ΝΔ: 11149
ΚΩΔ: 8072

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Φλώρου Γιαννούλα

2 8 ΜΑΙ. 1999

ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΤΗΣ ΙΔΑΝΙΚΟΤΕΡΗΣ ΜΕΤΡΙΚΗΣ ΑΠΟΣΤΑΣΗΣ ΚΑΙ ΚΑΛΥΤΕΡΟΥ ΤΡΟΠΟΥ ΟΜΑΔΟΠΟΙΗΣΗΣ ΣΤΙΣ ΔΙΑΦΟΡΕΣ ΜΕΘΟΔΟΥΣ ΤΗΣ ΑΥΤΟΜΑΤΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΚΑΤΑ ΑΥΞΟΥΣΑ ΙΕΡΑΡΧΙΑ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
ΠΟΥ ΥΠΟΒΛΗΘΗΚΕ ΣΤΟ ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΤΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΜΑΚΕΔΟΝΙΑΣ

ΘΕΣΣΑΛΟΝΙΚΗ 1997

Αφιερωμένο στον πατέρα μου και
στη μνήμη της μητέρας μου

Δεν έχουμε υποθέσεις. Εφαρμόζουμε το αντίστροφο σχήμα: Δεν είναι οι υποθέσεις που θεμελιώνουν την έρευνα αλλά είναι η έρευνα που εφευρίσκει νέες υποθέσεις. (COHEN Daniele, Les genes de l'espoir, 1993)

Ευχαριστίες

Παρουσιάζοντας το τελικό κείμενο, θα ήθελα να ευχαριστήσω όλους όσους συνέβαλλαν στην πραγματοποίησή του.

Ευχαριστώ ολόψυχα τον επιβλέποντα καθηγητή μου και πολύτιμο σύμβουλο, κ. Γιάννη Παπαδημητρίου, χωρίς την βοήθεια του οποίου δεν θα υπήρχε η παρούσα διατριβή. Εκφράζω την ευγνωμοσύνη μου, τόσο για την επιστημονική όσο και για την ανθρώπινη βοήθειά του, για την συνεχή παρακολούθηση και καθοδήγηση, για το κουράγιο και τις ιδέες που μου έδωσε σε όλη την διαδικασία για την εκπόνιση της διατριβής μου.

Επίσης ευχαριστώ θερμά τα άλλη δύο μέλη της τριμελούς Συμβουλευτικής επιτροπής, τον καθηγητή κ. Κωνσταντίνο-Κλαύδιο Τσούρο του Πανεπιστημίου Μακεδονίας και τον επίκουρο καθηγητή κ. Αλέξανδρο Καράκο του Πολυτεχνείου Ξάνθης για την πολύτιμη βοήθειά τους και τις χρήσιμες συμβουλές τους, σε όλα τα στάδια της έρευνας.

Θα ήταν παράλειψη να μην ευχαριστήσω όλα τα μέλη του τμήματος Εφαρμοσμένης Πληροφορικής, και ιδιαίτερα τους φίλους που με ενθάρρυναν με την κατανόηση και ηθική συμπαράστασή τους. Ακόμη ευχαριστώ όλους τους ανθρώπους του Πανεπιστημίου Μακεδονίας για την φιλική τους διάθεση να συζητήσουν οτιδήποτε με απασχολούσε.

Ευχαριστώ ακόμη όλους εκείνους που μου παρείχαν την τεχνική βοήθεια και τις γνώσεις, για την όσο γίνεται καλύτερη παρουσίαση του κειμένου, των σχημάτων και του προγράμματος. Βέβαια, σαν συγγραφέας της παρούσας μελέτης, είμαι αποκλειστικά υπεύθυνη, για ότι λάθη, παραλείψεις ή ασάφειες μπορεί να υπάρχουν στο κείμενο.

Τέλος, εκφράζω τις ευχαριστίες μου στα μέλη της εφταμελούς εξεταστικής επιτροπής για τον χρόνο που αφιέρωσαν στη μελέτη και κριτική της διατριβής.

ΦΛΩΡΟΥ ΓΙΑΝΝΟΥΛΑ

ΘΕΣΣΑΛΟΝΙΚΗ ΦΕΒΡΟΥΑΡΙΟΣ 1997

Πίνακας Περιεχομένων

ΣΥΝΤΟΜΗ ΠΑΡΟΥΣΙΑΣΗ ΔΙΑΤΡΙΒΗΣ

Εισαγωγή

ΑΝΤΙΚΕΙΜΕΝΟ: ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

1. Γενικά - Θεωρητική προσέγγιση 15

2. Σκοπός έρευνας 20

3. Διάρθρωση διατριβής 21

Κεφάλαιο 1

ΚΥΡΙΩΤΕΡΕΣ ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

1. Ανάλυση σε κύριες συνιστώσες (Analyse en Composantes Pprincipales- A.C.P.) 25

1.1 Εισαγωγή 25

1.2 Βασικές αρχές 27

1.3 Διαδικασία 30

1.4 Νέφος n σημείων στις p διαστάσεις 32

1.5 Νέφος p σημείων στις n διαστάσεις 33

1.6 Εύρεση των κυρίων συνιστωσών 34

1.7 Ερμηνεία των αποτελεσμάτων 34

1.8 Ποιότητα αναπαράστασης 37

2. Παραγοντική ανάλυση των αντιστοιχιών (Analyse Factorielle des Correspondances-A.F.C.) 38

2.1 Εισαγωγή 38

2.2 Μετασχηματισμοί των δεδομένων 40

2.3 Ομοιότητα μεταξύ γραμμών ή στηλών 41

2.4 Ανάλυση πίνακα για πολλές ποιοτικές μεταβλητές 41

2.4.1 Παράδειγμα 43

2.5 Ερμηνεία των αποτελεσμάτων 44

3. Ταξινόμηση (classification) 46

4. Επίλογος 47

Κεφάλαιο 2

ΠΕΡΙΓΡΑΦΗ ΜΕΤΑΒΛΗΤΩΝ- ΜΕΤΑΤΡΟΠΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΠΙΝΑΚΕΣ ΚΑΤΑΛΛΗΛΟΥΣ ΝΑ ΔΕΧΘΟΥΝ ΕΠΕΞΕΡΓΑΣΙΑ ΜΕ ΜΕΘΟΔΟΥΣ ΤΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ - ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

1. Εισαγωγή 49

2. Ορισμοί 50

3. Πίνακες δεδομένων 51

4. Πίνακες μεταβλητών 53

4.1 Κατασκευή πίνακα Burt 56

5. Επεξεργασία πίνακα συμπτώσεων 57

5.1 Υπό συνθήκη κατανομή μιας μεταβλητής σε σχέση με μια άλλη 57

5.1.1 Παράδειγμα 57

6. Έλεγχος συσχέτισης-ανεξαρτησίας μεταβλητών 59

7. Πρόταση δημιουργίας πίνακα Burt υπό συνθήκη σχετικών συχνοτήτων 60

8. Γενίκευση του πίνακα Burt για ποσοτικές μεταβλητές 62
8.1 Παράδειγμα 63

9. Τρισδιάστατοι πίνακες 65

9.1 Μεθοδολογία για την επεξεργασία τρισδιάστατου πίνακα 67
9.1.1 Ανάλυση των τριών περιθωριακών πινάκων 68
9.1.2 Ανάλυση όλων των πινάκων-φέτες 72
9.1.3 Συμπεράσματα 73

9.2 Εφαρμογή-Στατιστική επεξεργασία του τρισδιάστατου πίνακα δεδομένων και τρόπος ερμηνείας των αποτελεσμάτων 75
9.2.1 Εισαγωγή 75
9.2.2 Ανάλυση όλων των τμημάτων 76
9.2.3 Ανάλυση του πίνακα άθροισμα 77
9.2.4 Ανάλυση πίνακα άθροισμα με όλα τα σημεία των 5 τμημάτων 79
9.2.5 Συμπεράσματα εφαρμογής 80

10. Συμπεράσματα 82

Κεφάλαιο 3

ΜΕΤΡΙΚΕΣ ΚΑΙ ΔΕΙΚΤΕΣ ΑΠΟΣΤΑΣΗΣ-ΟΜΟΙΟΤΗΤΑΣ

1. Αποστάσεις για αντικείμενα 83

1.1 Αντικείμενα που χαρακτηρίζονται από ποιοτικές μεταβλητές 84
1.1.1 Εισαγωγή 84
Διατεταγμένες ποιοτικές μεταβλητές 85
Μη διατεταγμένες ποιοτικές μεταβλητές 86
Σύγκριση των τριών μετρικών 86
1.1.2. Υπολογισμός αποστάσεων 87
α. Πλήρης πίνακας μιας μεταβλητής χωρίς επαναλήψεις (ανόμοια αντικείμενα) 87
β. Πίνακας μιας μεταβλητής με επαναλήψεις 88
γ. Πίνακας δύο μεταβλητών, χωρίς επαναλήψεις. 89
δ. Πίνακας m μεταβλητών, χωρίς επαναλήψεις. 90
ε. Πίνακας m μεταβλητών, με επαναλήψεις 91
1.1.3. Σύγκριση ευκλείδειας - Jaccard 92
Περίπτωση που τα αντικείμενα μπορούν να ανήκουν σε περισσότερες από μία κλάσεις της ίδιας μεταβλητής 92
1.1.4. Συμπεράσματα 93

1.2 Αντικείμενα που χαρακτηρίζονται από ποσοτικές μεταβλητές 94

1.3 Απόσταση για αντικείμενα με διαφορετικού τύπου μεταβλητές 95

2. Αποστάσεις για μεταβλητές 96

2.1 Εισαγωγή 96

2.2 Ποσοτικές μεταβλητές 96

2.3 Ποιοτικές μεταβλητές 99
Πίνακες συμπτώσεων πολλών μεταβλητών (Burt) 101

2.4 Δυαδικές μεταβλητές 102

2.5 Διαφορετικού τύπου μεταβλητές 103

2.6 Πορεία μετά τον υπολογισμό των αποστάσεων (τρόπος υπολογισμού αποστάσεων μεταξύ ομάδων) 103

2.7 Συμπεράσματα 105

Κεφάλαιο 4

ΤΑΞΙΝΟΜΗΣΗ ΑΡΧΕΣ-ΜΕΘΟΔΟΙ

1. Εισαγωγή-Βασικές αρχές 107

2. Σχηματική απεικόνιση της ταξινόμησης 111

3. Στάδια ταξινόμησης 112

4. Επίδραση των δεδομένων στην επιλογή μετρικής 112

5. Μετρικές για την ένωση δύο ομάδων 113

6. Διαδικασία δημιουργίας της ιεραρχίας 116

6.1 Παράδειγμα 116

7. Δενδρόγραμμα ιεραρχίας 123

8. Οι σχηματιζόμενοι διαμελισμοί 125

9. Τομή του δενδρογράμματος 128

10. Κριτήριο για κάθε διαμελισμό σε r κλάσεις 128

11. Κριτήριο συνολικής ιεραρχικής ταξινόμησης 129

12. Επιλογή διαμελισμού για συγκεκριμένο πλήθος κλάσεων 132

13. Παράδειγμα: Ταξινόμηση 19 χωρών ως προς τις αιτίες θανάτου (ευκλείδεια μετρική) . 133

14. Παράδειγμα: Ταξινόμηση των 10 περιοχών της Ελλάδας ως προς την εκπαίδευση (X^2 μετρική) 136

15. Ταξινόμηση μεταβλητών 140

16. Συμπεράσματα 142

Κεφάλαιο 5

ΕΡΜΗΝΕΙΑ ΤΟΥ ΔΕΝΔΡΟΓΡΑΜΜΑΤΟΣ ΤΗΣ ΤΑΞΙΝΟΜΗΣΗΣ (ΜΕΘΟΔΟΣ VACOR)

1. Εισαγωγή 145

2. Ερμηνεία δενδρογράμματος ταξινόμησης 145

3. Βασικές αρχές 146

4. Υπολογισμοί δεικτών (Βοηθήματα εμηνείας του δενδρογράμματος) 148

5. Τα δίπολα 151

6. Τρόπος παρουσίασης των αποτελεσμάτων του Vacor 152

7. Στάδια της ερμηνείας του δενδρογράμματος 154

7.1 Παράδειγμα ερμηνείας του δενδρογράμματος 154

8. Ομαδοποιημένες μεταβλητές 155

9. Συμπεράσματα 156

Κεφάλαιο 6

ΣΥΝΕΠΑΓΩΓΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΜΕΤΑΞΥ ΔΥΑΔΙΚΩΝ Η ΠΟΙΟΤΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ

1. Εισαγωγή-Ανάγκη χρήσης Συνεπαγωγικής Ταξινόμησης 157

2. Μέτρηση της συνεπαγωγής μεταξύ δύο δυαδικών μεταβλητών 158

3. Ορισμοί - Μαθηματική Ανάλυση 160

3.1 Ιδιότητες 162

3.2 Σχέση της έντασης συνεπαγωγής με άλλα μέτρα συσχέτισης 163

4. Αριθμητικές Μεταβλητές 164

4.1 Παράδειγμα 164

5. Αναπαράσταση με γράφημα 166

6. Ανάλυση και ερμηνεία Συνεπαγωγικού Γραφήματος 166

7. Ομάδες μεταβλητών 167

7.1 Μέτρηση της συνοχής 168

7.2 Συνεπαγωγή μεταξύ δύο ομάδων μεταβλητών 169

8. Αλγόριθμος ιεραρχικής ταξινόμησης βασισμένος στην συνοχή 170

8.1 Παράδειγμα 171

9. Προσδιορισμός των σημαντικών κόμβων 173

10. Συμβολή των αντικειμένων στους κόμβους 173

10.1 Διάνυσμα έντασης συνεπαγωγής 174

10.2 Συνεισφορά αντικειμένου σε κόμβο 174

11. Συμπεράσματα 175

Κεφάλαιο 7

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΤΗΝ ΣΥΜΒΟΛΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (SYMBOLIC DATA ANALYSIS)

1. Γενικά 177

2. Εισαγωγή 178

3. Ορισμός συμβολικών αντικειμένων 178

4. Δεδομένα φωτογραφιών 180

5. Αρχικός διαμελισμός 182

6. Ιεραρχική ταξινόμηση 182

7. Μετρικές - Μέτρα ή δείκτες ομοιότητας 183

7.1 Νέο μέτρο ομοιότητας 184

7.2 Εφαρμογή του νέου μέτρου ομοιότητας για pixels 186

8. Αποτελέσματα 187

9. Συμπεράσματα 188

Κεφάλαιο 8

ΑΝΤΙΣΤΟΙΧΙΣΗ ΑΚΕΡΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ ΣΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΙΕΡΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ

1. Τοποθέτηση του προβλήματος 189

2. Εισαγωγή 190

3. Κριτήριο βασισμένο στη διατήρηση της αδράνειας (διασποράς) 190

3.1 Υπόθεση αριθμητικής προόδου 191

4. Αδράνεια (διασπορά) ενός διαμελισμού δεδομένων 192

5. Κριτήριο βασισμένο στο δενδρόγραμμα της ταξινόμησης 194

5.1 Παράδειγμα 194

6. Συμπεράσματα 196

ΕΠΙΛΟΓΟΣ 197

ΒΙΒΛΙΟΓΡΑΦΙΑ 199

ΠΑΡΑΡΤΗΜΑ
ΠΑΡΟΥΣΙΑΣΗ ΤΟΥ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΤΗΝ ΤΑΞΙΝΟΜΗΣΗ

Γενικά 207

Εγκατάσταση του προγράμματος στον υπολογιστή 207

Διαδικασίες του προγράμματος 208

Εκτέλεση ("τρέξιμο") του προγράμματος 208

Διαδικασία εισαγωγής δεδομένων 209

ΕΙΣΑΓΩΓΗ ΑΠΟ ΠΛΗΚΤΡΟΛΟΓΙΟ 209

ΕΙΣΑΓΩΓΗ ΑΠΟ ΑΡΧΕΙΟ ASC 211

Διαδικασία επεξεργασίας δεδομένων 211

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ 212

ΣΥΝΕΠΑΓΩΓΙΚΗ ΤΑΞΙΝΟΜΗΣΗ 216

ΕΠΕΞΕΡΓΑΣΙΑ ΤΡΙΣΔΙΑΣΤΑΤΟΥ ΠΙΝΑΚΑ 218

ΣΥΝΤΟΜΗ ΠΑΡΟΥΣΙΑΣΗ ΔΙΑΤΡΙΒΗΣ

Σήμερα με τον μεγάλο αριθμό στατιστικών ερευνών ή αυτομάτων συστημάτων καταμέτρησης, έχουμε πάρα πολλά δεδομένα, αλλά όχι αρκετή πληροφορία. Πολλά καθημερινά προβλήματα περιμένουν απάντηση από τον Στατιστικό Επιστήμονα που καλείται να δώσει φωνή στα δεδομένα και να ανακαλύψει τις υπάρχουσες σχέσεις τους.

Η Ανάλυση Δεδομένων υπάγεται στις μεθόδους της Πολυδιάστατης Στατιστικής Ανάλυσης, δηλαδή στις μεθόδους εκείνες που αναλύουν δεδομένα και παρατηρήσεις που περιγράφονται από πολλές και διαφορετικού τύπου μεταβλητές.

Πατέρας της Ανάλυσης Δεδομένων είναι ο Benzecri, ο οποίος ήταν και ο πρώτος που μελέτησε και θεμελίωσε συστηματικά τις μεθόδους της Ανάλυσης Δεδομένων, έχοντας σαν βασική αρχή: "Το μοντέλο οφείλει να προσαρμόζεται στα δεδομένα και όχι το αντίστροφο". Ίδρυσε την λεγόμενη "γαλλική" σχολή της Ανάλυσης Δεδομένων, η οποία δέχεται τα δεδομένα όπως είναι, χωρίς να κάνει καμιά εκ των προτέρων υπόθεση για αυτά.

Τα αρχικά δεδομένα αναλύονται έτσι ώστε να αποκαλυφθούν οι τάσεις τους (δηλαδή οι αιτίες απομάκρυνσης από τη θέση ισορροπίας, από το μέσο όρο), και οι νόμοι στους οποίους υπακούουν. Από κει και πέρα, για την ερμηνεία και εξαγωγή συμπερασμάτων απαιτείται η συνεργασία του Στατιστικού επιστήμονα με τον γνώστη και χρήστη των δεδομένων.

Ας περάσουμε τώρα ειδικότερα στον σκοπό και την έρευνα της παρούσας διατριβής. Ασχολούμαστε με τις μεθόδους της Ανάλυσης Δεδομένων για την επεξεργασία ενός αρχικού συνόλου δεδομένων, όπως αυτό προκύπτει από μια έρευνα.

Παρουσιάζουμε τις μεθόδους της Παραγοντικής Ανάλυσης των Αντιστοιχιών, της Ανάλυσης σε Κύριες Συνιστώσες και αναλυτικότερα ασχολούμαστε με την ταξινόμηση κατ' αύξουσα ιεραρχία. (Αυτόματη ταξινόμηση, Συνεπαγωγική ταξινόμηση και Συμβολική ταξινόμηση).

Για να κατανοήσουμε καλύτερα τη συνοχή της διατριβής, ας θεωρήσουμε ότι διαθέτουμε και θέλουμε να αναλύσουμε τα στοιχεία ενός οποιουδήποτε συνόλου δεδομένων.

Υπάρχουν πολλοί τρόποι παρουσίασης τέτοιων αρχικών δεδομένων. Ο συνηθέστερος και πιο απλός, είναι η παράσταση με ένα πίνακα *αντικείμενα (παρατηρήσεις) x μεταβλητές.*

Άλλοι συχνά χρησιμοποιούμενοι πίνακες, είναι οι πίνακες συχνοτήτων, οι πίνακες συμπτώσεων (Burt) ή ακόμη οι τρισδιάστατοι πίνακες. Καταρχήν

παρουσιάζουμε όλους αυτούς τους πίνακες με τις ιδιότητές τους και τους τρόπους χρήσης τους.

Σε μια στατιστική έρευνα, ξεκινώντας από ένα πίνακα αρχικών δεδομένων αντικείμενα x μεταβλητές, επιθυμούμε να ανακαλύψουμε και να παρουσιάσουμε με εύκολο και κατανοητό τρόπο, τις σχέσεις μεταξύ αντικειμένων, μεταξύ μεταβλητών, καθώς και τις σχέσεις συνδυασμού αντικειμένων-μεταβλητών.

Το πρώτο βήμα για την επιτυχία του σκοπού αυτού, είναι η εύρεση κάποιου μέτρου (μετρικής ή δείκτη απόστασης ή ομοιότητας), για την ποσοτική μέτρηση της σχέσης μεταξύ δύο αντικειμένων ή δύο μεταβλητών.

Αναφέρουμε τα προταθέντα μέτρα, αναλύοντας τις ιδιότητες τους, και προτείνουμε το καταλληλότερο κάθε φορά, ανάλογα με τη φύση των δεδομένων που αναλύουμε και τους σκοπούς που επιδιώκουμε.

Στο τέλος αυτού του σταδίου έχουμε στην διάθεσή μας ένα πίνακα αποστάσεων που "μετράει" τις σχέσεις ανά δύο, αντικειμένων ή μεταβλητών.

Στη συνέχεια, για να βρούμε τις συνολικές σχέσεις αντικειμένων-μεταβλητών, χρησιμοποιώντας τον πίνακα αυτό, προχωρούμε σε μια ανάλυση των δεδομένων. (Παραγοντική Ανάλυση, Ανάλυση σε Κύριες Συνιστώσες, Ιεραρχική Ταξινόμηση, κ.α. ή συνδυασμό των αναλύσεων αυτών).

Ειδικότερα με την Ταξινόμηση, επιθυμούμε να διακρίνουμε τις υπάρχουσες ομάδες δεδομένων, που περιλαμβάνουν αντικείμενα με παρόμοια χαρακτηριστικά καθώς και τις ιεραρχήσεις των ομάδων αυτών. Επίσης αναζητούμε τις αιτίες που δημιουργούν και χαρακτηρίζουν τις ομάδες αυτές.

Υπάρχουν γενικά τρεις κατηγορίες ταξινόμησης:

α) Ταξινόμηση κατ' αύξουσα ιεραρχία

β) Ταξινόμηση κατά φθίνουσα ιεραρχία

γ) Εύρεση μόνο ενός διαμελισμού των δεδομένων σε καθορισμένο αριθμό κλάσεων.

