σημείωση: L: Low, M: Medium και H: High

Οι χώρες της ομάδας R- περιλαμβάνουν χώρες με τα εξής χαρακτηριστικά:

• Υψηλή πολιτική σταθερότητα (υψηλή τιμή του δείκτη ICRG), εξωτερική πιστοληπτική διαβάθμιση και υψηλό κατά κεφαλή ΑΕΠ (σε ισοδύναμα USD αγοραστικής δύναμης)

• Μέσο δημοσιονομικό ισοζύγιο, κεφαλαιακή επάρκεια και πρόβλεψη για την εξωτερική πιστοληπτική διαβάθμιση

• Χαμηλά συναλλαγματικά διαθέσιμα, εξωτερικό χρέος και ισοζύγιο τρεχουσών συναλλαγών

Οι χώρες της ομάδας R χαρακτηρίζονταν το 2007 από:

• Μεσαίες τιμές για το κατά κεφαλή ΑΕΠ, την πολιτική κατάσταση και την εξωτερική διαβάθμιση

- Υψηλές τιμές για το ισοζύγιο τρεχουσών συναλλαγών, το δημοσιονομικό ισοζύγιο, τα συναλλαγματικά διαθέσιμα, το εξωτερικό χρέος και την πρόβλεψη για την εξωτερική πιστοληπτική διαβάθμιση

• Χαμηλή κεφαλαιακή επάρκεια του τραπεζικού συστήματος

Το ίδιο έτος, οι χώρες της ομάδας R+ είχαν:

• Χαμηλές τιμές για την εξωτερική πιστοληπτική διαβάθμιση και την πρόβλεψη της, το κατά κεφαλή ΑΕΠ και το δημοσιονομικό ισοζύγιο

• Μεσαίες τιμές για τα συναλλαγματικά διαθέσιμα, το εξωτερικό χρέος και το ισοζύγιο τρεχουσών συναλλαγών

• Υψηλή κεφαλαιακή επάρκεια των τραπεζών

Πίνακας ΙΙΙ: Χαρακτηριστικά Ομάδων, 2008			
	R-	R	R+
Συναλλαγματικά Διαθέσιμα	L	H	M
Εξωτερικό Χρέος	L	H	M
ΑΕΠ	H	M	L
Ισοζύγιο Τρεχουσών Συναλλαγών	M	H	L
Δημοσιονομικό Ισοζύγιο	M	H	L
Κεφαλαιακή Επάρκεια	L	M	H
ICRG	H	M	L
LCLT	H	M	L
LCO	H	M	L
LCST	H	M	L
FCLT	H	M	L
FCO	H	M	L
FCST	H	M	L
TCR	H	M	L

Σημείωση: L: Low, M: Medium και H: High

Πηγή: Piraeus Bank Research

Σύμφωνα με την ομαδοποίηση των χωρών για το 2008 (Πίνακας ΙΙΙ), οι χώρες της ομάδας R- έχουν:

• Υψηλή πολιτική σταθερότητα, κατά κεφαλή ΑΕΠ, εξωτερική πιστοληπτική διαβάθμιση κυβερνητικού χρέους σε εγχώριο και ξένο νόμισμα, μακροχρόνιας και βραχυχρόνιας διάρκειας και διαβάθμιση αναφορικά με τον κίνδυνο μετατροπής και μεταφοράς συναλλάγματος

• Χαμηλή κεφαλαιακή επάρκεια του τραπεζικού συστήματος, εξωτερικό χρέος και συναλλαγματικά διαθέσιμα

• Μεσαίου μεγέθους ισοζύγιο τρεχουσών συναλλαγών και δημοσιονομικό ισοζύγιο

Το 2008 οι χώρες της ομάδας R είχαν:

• Μεσαία εξωτερική πιστοληπτική διαβάθμιση και πρόβλεψη της, πολιτική σταθερότητα, κεφαλαιακή επάρκεια τραπεζικού συστήματος και κατά κεφαλή ΑΕΠ

• Υψηλά συναλλαγματικά αποθέματα, εξωτερικό χρέος, ισοζύγιο τρεχουσών συναλλαγών και δημοσιονομικό ισοζύγιο

Το ίδιο έτος τα μέλη της ομάδας R+ είχαν:

• Χαμηλή εξωτερική πιστοληπτική διαβάθμιση και πρόβλεψη της, πολιτική σταθερότητα, κατά κεφαλή ΑΕΠ, ισοζύγιο τρεχουσών συναλλαγών και δημοσιονομικό ισοζύγιο

• Μεσαίο μέγεθος συναλλαγματικών διαθεσίμων και εξωτερικού χρέους

• Υψηλή κεφαλαιακή επάρκεια

Παράδειγμα 2

Στην ενότητα αυτή γίνεται μια προσπάθεια να παρουσιαστεί από ένα σπουδαστεί (Κάλτσας, 2008) η κατάσταση δυναμικού των ΤΕΕ σε κάθε νομό της Ελλάδας για το Σχολικό έτος 2004-2005. Ο λόγος που επιλέχθηκε σύμφωνα με τον συγγραφέα το συγκεκριμένο σχολικό έτος είναι γιατί αυτό είχε προαναγγελθεί από την νέα ηγεσία του Υπουργείου Παιδείας ότι θα είναι το τελευταίο μιας εκπαιδευτικής μεταρρύθμισης στην Τεχνική Επαγγελματική Εκπαίδευση που ξεκίνησε το 1997.

Πιο συγκεκριμένα θέλουμε να δούμε σε ποιους νομούς έχουμε παρόμοιο ποσοστό μαθητών που φοιτούν στα ΤΕΕ. Δηλαδή εξετάζουμε ποιες συστάδες θα δημιουργηθούν που θα περιέχουν νομούς με παρόμοια ποσοστά φοίτησης στα ΤΕΕ. Στο συγκεκριμένο παράδειγμα εφαρμόζονται τρεις διαφορετικές μέθοδοι ανάλυσης συνιστωσών. Αυτές είναι: η ιεραρχική ανάλυση, η ανάλυση K-means και η ανάλυση two stage (δύο βημάτων). Η πρώτη χρησιμοποιείται όταν έχουμε μικρό δείγμα, δεύτερη όταν έχουμε μεγάλο δείγμα αλλά πρέπει να προκαθοριστεί το πλήθος των ομάδων και η τρίτη μπορεί να χειρίζεται μεγάλα σύνολα δεδομένων ενώ επιλέγει αυτόματα το πλήθος των ομάδων.

Ιεραρχική ανάλυση

Το καίριο σημείο για τον αλγόριθμο είναι πως θα υπολογιστεί την απόσταση της ομάδας που έχουν γίνει (ή από συγχώνευση άλλων ομάδων είτε αποσυγχώνευση παρατηρήσεων). Σε κάθε παράδειγμα πρέπει να αποφασίσουμε ποια μέθοδος μας συμφέρει να ακολουθήσουμε, για να βγουν πιο έγκυρα και γρήγορα αποτελέσματα. Υπάρχουν πολλές μέθοδοι, όπως:

• Η μέθοδος του κοντινότερου γείτονα (nearest neighbor or single linkage)