Παρακάτω θεωρήσαμε σκόπιμο να ασχοληθούμε μόνο με την διαδικασία της προς τα πάνω ταξινόμησης. Οι μέθοδοι που εφαρμόσαμε σ' αυτή, μπορούν να εφαρμοσθούν και στις υπόλοιπες κατηγορίες, με κατάλληλες προσαρμογές. Η βασική αρχή της προς τα πάνω ιεραρχικής ταξινόμησης, είναι: Διαμελίζουμε ένα σύνολο δεδομένων σε όσο το δυνατό ομοιογενείς ομάδες, και κατόπιν οι ομάδες αυτές συνενώνονται διαδοχικά, πάντα με επιδίωξη τον σχηματισμό όσο γίνεται

ομοιογενών ομάδων, μέχρι να φθάσουμε σε μια μόνο ομάδα που περιέχει όλα τα δεδομένα..

Η ταξινόμηση δημιουργεί διαδοχικούς διαμελισμούς των δεδομένων. Οι διαμελισμοί σχηματικά παριστάνονται με ένα δενδρόγραμμα Η διαδικασία αυτή παρουσιάζεται με τον σχηματισμό του δενδρογράμματος. Στη βάση υπάρχουν τα αρχικά δεδομένα, στην κορυφή η μοναδική ομάδα που περιλαμβάνει όλα τα δεδομένα, ενώ κάθε ενδιάμεσος κόμβος αντιστοιχεί σε μια σχηματιζόμενη ομάδα. Η ομοιογένεια του, είναι ανάλογη με το ύψος του κόμβου από τη βάση του δένδρου. Όσο πλησιέστερα στη βάση βρίσκεται ένας κόμβος, τόσο η ομάδα στην οποία αντιστοιχεί, παρουσιάζει μεγαλύτερη ομοιογένεια.

Κατά τη διαδικασία της δημιουργίας του δενδρογράμματος απαιτείται ο υπολογισμός της "απόστασης" μεταξύ δύο ομάδων. Υπάρχουν γι' αυτό, τρεις βασικές "μετρικές":

α) MIN

β) MAX

γ) MOYENNE

Η επιλογή της καλύτερης εξαρτάται από το είδος των δεδομένων. Προτείνουμε ένα κριτήριο που θα βοηθήσει στην επιλογή της καταλληλότερης, το οποίο βασίζεται στον λόγο της ερμηνευόμενης αδράνειας (διασποράς). Καλύτερη μέθοδος για την ιεράρχηση, είναι εκείνη που έχει υψηλότερη τιμή του λόγου ερμηνευόμενης αδράνειας για κάθε πλήθος κλάσεων. Το προτεινόμενο κριτήριο επίσης βοηθάει να επιλεγεί ο διαμελισμός με πλήθος κλάσεων καλύτερα προσαρμοζόμενο στα αρχικά δεδομένα, και επίσης ο καλύτερος διαμελισμός για καθορισμένο πλήθος κλάσεων.

Έχοντας επιλέξει τον καταλληλότερο διαμελισμό των δεδομένων, πρέπει να βρούμε και τις μεταβλητές εκείνες που συμβάλλουν στην δημιουργία των ομάδων (κλάσεων) ενός διαμελισμού, που ερμηνεύουν τις κλάσεις αυτές, αλλά και τις μεταβλητές εκείνες που συμβάλλουν στη διάσπαση των κλάσεων και ερμηνεύουν τις βασικές διαφορές μεταξύ κλάσεων.

Για το σκοπό αυτό, χρησιμοποιούμε τη μέθοδο Vacor, ή και τη Συνεπαγωγική Ταξινόμηση. Με τη μέθοδο Vacor, για κάθε κόμβο (ομάδα) εντοπίζουμε τις μεταβλητές που χαρακτηρίζουν περισσότερο τον κόμβο, καθώς και τις μεταβλητές που είναι υπεύθυνες για το διαχωρισμό κάθε κόμβου σε δύο υποκόμβους. Με τον τρόπο αυτό, μπορούμε να έχουμε τα πρώτα χρήσιμα συμπεράσματα για τη δομή και τις σχέσεις του συνόλου των παρατηρήσεων μας.

Η Συνεπαγωγική Ταξινόμηση εντοπίζει τις σχέσεις συνεπαγωγής σε ένα σύνολο δεδομένων μεταβλητών, εξετάζοντας τις μεταβλητές συνολικά. Προτάθηκε πρόσφατα από τον Gras και είναι πολύ χρήσιμη σε τομείς όπου οι σχέσεις μεταξύ μεταβλητών είναι μη συμμετρικές και επιδιώκουμε την εύρεση εκείνων των μεταβλητών που γίνονται αιτία για να προκαλέσουν τις υπόλοιπες.

Η μέθοδος βασίζεται στη μέτρηση όχι πλέον της απόστασης ή ομοιότητας μεταβλητών αλλά στη μέτρηση της σχέσης συνεπαγωγής μιας μεταβλητής σε κάποια άλλη. Η σχέση αυτή μετριέται με την ένταση συνεπαγωγής που είναι ένα μέτρο πιθανότητας.

Οι ομάδες που προκύπτουν από μια Συνεπαγωγική Ταξινόμηση παρουσιάζονται με ένα δενδρόγραμμα, το οποίο έχει προσανατολισμένους κόμβους.

Η Συνεπαγωγική Ταξινόμηση όπως προαναφέραμε είναι πολύ χρήσιμη σαν συμπληρωματική της μεθόδου vacor για την ερμηνεία των κόμβων μιας ιεράρχησης. Χρησιμοποιώντας την, μπορούμε να εντοπίσουμε τις σχέσεις συνεπαγωγής μεταξύ των σημαντικότερων μεταβλητών κάθε κόμβου όπως αυτές προκύπτουν από τη μέθοδο vacor.

Τέλος όταν τα δεδομένα δε μπορούν να παρασταθούν σωστά από ένα πίνακα αντικείμενα x μεταβλητές με τον κλασσικό τρόπο, χρησιμοποιούμε ένα άλλο τρόπο παρουσίασης για τα αντικείμενα, χρησιμοποιώντας τις λογικές σχέσεις της σύζευξης, του περιέχεσθαι, της ένωσης και της τομής. Αντικείμενα που περιγράφονται από τέτοιες σχέσεις ονομάζονται συμβολικά αντικείμενα και η μέθοδος της ταξινόμησης γι' αυτά, Συμβολική Ταξινόμηση.

Όπως και στην απλή ταξινόμηση, για την μέτρηση της σχέσης μεταξύ αντικειμένων απαιτείται ο υπολογισμός ενός μέτρου απόστασης ή ομοιότητας μεταξύ τους. Προτείνουμε ένα νέο τέτοιο μέτρο για συμβολικά αντικείμενα που περιγράφονται από ποσοτικές μεταβλητές, το οποίο βασίζεται στην μέτρηση και σύγκριση των "διαγωνίων" των αντικειμένων αλλά και στη σύγκριση των τιμών των αντικειμένων για τις μεταβλητές που τα περιγράφουν. Το υπόλοιπο της διαδικασίας της Συμβολικής Ταξινόμησης είναι παρόμοιο με τη διαδικασία της απλής ταξινόμησης.

Μια άλλη πολύ χρήσιμη διαδικασία προκειμένου να αναλύσουμε περαιτέρω τις κλάσεις-ομάδες που προκύπτουν από μια ταξινόμηση, είναι η αντιστοίχηση μιας νέας μεταβλητής στα αντικείμενα κάθε κλάσης, ανάλογα με τα βασικά τους χαρακτηριστικά. Παρουσιάζουμε την μαθηματική ανάλυση για την εύρεση μιας βέλτιστης τέτοιας μεταβλητής (με την έννοια διατήρησης της διασποράς), η οποία

ΤΕΚΜΗΡΙΩΣΗ ΕΘΝΙΚΟ ΒΙΒΛΙΟΘΗΚΗ

βασίζεται στις μεθόδους του γραμμικού προγραμματισμού. Επίσης προτείνουμε ένα εύκολο και πρωτότυπο τρόπο για την εύρεση μιας τέτοιας μεταβλητής, χρησιμοποιώντας τα αποτελέσματα του δενδρογράμματος μιας ταξινόμησης.

Νέες προτάσεις που προτείνονται

- Τρόπος παρουσίασης του πίνακα συμπτώσεων χρησιμοποιώντας τους πίνακες υπό συνθήκη σχετικών συχνοτήτων, και μέτρηση της σχέσης δύο μεταβλητών από τον πίνακα αυτό. (κεφάλαιο 2).
- Γενίκευση του πίνακα Burt για ποσοτικές μεταβλητές. (κεφάλαιο 2).
- Μεθοδολογία για την ανάλυση τριδιάστατων πινάκων. (κεφάλαιο 2).
- Τρόποι για την επιλογή της κατάλληλης μετρικής απόστασης, ανάλογα με το είδος των δεδομένων και τις ανάγκες της έρευνας. (κεφάλαιο 3).
- Κριτήριο για την επιλογή του καλύτερου τρόπου ομαδοποίησης στην Ταξινόμηση κατά αύξουσα ιεραρχία. (κεφάλαιο 4).
- Κριτήριο για την επιλογή βέλτιστου διαμελισμού των δεδομένων. (κεφάλαιο 4).
- Κριτήριο για την επιλογή βέλτιστου διαμελισμού των δεδομένων σε ένα επιθυμητό πλήθος κλάσεων. (κεφάλαιο 4).
- Τρόπος παρουσίασης των αποτελεσμάτων της ταξινόμησης. (κεφάλαιο 5).
- Χρήση της Συνεπαγωγικής Ταξινόμησης για την παραπέρα ανάλυση και ερμηνεία των αποτελεσμάτων της Ταξινόμησης. (κεφάλαιο 6).
- Παρουσίαση των Συμβολικών αντικειμένων και της Συμβολικής Ανάλυσης δεδομένων. (κεφάλαιο 7).
- Χρήση Ιεραρχικής Ταξινόμησης στη Συμβολική Ανάλυση δεδομένων. (κεφάλαιο 7).
- Νέο μέτρο για την μέτρηση της απόστασης μεταξύ συμβολικών αντικειμένων που περιγράφονται από ποσοτικές μεταβλητές. (κεφάλαιο 7).
- Χρήση της Συμβολικής Ταξινόμησης στην επεξεργασία εγχρώμων φωτογραφιών. (κεφάλαιο 7).
- Τρόπος αντιστοίχησης νέας μεταβλητής στις κλάσεις που προκύπτουν από μια ταξινόμηση. (κεφάλαιο 8).

ΕΙΣΑΓΩΓΗ
ΑΝΤΙΚΕΙΜΕΝΟ: ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

1. Γενικά - Θεωρητική προσέγγιση

Στην πρώτη εμφάνισή της η Στατιστική χρησιμοποιήθηκε μόνο για την καταμέτρηση και παρουσίαση στοιχείων. Με την ανάπτυξή της και τη θεσμοθέτησή της σαν επιστήμη προσπάθησε εκτός από την συλλογή και παρουσίαση των στοιχείων να προχωρήσει στην ανάλυσή τους και στην εξαγωγή συμπερασμάτων και προβλέψεων.

Η ανάλυση των στοιχείων γίνεται με την βοήθεια των Μαθηματικών και τα αποτελέσματα εμφανίζονται με όσο γίνεται πιο άμεσο και εποπτικό τρόπο. Η παρουσίαση μιας ή δύο μεταβλητών για τα ίδια δεδομένα είναι εύκολη και πραγματοποιείται με πολλούς τρόπους όπως γραφήματα, πίνακες, διαγράμματα κλπ. Ακόμη η αναζήτηση των σχέσεων μεταξύ δύο μεταβλητών, γίνεται εύκολα με τον υπολογισμό της μεταξύ τους συσχέτισης. Όταν όμως θέλουμε την ταυτόχρονη παρουσίαση και ανάλυση πολλών μεταβλητών, πρέπει να αναζητήσουμε την καταλληλότερη μέθοδο. Τέτοιες μέθοδοι που επιτρέπουν την ταυτόχρονη παρουσίαση και αναζήτηση σχέσεων πολλών ετερογενών ή ομογενών μεταβλητών, είναι οι μέθοδοι της Πολυδιάστατης Στατιστικής Ανάλυσης. [Aderson,1958] Πρόκειται για μεθόδους που επιτρέπουν την ταυτόχρονη ανάλυση μεγάλου όγκου δεδομένων, με την βοήθεια των ηλεκτρονικών υπολογιστών. [Bertier,1975]

Ένας από τους κλάδους της Πολυδιάστατης Στατιστικής Ανάλυσης είναι και η Ανάλυση Δεδομένων. Λέγοντας μέθοδοι Ανάλυσης Δεδομένων, εννοούμε διάφορες μεθόδους που μας επιτρέπουν να ομαδοποιήσουμε, να κατατάξουμε, να οπτικοποιήσουμε και γενικά να περιγράψουμε δεδομένα, τα οποία δεν μπορούμε να τα αντιμετωπίσουμε ικανοποιητικά με άλλες στατιστικές μεθόδους, είτε γιατί είναι τελείως ανομοιογενή, είτε γιατί δεν ικανοποιούν κάποιες συνθήκες απαραίτητες για αυτές τις μεθόδους.

Η Ανάλυση Δεδομένων προέκυψε από την μεγάλη ανάγκη που υπάρχει σε πολλές επιστήμες, για επεξεργασία πάρα πολλών δεδομένων, προκειμένου να καταλήξουμε σε χρήσιμα συμπεράσματα. Μια φράση συνοψίζει το πρόβλημα του καλούνται να επιλύσουν σήμερα οι Στατιστικοί: "*Έχουμε πάρα πολλά δεδομένα αλλά όχι αρκετή πληροφορία*". Τεράστιος όγκος δεδομένων περιμένει να επεξεργαστεί. Τα δεδομένα αυτά προκύπτουν από στατιστικές έρευνες, ή αυτόματα από διάφορα συστήματα καταμέτρησης, όπως bar-code, πιστωτικές ή τηλεφωνικές

κάρτες, και προβλέπεται να αυξηθούν ακόμη περισσότερο στα επόμενα χρόνια. Ο ρόλος του Στατιστικού είναι να δώσει φωνή στα δεδομένα αυτά, να ανακαλύψει το πως τα δεδομένα μπορούν να απαντήσουν στα διάφορα καθημερινά προβλήματα. [Panelist,1996].

Σήμερα που η πληροφορία κυκλοφορεί πολύ εύκολα και γρήγορα, έχουμε μεγάλη ανάγκη από την άμεση επεξεργασία στοιχείων-δεδομένων διαφόρων ερευνών, ερωτηματολογίων, συνεντεύξεων κλπ. Τα δεδομένα αυτά είναι συνήθως ετερογενή, το πλήθος τους είναι συχνά τεράστιο, και επιπλέον δεν ακολουθούν κάποια από τις γνωστές κατανομές πιθανοτήτων. Πώς λοιπόν μπορούμε να έχουμε μια σαφή εικόνα αυτού του μεγάλου ετερογενή όγκου δεδομένων, που πολλές φορές οι υποθέσεις των στατιστικών ότι ακολουθούν κάποιες κατανομές πιθανοτήτων είναι λανθασμένες; [Lagarde,1983] [Droesbeke,1992].

Η Ανάλυση Δεδομένων χαρακτηρίζεται από τη βασική αρχή να μείνουμε όσο γίνεται πλησιέστερα στα αρχικά δεδομένα (χωρίς τροποποιήσεις και a priori υποθέσεις), μετασχηματίζοντας τα το λιγότερο δυνατό. Προσπαθούμε δηλαδή να αναλύσουμε το φαινόμενο, όπως αυτό παρουσιάζεται μόνο από τον πίνακα δεδομένων. Χαρακτηριστική είναι η φράση του J.P. Benzecri (1973), ενός από τους πρώτους και σημαντικότερους ερευνητές της Ανάλυσης Δεδομένων: *"Το μοντέλο οφείλει να προσαρμόζεται στα δεδομένα και όχι το αντίστροφο"*. [Diday,1982]. Χρησιμοποιώντας την Ανάλυση Δεδομένων, δεν επιθυμούμε να προσαρμόσουμε τα δεδομένα στα εργαλεία της ανάλυσης, αλλά προσπαθούμε να βρούμε μεθόδους ανάλυσης που να ταιριάζουν το μέγιστο δυνατό, στα δεδομένα. *"Μελετούμε την επιστήμη, αλλά δεν έχουμε υποθέσεις. Εφαρμόζουμε το αντίστροφο σχήμα: δεν είναι η υπόθεση που θεμελιώνει την έρευνα, αλλά η έρευνα είναι εκείνη που θέτει νέες υποθέσεις. Τέτοια τολμήματα προκάλεσαν τις ανακαλύψεις της φυσικής, τους πλανήτες, τα χημικά στοιχεία κλπ. Και οι εφευρέτες τους διατύπωσαν εκπληκτικές υποθέσεις και θεωρίες που άνοιξαν το δρόμο στα μυστικά της φύσης."* [Coen,1993]

Η Ανάλυση Δεδομένων δημιούργησε νέες ευκαιρίες για την μελέτη και ανάλυση στατιστικών ερευνών, και θεμελιώνοντας καινούριες μεθόδους για την εξαγωγή συμπερασμάτων, μείωσε το ενδιαφέρον για πιθανολογικά συμπεράσματα. Ανοίγοντας στους στατιστικούς νέες δυνατότητες πρόσβασης στα δεδομένα, και επιτρέποντας τους να τα "δουν" σε ένα γράφημα, η Ανάλυση Δεδομένων παρέχει επιχειρήματα σ'όσους αμφιβάλλουν για την καλή θεμελίωση των συνηθισμένων υποθέσεων της μαθηματικής στατιστικής: Όταν "παρατηρούμε" ένα νέφος δεδομένων (με την βοήθεια μιας παραγοντικής ανάλυσης), ανακαλύπτουμε ότι δεν υπάρχει σχεδόν ποτέ το ελλειψοειδές σχήμα που έπρεπε να είχαν, αν ακολουθού-

σαν την πολυδιάστατη κανονική κατανομή. Πολύ συχνά έχουν το σχήμα του τετραέδρου, του μανιταριού, βρίσκονται γύρω από μια παραβολή κτλ. [Volle,1978].

Αν θέλουμε να ορίσουμε την Ανάλυση Δεδομένων σε λίγες γραμμές, θα λέγαμε ότι είναι ένα σύνολο μεθόδων, των οποίων η βασική αρχή είναι να ανακαλύψει όλες τις υπάρχουσες σχέσεις μεταξύ των αντικειμένων (ή αλλιώς παρατηρήσεων ή στοιχείων ή ατόμων), μεταξύ των μεταβλητών που τα χαρακτηρίζουν, καθώς και όλες τις σχέσεις μεταξύ αντικειμένων και μεταβλητών. Σκοπός τους είναι να αναδειχθούν οι επικρατέστερες σχέσεις, χωρίς φυσικά να παραληφθεί η περιγραφή και των υπολοίπων. Συνίσταται κυρίως στο να βρούμε τις σχέσεις μεταξύ των αντικειμένων, και να μετρήσουμε τη σημαντικότητα καθεμιάς μεταβλητής στις σχέσεις αυτές.

Η Ανάλυση Δεδομένων είναι ένας κλάδος που κυρίως αναπτύχθηκε στην Γαλλία από το 1960 και μετά. Μεγαλύτερος εκφραστής της ήταν ο Benzecri, ακολούθησαν Diday, Cazes, Roux, Pages, Lerman, Guigou, Jambu και άλλοι. Και Άγγλοι ασχολήθηκαν από πολύ νωρίς με τις πολυδιάστατες μεθόδους Ανάλυσης Δεδομένων. Ο Hotteling στην δεκαετία του 30 έθεσε της βάσεις για την ανάλυση σε κύριες συνιστώσες [Hotteling,1933], αναπτύσσοντας τις προηγούμενες εργασίες των Spearman [Spearman,1904] και Pearson [Pearson,1901]. Στη συνέχεια οι Sneath, Sokal και Anderson [Anderson,1958], [Sneath & Sokal,1972]. Η μεθοδολογία όμως που ακολούθησαν είναι προσανατολισμένη κυρίως στην μαθηματική Στατιστική και όχι στις εφαρμογές. Για την ανάλυση ενός συνόλου δεδομένων, γίνονται πολλές υποθέσεις, για το τι κατανομή ακολουθούν, σε τι νόμους υπακούουν κτλ.

Σε αντίθεση μ'αυτούς, η "γαλλική σχολή" δεν κάνει καμιά υπόθεση για τα δεδομένα. Τα δέχεται όπως έρχονται, και προσπαθεί να εξάγει τις σχέσεις που τυχόν υπάρχουν μεταξύ τους, ή μεταξύ των αντικειμένων που χαρακτηρίζουν, μόνο από τα ίδια τα δεδομένα. Σε ένα κόσμο όπου σχεδόν τα πάντα είναι τυχαία, δεν μπορούμε να ξέρουμε τους νόμους στους οποίους υπακούουν τα φαινόμενα και ίσως δεν είναι βέβαιο ότι οι νόμοι αυτοί υπάρχουν πράγματι. Έτσι οι στατιστικοί επιστήμονες της σχολής αυτής, αρνούμενοι να κάνουν αμφίβολες υποθέσεις, παρουσιάζουν τα αρχικά δεδομένα και τις τάσεις που εξάγονται απ' αυτά. Αφήνουν στον χρήστη τη φροντίδα και την υπευθυνότητα να εξάγει τις ερμηνείες των γεγονότων και τις συνέπειές τους. [Lagarde,1983]. Για την εξαγωγή των συμπερασμάτων χρειάζεται απαραιτήτως η συνεργασία του έμπειρου στατιστικού με τον χρήστη και γνώστη των ιδιαιτεροτήτων των δεδομένων.