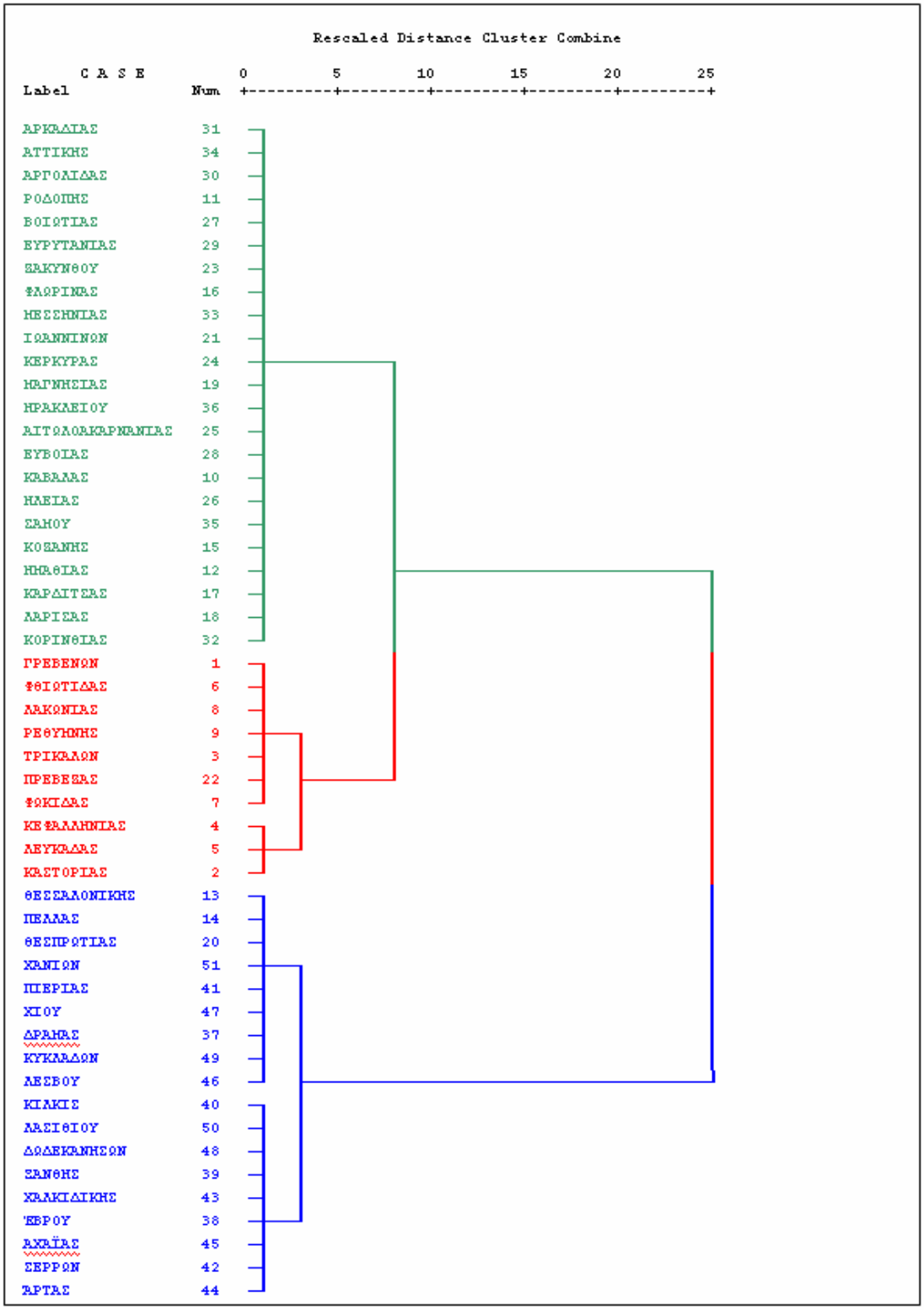
• Η μέθοδος του μακρινότερου γείτονα (furthest neighbor or complete linkage)

• Η μέθοδος του μέσου ανάμεσα στις ομάδες (Average between groups)

• Η μέθοδος του μέσου μέσα στις ομάδες (Average within groups)

• Η μέθοδος του Ward's και άλλες

Από αυτές η πιο απλή είναι η μέθοδος του κοντινότερου γείτονα η οποία όμως είχε το μειονέκτημα πως δίνει ομάδες με μεγάλες διαφορές ως προς το μέγεθος τους. Η μέθοδος του Ward έχει το πλεονέκτημα ότι μας δίνει περίπου ισοπληθείς ομάδες και για αυτό καλό είναι να την προτιμάμε. Στην περίπτωσή αυτή θα χρησιμοποιήσουμε το συνδυασμό των μεθόδων που δίνει την καλύτερη ομαδοποίηση και είναι η χρήση του μέτρου της τετραγωνικής ευκλείδειας απόστασης καθώς και τη μέθοδο του Ward για την ένωση των ομάδων. Έχουμε λοιπόν:



Από το παραπάνω δενδρόγραμμα δημιουργούνται τρεις ομάδες.