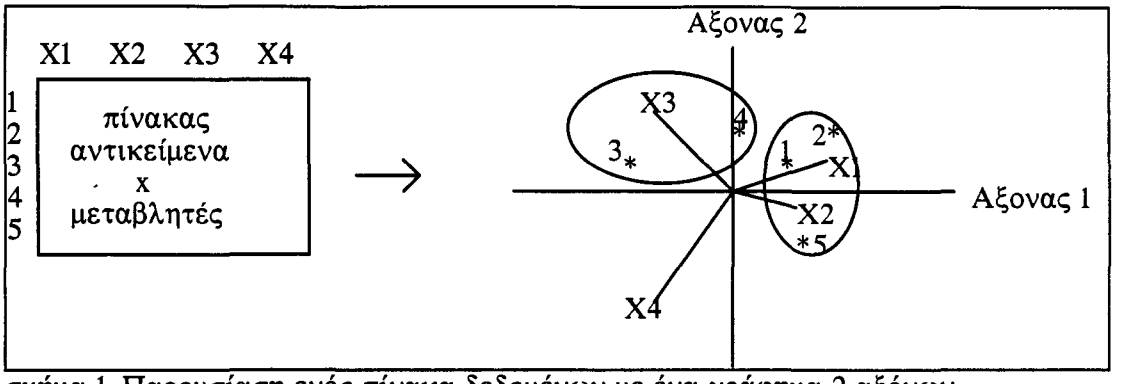
Για να είναι αξιόπιστα τα συμπεράσματα, απαιτείται ένας μεγάλος όγκος δεδομένων, και πολλές αριθμητικές πράξεις. Αυτή ήταν και η μεγαλύτερη δυσκολία στο ξεκίνημα της Ανάλυσης Δεδομένων. Με την μεγάλη ανάπτυξη των υπολογιστών, η δυσκολία αυτή ξεπεράστηκε. Με την βοήθεια των υπολογιστών και των νέων μεθόδων Ανάλυσης Δεδομένων, γίνονται άμεσα ορατές όλες οι υπαρκτές σχέσεις και συνάγονται γρήγορα πολλά συμπεράσματα. Σήμερα έχουν αναγνωριστεί και από άλλες χώρες οι μέθοδοι Ανάλυσης Δεδομένων, έχουν μπει σε μεγάλα πακέτα στατιστικής (SAS, Statgraphics, SPSS), βρίσκουν εφαρμογές σχεδόν σ'όλες τις επιστήμες και περιγράφονται σε διδακτικά βιβλία.

Το πρόβλημα που καλούνται να λύσουν οι υπολογιστές σήμερα, είναι το πρόβλημα της μηχανικής αναγνώρισης και αντιμετώπισης, πληροφοριών που δεν μπορούν να παρασταθούν με αριθμούς ή με μια απλή λογική έκφραση (τεχνητή νοημοσύνη). Π.χ. η διαφορά μεταξύ δύο προσώπων δεν είναι της ίδιας φύσης με την διαφορά μεταξύ δύο δεκαδικών αριθμών ούτε με την διαφορά δύο μηκών. Ένα πρόσωπο βρίσκεται στο χώρο, είναι απτό, μπορούμε να μετρήσουμε αποστάσεις, χρώματα κ.α. αλλά τα μεγέθη αυτά διαφέρουν πάρα πολύ και δεν είναι οι ακριβείς μετρήσεις τους που διαχωρίζουν ένα πρόσωπο από ένα άλλο, αλλά οι αναλογίες μεταξύ των μετρήσεων αυτών. Η μεγαλύτερη δυνατότητα της ανθρώπινης νοημοσύνης είναι η σύνθεση: Βλέποντας ένα μεγάλο όγκο δεδομένων μπορούμε να αναγνωρίσουμε τις βασικότερες τάσεις τους. Η προσπάθεια λοιπόν επικεντρώνεται στο να κατορθώσουμε να αποδείξουμε, ότι αντιλαμβανόμαστε και κατανοούμε αυτό, που ένας καλός γνώστης του αντικειμένου έχει ήδη αντιληφθεί.

Μπορούμε να θεωρήσουμε τις μεθόδους της Ανάλυση Δεδομένων σαν ένα σύνολο μεθόδων για την αυτόματη σύνθεση της πληροφορίας. Διακρίνουμε γενικά δύο τύπους προσέγγισης: [Tenenbaus,1994]

- Η παραγοντική ανάλυση και η ανάλυση σε κύριες συνιστώσες, έχουν σαν σκοπό την αναγνώριση των αξόνων, γύρω από τους οποίους οργανώνονται τα δεδομένα, και την παρουσίαση μιας "γεωμετρικής" εικόνας τους. [Greenacre,1993] [Παπαδημητρίου,1994]
- Η ταξινόμηση διαχωρίζει σε κλάσεις ένα σύνολο δεδομένων, ανάλογα με τις ιδιότητες που κατέχει, και δομεί τις σχέσεις μεταξύ των παρατηρήσεων. [Benzecri,1973] [Diday,1980]

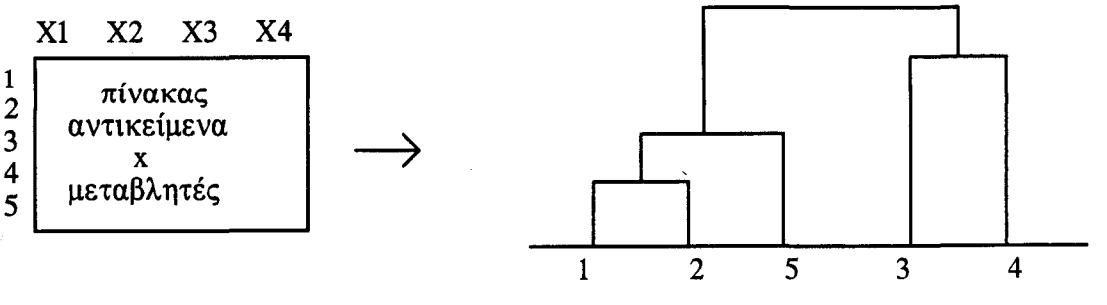
Δύο παραδείγματα των αποτελεσμάτων των μεθόδων φαίνονται στο σχήμα 1 και στο σχήμα 2.



σχήμα 1. Παρουσίαση ενός πίνακα δεδομένων με ένα γράφημα 2 αξόνων

Στο σχήμα 1, εμφανίζονται τόσο τα αντικείμενα όσο και οι μεταβλητές. Ανάλογα με το που βρίσκονται μπορούμε να βγάλουμε τα αντίστοιχα συμπεράσματα:

- Η θέση των αντικειμένων μεταφράζεται σε ομοιότητες μεταξύ τους. Π.χ. Τα αντικείμενα 1 και 2 μοιάζουν πολύ μεταξύ τους και είναι τελείως διαφορετικά από το αντικείμενο 3. [Pages,1976] [Greenacre,1993]
- Η θέση των μεταβλητών στο γράφημα παρουσιάζει τις συσχετίσεις των μεταβλητών μεταξύ τους. Π.χ. οι μεταβλητές X1 και X2 είναι ισχυρά συσχετισμένες θετικά, οι X3 και X4 είναι συσχετισμένες αρνητικά και οι X1, X2 είναι λίγο συσχετισμένες με τις X3, X4. [Bouroche,1980]
- Οι προβολές των αντικειμένων στους άξονες των μεταβλητών αντανακλούν τις τιμές των μεταβλητών στα αντικείμενα. Π.χ. Οι τιμές των X1, X2 είναι μεγάλες για τα αντικείμενα 1 και 2, ενώ είναι μικρές για το αντικείμενο 3. [Escofier,1990]



σχήμα 2. Παρουσίαση ενός πίνακα δεδομένων με ένα δενδρόγραμμα

Στο σχήμα 2, παρουσιάζονται τα αποτελέσματα μιας ιεραρχικής ταξινόμησης, με ένα δενδρόγραμμα. Στο δενδρόγραμμα αυτό, μπορούμε να δούμε άμεσα τις ομάδες αντικειμένων και τις σχέσεις μεταξύ τους. Αντικείμενα που ανήκουν στην ίδια ομάδα (δηλαδή είναι παρόμοια), βρίσκονται στον ίδιο κλάδο του δένδρου, ενώ αυτά που ανήκουν σε διαφορετικές ομάδες (δηλαδή διαφέρουν μεταξύ τους), είναι απομακρυσμένα. Το μέγεθος της ομοιότητας μεταξύ των αντικειμένων

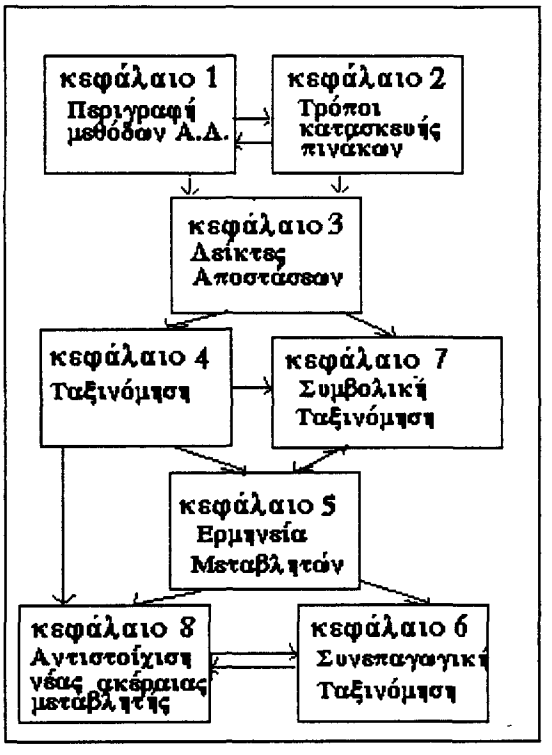
μιας ομάδας, παριστάνεται από το ύψος του κόμβου που αντιστοιχεί στην κλάση αυτή. [Roux,1985] Π.χ. Τα αντικείμενα 1 και 2 μοιάζουν πολύ μεταξύ τους, και μοιάζουν αρκετά με το αντικείμενο 5. Τα αντικείμενα 3 και 4 μοιάζουν μεταξύ τους αλλά λιγότερο απ' ότι τα 1, 2 και 5. Στην ερμηνεία του δενδρογράμματος μπορούμε να συμπεριλάβουμε και τις μεταβλητές όπως θα δούμε στο κεφάλαιο 5. Μπορούμε δηλαδή να εντοπίσουμε τις μεταβλητές εκείνες που συμβάλλουν στην ομαδοποίηση ή τη διαφοροποίηση των ομάδων των αντικειμένων. [Benzecri,1980]

2. Σκοπός έρευνας

Ο σκοπός της παρούσας διατριβής, είναι καταρχήν η μορφοποίηση των δεδομένων, ώστε να είναι κατάλληλα για να υποστούν επεξεργασία με τις μεθόδους της Ανάλυσης Δεδομένων και ειδικότερα με την ταξινόμηση. Ασχολούμαστε ιδιαίτερα με την ταξινόμηση κατ' αύξουσα ιεραρχία, και ζητάμε να προσδιορίσουμε την κατάλληλη μετρική απόστασης για την φύση των δεδομένων μας, καθώς και την μέθοδο ομαδοποίησης που οδηγεί στην ιεράρχηση την καλύτερα προσαρμοζόμενη στα δεδομένα. Επίσης μελετούμε τις νέες τάσεις στην ταξινόμηση, όπως η Συνεπαγωγική Ταξινόμηση και η Ταξινόμηση Συμβολικών αντικειμένων. Τέλος προσδιορίζεται ο τρόπος αντιστοίχησης μιας νέας μεταβλητής, στα αποτελέσματα της ταξινόμησης, ώστε να μπορούμε να εφαρμόσουμε σ' αυτά μια άλλη ανάλυση.

Όλα τα παραπάνω μας οδήγησαν στην κατασκευή ενός εύχρηστου λογισμικού το οποίο επιτρέπει να πραγματοποιούνται μέσω Η/Υ, όλες οι διαδικασίες που αναφέρονται στην διατριβή.

3. Διάρθρωση διατριβής



σχήμα 3. Σχέσεις μεταξύ των κεφαλαίων της διατριβής

Η παρούσα μελέτη περιλαμβάνει δύο μέρη. Το πρώτο μέρος αποτελείται από τα κεφάλαια 1, 2 και 3 και αναφέρεται κυρίως στην Κλασσική Στατιστική και στην παρουσίαση των μεθόδων Ανάλυσης Δεδομένων, ενώ το δεύτερο μέρος αποτελείται από τα κεφάλαιο 4, 5, 6, 7 και 8, που αναφέρονται στα διάφορα είδη ταξινόμησης. Έγινε προσπάθεια για την όσο γίνεται μεγαλύτερη συνοχή μεταξύ των διαφόρων κεφαλαίων. Έτσι ώστε κάθε κεφάλαιο να αποτελεί συνέχεια και επέκταση του προηγούμενου. Ωστόσο ένα διάγραμμα της έρευνας φαίνεται στο παραπάνω σχήμα 3.

Γενικά έχοντας ένα σύνολο δεδομένων και επιθυμώντας να ανακαλύψουμε τις υπάρχουσες σχέσεις, ακολουθούμε τα επόμενα βήματα:

1. Χρησιμοποιώντας την κλασσική Στατιστική, αρχικά περιγράφουμε τις μεταβλητές, με την κατανομή συχνοτήτων τους και τις διάφορες στατιστικές παραμέτρους, καθώς και τις σχέσεις μεταξύ τους ανά δύο. (κεφάλαιο 2).
2. Κατόπιν βρίσκουμε τις σχέσεις μεταξύ των αντικειμένων, χρησιμοποιώντας ένα κατάλληλο δείκτη ή μετρική απόστασης. (κεφάλαιο 3).
3. Με την χρήση μιας μεθόδου Ανάλυσης Δεδομένων, προβαίνουμε σε μια πρώτη Στατιστική Ανάλυση του φαινομένου. (κεφάλαιο 1).
4. Προχωρούμε στην δημιουργία της ιεραρχικής ταξινόμησης των αντικειμένων. Όταν πρόκειται για απλά αντικείμενα, εφαρμόζουμε την κλασσική μέθοδο

ταξινόμησης (κεφάλαιο 4), ενώ όταν πρόκειται για συμβολικά αντικείμενα την μέθοδο της Συμβολικής ταξινόμησης (κεφάλαιο 7).

5. Ερμηνεύουμε τα αποτελέσματα της ταξινόμησης, βρίσκοντας το βέλτιστο πλήθος ομάδων για τα δεδομένα μας, και τις μεταβλητές που χαρακτηρίζουν κάθε ομάδα. (κεφάλαια 4 και 5).
6. Στην περίπτωση που θέλουμε περαιτέρω ανάλυση των δεδομένων χρησιμοποιώντας τις κλάσεις που προκύπτουν από μια ταξινόμηση, απαιτείται ο ορισμός μιας νέας μεταβλητής, αντιπροσωπευτικής των κλάσεων και των μεταβλητών που τις χαρακτηρίζουν (κεφάλαιο 8).
7. Τέλος βρίσκουμε τις τυχόν σχέσεις συνεπαγωγής μεταξύ των σημαντικότερων μεταβλητών που χαρακτηρίζουν τις ομάδες (κεφάλαιο 6).

Η παρούσα διατριβή, ξεκινά με την παρουσίαση της ιστορίας και εξέλιξης της Ανάλυσης Δεδομένων στην εισαγωγή, ενώ στο πρώτο κεφάλαιο περιγράφονται οι βασικές μέθοδοι της.

Στο δεύτερο κεφάλαιο αναφέρονται οι βασικές έννοιες της Περιγραφικής Στατιστικής και περιγράφεται η κατασκευή και ερμηνεία του πίνακα συμπτώσεων. Παρουσιάζεται ένας νέος τρόπος παρουσίασης του πίνακα συμπτώσεων πολλών μεταβλητών, ώστε να είναι εύκολη η ερμηνεία της σχέσης μεταξύ των μεταβλητών, καθώς και η εξαγωγή συμπερασμάτων για την ανεξαρτησία ή την συσχέτισή τους.

Στο τρίτο κεφάλαιο περιγράφονται οι μετρικές και δείκτες αποστάσεων που χρησιμοποιούνται για την εύρεση της προσέγγισης μεταξύ αντικειμένων, προκειμένου να ακολουθήσει κατόπιν μια ταξινόμησή τους. Επίσης γίνεται μια εκτενής αναφορά στις μετρικές και στους διάφορους συντελεστές συσχέτισης που χρησιμοποιούνται για την μέτρηση της σχέσης μεταξύ δύο μεταβλητών. Ακολουθεί μια σύγκριση των μετρικών αυτών προκειμένου να βρεθούν οι συνθήκες εκείνες που θα μας βοηθήσουν να επιλέξουμε κάθε φορά την μετρική, την καλύτερα προσαρμοζόμενη στα δεδομένα μας.

Στο τέταρτο κεφάλαιο περιγράφεται η ιεραρχική ταξινόμηση. Αναλύονται οι διάφορες μέθοδοι ταξινόμησης, περιγράφεται το δενδρόγραμμα των αποτελεσμάτων της, ο τρόπος επιλογής διαμελισμού με συγκεκριμένο ή βέλτιστο αριθμό κλάσεων και παρουσιάζεται ένα κριτήριο για την εύρεση της καλύτερης μεθόδου ανάλογα με τα δεδομένα μας.

Στο επόμενο κεφάλαιο περιγράφεται ο τρόπος ερμηνείας των αποτελεσμάτων της ταξινόμησης αντικειμένων. Εξάγονται οι σημαντικές μεταβλητές για κάθε κόμβο του δενδρογράμματος, για τις κλάσεις του επιλεγόμενου διαμελισμού όπως και για την αιτία διάσπασης κάθε κόμβου. Τέλος προτείνεται ένας νέος

τρόπος παρουσίασης των αποτελεσμάτων της ταξινόμησης, που κάνει όλα τα παραπάνω ορατά γρήγορα και εύκολα.

Στο έκτο κεφάλαιο αναλύεται η μέθοδος της συνεπαγωγικής ταξινόμησης, πολύ χρήσιμης για την εύρεση σχέσεων συνεπαγωγής μεταξύ των μεταβλητών, όσο και μεταξύ των κλάσεων που προκύπτουν από μια ταξινόμηση. Χρησιμοποιείται μετά την ερμηνεία ενός δενδρογράμματος για να βρούμε τις σχέσεις συνεπαγωγής μεταξύ των σημαντικών μεταβλητών για κάθε κλάση.

Στο έβδομο κεφάλαιο περιγράφεται η Συμβολική Ταξινόμηση, η οποία είναι μια γενίκευση της απλής ταξινόμησης, που διαχειρίζεται όχι απλά δεδομένα, αλλά πολύπλοκα και σύνθετα αντικείμενα (τα συμβολικά αντικείμενα). Η συμβολή μας στην μέθοδο αυτή είναι η εύρεση μιας νέας μετρικής απόστασης μεταξύ συμβολικών αντικειμένων.

Στο όγδοο κεφάλαιο παρουσιάζεται ένας πρωτότυπος τρόπος αντιστοίχησης μιας νέας ακέραιας μεταβλητής για τα αντικείμενα που ανήκουν στις δημιουργούμενες κλάσεις μιας ταξινόμησης. Η αντιστοίχηση αυτή στηρίζεται στις ιδιότητες των κλάσεων και είναι τέτοια ώστε να μην χάνουμε σημαντική πληροφορία με την ομαδοποίηση των αντικειμένων.

Τέλος στον επίλογο υπάρχουν τα συμπεράσματα της όλης έρευνας καθώς και οι σκέψεις για μελλοντική έρευνα. Και αυτό γιατί πιστεύουμε ότι η επιστημονική έρευνα δεν σταματά ποτέ. Κάθε προσπάθεια δεν είναι παρά μόνο μια αρχή για τις επόμενες προσπάθειες. Επομένως μια έρευνα θεωρείται πετυχημένη, όταν δίνει κίνητρα, τόσο με τα συμπεράσματα της, αλλά κυρίως με τις ελλείψεις της, σε νέες έρευνες. Στο σημείο αυτό θα ήθελα να αναφέρω τον επίκουρο καθηγητή του Πανεπιστημίου Ιωαννίνων κ. Κων. Καρακώστα, ο οποίος διακρίνοντας την έλλειψη κάποιου τρόπου αξιολόγησης των διαφόρων μεθόδων της Ανάλυσης Δεδομένων, έδωσε την αφορμή για το ξεκίνημα της έρευνας αυτής. Ευχαριστώ επίσης ιδιαίτερα τον καθηγητή μου και επιβλέποντα της παρούσας διατριβής κ. Γιάννη Παπαδημητρίου, για την πολύτιμη βοήθεια του. Τελειώνοντας, θέλω να ελπίζω ότι και η εργασία αυτή θα αποτελέσει ένα έναυσμα για μελλοντικές καλύτερες έρευνες στην ταξινόμηση.

ΚΕΦΑΛΑΙΟ 1
ΚΥΡΙΩΤΕΡΕΣ ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

Στο κεφάλαιο αυτό, θα περιγράψουμε πολύ σύντομα τις κυριότερες μεθόδους Ανάλυσης Δεδομένων. Πρόκειται για μεθόδους που επεξεργάζονται πολυδιάστατα δεδομένα. Δηλαδή δεδομένα που χαρακτηρίζονται από πολλές μεταβλητές. Αναφέρουμε εδώ την Ανάλυση σε Κύριες Συνιστώσες, την Παραγοντική Ανάλυση των Αντιστοιχιών και την Ταξινόμηση. Θα παρουσιάσουμε την ταξινόμηση με περισσότερες λεπτομέρειες στο κεφάλαιο 4.

1. Ανάλυση σε κύριες συνιστώσες (Analyse en Composantes Principales- A.C.P.)

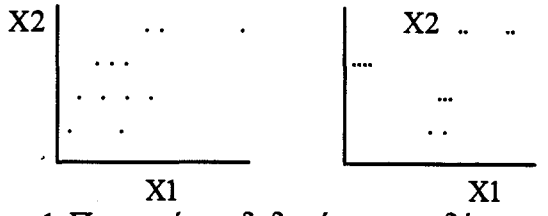
1.1 Εισαγωγή

Η μέθοδος της Ανάλυσης σε Κύριες Συνιστώσες, ξεκίνησε από τον Hotelling, ο οποίος βασίστηκε στις εργασίες του Pearson. Με την μέθοδο αυτή ένα σύνολο μεταβλητών δομείται σύμφωνα με κάποιες συντεταγμένες τους σε ένα άλλο σύνολο αξόνων (composantes principales), που επιλέγονται έτσι ώστε να είναι πιο αντιπροσωπευτικοί. [Bertier,1975] [Παπαδημητρίου,1994]

Σαν σκοπό της έχει να περιγράψει ένα πίνακα δεδομένων ποσοτικών μεταβλητών. Η ανάλυση γίνεται με δυϊκό τρόπο. Οι n παρατηρήσεις εξετάζονται ως προς p σε πλήθος μεταβλητές (p>3), και μπορούν να θεωρηθούν σαν n σημεία στο χώρο των p διαστάσεων, με συντεταγμένες τις τιμές των μεταβλητών, αλλά και οι μεταβλητές μπορούν να θεωρηθούν σαν p σημεία στο χώρο των n διαστάσεων. [Pages,1976]

Προσπαθούμε λοιπόν να παραστήσουμε όσο γίνεται καλύτερα τα σημεία αυτά, (χάνοντας όσο το δυνατό λιγότερη πληροφορία), σε ένα χώρο πολύ λιγότερων διαστάσεων. Προσπαθούμε να πετύχουμε την οπτικοποίηση των σημείων αυτών στο χώρο των τριών ή δύο διαστάσεων, σε συνάρτηση με τις προσεγγίσεις μεταξύ τους. Με την μείωση των διαστάσεων κατορθώνουμε να διακρίνουμε τις σχέσεις μεταξύ των παρατηρήσεων, τις σχέσεις μεταξύ των μεταβλητών, καθώς και την επίδραση των μεταβλητών στις παρατηρήσεις.

Όταν τα δεδομένα μας δεν αφορούν παρά μόνο δύο μεταβλητές X1 και X2, είναι εύκολο να παραστήσουμε σε ένα γράφημα το σύνολο των δεδομένων. Κάθε αντικείμενο i, έχει δύο συντεταγμένες $X1_i$ και $X2_i$. Παριστάνοντας όλα τα σημεία στο επίπεδο που σχηματίζουν οι άξονες X1 και X2 (σχήμα 1), μπορούμε να μελετήσουμε τη σχέση των δύο μεταβλητών αν υπάρχει ή να ομαδοποιήσουμε γειτονικά αντικείμενα.



σχήμα 1. Παρουσίαση δεδομένων για δύο μεταβλητές X1 και X2

Αν υπάρχουν 3 μεταβλητές η οπτική αναπαράσταση είναι ακόμη δυνατή, στο χώρο των 3 διαστάσεων. Αλλά όταν ο αριθμός p των μεταβλητών είναι μεγαλύτερος ή ίσος του 4 η οπτική αναπαράσταση γίνεται αδύνατη.

Η μέθοδος της Ανάλυσης σε Κύριες Συνιστώσες (Analyse en Composantes Principales-A.C.P.), όπως και η Παραγοντική Ανάλυση των Αντιστοιχιών (Analyse Factorielle des Correspondances-A.F.C.), συνίσταται στην μετάβαση από ένα χώρο p διαστάσεων όπου βρίσκονται τα δεδομένα, στο χώρο των 3 ή 2 διαστάσεων, ώστε να είναι δυνατή η οπτική παρουσίαση, αλλά και να απεικονίζονται όσο γίνεται καλύτερα τα αρχικά δεδομένα.

Ο πίνακας δεδομένων που μελετάμε, δεν είναι πίνακας συμπτώσεων. Περιέχει ποσοτικά δεδομένα τα οποία δεν είναι της ίδιας τάξης μεγέθους, (π.χ. μια μεταβλητή μετριέται σε χιλιάδες ενώ μια άλλη σε εκατοστά), ούτε και έχουν ίδιες μονάδες μέτρησης. Αυτή είναι και η διαφορά της A.C.P. από την A.F.C.

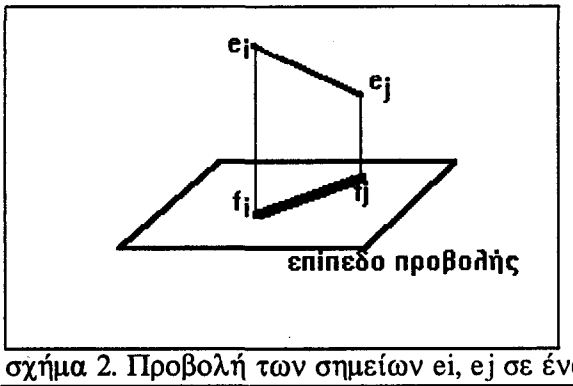
Πίνακας 1

Μαθητές	Μαθηματικά	Φυσική	Γαλλικά	Αγγλικά	Γυμναστική
1	75	60	50	55	70
2	90	60	40	50	30
3	30	25	80	90	40
4	35	35	40	30	90
5	80	85	90	85	75
6	50	95	60	70	35
7	15	20	90	80	50
8	75	80	35	45	80
9	80	90	85	75	90
10	70	50	65	60	60
11	20	30	40	35	85
12	75	60	65	65	20
13	55	60	65	30	80
14	95	90	45	55	50
15	35	40	80	90	45
Σύνολα	880	880	930	915	900
μέση τιμή	58.67	58.67	62	61	60
τυπ. απόκλιση	26.15	25.17	19.62	20.63	23.14

Στον πίνακα 1, κάθε αντικείμενο χαρακτηρίζεται από 5 μεταβλητές (βαθμούς σε 5 μαθήματα). Τα 15 αντικείμενα (μαθητές) σχηματίζουν ένα νέφος σημείων (δεν είναι βέβαια άμεσσα ορατό), στο χώρο των 5 διαστάσεων, καθώς υπάρχουν 5 συντεταγμένες. Την πλήρη εικόνα του φαινομένου, μπορούμε να οπτικοποιήσουμε αν μπορούσαμε να το παρουσιάσουμε στο χώρο των 5 διαστάσεων.

Υποθέτουμε ωστόσο ότι μπορούμε να δούμε τα 15 αντικείμενα σε ένα επίπεδο. Η αναπαράσταση που παίρνουμε διαφέρει αρκετά από την πραγματικότητα. Οι αποστάσεις μεταξύ των 15 αντικειμένων στο επίπεδο δεν μπορούν να είναι όλες ίδιες με τις αποστάσεις μεταξύ των 15 αντικειμένων στο χώρο των 5 διαστάσεων. Θα υπάρχουν λοιπόν αποκλίσεις τις οποίες προσπαθούμε να ελαχιστοποιήσουμε.

Γεωμετρικά η αναπαράσταση στο επίπεδο επιτυγχάνεται, προβάλλοντας τα αντικείμενα στο επίπεδο αυτό, όπως φαίνεται στο παρακάτω σχήμα 2. Τα σημεία e_i, e_j, παριστάνουν δύο αντικείμενα στο χώρο, με απόσταση μεταξύ τους $d(e_i,e_j)$. Οι προβολές τους στο επίπεδο είναι f_i, f_j αντίστοιχα. Η απόσταση των δύο προβολών στο επίπεδο, αντιπροσωπεύεται από το ευθύγραμμο τμήμα f_if_j.

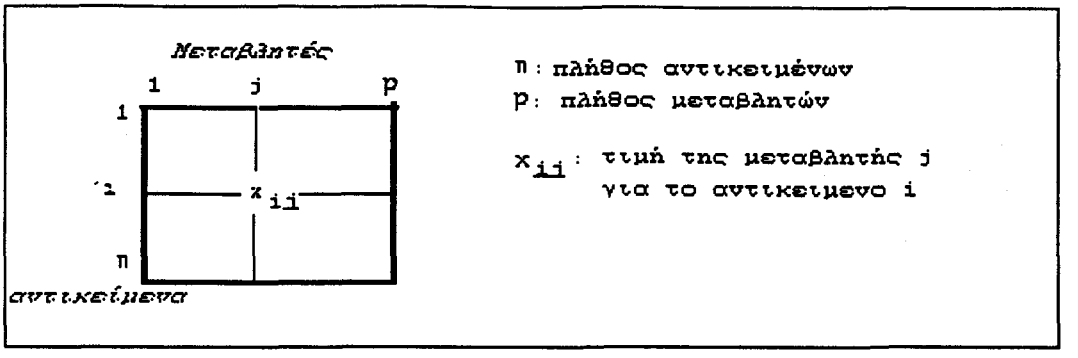


σχήμα 2. Προβολή των σημείων ei, ej σε ένα επίπεδο

Πρέπει να διαλέξουμε σαν επίπεδο προβολής, αυτό στο οποίο το σύνολο των αποστάσεων, θα διατηρείται καλύτερα. Καθώς με την προβολή πάντα έχουμε $d(f_i,f_j)\leq d(e_i,e_j)$, θα θεωρήσουμε σαν κριτήριο, την μεγιστοποίηση, του μέσου των τετραγώνων των αποστάσεων, μεταξύ των προβολών f_1, f_2,...f_n. Δηλαδή καλύτερο επίπεδο προβολής είναι το επίπεδο για το οποίο έχουμε max $\Sigma d^2(f_i,f_j)$. [Escofier,1990]

1.2 Βασικές αρχές

Μελετάμε ένα σύνολο n δεδομένων ατόμων ή αντικειμένων για p μεταβλητές. Οι μεταβλητές είναι ποσοτικές, δηλαδή η τιμή κάθε μεταβλητής για κάθε αντικείμενο είναι ένας αριθμός.



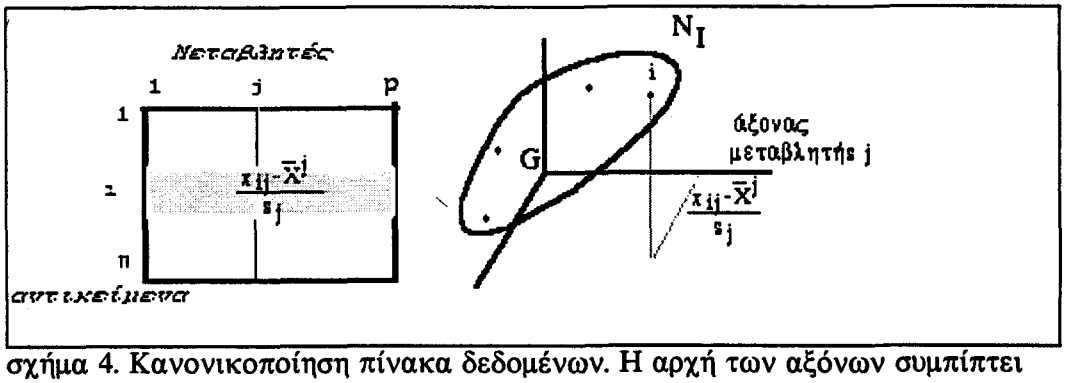
σχήμα 3. Πίνακας δεδομένων nxp

Συμβολίζουμε με X τον πίνακα nxp των δεδομένων, με x_{ij} την τιμή της j μεταβλητής για την i παρατήρηση. Κάθε στήλη του πίνακα αυτού αντιστοιχεί σε μια μεταβλητή X^j. (σχήμα 3). Τα στοιχεία του πίνακα αυτού είναι ανομοιογενείς ποσότητες. Δηλαδή κάθε μεταβλητή είναι δυνατό να μετριέται με διαφορετικές μονάδες μέτρησης.

Σε κάθε αντικείμενο αντιστοιχεί ένα "βάρος" p_i. Συνήθως $p_i=1$. Συμβολίζουμε ακόμη:

$\overline{X}^j$= την μέση τιμή της μεταβλητής X^j
s_j^2= τη διακύμανση της μεταλητής X^j
s_{jk}= τη συνδιακύμανση των X^j και X^k

Για να μπορέσουμε να μελετήσουμε ένα τέτοιο πίνακα δεδομένων, θα πρέπει να προσαρμόσουμε την κάθε μεταβλητή στην τυπική κανονική κατανομή N(0,1). Πρώτα λοιπόν κεντροποιούμε τις τιμές της κάθε μεταβλητής, αφαιρώντας απ' αυτές την αντίστοιχη μέση τιμή τους και στη συνέχεια τις προσαρμόζουμε στην N(0,1), διαιρώντας την διαφορά $X_{ij}-\overline{X}^j$ με την τυπική απόκλιση s_j (σχήμα 4). Το κέντρο βάρους G των αντικειμένων μετά την κανονικοποίηση είναι η αρχή των νέων αξόνων, που ονομάζονται παράγοντες.



σχήμα 4. Κανονικοποίηση πίνακα δεδομένων. Η αρχή των αξόνων συμπίπτει με το κέντρο βάρους G του νέφους

Ο πίνακας 1 μετά την κανονικοποίηση, μετατρέπεται στον πίνακα 2.

Πίνακας 2

Μαθητές	Μαθηματικά	Φυσική	Γαλλικά	Αγγλικά	Γυμναστική
1	0,577	0,094	-0,566	-0,242	0,566
2	1,108	0,094	-1,037	-0,444	-1,697
3	-1,014	-2,381	0,849	1,172	-1,131
4	-0,837	-1,673	-1,037	-1,253	1,697
5	0,754	1,862	1,32	0,97	0,849
6	-0,306	2,569	-0,094	0,364	-1,414
7	-1,544	-2,734	1,32	0,768	-0,566
8	0,577	1,508	-1,273	-0,646	1,131
9	0,754	2,216	1,084	0,566	1,697
10	0,401	-0,613	0,141	-0,04	0
11	-1,367	-2,027	-1,037	-1,051	1,414
12	0,577	0,094	0,141	0,162	-2,263
13	-0,13	0,094	0,141	-1,253	1,131
14	1,285	2,216	-0,801	-0,242	-0,566
15	-0,837	-1,32	0,849	1,172	-0,849
Σύνολα	0	0	0	0	0

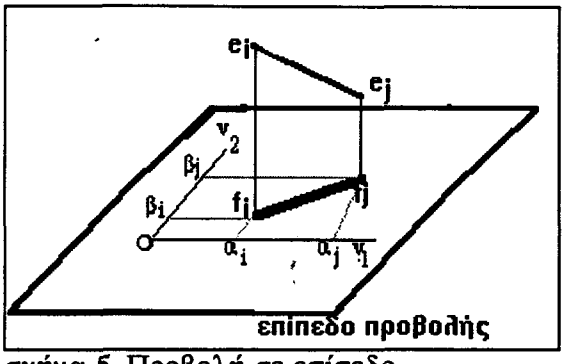
Κάθε μεταβλητή X^j (στήλη του πίνακα X), παριστάνει ένα διάνυσμα στο χώρο των n διαστάσεων. Οι τιμές X_i^j i=1..n, θεωρούνται σαν n συντεταγμένες. Κάθε γραμμή αντιστοιχεί σε μια παρατήρηση X_i και παριστάνει ένα διάνυσμα στο χώρο των p διαστάσεων. Οι τιμές X_i^j j=1..n, θεωρούνται σαν p συντεταγμένες.

Το σύνολο λοιπόν των πληροφοριών που μας παρέχει ο πίνακας X, παριστάνεται με δύο τρόπους. Σαν νέφος (σύνολο) n σημείων στο χώρο των p διαστάσεων, αλλά και σαν νέφος p σημείων στο χώρο των n διαστάσεων. Σκοπός μας είναι να παραστήσουμε τα δύο νέφη σημείων σε ένα χώρο πολύ λιγότερων διαστάσεων έστω m (m<p και m<n).

Το επίπεδο προβολής, όπου οι αποστάσεις των αρχικών αντικειμένων διατηρούνται όσο γίνεται καλύτερα, το ονομάζουμε **κύριο** (principal) επίπεδο. Για να καθορίσουμε το επίπεδο αυτό, αρκεί να βρούμε δύο άξονες v_1 και v_2. Αν αυτοί είναι κάθετοι, τότε έχουμε $d^2(f_i,f_j)=d^2(\alpha_i,\alpha_j)+d^2(\beta_i,\beta_j)$ όπου α_i και β_i είναι οι προβολές των e_i (αλλά και των f_i) στους v_1 και v_2 αντίστοιχα. (σχήμα 5). Ο μέσος των τετραγώνων των αποστάσεων μεταξύ των f_i είναι ίσος με το άθροισμα του μέσου των τετραγώνων των αποστάσεων μεταξύ των α_i και του μέσου των τετραγώνων των αποστάσεων των β_i.

Η μέθοδος επομένως, συνίσταται στην εύρεση καταρχήν, του άξονα v_1 που μεγιστοποιεί τον μέσο των $d^2(\alpha_i,\alpha_j)$, και κατόπιν του άξονα v_2 κάθετο στον v_1, ο οποίος μεγιστοποιεί τον μέσο των $d^2(\beta_i,\beta_j)$. Μπορούμε να συνεχίσουμε έξω από το

επίπεδο βρίσκοντας και άλλους άξονες v_3, v_4, ...v_p, κάθετους μεταξύ τους. Οι v_i ονομάζονται **κύριοι** (principaux) άξονες του νέφους. [Pages,1976] [Volle,1981]



σχήμα 5. Προβολή σε επίπεδο

Προβάλλοντας τα e_i που έχουν για αρχικές συντεταγμένες (x^1_i, x^2_i,...x^p_i) πάνω στους κύριους άξονες, λαμβάνουμε νέες συντεταγμένες (c^1_i, c^2_i,...c^p_i). Σχηματίζουμε λοιπόν νέα διανύσματα (c^1,c^2,....c^p), τα οποία ονομάζουμε **κύριες συνιστώσες** (composantes principales). Κάθε κύρια συνιστώσα c^k αποτελείται από τις συντεταγμένες των n αντικειμένων πάνω στον άξονα v_k και είναι ένας γραμμικός συνδυασμός των αρχικών διανυσμάτων:

$$c^k = u^k_1 x^1 + u^k_2 x^2 + \ldots\ldots u^k_p x^p$$

Οι συντελεστές (u^k_1,u^k_2,...u^k_p) σχηματίζουν τον k-οστό κύριο παράγοντα (facteur principal) u^k.

Η καλύτερη αναπαράσταση των δεδομένων μόνο με m διαστάσεις (m‹p), λαμβάνεται παίρνοντας τις m πρώτες κύριες συνιστώσες.

Η ACP λοιπόν είναι μια μέθοδος μείωσης του αριθμού των χαρακτηριστικών μεταβλητών (διαστάσεων), και επιτρέπει την γεωμετρική αναπαράσταση των αντικειμένων και των μεταβλητών. Η μείωση αυτή είναι δυνατή, αν οι p αρχικές μεταβλητές είναι εξαρτημένες δηλαδή έχουν μη μηδενικό συντελεστή συσχέτισης.

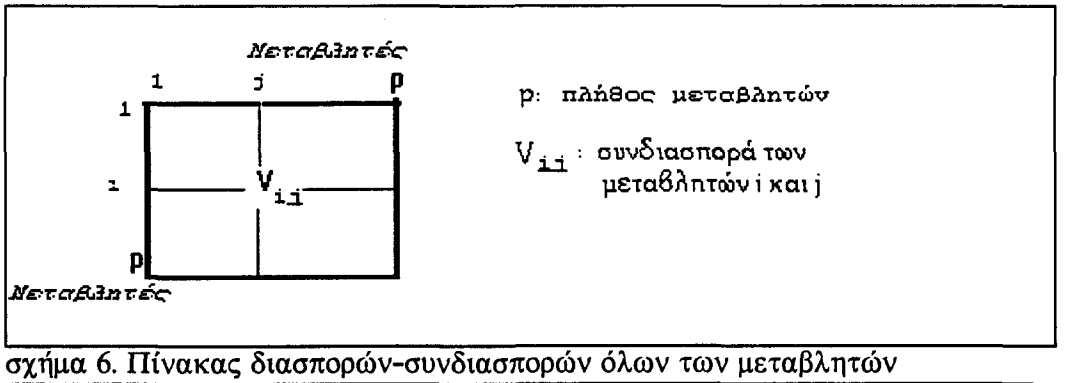
Η ACP είναι μια μέθοδος παραγοντική, καθώς η μείωση του αριθμού των μεταβλητών δεν γίνεται με μια απλή επιλογή ορισμένων απ' αυτές αλλά με την κατασκευή νέων μεταβλητών, από γραμμικούς συνδυασμούς των αρχικών με την βοήθεια των "παραγόντων" και έτσι είναι μια μέθοδος γραμμική. [Diday,1982]

1.3 Διαδικασία

Κάθε μία από τις p μεταβλητές, χαρακτηρίζεται από την μέση τιμή της $\overline{X}^j$ και την τυπική της απόκλιση s_j. Οι σχέσεις των μεταβλητών ανά δύο

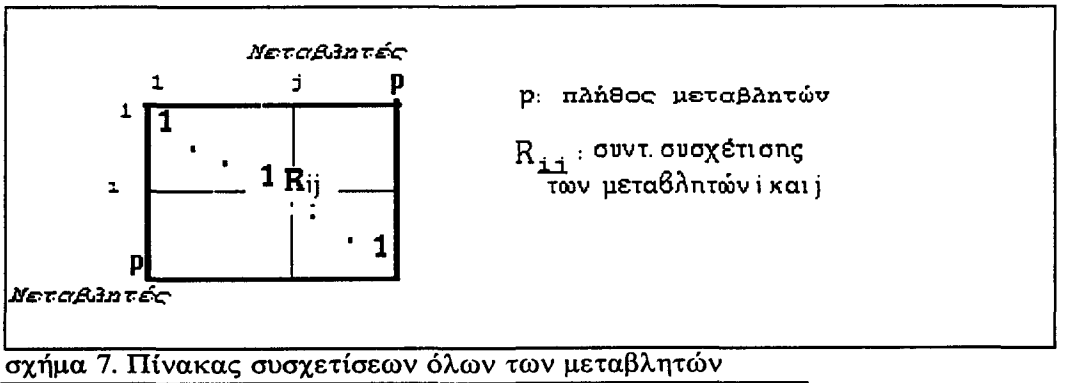
παριστάνονται, με την συνδιασπορά (συνδιακύμανση) τους ή με τον συντελεστή συσχέτισή τους.

Πρέπει να υπολογίσουμε p(p-1)/2 συντελεστές. Ο πίνακας διασπορών-συνδιασπορών συμβολίζεται με V, είναι συμμετρικός και έχει διαστάσεις pxp. (σχήμα 6).



σχήμα 6. Πίνακας διασπορών-συνδιασπορών όλων των μεταβλητών

Παρόμοια οι συντελεστές συσχέτισης παριστάνονται με τον συμμετρικό πίνακα R, που έχει διαστάσεις pxp. (σχήμα 7). Στην κύρια διαγώνιο υπάρχουν μονάδες.