Βλέπουμε ότι η ομάδα των Νομών με το πράσινο χρώμα να είναι η πολυπληθέστερη (23 Νομοί). Το ποσοστό των μαθητών που φοιτούν στα ΤΕΕ (34,73%) είναι σχεδόν ίδιο με το πανελλαδικό ποσοστό (34%)ν ενώ η τυπική απόκλιση είναι μικρή (1,43%) κάτι που σημαίνει ότι έχουμε μια πολύ ομοιογενή ομάδα. Ακόμα η επόμενη ομάδα περιλαμβάνει του Νομούς με το μικρότερο ποσοστό μαθητών που φοιτούν στα ΤΕΕ (27,64%). Η ομάδα με το μεγαλύτερο ποσοστό μαθητών που φοιτούν στα ΤΕΕ εμφανίζει ποσοστό της τάξης του 42,32%.

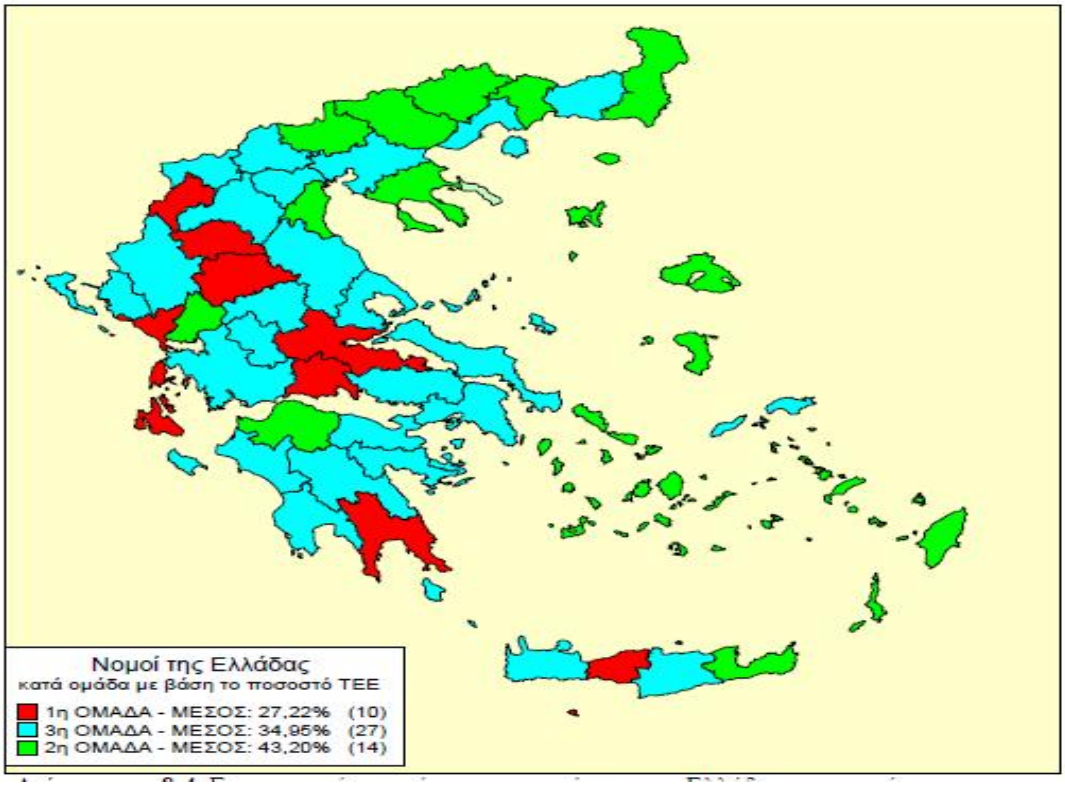
Μέθοδοσ K-means

Εφαρμόζεται η μέθοδος K-Means επιλέγοντας να μας κατασκευάσει 3 ομάδες για να πάρουμε συγκριτικά αποτελέσματα με την Ιεραρχική Ομαδοποίηση. Για κάθε παρατήρηση υπολογίστηκε η Ευκλείδεια απόστασή της από τα κέντρα των ομάδων και την κατατάσσουμε στην ομάδα που είναι πιο κοντά. Αφού κατατάξουμε όλες τις παρατηρήσεις, υπολογίζουμε εκ νέου τα κέντρα και η διαδικασία επαναλαμβάνεται μέχρις ότου δεν υπάρχουν διαφορές ανάμεσα σε δυο διαδοχικές επαναλήψεις.