σχήμα 7. Πίνακας συσχετίσεων όλων των μεταβλητών

Γνωρίζοντας τον πίνακα V μπορούμε να υπολογίσουμε αμέσως τον πίνακα R από τη σχέση:

$R=D_{1/s}\ V\ D_{1/s}$

όπου $D_{1/s}$ είναι ο επόμενος διαγώνιος πίνακας:

$$D_{1/s}=\begin{pmatrix} 1/s_1 & & 0 \\ & \cdots & \\ 0 & & 1/s_p \end{pmatrix}$$

Αν τα δεδομένα είναι κεντραρισμένα, ο πίνακας V υπολογίζεται αμέσως από τον αρχικό πίνακα δεδομένων από τη σχέση: [Pages,1976]

$V=X' D X$

όπου D είναι ο n τάξης διαγώνιος πίνακας με τα βάρη κάθε αντικειμένου:

$$D = \begin{pmatrix} p_1 & & 0 \\ & \ldots & \\ 0 & & p_n \end{pmatrix}$$

1.4 Νέφος n σημείων στις p διαστάσεις

Θεωρούμε καταρχήν το νέφος των n αντικειμένων, τα οποία μπορούν να αντιμετωπιστούν σαν n σημεία με p συντεταγμένες το καθένα, δηλαδή σαν σημεία σε ένα χώρο p διαστάσεων. Το κέντρο μάζας του νέφους είναι το διάνυσμα G που ορίζεται από τις μέσες τιμές όλων των μεταβλητών. $G=(\bar{X}^1, \bar{X}^2..., \bar{X}^p)$

Η διασπορά (αλλιώς **αδράνεια**) του νέφους σημείων, γύρω από το κέντρο μάζας G είναι $I_{N,G}=\sum_{i-1}^{n} p_i d^2(X_i,G)$. [Tenenbaus,1994]

Με $d^2(X_i,G)=\sum_{j-1}^{p}\left(X_{ij}-\bar{X}^j\right)^2$ συμβολίζουμε το τετράγωνο της απόστασης του σημείου X_i από το G.

Η ποσότητα $I_{N,G}$ ονομάζεται ολική αδράνεια του νέφους ως προς το κέντρο μάζας G. Όταν οι μεταβλητές X^j ακολουθούν την N(0,1) κατανομή ($S_j^2=1$) η ολική αδράνεια του νέφους είναι ίση με το πλήθος των μεταβλητών p.

Για την μέτρηση της απόστασης μεταξύ δύο σημείων (αντικειμένων) χρησιμοποιείται η ευκλείδεια μετρική και η απόσταση υπολογίζεται από τον τύπο:

$$d^2(i,i')=\sum_{j-1}^{p}\left(X_{ij}-X_{i'j}\right)^2$$

Μπορούμε να δείξουμε ότι η αδράνεια είναι ίση με το μέσο των τετραγώνων των n(n-1)/2 διαφορών απόστασης μεταξύ των σημείων του νέφους. Ερμηνεύουμε λοιπόν το κύριο επίπεδο (principal) του νέφους, σαν το επίπεδο εκείνο που μεγιστοποιεί την αδράνεια του συνόλου των n σημείων προβαλλομένων πάνω του. Το επίπεδο αυτό είναι και το "πιο κοντινό" στο νέφος των σημείων, με την έννοια ότι ο μέσος όρος των τετραγώνων των αποστάσεων των σημείων του νέφους από το επίπεδο είναι ελάχιστος.

Σύμφωνα με το θεώρημα του Huyghens, αποδεικνύεται ότι τόσο το κύριο επίπεδο όσο και οι κύριοι άξονες του νέφους, διέρχονται από το κέντρο βάρους του νέφους των σημείων.

Ακόμη αποδεικνύεται ότι η αδράνεια του νέφους ισούται με το ίχνος trace(V), του πίνακα διασπορών-συνδιασπορών, όταν τα σημεία είναι

κεντραρισμένα. (Η αδράνεια είναι ίση με το άθροισμα των διασπορών των p μεταβλητών). [Pages,1976]

1.5 Νέφος p σημείων στις n διαστάσεις

Μπορούμε ακόμη να θεωρήσουμε το νέφος των p μεταβλητών, σαν νέφος p σημείων με n συντεταγμένες το καθένα, δηλαδή σαν νέφος σημείων στο χώρο των n διαστάσεων.

Κάθε μεταβλητή μπορεί να παρασταθεί σαν ένα διάνυσμα X^j, με n συντεταγμένες που αντιστοιχούν στις τιμές της για κάθε αντικείμενο. Θα θεωρούμε λοιπόν τα διανύσματα αυτά, σε χώρο n διαστάσεων,τον οποίο ονομάζουμε χώρο των μεταβλητών και συμβολίζουμε με R^n.

Το "μήκος" μιας μεταβλητής, δεν είναι παρά η τυπική της απόκλιση s_j. Η "απόσταση" μεταξύ δύο μεταβλητών, μετριέται με τον συντελεστή συσχέτισης τους. [Pottier,1994] Ο παρακάτω τύπος χρησιμοποιείται για τη μέτρηση της προσέγγισης (συσχέτισης) μεταξύ δύο μεταβλητών k και h:

$$r_{kh}=\frac{\text{covariance(k,h)}}{\sqrt{\text{variance(k) variance(h)}}}=\sum_{i\in I}\left(\frac{x_{ik}-\bar{x}_k}{s_k}\right)\left(\frac{x_{ih}-\bar{x}_h}{s_h}\right)=\frac{s_{kh}}{s_k\, s_h}$$

Παρατηρούμε ότι αν οι μεταβλητές είναι κεντραρισμένες $r_{kh}=s_{kh}$. Δηλαδή ο συντελεστής συσχέτισής τους είναι ίσος με την συνδιασπορά τους, και επομένως η μέτρηση της "απόστασης" δύο μεταβλητών μετριέται με την συνδιασπορά τους. Η συνδιασπορά πάλι μπορεί να εκφραστεί με την γωνία των δύο διανυσμάτων που παριστάνουν τις μεταβλητές, αφού για την γωνία θ μεταξύ δύο διανυσμάτων ισχύει:

$$\langle x_k,x_h\rangle=|x_k||x_h|\cos\theta_{kh} \Leftrightarrow \cos\theta_{kh}=s_{kh}/s_k\, s_h=r_{kh}$$

Ενώ λοιπόν στο χώρο των αντικειμένων, ενδιαφερόμαστε για αποστάσεις μεταξύ σημείων, στο χώρο των μεταβλητών θα ενδιαφερόμαστε περισσότερο για τις γωνίες μεταξύ των διανυσμάτων. [Sneath,1972]

Όσο μικρότερη είναι η γωνία μεταξύ δύο μεταβλητών (διανυσμάτων), τόσο μεγαλύτερο είναι το συνημίτονό της, και επομένως τόσο μεγαλύτερος και ο συντελεστής συσχέτισης μεταξύ των μεταβλητών. Τότε οι μεταβλητές θεωρούνται "πολύ κοντά".

1.6 Εύρεση των κυρίων συνιστωσών

Οι κύριοι άξονες (οι κύριοι συνιστώσες) $v_1,v_2,...v_p$ του κεντραρισμένου νέφους σημείων, αποδεικνύεται ότι είναι τα χαρακτηριστικά διανύσματα, του πίνακα διασπορών V, που αντιστοιχούν στις χαρακτηριστικές του τιμές $\lambda_1,\lambda_2,...\lambda_p$ γραμμένες σε φθίνουσα σειρά μεγέθους.

Ο πίνακας διασποράς-συνδιασποράς, των κυρίων συνιστωσών γράφεται:

$$V_c = \begin{pmatrix} \lambda_1 & & 0 \\ & \cdots & \\ 0 & & \lambda_p \end{pmatrix}$$

Παρατηρούμε ότι οι κύριες συνιστώσες είναι μη συσχετισμένες ανά δύο. Άρα με την A.C.P. αντικαθιστούνται οι p αρχικές μεταβλητές με νέες μη συσχετισμένες μεταξύ τους μεταβλητές, που έχουν μέγιστη διασπορά διατεταγμένες κατά φθίνουσα σειρά.

Το άθροισμα των χαρακτηριστικών τιμών $I=\lambda_1+\lambda_2+...\lambda_p$, είναι σταθερό ίσο με το ίχνος του πίνακα V_c και ίσο με το ίχνος του πίνακα V, που όπως έχουμε δει ισούται με την ολική αδράνεια του νέφους.

Το πηλίκο λ_k/I ονομάζεται ποσοστό της αδράνειας (ή της διασποράς) που ερμηνεύεται από τον k-οστό άξονα.

Το πηλίκο $(\lambda_1+\lambda_2)/I$ είναι το ποσοστό αδράνειας που ερμηνεύεται από τους 2 πρώτους κύριους άξονες μαζί και μετράει την προσαρμογή του νέφους στο κύριο επίπεδο 1x2. Οσο μεγαλύτερο είναι το ποσοστό αυτό, τόσο καλύτερη είναι η αναπαράσταση του νέφους στο επίπεδο των αξόνων 1 και 2.

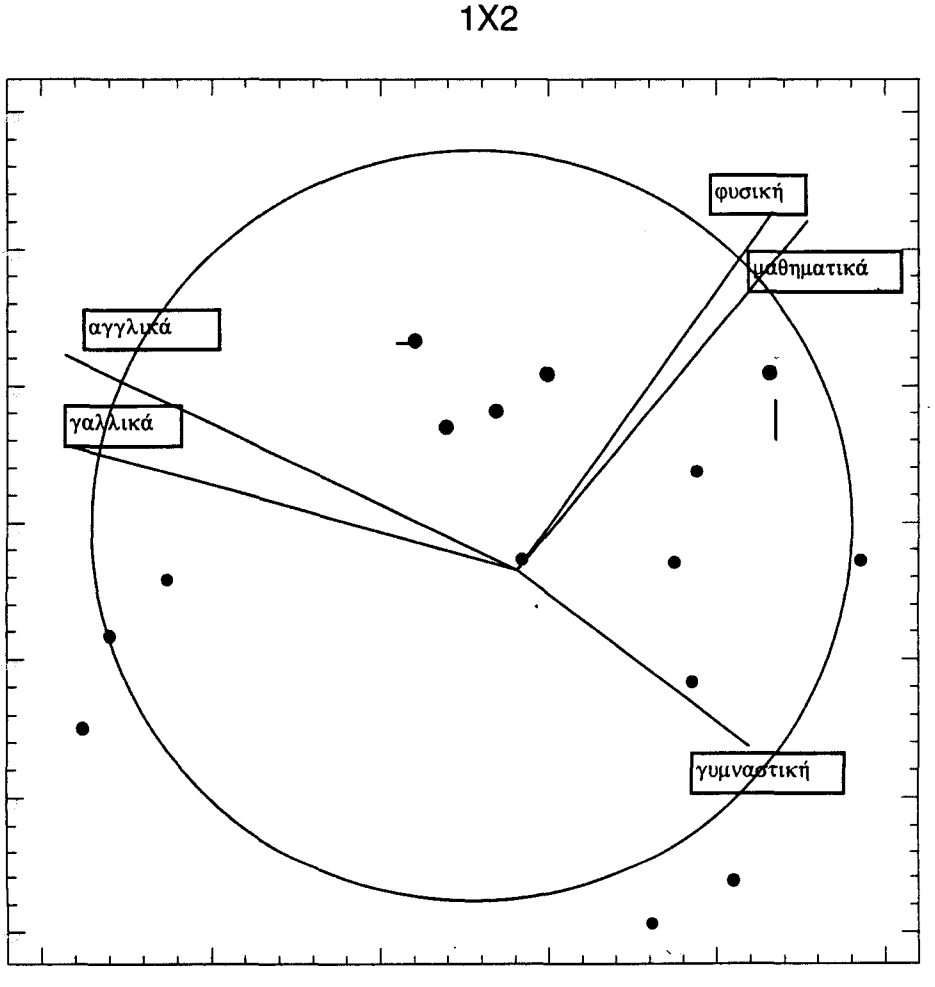
1.7 Ερμηνεία των αποτελεσμάτων

Εφαρμόζοντας την ανάλυση σε κύριες συνιστώσες για το παράδειγμα του πίνακα 1, παίρνουμε την αναπαράσταση των σημείων στο πρώτο κύριο επίπεδο, όπως φαίνεται στο σχήμα 8. Επίσης οι συντεταγμένες των μεταβλητών και των αντικειμένων στο σύστημα αξόνων των κυρίων συνιστωσών, παρουσιάζονται στους πίνακες 3 και 4 αντίστοιχα. Στην τελευταία γραμμή του πίνακα 4 παρουσιάζεται και το ποσοστό αδράνειας που ερμηνεύει ο άξονας κάθε κύριας συνιστώσας.

Πίνακας 3 Χαρακτηριστικά διανύσματα

Μεταβλητές	1η συνιστ.	2η συνιστ	3η συνιστ	4η συνιστ	5η συνιστ
Μαθηματικά	0.374	0.596	-0.030	0.702	0.095
Φυσική	0.329	0.610	0.230	-0.660	-0.171
Γαλλικά	-0.572	0.212	0.456	0.226	-0.605
Αγγλικά	-0.578	0.368	0.099	-0.098	0.714
Γυμναστική	0.298	-0.300	0.853	0.093	0.290

Σημείωση: Η γραφική παράσταση που παίρνουμε είναι μια προβολή (σχήμα 8). Δεν πρέπει να συγχέουμε τις προσεγγίσεις στο επίπεδο με τις προσεγγίσεις στο χώρο. Λάθος είναι δυνατό να γίνει αν δύο σημεία βρίσκονται το ένα πάνω από το επίπεδο και το άλλο από κάτω. Τότε οι προβολές τους είναι κοντά αλλά τα σημεία δεν είναι κοντινά στο χώρο. Αυτό φαίνεται καθαρά στο σχήμα 9. Αν και οι προβολές f_i f_j, των σημείων e_i e_j βρίσκονται αρκετά κοντά μεταξύ τους, στο επίπεδο, εντούτοις τα σημεία είναι πολύ απομακρυσμένα στο χώρο.[Escofier,1990]

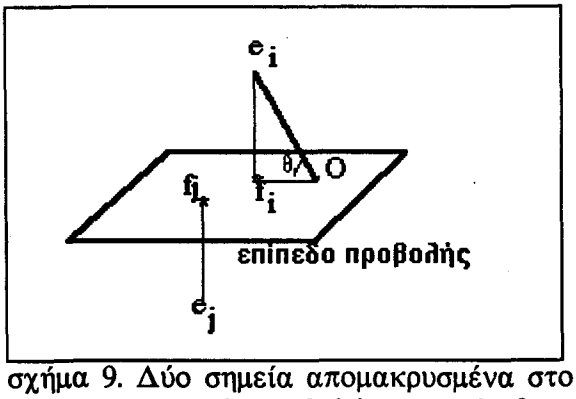


σχήμα 8. Επίπεδο κυρίων παραγοντικών αξόνων 1 και 2

Πίνακας 4 Κύριες συνιστώσες

Μαθητές	1η συνιστ.	2η συνιστ	3η συνιστ	4η συνιστ	5η συνιστ
1	0.898	0.038	0.053	0.334	0.338
2	1.029	0.701	1.694	0.483	0.027
3	-2.446	0.498	-0.453	0.102	0.321
4	1.249	2.295	0.255	0.001	0.057
5	-0.646	16.628	1.536	0.152	0.053
6	-0.164	11.466	0.581	-1.353	-0.219
7	-2.609	11.619	0.071	0.034	-0.22
8	2.007	0.052	0.207	-0.276	0.445
9	0.038	13.567	1.970	0.071	0.015
10	-0.010	0.063	-0.027	0.571	-0.026
11	0.764	-2.604	0.067	-0.316	0.146
12	-0.464	1.027	-1.392	0.257	-0.405
13	1.004	-0.8322	0.673	0.129	-0.937
14	1.465	1.427	-0.548	-0.054	0.111
15	-2.114	-0.085	-0.137	-0.136	0.300
Ποσοστό αδράνειας	41.6106%	34.093%	17.8447%	4.0352%	2.4167%

Πρέπει λοιπόν να εξετάζουμε την ποιότητα της αναπαράστασης κάθε σημείου, στο επίπεδο. Αυτό γίνεται με την γωνία θ μεταξύ του διανύσματος e_i και της προβολής του f_i. (σχήμα 9). Το κριτήριο γι' αυτό, είναι το τετράγωνο του συνημιτόνου της γωνίας με το επίπεδο. Μια τιμή κοντά στο 1, δείχνει ότι τα e_i και f_i συμπίπτουν. Αντίθετα μια τιμή κοντά στο μηδέν, σημαίνει ότι το αντικείμενο e_i, δεν αναπαριστάται καλά στο επίπεδο, εκτός και αν η απόστασή του από το κέντρο βάρους είναι μικρή. Για τα σημεία αυτά πρέπει να εξετάσουμε τα κύρια επίπεδα που σχηματίζουν οι 1 και 3 ή οι 2 και 3 κύριοι άξονες.



σχήμα 9. Δύο σημεία απομακρυσμένα στο χώρο, με κοντινές προβολές στο επίπεδο

1.8 Ποιότητα αναπαράστασης

Υπάρχουν διάφοροι δείκτες με τους οποίους μπορούμε να μετρήσουμε την ποιότητα της αναπαράστασης ενός αντικειμένου ή μεταβλητής, από ένα άξονα κύριας.συνιστώσας. [Guigou,1977]

1. Ποιότητα αναπαράστασης αντικειμένου i από τον άξονα s

$$QLTs(i)=\frac{\text{αδράνεια της προβολής του στοιχείου i στον άξονα s}}{\text{ολική αδράνεια του i}}$$

$$QLTs(i)=\frac{(OH_i)^2}{(Oi)^2}=\cos^2\theta$$

2. Ποιότητα αναπαράστασης όλου του νέφους από τον άξονα s

$$QLTs=\frac{\text{αδράνεια της προβολής του νέφους στον άξονα s}}{\text{ολική αδράνεια του νέφους}}$$

όπου η ολική Αδράνεια του νέφους= $\frac{1}{n}\sum_k\sum_i\left(\frac{x_{ik}-\bar{x}_k}{s_k}\right)^2=p$ (αριθμός μεταβλητών)

3. Συνεισφορά (Contribution) του αντικειμένου i στην δημιουργία του άξονα s

Μπορούμε να αναζητήσουμε ποια είναι τα αντικείμενα που χαρακτηρίζουν περισσότερο ένα κύριο άξονα, επειδή συμβάλλουν στην δημιουργία του, υπολογίζοντας την "συνεισφορά" του σημείου στον k-οστό άξονα, που ορίζεται σαν Ctrs(i). Αυτό είναι το μέρος της διασποράς του c_k, που οφείλεται στο αντικείμενο i. Η θεώρηση αυτή ισχύει μόνο αν τα αντικείμενα από μόνα τους παρουσιάζουν ενδιαφέρον. Αν έχουν ληφθεί τυχαία σε μια έρευνα, το ενδιαφέρον της ACP βρίσκεται στην μελέτη του κύκλου συσχετίσεων.

$$Ctrs(i)=\frac{\text{αδράνεια προβολής του i στον άξονα s}}{\text{αδράνεια προβολής όλου του νέφους στον s}}=p_i\,c^2_{ik}/\lambda_k$$

4. Ο κύκλος των συσχετίσεων

Ο υπολογισμός των συσχετίσεων μεταξύ των κυρίων συνιστωσών και των αρχικών μεταβλητών, είναι πολύ εύκολο να γίνει. Ο συντελεστής γραμμικής συσχέτισης μεταξύ x_j και c_k είναι ίσος με την j-οστή συνιστώσα του k-οστου χαρακτηριστικού διανύσματος v_k πολλαπλασιασμένη με ρίζα λ_k. (Το άθροισμα των τετραγώνων των συσχετίσεων του c_k με τα x^j είναι λ_k).

Αν παραστήσουμε κάθε μεταβλητή με ένα σημείο του οποίου οι συντεταγμένες είναι οι συσχετίσεις του με τις κύριες συνιστώσες c_1 και c_2, οι αρχικές μεταβλητές τότε περιέχονται σε ένα κύκλο με ακτίνα την μονάδα, ο οποίος ονομάζεται κύκλος συσχετίσεων. Καθώς οι c_1 και c_2, είναι ασυσχέτιστες, αποδεικνύεται ότι $r^2(c_1,x^j)+r^2(c_2,x^j)=1$

Η εξέταση αυτού του κύκλου (σχήμα 8), επιτρέπει την ερμηνεία των κυρίων συνιστωσών και την γρήγορη εύρεση των ομάδων μεταβλητών που βρίσκονται μαζί ή αντίθετα, με την προϋπόθεση πάντα ότι τα σημεία βρίσκονται κοντά στην περιφέρεια του κύκλου (καλή η αναπαράσταση).

Αυτή η αναπαράσταση, παίζει για τις μεταβλητές τον ίδιο ρόλο που παίζει το κύριο επίπεδο για τα αντικείμενα.

2. Παραγοντική ανάλυση των αντιστοιχιών (Analyse Factorielle des Correspondances-A.F.C.)

2.1 Εισαγωγή

Ο πρώτος που εισήγαγε για πρώτη φορά την έννοια των παραγόντων ήταν ο Spearman το 1904. Κατόπιν το 1930 ο Thurstone έθεσε για πρώτη φορά το πρόβλημα της αναζήτησης πολλών παραγόντων [Diday,1982]. Η πρώτη μαθηματική τυποποίηση της παραγοντικής ανάλυσης των αντιστοιχιών έγινε το 1935 από τον Hirschfeld [Hirschfeld,1935] και στη συνέχεια από τον Guttman [Guttman,1941] το 1941. Στο άρθρο του Nishisato [Nishisato,1980], μπορούμε να βρούμε μια ιστορική αναδρομή από την μεριά της "αγγλικής σχολής".

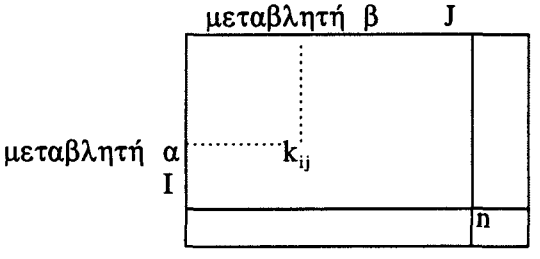
Η Παραγοντική Ανάλυση των Αντιστοιχιών είναι μια γενίκευση της Ανάλυσης σε κύριες Συνιστώσες, προσαρμοσμένη κυρίως σε ποιοτικές μεταβλητές ή σε ποσοτικές που είναι χωρισμένες σε κλάσεις.