Το αποτέλεσμα της μεθόδου αυτής δίνει ότι η τρίτη ομάδα των Νομών είναι η πολυπληθέστερη (27 Νομοί). Το ποσοστό των μαθητών που φοιτούν στα ΤΕΕ (35%) είναι σχεδόν ίδιο με το πανελλαδικό ποσοστό (34%). Ακόμα η πρώτη ομάδα περιλαμβάνει του Νομούς με το μικρότερο ποσοστό μαθητών που φοιτούν στα ΤΕΕ (28%). Η ομάδα με το μεγαλύτερο ποσοστό μαθητών που φοιτούν στα ΤΕΕ είναι η δεύτερη (44%).

Μέθοδος Two Stage

Η μέθοδος αυτή όπως έχουμε αναφέρει, προσδιορίζει από μόνη της τον βέλτιστο αριθμό ομάδων. Στη συγκεκριμένη περίπτωση προκύπτει ο αριθμός των ομάδων να είναι τρεις.



Από το παραπάνω γράφημα παρατηρούμε τρεις ομάδες με ποσοστά της τάξης του 27,2%, 34,9% και 43,2% αντίστοιχα.

Με βάσει τις τρεις μεθόδους η έρευνα κατέληξε στο ότι οι μέθοδοι K-Means και TwoStage cluster δίνουν ακριβώς την ίδια ομαδοποίηση. Όμως η ομαδοποίηση που έγινε χρησιμοποιώντας το δενδρόγραμμα διαφέρει από τα αποτελέσματα των προηγούμενων δυο μεθόδων στους νομούς Θεσσαλονίκης, Πέλλας, Θεσπρωτίας και Χανίων. Επομένως η έρευνα καταλήγει στο ότι η σωστότερη ομαδοποίηση είναι αυτή που δίνουν οι μέθοδοι K-Means για 3 ομάδες και η μέθοδος TwoStage cl

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Hartigan, J. (1975) Clustering Algorithms. Wiley, New York, NY.
2. Tan Steinbach Kumar, ‘Data Mining, Cluster Analysis: Basic Concepts and Algorithms’, Lecture notes, 1994
3. Tryon, R. C. (1939) *Cluster analysis.* Ann Arbor: Edwards Brothers
4. http://www.statsoft.com/textbook/cluster-analysis/
5. Aldenderfer s. MARK, Blashfield kRoger. Cluster Analysis. Quantitative Applications in the Social Sciences. SAGE publications, INC. , 1984
6. http://stat-athens.aueb.gr/~jbn/courses/diplomatikes/stats_part/Kaltsas%282008%29.pdf
7. http://www.piraeusbank.gr/Documents/internet/Economic_Research/SEE&Egypt_Quarterly/eidiko_thema.pdf
8. Dillon,W., Goldstein, M. (1984). Multivariate Analysis. Methods and Application, John Wiley & Son, New York
9. W. M. Rand (1971). "Objective criteria for the evaluation of clustering methods"
10. Chris Ding and Xiaofeng He. "K-means Clustering via Principal Component Analysis". Proc. of Int'l Conf. Machine Learning (ICML 2004), pp 225-232. July 2004.
11. Tuffery, Stéphane (2011), "9.10 Agglomerative hierarchical clustering", *Data Mining and Statistics for Decision Making*, Wiley Series in Computational Statistics
12. Day, William H. E.; Edelsbrunner, Herbert (1984),"Efficient algorithms for agglomerative hierarchical clustering methods", *Journal of Classification*
13. Everitt B.S. (2002) The Cambridge Dictionary of Statistics, CUP

14. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "Hierarchical clustering" (PDF). The Elements of Statistical Learning (2nd ed.). New York

15. Trevor Hastie, Robert Tibshirani, Jerome Friedman (2001), The Elements of Statistical Learning, Data Mining,Inference and Prediction, Springer ,Series in Statistics.

16. Hall P, Park BU, Samworth RJ (2008). "Choice of neighbor order in nearest-neighbor classification". Annals of Statistics