Είναι η πιο συχνά χρησιμοποιούμενη μέθοδος είτε μόνη της, είτε σε συνδυασμό με μια μέθοδο ταξινόμησης. Αυτό οφείλεται κυρίως στη μεγάλη της δυνατότητα να αναπαριστάνει γεωμετρικά τα δεδομένα, μετασχηματίζοντας τις στατιστικές προσεγγίσεις σε ευκλείδειες αποστάσεις. Στα γραφικά που προκύπτουν από μια παραγοντική ανάλυση, βλέπουμε (με τα μάτια μας) την εικόνα των ομαδοποιήσεων, των τάσεων, των αντιθέσεων ανάμεσα στα δεδομένα μας, κάτι που είναι δύσκολο αν όχι αδύνατο να προκύψει κατευθείαν από την εξέταση ενός μεγάλου πίνακα δεδομένων. [Escofier,1990].

Ονομάζεται "παραγοντική" ανάλυση των Αντιστοιχιών γιατί πετυχαίνει την περιγραφή ενός φαινομένου, χρησιμοποιώντας ένα μικρό αριθμό νέων μεταβλητών που συντίθενται από τις αρχικές μεταβλητές και ονομάζονται "παράγοντες". Οι παράγοντες αυτοί ανακαλύπτουν και αναδεικνύουν τα

ουσιαστικώτερα κρυμμένα χαρακτηριστικά των δεδομένων. Εδώ βρίσκεται και η ομοιότητα της μεθόδου με την Ανάλυση σε κύριες Συνιστώσες.

Στην πρώτη εμφάνιση της η μέθοδος αφορούσε μόνο στην μελέτη πίνακα συμπτώσεων δύο μεταβλητών με πολλές κατηγορίες (κλάσεις). Ένας πίνακας συμπτώσεων (σχήμα 10) διασταυρώνει τις κατηγορίες δύο ποιοτικών μεταβλητών (I και J). Θεωρούμε ότι η μεταβλητή I αντιστοιχεί στις γραμμές και η J στις στήλες του πίνακα συμπτώσεων. (Περισσότερα για τα χαρακτηριστικά και τον τρόπο δημιουργίας ενός τέτοιου πίνακα θα αναφέρουμε στο κεφάλαιο 2). Αργότερα η μέθοδος επεκτάθηκε και για οποιοδήποτε αριθμό ποιοτικών ή και ποσοτικών μεταβλητών χωρισμένων σε κλάσεις.



I: πλήθος των γραμμών
J: πλήθος των στηλών
k_{ij}: αριθμός αντικειμένων που ανήκουν ταυτόχρονα στην κατηγορία i της α μεταβλητής και στην κατηγορία j της β μεταβλητής.
$\Sigma_i \Sigma_j k_{ij}=n$ πλήθος αντικειμένων

σχήμα 10. Ο πίνακας των δεδομένων

Η Παραγοντική Ανάλυση εφαρμόζεται γενικώτερα σε οποιοδήποτε πίνακα παρουσιάζει ομοιογένεια, δηλαδή οι αριθμοί που εμφανίζονται στον πίνακα μετριώνται με την ίδια μονάδα μέτρησης. Η μέθοδος εφαρμόζεται σε οποιοδήποτε πίνακα θετικών αριθμών όπου τα αθροίσματα γραμμών και στηλών έχουν νόημα. Συνήθως ασχολούμαστε με πίνακα σχετικών συχνοτήτων που προκύπτει από ένα πίνακα συχνοτήτων διαιρώντας κάθε στοιχείο του με το άθροισμα των τιμών όλου του πίνακα. (βλέπε κεφάλαιο 2). [Παπαδημητρίου,1994]

Ο πίνακας των δεδομένων, όταν παρουσιάζει δύο μεταβλητές, είναι πίνακας IxJ, όπου I είναι το πλήθος των κατηγοριών της α μεταβλητής και J το πλήθος των κατηγοριών της β μεταβλητής. Από τον αρχικό πίνακα δεδομένων, δημιουργούμε δύο νέφη σημείων. Το νέφος γραμμών (N_I) των I σημείων στο χωρο των J διαστάσεων και το νέφος στηλών (N_J) των J σημείων στο χώρο των I διαστάσεων.

Προσπαθώντας να διακρίνουμε τη σχέση μεταξύ των κατηγοριών (κλάσεων) των 2 μεταβλητών, αναπαριστούμε τα δύο νέφη δεδομένων στο ίδιο χώρο 2 ή 3 διαστάσεων, και ανάλογα του πόσο "κοντά" βρίσκονται οι εικόνες των δεδομένων, βγάζουμε συμπεράσματα για τις σχέσεις μεταξύ των μεταβλητών και των κατηγοριών τους. [Torrens,1972]. Η Παραγοντική Ανάλυση των Αντιστοιχιών, εφαρμόζεται και για τη στατιστική ανάλυση πινάκων που περιγράφουν αντικείμενα και μεταβλητές. Αρκεί όλες οι μεταβλητές να είναι ποιοτικές, ή αν

πρόκειται για ποσοτικές, να έχουν χωριστεί σε κλάσεις. Με την ανάλυση ενός τέτοιου πίνακα, προσδιορίζουμε τις σχέσεις μεταξύ αντικειμένων, μεταξύ των κλάσεων όλων των μεταβλητών, και μεταξύ αντικειμέων και κλάσεων μεταβλητών. Παράδειγμα αποτελεσμάτων της μεθόδου αυτής φαίνεται στο σχήμα 13.

2.2 Μετασχηματισμοί των δεδομένων

Σκοπός της Παραγοντικής Ανάλυσης των Αντιστοιχιών, είναι η εύρεση των σχέσεων μεταξύ των γραμμών, μεταξύ των στηλών και ακόμη μεταξύ γραμμών-στηλών.

Δύο γραμμές θεωρούνται σαν κοντινές αν οι τιμές τους, κατανέμονται με τον ίδιο τρόπο στο σύνολο των στηλών. Δηλαδή αν παρουσιάζουν συγχρόνως σχετικά μικρές ή μεγάλες τιμές στις ίδιες στήλες. Συμμετρικά δύο στήλες θεωρούνται όμοιες αν συμμετέχουν με τον ίδιο τρόπο στο σύνολο των γραμμών.

Έστω ότι σε κάποιο πίνακα δεδομένων έχουμε τις επόμενες δύο γραμμές.

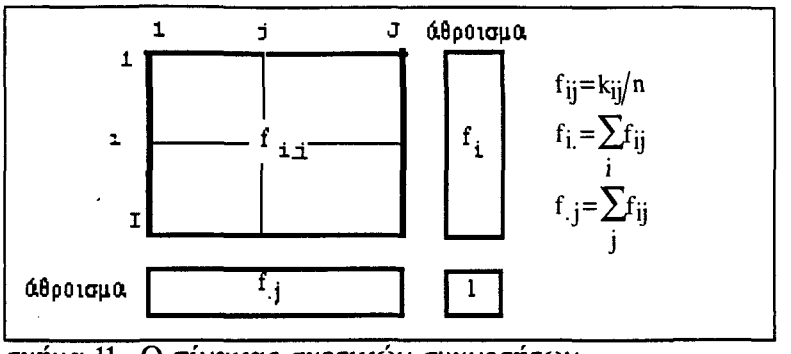
					άθροισμα
5	7	10	53	25	100
15	21	30	159	75	300

Οι αριθμοί των δύο γραμμών διαφέρουν πολύ μεταξύ τους. Παρατηρούμε όμως ότι οι δύο γραμμές είναι ανάλογες. Οι αριθμοί της δεύτερης, προκύπτουν από τους αντίστοιχους της πρώτης, αν τους πολλαπλασιάσουμε με 3. Αν διαιρέσουμε κάθε γραμμή με το άθροισμά της, παίρνουμε τις παρακάτω δύο νέες γραμμές:

					άθροισμα
5/100	7/100	10/100	53/100	25/100	1
15/300	21/300	30/300	159/300	75/300	1

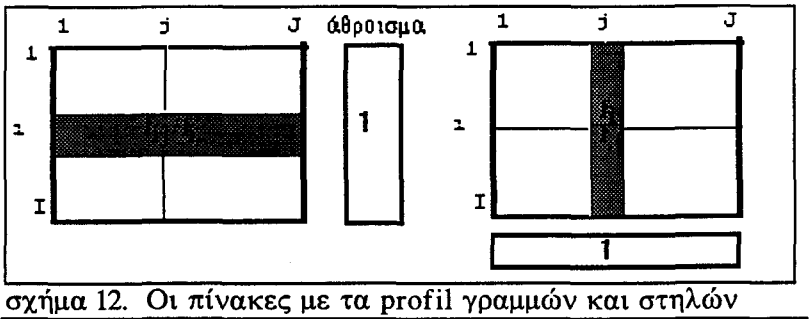
Όπως βλέπουμε μετά την διαίρεση οι δύο γραμμές είναι όμοιες. Οι σχετικές κατανομές τους ως προς το σύνολο των στηλών είναι ίδιες.

Εξαιτίας της συμμετρίας κάτι ανάλογο συμβαίνει και στις στήλες του πίνακα. Αν ενδιαφερόμαστε για την ανάλυση των στηλών, εξετάζουμε ένα άλλο μετασχηματισμό του ίδιου πίνακα. Κάθε στήλη διαιρείται με το άθροισμα των στοιχείων της. Πάντως και οι δύο μετασχηματισμοί έχουν την ίδια σημαντικότητα για τους επιδιωκόμενους σκοπούς. [Greenacre,1993]



σχήμα 11. Ο πίνακας σχετικών συχνοτήτων

Πριν προχωρήσουμε λοιπόν στην Παραγοντική Ανάλυση των Αντιστοιχιών, σχηματίζουμε τον πίνακα σχετικών συχνοτήτων (σχήμα 11) και κατόπιν μετασχηματίζουμε τα δεδομένα διαιρώντας κάθε γραμμή με το άθροισμα της (ή κάθε στήλη με το άθροισμά της). Η νέα γραμμή (ή στήλη) που δημιουργείται ονομάζεται profil-γραμμή (ή profil-στήλη) (σχήμα 12), και είναι η σχετική κατανομή της αντίστοιχης γραμμής ή στήλης. Κάθε στοιχείο της $f_{ij}/f_{.j}$ εκφράζει την υπό συνθήκη κατανομή σχετικών συχνοτήτων των J κατηγοριών, όταν γνωρίζουμε την κατηγορία I. Έχουμε δηλαδή εκτίμηση της πιθανότητας, κάποιο αντικείμενο να ανήκει στην j κατηγορία της β μεταβλητής, γνωρίζοντας ότι ανήκει στην I κατηγορία της α μεταβλητής. [Benzecri,1973b]



σχήμα 12. Οι πίνακες με τα profil γραμμών και στηλών

2.3 Ομοιότητα μεταξύ γραμμών ή στηλών

Η ομοιότητα μεταξύ γραμμών ή στηλών ορίζεται από μια μετρική απόστασης μεταξύ των profil τους. Η μετρική αυτή είναι η X^2. Περισσότερα για τον τρόπο υπολογισμού και τις ιδιότητες της X^2 θα αναφέρουμε στο κεφάλαιο 2. Η μετρική X^2 ικανοποιεί την αρχή της "κατανεμημένης ισοδυναμίας". Αν δύο ανάλογες γραμμές ή στήλες του πίνακα δεδομένων, ενωθούν σε μία, δεν αλλάζει η απόσταση μεταξύ των κατανομών στηλών ή γραμμών. Έτσι το αν μια μεταβλητή χωριστεί σε λίγες ή πολλές κλάσεις, δεν έχει επίδραση στην μέτρηση των αποστάσεων. [Παπαδημητρίου-Φλώρου,1994]

Οι δύο αναλύσεις των νεφών γραμμών N_I και στηλών N_J δίνουν τον ίδιο αριθμό παραγοντικών αξόνων, οι οποίοι δημιουργούνται από τις ίδιες ιδιοτιμές. Οι συντεταγμένες όλων των σημείων του νέφους N_I (αντίστοιχα του N_J) στον παραγοντικό άξονα τάξης s, συνιστούν μια νέα κατηγορία F_s (αντίστοιχα G_s), που οναμάζεται παράγοντας. Υπάρχει στενή σχέση μεταξύ των παραγόντων των δύο νεφών και μπορούμε να παραστήσουμε τα δύο νέφη ταυτόχρονα.

Οι παραγοντικοί άξονες είναι οι άξονες εκείνοι που διέρχονται όσο γίνεται πλησιέστερα από το νέφος σημείων. Από το σύνολο των παραγοντικών αξόνων, ερμηνεύεται η συνολική αδράνεια του νέφους. Αν θεωρήσουμε τον χώρο των m πρώτων παραγοντικών αξόνων, έχουμε τον καλύτερο χώρο m διαστάσεων, με την έννοια της μέγιστης πληροφορίας, αφού στο χώρο αυτό παρουσιάζεται το μέγιστο ποσοστό ερμηνευόμενης αδράνειας του νέφους. [Escofier,1990].

Αποδεικνύεται ότι οι παραγοντικοί άξονες, προκύπτουν από τη διαγωνιοποίηση του πίνακα X'DXM όπου M και D είναι διαγώνιοι πίνακες JxJ και IxI αντίστοιχα, με στοιχεία $f_{.j}$ και $f_{i.}$. Ο πίνακας X έχει για στοιχεία:

$$x_{ij} = \frac{f_{ij}}{f_{i.}f_{.j}} - 1 = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}}.$$

$$X = \begin{pmatrix} x_{11} & & 0 \\ & \dots & \\ 0 & & x_{IJ} \end{pmatrix} \quad D = \begin{pmatrix} f_{.1} & & 0 \\ & \dots & \\ 0 & & f_{.I} \end{pmatrix} \quad M = \begin{pmatrix} f_{1.} & & 0 \\ & \dots & \\ 0 & & f_{J.} \end{pmatrix}$$

Οι παραγοντικοί άξονες είναι τα χαρακτηριστικά διανύσματα που αντιστοιχούν στις $\lambda_1,\lambda_2,\dots\lambda_K$ ιδιοτιμές του πίνακα X'DXM. Το άθροισμα $\lambda_1+\lambda_2+\dots+\lambda_K$ είναι ίσο με την ολική αδράνεια (διασπορά) του νέφους. [Lebart,1982]

2.4 Ανάλυση πίνακα για πολλές ποιοτικές μεταβλητές

Όταν επιθυμούμε την ανάλυση ενός πίνακα *αντικείμενα x μεταβλητές*, πολλών μεταβλητών, μετατρέπουμε τις ποσοτικές μεταβλητές σε ποιοτικές, χωρίζοντας τις σε κλάσεις, και κατόπιν σχηματίζουμε τον λογικό πίνακα 0-1 και τον πίνακα συμπτώσεων Burt, όπως θα περιγράψουμε στο κεφάλαιο 2.

Η Παραγοντική Ανάλυση ενός πίνακα 0-1 είναι ισοδύναμη με την ανάλυση του αντίστοιχου πίνακα Burt. Αν μας ενδιαφέρουν οι σχέσεις μεταξύ των αντικειμένων και μεταβλητών, εφαρμόζουμε την Παραγοντική Ανάλυση στον πίνακα 0-1. Αν όμως δεν μας ενδιαφέρουν τα αντικείμενα αλλά μόνο οι σχέσεις μεταξύ των μεταβλητών εφαρμόζουμε την Παραγοντική Ανάλυση στον πίνακα Burt. Συνήθως για μια έρευνα με ανώνυμα ερωτηματολόγια αναλύουμε τον πίνακα Burt. Εξάλλου και η ποιότητα της αναπαράστασης στα παραγοντικά επίπεδα, είναι

καλύτερη στην περίπτωση ανάλυσης του πίνακα Burt. Περισσότερο αναλυτικά για τον τρόπο κατασκευής του πίνακα αυτού αλλά και για προτάσεις κατασκευής νέων πολύ χρήσιμων πινάκων που αναλύονται με την Παραγοντική Ανάλυση των Αντιστοιχιών, θα περιγράψουμε στο επόμενο κεφάλαιο.

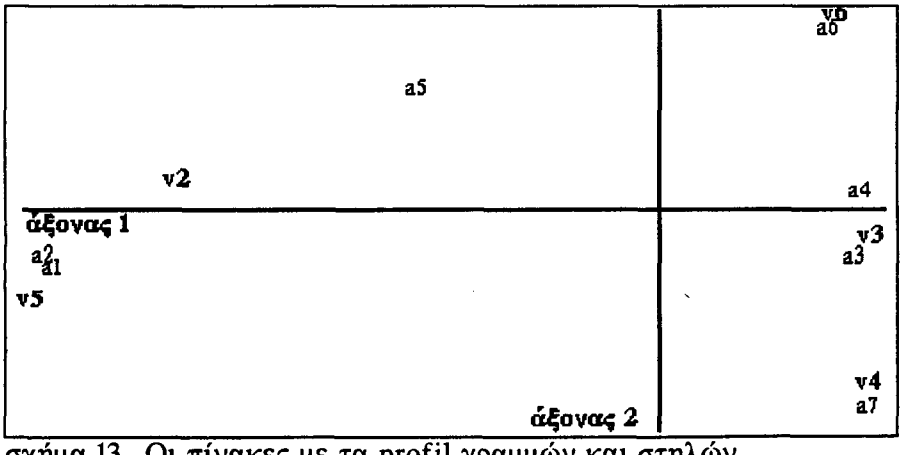
2.4.1 Παράδειγμα

Έστω ότι έχουμε τον παρακάτω πίνακα συμπτώσεων, για δεδομένα που αφορούν στη χρήση κάποιων φαρμάκων (στήλες) για ορισμένες ασθένειες (γραμμές). Τα δεδομένα προέρχονται από το άρθρο της F. Benzecri [Benzecri F.,1980] και αφορούν στην χρήση φαρμάκων από γιατρούς για την αντιμετώπιση κάποιων ασθενειών.

Πίνακας 5

	penicilline	tifomycine	tetracyline	erythromycine	tiophenicol	gentalline
TYPHOIDE	0	4	0	0	2	0
SALMONELLOSE DIGESTIVE	0	2	0	0	1	0
AR.ORL	8	0	5	3	0	3
PENUMOPATHIE	7	0	5	2	0	3
MENINGITE	2	2	0	0	0	1
AFFECTION DES VOIES URINAIRES	4	0	2	0	0	6
STAPHYLOCOCCIE	3	0	1	3	0	0

Εφαρμόζοντας την Παραγοντική Ανάλυση των Αντιστοιχιών, παίρνουμε τα αποτελέσματα του πίνακα 6 και 7 και το παραγοντικό επίπεδο 1x3, όπως φαίνεται στο σχήμα 13.



σχήμα 13. Οι πίνακες με τα profil γραμμών και στηλών

Πίνακας 6

Α/Α	ΙΔΙΟΤΙΜΕΣ	ερμ. αδρ. %	ΑΘΡΟΙΣΤΙΚΑ	ΙΣΤΟΓΡΑΜΜΑ ΤΩΝ ΙΔΙΟΤΙΜΩΝ
2	.87831430	75.726	75.726	**********
3	.20369690	17.562	93.288	********
4	.04922439	4.244	97.532	***
5	.02857669	2.464	99.996	**

Πίνακας 7

Α/Α	ΕΝΕΡΓΑ ΑΝΤΙΚ	1ΠΑΡΑΓ	ΣΥΝ	ΣΥΜ	2ΠΑΡΑΓ	ΣΥΝ	ΣΥΜ
1	TYPHOIDE	-2314	994	530	-83	1	3
2	SALMONEL	-2314	994	265	-83	1	1
3	AR.ORL	440	861	61	-144	93	28
4	PENUMOPATH	438	831	54	-35	5	1
5	MENINGITE	-625	371	32	312	93	35
6	AF.V.URIN	411	213	33	758	724	491
7	STAPHYLOCOC	459	172	24	-941	725	441
	ΣΥΝΟΛΟ=			1000			1000
Α/Α	ΕΝΕΡΓΕΣ ΜΕΤΑΒ	1ΠΑΡΑΓ	ΣΥΝ	ΣΥΜ	2ΠΑΡΑΓ	ΣΥΝ	ΣΥΜ
1	penicilline	372	872	55	-52	17	5
2	tifomycine	-2019	976	538	36	0	1
3	tetracyline	466	653	46	-54	9	3
4	erythromyc	477	197	30	-921	737	482
5	tiophenicol	-2470	915	302	-183	5	7
6	gentalline	367	188	29	737	759	502
	ΣΥΝΟΛΟ=			1000			1000

2.5 Ερμηνεία των αποτελεσμάτων

Όπως και στην A.C.P. υπάρχουν διάφοροι δείκτες για να ερμηνεύσουμε τα αποτελέσματα μιας Παραγοντικής Ανάλυσης των Αντιστοιχιών. [Escofier,1990]

1. Ο λόγος της αδράνειας που εξηγεί ο παραγοντικός άξονας τάξης s (δηλαδή η χαρακτηριστική τιμή λ_s) προς την συνολική αδράνεια (άθροισμα όλων των χαρακτηριστικών τιμών), δίνει το ποσοστό της συνολικής αδράνειας που εξηγεί αυτός ο άξονας.

$$\text{ερμηνευόμενη αδράνεια}=\frac{\lambda_s}{\lambda_1+\lambda_2+...+\lambda_p}$$

2. Η συνεισφορά (Contribution) του αντικειμένου i στον άξονα s εκφράζει το ποσοστό συμμετοχής του i στο μέρος της αδράνειας που εξηγεί ο άξονας s.

$$Ctr_s(i)=\frac{f_i . F_s^2(i)}{\lambda_s}$$

όπου $F_s(i)$ είναι η προβολή του i στον άξονα s

Ο δείκτης αυτός, δείχνει ποια είναι τα σημεία που είναι κυρίως υπεύθυνα για την δημιουργία του παραγοντικού άξονα. Το άθροισμα των Ctrs για όλα τα σημεία είναι 1 (100%).

3. Η σχετική συνεισφορά του άξονα s στο σημείο I, εκφράζει μια μορφή συσχέτισης του σημείου i με τον παραγοντικό άξονα τάξης s. Αυτή εκφράζεται με την ποσότητα COR, που είναι ίση με το τετράγωνο του συνημιτόνου της γωνίας ω που σχηματίζει ο άξονας με την ευθεία που ενώνει το σημείο i με το κέντρο βάρους του νέφους.

$$Cor_s(i) = \frac{F_s^2(i)}{r^2(i)} = \cos^2\omega$$

όπου:
$F_s(i)$ είναι η προβολή του i στον άξονα s
r(i) η απόσταση του i από το κέντρο βάρους

Μια μεγάλη τιμή του Cors(i) μεταφράζεται σε μια μικρή γωνία ω, δηλαδή το i πλησιάζει τον άξονα s, ή λέμε ότι είναι υψηλά συσχετισμένο με τον άξονα αυτό. Επίσης ο δείκτης Cors(i) εκφράζει και το ποσοστό της αδράνειας του σημείου i που ερμηνεύεται από τον παραγοντικό άξονα s.

4. Για κάθε σημείο i, και για το πλήθος των πρώτων παραγοντικών αξόνων, (συνήθως 2 ή 3) ο δείκτης QLT, είναι ίσος με το άθροισμα των COR για τους άξονες αυτούς. Ο δείκτης αυτός εκφράζει την ποιότητα αναπαράστασης του σημείου i στο σύστημα των πρώτων παραγοντικών αξόνων. Όπως είναι φυσικό, για το σύνολο όλων των παραγοντικών αξόνων, έχουμε QLT =1 (100%).

Ο δείκτης CTR βοηθά στην αναζήτηση των σημείων που έχουν μεγάλη συμμετοχή στην δημιουργία των παραγοντικών αξόνων. Ο δείκτης COR βοηθά στον εντοπισμό των σημείων που παρουσιάζουν την καλύτερη αναπαράσταση από ένα παραγοντικό άξονα. Στις περισσότερες περιπτώσεις, σημεία με υψηλή τιμή CTR έχουν και υψηλή τιμή COR. Μπορεί όμως να συμβεί να υπάρχουν σημεία με χαμηλή τιμή CTR και υψηλή τιμή COR για κάποιο παραγοντικό άξονα. Τα σημεία αυτά παρόλο που δεν συμμετέχουν στην κατασκευή του παραγοντικού άξονα, ερμηνεύονται πολύ καλά απ' αυτόν. Ακόμη μπορεί να υπάρχουν σημεία με υψηλή τιμή CTR και χαμηλή τιμή COR για κάποιο παραγοντικό άξονα. Τα σημεία αυτά συνεισφέρουν στην δημιουργία του άξονα, αλλά η χαμηλή τιμή του COR, εξηγείται από το γεγονός ότι υπάρχουν και άλλοι παραγοντικοί άξονες με τους οποίους συσχετίζονται τα σημεία και για την κατασκευή των οποίων συνεισφέρουν.

Έτσι για την ερμηνεία του παραδείγματός μας (σχήμα 13), θα έχουμε:

Ο πρώτος παραγοντικός άξονας ερμηνεύει καλύτερα τις αρρώστιες typhoide και salmonellose (έχει μεγάλη τιμή του συνημιτόνου και αρνητική συντεταγμένη)

από τα αριστερά και ar.orl penumopath στα δεξιά του (παρουσιάζει επίσης μεγάλη τιμή του συνημιτόνου με θετική συντεταγμένη). Για τα φάρμακα, στα αριστερά του άξονα αντιπροσωπεύονται καλύτερα tifomycine tiophenicol (αφού έχουν αρνητικές συντεταγμένες με υψηλή τιμή συνημιτόνου), ενώ στα δεξιά του penicilline tetracyline (παρουσιάζουν υψηλές τιμές συνημιτόνου με θετικές συντεταγμένες).

Ο δεύτερος παραγοντικός άξονας ερμηνεύει καλύτερα τις αρρώστιες af.v.urin στο πάνω μέρος του (τιμή συνημιτόνου 724 με συντεταγμένη 758) και staphylococ στην αντίθετη πλευρά του (αντίστοιχη τιμή συνημιτόνου 725 με συντεταγμένη -941). Στο κάτω μέρος παρουσιάζεται το φάρμακο erythromyc (συνημίτονο 737 με συντεταγμένη -921) ενώ στο πάνω μέρος το gentalline (συνημίτονο 759 με συντεταγμένη 737).

Παρατηρώντας το επίπεδο που σχηματίζεται από τους 1 και 2 παραγοντικούς άξονες διακρίνουμε αμέσως 2 ομάδες. Μία στα αριστερά του επιπέδου όπου υπάρχουν οι αρρώστιες typhoide και salmonel με τα φάρμακα tifomycine tiophenicol, ενώ στη δεξιά μεριά η άλλη ομάδα περιλαμβάνει τις αρρώστιες ar.orl penumopath af.v.urin staphylococ με αντίστοιχα φάρμακα erythromyc gentalline.

3. Ταξινόμηση (classification)

Η ταξινόμηση χρησιμοποιήθηκε πολύ παλιά από διάφορες επιστήμες. Αυτό οφείλεται στο γεγονός ότι προσαρμόζεται καλύτερα στην ανθρώπινη μνήμη και στις διαδικασίες ανάλυσης και σύνθεσης του εγκεφάλου μας. [Benzecri,1973]

Με την ταξινόμηση όπως και με τις άλλες μεθόδους της Ανάλυσης Δεδομένων, επιχειρείται η απεικόνιση σε όσο το δυνατό απλούστερο σχήμα ενός πίνακα δεδομένων του οποίου οι γραμμές είναι οι παρατηρήσεις (ή αντικείμενα) που περιγράφονται από ένα σύνολο μεταβλητών που αντιστοιχούν στις στήλες του.

Αντικειμενικός σκοπός της ταξινόμησης είναι να επιτευχθεί διαμελισμός του συνόλου των παρατηρήσεων σε ομογενείς ομάδες, ως προς το σύνολο των μεταβλητών, κάθε μια των οποίων να διαφέρει σημαντικά από τις υπόλοιπες. Μ' αυτό τον τρόπο επιτυγχάνεται μια ιεράρχηση των παρατηρήσεων, δηλαδή, μια σειρά διαμελισμών "ο ένας μέσα στον άλλο", που όσο προχωρά, τόσο πιο λεπτομερής γίνεται. Η ταξινόμηση κατ' αύξουσα ιεραρχία ξεκινά από το γεγονός ότι κάθε αντικείμενο της ανάλυσης αποτελεί μια διακεκριμένη ομάδα και ενώνοντας τις κοντινότερες ομάδες, καταλήγει σε μια μόνο ομάδα που περιλαμβάνει όλα τα αντικείμενα. Παραδείγματα αποτελεσμάτων της μεθόδου αυτής θα δούμε σε επόμενα κεφάλαια.

Η βασική διαφορά μεταξύ ταξινόμησης και κατάταξης είναι ο τρόπος δημιουργίας των διαδοχικών διαμελισμών των δεδομένων. Στην κατάταξη τα αντικείμενα τοποθετούνται σε ομάδες καθορισμένες εκ των προτέρων, με βάση κάποια χαρακτηριστικά τους. Αντίθετα στην ταξινόμηση αναζητούνται και προσδιορίζονται οι ομάδες των διαμελισμών λαμβάνοντας υπόψιν τα δεδομένα και τις μεταβλητές που τα χαρακτηρίζουν.

Σε πολλές επιστήμες επιθυμούμε να έχουμε διαδοχικές ομαδοποιήσεις των αρχικών δεδομένων, ώστε να τα μελετάμε καλύτερα και ευκολότερα. Στην ιατρική για τις αρρώστιες και τα συμπτώματά τους, στη βοτανική και ζωολογία για την κατάταξη των φυτών και ζώων. Στο marketing και την έρευνα αγοράς για την συμπεριφορά των ατόμων.

Στην τεχνητή νοημοσύνη, ο υπολογιστής "μαθαίνει" και προχωράει στο επόμενο βήμα, αφού ομαδοποιήσει μια σειρά από πληροφορίες. Η ταξινόμηση (classification) χρησιμοποιείται ακόμη στην γεωλογία, γεωγραφία, γλωσσολογία, αρχαιολογία, οικονομία και αλλού.

Η ταξινόμηση συχνά είναι συμπληρωματική σ' άλλες μεθόδους της στατιστικής. Βοηθά στην ερμηνεία του γραφήματος που προκύπτει από μια παραγοντική ανάλυση των αντιστοιχιών, ή ορίζει καλύτερα τις ομογενείς ομάδες πριν από μια πολλαπλή γραμμική παλινδρόμηση.

Πιο λεπτομερής ανάλυση της μεθόδου της ταξινόμησης θα γίνει στο κεφάλαιο 4.

4. Επίλογος

Παρουσιάσαμε στο κεφάλαιο αυτό τρεις από τις κυριότερες μεθόδους της Ανάλυσης Δεδομένων. Με τις μεθόδους αυτές μπορούμε να αναλύσουμε δεδομένα που χαρακτηρίζονται τόσο από ποσοτικές, όσο και από ποιοτικές μεταβλητές. Η παρουσίαση των αποτελεσμάτων γίνεται με ένα γράφημα, στο οποίο μπορούμε να δούμε την όσο το δυνατόν πληρέστερη εικόνα του φαινομένου που μελετάμε. Παρόλα αυτά για την καλύτερη ερμηνεία και κατανόηση των αποτελεσμάτων υπάρχουν και διάφοροι ποσοτικοί δείκτες, πολύ χρήσιμοι στην εξαγωγή συμπερασμάτων.

Στην συνέχεια της διατριβής αυτής θα ασχοληθούμε ιδιαίτερα με την μέθοδο της Ταξινόμησης κατ' αύξουσα ιεραρχία, η οποία χρησιμοποιείται για την ανάλυση δεδομένων, τόσο σαν αυτόνομη μέθοδος, όσο και σαν συμπληρωματική των μεθόδων της Παραγοντικής Ανάλυσης των Αντιστοιχιών και της Ανάλυσης σε Κύριες Συνιστώσες, που παρουσιάστηκαν στο κεφάλαιο αυτό.

ΚΕΦΑΛΑΙΟ 2
ΠΕΡΙΓΡΑΦΗ ΜΕΤΑΒΛΗΤΩΝ- ΜΕΤΑΤΡΟΠΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΠΙΝΑΚΕΣ ΚΑΤΑΛΛΗΛΟΥΣ ΝΑ ΔΕΧΘΟΥΝ ΕΠΕΞΕΡΓΑΣΙΑ ΜΕ ΜΕΘΟΔΟΥΣ ΤΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ - ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

Στο κεφάλαιο αυτό παρουσιάζουμε τους στατιστικούς πίνακες που χρησιμοποιούνται από τις μεθόδους της Ανάλυσης Δεδομένων. Περιγράψαμε τον τρόπο επεξεργασίας ενός απλού πίνακα με τις μεθόδους της Παραγοντικής Ανάλυσης των Αντιστοιχιών (AFC) και της Ανάλυσης σε Κύριες Συνιστώσες (ACP), σε προηγούμενο κεφάλαιο. Προτείνουμε εδώ τη δημιουργία και επεξεργασία νέων πινάκων και παρουσιάζουμε μια εφαρμογή-παράδειγμα επεξεργασίας τρισδιάστατου πίνακα με την μέθοδο της Παραγοντικής Ανάλυσης των Αντιστοιχιών (AFC). Η επεξεργασία πινάκων με την μέθοδο της Αυτόματης Ταξινόμησης (CAH), δεν θα περιγραφεί εδώ αφού θα παρουσιαστεί αναλυτικά στο κεφάλαιο 4. Η τελευταία αυτή μέθοδος είναι απαραίτητη και σαν συμπληρωματική των δύο άλλων μεθόδων δίνοντας πολύ χρήσιμα συμπεράσματα.

1. Εισαγωγή

Για τη μελέτη διαφόρων φαινομένων και προκειμένου να λάβουμε σημαντικές αποφάσεις, πολλές φορές χρειάζεται να προβούμε σε μια πρώτη κλασσική στατιστική ανάλυση. Αναλύοντας μ' αυτή τα δεδομένα της έρευνας μας, εξάγουμε διάφορα αποτελέσματα και συμπεράσματα, τα οποία μας βοηθούν και μας οδηγούν στη λήψη των καταλληλότερων αποφάσεων, περαιτέρω στατιστικής διερεύνησης.

Τα βήματα για μια στατιστική έρευνα με την Ανάλυση Δεδομένων [Diday,1982] είναι:

- Συλλογή των απαραίτητων στοιχείων της έρευνας.
- Κωδικοποίηση των μελετώμενων μεταβλητών για την κατασκευή του πίνακα δεδομένων πάνω στον οποίο θα γίνει η ανάλυση. (Η φάση αυτή συνίσταται στην κατάργηση κάποιων μεταβλητών, στη διαχείριση των ελλειπών δεδομένων, στη διόρθωση λαθών, ομογενοποίηση μεταβλητών κλπ.)
- Επιλογή μετρικής για τη μέτρηση της ομοιότητας μεταξύ δύο παρατηρήσεων.
- Επιλογή της καταλληλότερης μεθόδου ανάλυσης του πίνακα δεδομένων.
- Παρουσίαση των αποτελεσμάτων της έρευνας με όσο γίνεται κατανοητό τρόπο, ώστε να εξάγουμε εύκολα τα συμπεράσματα για την λήψη των ορθότερων αποφάσεων.

Στο κεφάλαιο αυτό, περιγράφουμε το στάδιο της στατιστικής έρευνας, που ακολουθεί την συλλογή στοιχείων, κατά την διάρκεια του οποίου, δημιουργείται ο πίνακας δεδομένων της έρευνας, ο οποίος θα επεξεργαστεί αργότερα με μεθόδους της Ανάλυσης Δεδομένων όπως θα δούμε στα επόμενα κεφάλαια.

Ένα πρώτο βήμα για την αξιοποίηση των δεδομένων μιας έρευνας, γίνεται με τη βοήθεια της Περιγραφικής Στατιστικής. Η Περιγραφική Στατιστική ασχολείται με τις διάφορες μεθόδους που μας επιτρέπουν να οργανώσουμε, να ταξινομήσουμε, να συμπτήξουμε και να συνοψίσουμε τις παρατηρήσεις μας. [Παπαδημητρίου,1990]

Σκοπός της Περιγραφικής Στατιστικής είναι:

- Η περιληπτική παρουσίαση ενός συνόλου στατιστικών δεδομένων σε τέτοια μορφή, ώστε να προκύπτουν με τρόπο άμεσο και πλήρη τα βασικά τους χαρακτηριστικά.
- Ο υπολογισμός διαφόρων παραμέτρων οι οποίες μεμονωμένα ή σε συνδυασμό αντικαθιστούν (παραμετρούν) ένα σύνολο δεδομένων χωρίς να χάνονται ουσιαστικές πληροφορίες τους.

2. Ορισμοί

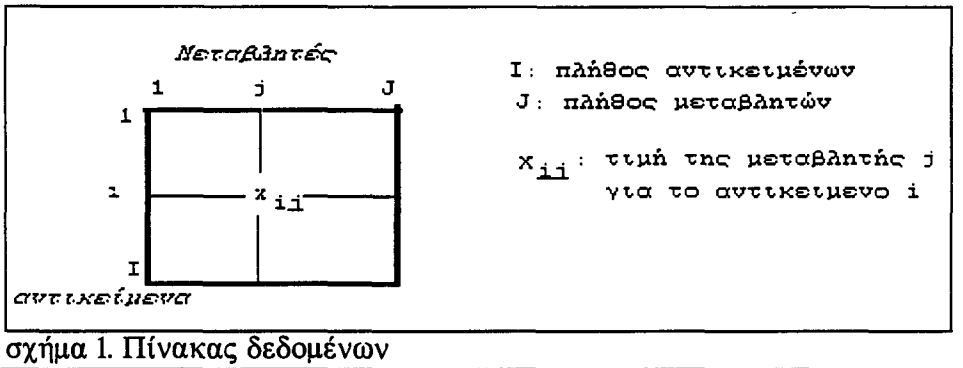
Στο κεφάλαιο αυτό θα ασχοληθούμε με τους τρόπους κατασκευής και μορφοποίησης καταλλήλων πινάκων δεδομένων και τις δυνατότητες ανάλυσης και ερμηνείας τέτοιων πινάκων. Πριν προχωρήσουμε, παραθέτουμε τον ορισμό διαφόρων εννοιών, χρήσιμων για την κατανόηση όσων ακολουθούν.

Πληθυσμός ονομάζεται στην Στατιστική το σύνολο των ατόμων ή αντικειμένων, στα οποία αναφέρονται οι παρατηρήσεις μας. Τα στοιχεία του συνόλου αυτού λέγονται **στατιστικές μονάδες** ή **άτομα**, ή **αντικείμενα**. Η Στατιστική ασχολείται με την μελέτη των ιδιοτήτων των στατιστικών μονάδων. Οι ιδιότητες αυτές ονομάζονται **μεταβλητές**. Οι αριθμοί ή οι συμβολικές εκφράσεις, που αντιπροσωπεύουν τις καταστάσεις μιας μεταβλητής ονομάζονται **τιμές** της μεταβλητής. Ανάλογα με το είδος των τιμών τους, οι μεταβλητές χωρίζονται σε δύο μεγάλες κατηγορίες. [Ζαχαροπούλου, 1989]

- Ποιοτικές μεταβλητές οι οποίες δεν μπορούν να μετρηθούν και οι τιμές τους εκφράζονται με λέξεις. Π.χ. Η μεταβλητή "το χρώμα ματιών ενός ανθρώπου", μπορεί να πάρει τις τιμές μαύρο, καστανό, κυανούν, πράσινο κτλ.
- Ποσοτικές μεταβλητές οι οποίες μπορούν να μετρηθούν και οι τιμές τους είναι αριθμοί αναφερόμενοι σε συγκεκριμένες μονάδες μέτρησης. Π.χ. Η μεταβλητή "ηλικία", μπορεί να έχει ως τιμές 25 χρόνια, 30 χρόνια, κτλ. Η μεταβλητή ύψος, παίρνει τιμές στο διάστημα [1m, 2.50m].

3. Πίνακες δεδομένων

Ο τρόπος παρουσίασης των στοιχείων κάποιας έρευνας, γίνεται συνήθως με έναν πίνακα. Ο απλούστερος τέτοιος πίνακας, για την συνολική περιγραφή ενός φαινομένου, είναι διαστάσεων IxJ, όπου I είναι το πλήθος των στατιστικών αντικειμένων και J το πλήθος των μελετώμενων μεταβλητών, που περιγράφουν τα αντικείμενα αυτά. (σχήμα 1). Οι μεταβλητές αυτές μπορεί να είναι ποσοτικές ή ποιοτικές. Όταν πρόκειται για ποσοτικές μεταβλητές, στα κελιά του πίνακα υπάρχουν αριθμοί που παριστάνουν τις πραγματικές τιμές των μεταβλητών. Στην περίπτωση των ποιοτικών μεταβλητών, στα κελιά του πίνακα υπάρχουν λέξεις που αντιστοιχούν στις κατηγορίες των μεταβλητών ή μπορεί να υπάρχουν ακέραιοι αριθμοί, οι οποίοι δεν αντιστοιχούν σε πραγματικές τιμές, αλλά απλώς σε μια κωδικοποίηση των κατηγοριών. Αυτό γίνεται ώστε να έχουμε καλύτερη επεξεργασία και αποθήκευση ενός πίνακα ποιοτικών μεταβλητών μέσω του ηλεκτρονικού υπολογιστή. [Παπαδημητρίου,1990]



σχήμα 1. Πίνακας δεδομένων

Ένα παράδειγμα ενός πίνακα δεδομένων, φαίνεται στον πίνακα 1. Τα αντικείμενα στα οποία αναφέρεται, είναι ένα δείγμα εφτά φοιτητών. Οι μελετώμενες μεταβλητές για κάθε φοιτητή είναι φύλο, καταγωγή, εισόδημα πατέρα, μέσος όρος βαθμολογίας. Οι δύο πρώτες είναι ποιοτικές μεταβλητές και οι άλλες δύο ποσοτικές.

Πίνακας 1

φοιτητές	φύλο	καταγωγή	εισόδημα πατέρα	μέσος όρος βαθμολογίας
1	άνδρας	Μακεδονία	2.300.000	8
2	γυναίκα	Ήπειρος	1.200.000	9
3	γυναίκα	Θράκη	3.800.000	7
4	άνδρας	Θράκη	4.500.000	6
5	άνδρας	Μακεδονία	10.600.000	4
6	γυναίκα	Θεσσαλία	7.000.000	5
7	γυναίκα	Ήπειρος	6.500.000	8

Από ένα αρχικό πίνακα δεδομένων, παράγουμε διάφορους άλλους πίνακες που μας βοηθούν στην επεξεργασία, στην ανάλυση δεδομένων και στην ερμηνεία αποτελεσμάτων.

Ένα είδος πίνακα που προκύπτει με μετασχηματισμό του αρχικού πίνακα δεδομένων, είναι ο πίνακας **0-1 ή "λογικός" ή διαζευκτικός** πίνακας. Χρησιμοποιείται για την ομογενοποίηση ανομοιογενών μεταβλητών και παρουσίαση ποιοτικών μεταβλητών ή ποσοτικών χωρισμένων σε κλάσεις. Οι γραμμές ενός τέτοιου πίνακα αντιστοιχούν στο πλήθος των παρατηρήσεων, ενώ ο αριθμός των στηλών του, είναι ίσος με το άθροισμα των κατηγοριών όλων των μελετώμενων μεταβλητών. Σε κάθε κελλί του πίνακα (διασταύρωση γραμμής και στήλης) υπάρχει ο αριθμός 1 ή ο αριθμός 0, ανάλογα με τον αν η παρατήρηση που αντιστοιχεί σ' αυτή την γραμμή, χαρακτηρίζεται ή όχι από την κατηγορία που αντιπροσωπεύει η αντίστοιχη στήλη. [Παπαδημητρίου,1990] [Escofier,1990]

Παράδειγμα τέτοιου πίνακα για την παρουσίαση των μεταβλητών φύλο (ποιοτική) και μέσος όρος βαθμολογίας (ποσοτική χωρισμένη σε 3 κλάσεις), είναι ο επόμενος πίνακας 2.

Πίνακας 2

	φύλο		μέσος όρος βαθμολογίας		
φοιτητές	άνδρας	γυναίκα	0-5	6-8	9-10
1	1	0	0	1	0
2	0	1	0	0	1
3	0	1	0	1	0
4	1	0	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	0	1	0	1	0

Όταν κάθε παρατήρηση μπορεί να χαρακτηρίζεται μόνο από μια κατηγορία κάθε μεταβλητής, το άθροισμα των μονάδων μιας γραμμής είναι ίσο με το πλήθος των μεταβλητών. Στο παραπάνω παράδειγμα το άθροισμα κάθε γραμμής είναι 2 αφού παρουσιάζουμε μόνο δύο μεταβλητές. Οι αντίστοιχες σχέσεις περιγράφονται από τον τύπο:

$\sum_{j=1}^{mJ} \mathbf{X}_{ij} = \mathrm{J} \quad \forall i \in \{1,\ldots,n\}$ όπου J είναι το πλήθος μεταβλητών, mJ το πλήθος των κατηγοριών όλων των μεταβλητών, n το πλήθος των παρατηρήσεων και x_{ij} η τιμή του λογικού πίνακα για την i γραμμή και j στήλη. [Παπαδημητρίου,1994]

Για την περίπτωση που μια παρατήρηση μπορεί να χαρακτηρίζεται από περισσότερες κατηγορίες της ίδιας μεταβλητής, το άθροισμα μπορεί να είναι

διάφορο από το πλήθος των μεταβλητών. Εκτενέστερη αναφορά για την περίπτωση αυτή θα κάνουμε στο τρίτο κεφάλαιο.

4. Πίνακες μεταβλητών

Οι πίνακες μεταβλητών, χρησιμοποιούνται για την εποπτικότερη παρουσίαση των μεταβλητών αλλά και για την καλύτερη και ευκολότερη ανάλυση των μεταβλητών και των σχέσεων μεταξύ τους. Ο απλούστερος πίνακας για την παραστατική παρουσίαση μιας μεταβλητής, είναι ο πίνακας συχνοτήτων. [Tenenbaus,1994]

Πρόκειται για ένα πίνακα απλής εισόδου με δύο στήλες, που αναφέρεται σε μια μόνο μεταβλητή. Στην πρώτη στήλη υπάρχουν οι τιμές ή διαστήματα τιμών της μεταβλητής και στην δεύτερη οι "συχνότητες" των τιμών αυτών. Δηλαδή το πλήθος των παρατηρήσεων που παρουσιάζουν (κατέχουν) την αντίστοιχη τιμή της μεταβλητής. Παράδειγμα για την μεταβλητή φύλο, οι συχνότητες των φοιτητών του παραπάνω παραδείγματος παρουσιάζονται στον παρακάτω πίνακα 3. Ρίχνοντας μια ματιά στον πίνακα αυτό παρατηρούμε εύκολα και γρήγορα πόσοι φοιτητές είναι άνδρες και πόσοι γυναίκες.

Πίνακας 3

τιμές μεταβλητής	συχνότητες
άνδρας	3
γυναίκα	4

Αν θέλουμε τις ίδιες πληροφορίες αλλά παρουσιασμένες με μορφή ποσοστών σε σχέση με το σύνολο των παρατηρήσεων, σχηματίζουμε τον πίνακα σχετικών συχνοτήτων, διαιρώντας κάθε συχνότητα με το πλήθος των παρατηρήσεων, όπως στον πίνακα 4. Από τον πίνακα αυτό, βλέπουμε ότι οι άνδρες αποτελούν ποσοστό 42.9%, ενώ οι γυναίκες 57.1%

Πίνακας 4

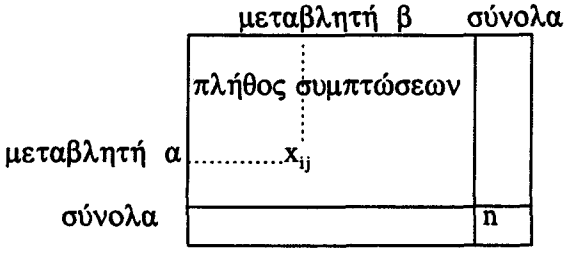
τιμές μεταβλητής	συχνότητες
άνδρας	3/7=42,9%
γυναίκα	4/7=57,1%

Όταν ζητάμε τον προσδιορισμό της σχέσης και επίδρασης των μεταβλητών μεταξύ τους, οι παραπάνω πίνακες δεν μας δίνουν μεγάλη βοήθεια. Το σύνολο των σχέσεων αυτών προκύπτει εύκολα από τους παρακάτω προτεινόμενους πίνακες.

Ένα είδος πίνακα για την παρουσίαση δύο μεταβλητών και της μεταξύ τους σχέσης είναι ο **πίνακας συμπτώσεων ή συνάφειας**. Ένας πίνακας

συμπτώσεων παριστάνει την σύγχρονη κατανομή δύο μεταβλητών στις παρατηρήσεις μας. Οι δύο μεταβλητές είναι ποιοτικές ή ποσοτικές αλλά χωρισμένες σε κλάσεις, οπότε αντιμετωπίζονται σαν ποιοτικές μεταβλητές. Ο πίνακας συμπτώσεων συνδέει δύο ποιοτικές μεταβλητές, ή δύο ποσοτικές και είναι ο μόνος πίνακας που μπορεί να παρουσιάσει ταυτόχρονα τη σχέση ανάμεσα σε μια ποιοτική και μια ποσοτική μεταβλητή, αφού αυτές έχουν κατηγοριοποιηθεί. [Escofier,1990]

Στη συνέχεια όταν αναφέρουμε μεταβλητή εννοούμε ποιοτική μεταβλητή που περιγράφεται από κάποιες τιμές, ονομαζόμενες κατηγορίες. Στην περίπτωση ποσοτικής μεταβλητής οι κλάσεις της θα ονομάζονται επίσης κατηγορίες. Ο πίνακας συμπτώσεων εμφανίζει την μορφή ενός πίνακα όπως στο σχήμα 2, για δύο μεταβλητές α και β.



σχήμα 2. Πίνακας συμπτώσεων δύο μεταβλητών

Οι γραμμές του είναι όσες και οι κατηγορίες της μεταβλητής α, και οι στήλες του, όσες οι κατηγορίες της β μεταβλητής. Στην διασταύρωση μιας γραμμής (κατηγορία i της α μεταβλητής) και μιας στήλης (κατηγορία j της β μεταβλητής), υπάρχει ο αριθμός x_{ij} που παριστάνει το πλήθος των παρατηρήσεων μας, που έχουν συγχρόνως την ιδιότητα να χαρακτηρίζονται από την i κατηγορία και την j κατηγορία των μεταβλητών α και β αντίστοιχα.

Το άθροισμα κάθε γραμμής είναι το πλήθος των παρατηρήσεων που έχουν την ιδιότητα να χαρακτηρίζονται από την i κατηγορία της α μεταβλητής που αντιστοιχεί στην γραμμή αυτή. Παρόμοια το άθροισμα κάθε στήλης παριστάνει το πλήθος των παρατηρήσεων που χαρακτηρίζονται από την αντίστοιχη j κατηγορία της β μεταβλητής. Τέλος το άθροισμα όλων των κελιών του πίνακα είναι ίσο με το πλήθος των παρατηρήσεων n.

Για το παράδειγμα μας, η σχέση των μεταβλητών φύλο και μέσος όρος βαθμολογίας, παριστάνεται από τον παρακάτω πίνακα 5.

Πίνακας 5

	0-5	6-8	9-10	σύνολο
άνδρας	1	2	0	3
γυναίκα	1	2	1	4
σύνολο	2	4	1	7

Παρατηρώντας τον πίνακα 5 διακρίνουμε αμέσως, ότι στο δείγμα μας έχουμε 3 άνδρες και 4 γυναίκες, 2 φοιτητές με βαθμολογία 0-5, 4 φοιτητές με βαθμολογία 6-8 και 1 με βαθμό άριστα 9-10. Επίσης με μια ματιά βλέπουμε ότι ένας άνδρας και μια γυναίκα έχουν βαθμό 0-5, δύο άνδρες και δύο γυναίκες έχουν βαθμολογία 6-8 και μόνο μια γυναίκα έχει βαθμό άριστα. Τις πληροφορίες αυτές δεν είναι εύκολο να τις διακρίνουμε άμεσα από τον αρχικό πίνακα.

Ο παραπάνω πίνακας 5, ονομάζεται ειδικότερα και πίνακας συμπτώσεων απολύτων συχνοτήτων αφού σε κάθε κελί εμφανίζεται η απόλυτη συχνότητα των συμπτώσεων των δύο κατηγοριών των δύο μεταβλητών. Υπάρχει και ο πίνακας συμπτώσεων σχετικών συχνοτήτων. Η δομή του είναι παρόμοια με τον πίνακα συμπτώσεων απολύτων συχνοτήτων, μόνο που η συχνότητα κάθε κελιού διαιρείται με το πλήθος των παρατηρήσεων και έτσι έχουμε τις σχετικές συχνότητες. Τέτοιου είδους πίνακες είναι χρήσιμοι, όταν θέλουμε να συγκρίνουμε ζεύγη μεταβλητών, γιατί δεν εξαρτώνται από το πλήθος των παρατηρήσεων που μετρήσαμε για κάθε ζεύγος. Ο προηγούμενος πίνακας 5 συμπτώσεων απολύτων συχνοτήτων, διαιρώντας με το σύνολο των φοιτητών του δείγματος, μετασχηματίζεται στον πίνακα 6, συμπτώσεων σχετικών συχνοτήτων.

Πίνακας 6

	0-5	6-8	9-10	σύνολο
άνδρας	14.3%	28.6%	0%	42.9%
γυναίκα	14.3%	28.6%	14.3%	57.1%
σύνολο	28.6%	57.1%	14.3%	100%

Ο πίνακας 6 είναι πολύ χρήσιμος, όταν μας ενδιαφέρει η κατανομή των ποσοστών των αντικειμένων του δείγματος μας. Φαίνεται τόσο η κατανομή ποσοστών για κάθε μεταβλητή όσο και τα ποσοστά των συμπτώσεων μεταξύ δύο κατηγοριών των μεταβλητών.

Βάζοντας στη σειρά (δίπλα-δίπλα και έναν κάτω από τον άλλο) πολλούς πίνακες συμπτώσεων, σχηματίζουμε ένα πίνακα συμπτώσεων πολλών μεταβλητών όπως τον επόμενο πίνακα 7. Ο πίνακας αυτός είναι γνωστός σαν **πίνακας Burt**. [Benzecri, 1973]. Οι γραμμές του και οι στήλες του είναι ίσες σε πλήθος και ίσες με το άθροισμα των κατηγοριών όλων των μελετώμενων μεταβλητών. Ο πίνακας είναι συμμετρικός, και στην κύρια διαγώνιο του βρίσκονται τετράγωνοι διαγώνιοι

υποπίνακες, που παριστάνουν την κατανομή των κατηγοριών μιας μεταβλητής, στις παρατηρήσεις μας. Οι υπόλοιποι υποπίνακες είναι πίνακες συμπτώσεων ανά δύο όλων των μεταβλητών.

Ο επόμενος πίνακας 7 είναι ο αντίστοιχος πίνακας Burt για το παράδειγμα μας. Αναλύοντας τον πίνακα αυτό, μπορούμε να παρατηρήσουμε τόσο την κατανομή κάθε μεταβλητής (διαγώνιοι πίνακες), όσο και τους πίνακες συμπτώσεων όλων των μεταβλητών ανά δύο.

Πίνακας 7

	άνδρας	γυναίκα	Μακ.	Θράκ.	Ηπειρ.	Θεσ.	<3000000	>3000000	0-5	6-8	9-10
άνδρας	3	0	2	1	0	0	1	2	1	2	0
γυναίκα	0	4	0	1	2	1	1	3	1	2	1
Μακ.	2	0	2	0	0	0	1	1	1	1	0
Θράκ.	1	1	0	2	0	0	0	2	0	2	0
Ηπειρ.	0	2	0	0	2	0	1	1	0	1	1
Θεσ.	0	1	0	0	0	1	0	1	1	0	0
<3000000	1	1	1	0	1	0	2	0	0	1	1
>3000000	2	3	1	2	1	1	0	5	2	3	0
0-5	1	1	1	0	0	1	0	2	2	0	0
6-8	2	2	1	2	1	0	1	3	0	4	0
9-10	0	1	0	0	1	0	1	0	0	0	1

Ο πίνακας Burt, χρησιμοποιείται στην παραγοντική ανάλυση (όπως αναφέραμε στο προηγούμενο κεφάλαιο), αλλά είναι πολύ χρήσιμος και στην Περιγραφική Στατιστική αφού βοηθάει να έχουμε μια γενική εποπτική εικόνα όλων των μεταβλητών και των σχέσεων μεταξύ τους.

4.1 Κατασκευή πίνακα Burt

Έστω το σύνολο I των αντικειμένων που χαρακτηρίζονται από ένα σύνολο J μεταβλητών. Κάθε μεταβλητή j είναι χωρισμένη σε ένα πλήθος κατηγοριών, που υποθέτουμε αριθμημένες από 1 έως n_j. Δηλαδή για το σύνολο J των μεταβλητών έχουμε το σύνολο mJ των κατηγοριών.

$mJ=\sum n_j$

Το σύνολο I των αντικειμένων μπορεί να περιγραφεί από ένα λογικό πίνακα D, IxmJ, με στοιχεία 0 και 1, όπου στην διασταύρωση της i γραμμής και j στήλης υπάρχει 1 αν το i αντικείμενο χαρακτηρίζεται από την j κατηγορία, και 0 διαφορετικά.

Από ένα τέτοιο λογικό πίνακα D, προκύπτει ένας τετράγωνος συμμετρικός πίνακας B, mJxmJ, που ονομάζεται πίνακας συμπτώσεων όλων των μεταβλητών ή πίνακας Burt. Η διανυσματική εξίσωση για την δημιουργία του πίνακα αυτού είναι:

$$\underset{(mJxmJ)}{B} = \underset{(mJxI)}{D'} \; \underset{(IxmJ)}{D}$$

όπου D′ είναι ο ανάστροφος του πίνακα D. Ο πίνακας B, δεν περιέχει τίποτε άλλο, παρά την απαρίθμηση όλων των συμπτώσεων των κατηγοριών. Στην k γραμμή και l στήλη του πίνακα αυτού B(k,l), περιέχεται ο αριθμός των αντικειμένων i που είναι τέτοια ώστε D(i,k)=D(i,l)=1.

Από τον ορισμό αυτό προκύπτει ότι:

B(k,k)=πλήθος των αντικειμένων που κατέχουν την ιδιότητα k, και

B(k,l)=0, αν k,l είναι δύο διαφορετικές κλάσεις της ίδιας μεταβλητής, αφού κάθε αντικείμενο ανήκει σε μια μόνο κλάση μιας μεταβλητής.

5. Επεξεργασία πίνακα συμπτώσεων

Αναλύουμε ένα πίνακα συμπτώσεων προκειμένου να βρούμε αν υπάρχει σχέση ή όχι μεταξύ των περιγραφόμενων μεταβλητών. Ακόμη ένας πίνακας συμπτώσεων είναι χρήσιμος στον υπολογισμό του συντελεστή συσχέτισης δύο μεταβλητών, αλλά και διάφορων άλλων μέτρων ομοιότητας μεταξύ μεταβλητών, όπως θα δούμε στο κεφάλαιο 2.

5.1 Υπό συνθήκη κατανομή μιας μεταβλητής σε σχέση με μια άλλη

Αν διαιρέσουμε κάθε γραμμή ενός πίνακα συμπτώσεων δύο μεταβλητών με το άθροισμα όλων των στοιχείων της, προκύπτει ένας νέος πίνακας που παριστάνει τα ποσοστά των παρατηρήσεων της κατηγορίας που αντιστοιχεί στη γραμμή αυτή, τα οποία κατανέμονται στις κατηγορίες της άλλης μεταβλητής (στήλες του πίνακα). Παρόμοια διαιρώντας κάθε στήλη του πίνακα με το άθροισμα των γραμμών της, προκύπτει ένας νέος πίνακας που παριστάνει την κατανομή των παρατηρήσεων μιας κατηγορίας (που αντιστοιχεί στην στήλη) στις κατηγορίες της άλλης μεταβλητής (γραμμές του πίνακα).

Με τον τρόπο αυτό ξεκινώντας από ένα πίνακα συμπτώσεων παίρνουμε δύο άλλους πίνακες, πολύ χρήσιμους για την εξαγωγή συμπερασμάτων μεταξύ των κατηγοριών των δύο μεταβλητών, τους οποίους ονομάζουμε **πίνακες υπό συνθήκη σχετικών συχνοτήτων**. [Benzecri,1973] [Greenacre,1993]

5.1.1 Παράδειγμα

Επανερχόμενοι πάλι στο παράδειγμα μας, και για τις δύο μεταβλητές φύλο και βαθμός, έχουμε τον παρακάτω πίνακα συμπτώσεων 8.

Πίνακας 8

	0-5	6-8	9-10	σύνολο
άνδρας	1	2	0	3
γυναίκα	1	2	1	4
σύνολο	2	4	1	7

Διαιρώντας κάθε γραμμή του πίνακα 8 με το αντίστοιχο σύνολο, λαμβάνουμε τον επόμενο πίνακα 9. Στον πίνακα 9, παρατηρούμε εύκολα τα εξής: Το ένα τρίτο των ανδρών (33.3%) έχει βαθμό μικρότερο ή ίσο με 5, ενώ τα δύο τρίτα (ποσοστό 66.7%) έχουν βαθμολογία μεταξύ 6 και 8. Για τις γυναίκες το 50% παρουσιάζει βαθμολογία 6-8, το 25% έχει βαθμό 0-5, ενώ το υπόλοιπο 25% έχει βαθμό άριστα 9-10. Τέλος στο δείγμα μας έχουμε 28.6% των φοιτητών με βαθμό 0-5, 52.1% με βαθμό 6-8, και μόνο 14.3% με βαθμό 9 ή 10.

Πίνακας 9

	0-5	6-8	9-10	σύνολο
άνδρας	33.3%	66.7%	0%	100%
γυναίκα	25%	50%	25%	100%
σύνολο	28.6%	52.1%	14.3%	100%

Αν τώρα διαιρέσουμε τον πίνακα 8 συμπτώσεων των δύο μεταβλητών με το άθροισμα κάθε στήλης, σχηματίζεται ο επόμενος πίνακας 10, στον οποίο παρατηρούμε ότι υπάρχει μια ισοκατανομή του φύλου τόσο για τους φοιτητές με βαθμό 0-5 όσο και για αυτούς που έχουν βαθμολογία 6-8 (ποσοστά 50% για τους άνδρες και 50% για τις γυναίκες). Αντίθετα το σύνολο των φοιτητών με βαθμό άριστα είναι γυναίκες (ποσοστό γυναικών 100%).

Πίνακας 10

	0-5	6-8	9-10	σύνολο
άνδρας	50%	50%	0%	42.9%
γυναίκα	50%	50%	100%	57.1%
σύνολο	100%	100%	100%	100%

Παρατηρούμε ότι οι δύο παραπάνω πίνακες υπό συνθήκη σχετικών συχνοτήτων, θεωρούμενοι σαν υποπίνακες ενός πίνακα Burt, δεν είναι συμμετρικοί ως προς την κύρια διαγώνιο του πίνακα Burt, αν και προκύπτουν από τον ίδιο πίνακα συμπτώσεων που είναι συμμετρικός ως προς την κύρια διαγώνιο, υποπίνακας του πίνακα Burt.

Μπορούμε να συμπτύξουμε τους 3 παραπάνω πίνακες μόνο σε ένα ώστε να εξάγουμε γρήγορα όλα τα χρήσιμα συμπεράσματα. Αυτό γίνεται φτιάχνοντας ένα πίνακα με 3 τιμές σε κάθε κελί. Μία για την απόλυτη συχνότητα (ή για τη σχετική συχνότητα), μία για το ποσοστό σε σχέση με το σύνολο της γραμμής, και μία για το ποσοστό σε σχέση με το σύνολο της στήλης. Έτσι θα έχουμε τον πίνακα 11:

Πίνακας 11

	0-5	6-8	9-10	σύνολο
άνδρας	1 33.3% 50%	2 66.7% 50%	0 0% 0%	3 (42.9%)
γυναίκα	1 25% 50%	2 50% 50%	1 25% 100%	4 (57.1%)
σύνολο	2 (28.6%)	4 (57.1%)	1 (14.3%)	7 (100%)

Προτείνουμε ο πίνακας συμπτώσεων δύο μεταβλητών να παρουσιάζεται με την παραπάνω μορφή που επιτρέπει μια εποπτική και άμεση εικόνα του φαινομένου. Επιπλέον ο πίνακας αυτός είναι χρήσιμος στην επεξεργασία των δεδομένων με την Παραγοντική Ανάλυση των Αντιστοιχιών.

6. Έλεγχος συσχέτισης-ανεξαρτησίας μεταβλητών

Μέχρι τώρα, ασχοληθήκαμε μόνο με την καλύτερη παρουσίαση των μεταβλητών. Μόνο η παρουσίαση όμως, δεν είναι αρκετή για την μελέτη της επίδρασης και εξάρτησης των μεταβλητών μεταξύ τους. Χρειάζεται να ελέγξουμε την ύπαρξη ή όχι της σχέσης ανάμεσα σε δύο μεταβλητές, υπολογίζοντας τον συντελεστή συσχέτισής τους.

Οι πίνακες συμπτώσεων είναι πολύ χρήσιμοι στον έλεγχο ανεξαρτησίας και συσχέτισης δύο μεταβλητών. Υποθέτοντας ανεξαρτησία των δύο μεταβλητών υπολογίζουμε τις αναμενόμενες θεωρητικές συχνότητες f_θ κάθε κελιού, και κατόπιν την τιμή της διαφοράς με την πραγματική συχνότητα f_π. Αν οι διαφορές είναι μικρές, παραδεχόμαστε την υπόθεση ανεξαρτησίας. Αν όχι την απορρίπτουμε.

Για να δούμε αν οι δύο μελετώμενες μεταβλητές σχετίζονται ή όχι, υπολογίζουμε την τιμή:

$$\sum_{\substack{\text{κελλια} \\ ij}} \frac{(f_\theta - f_\pi)^2}{f_\theta}$$

όπου $f_{\theta ij} = \dfrac{(\text{άθροισμα i γραμμής}) \times (\text{άθροισμα j στήλης})}{\text{συνολικό άθροισμα}}$

Η τιμή αυτή ακολουθεί την κατανομή X^2 με βαθμούς ελευθερίας ανάλογους με το πλήθος των κελιών. Εφαρμόζουμε λοιπόν το test ανεξαρτησίας X^2. Ανάλογα με την τιμή του στατιστικού αυτού μπορούμε να συμπεράνουμε αν οι μεταβλητές είναι ανεξάρτητες ή όχι.

Εφαρμόζουμε την ίδια διαδικασία και σε ένα πίνακα Burt. Για κάθε μεταβλητή βρίσκουμε τις τιμές του X^2 σε σχέση μ' όλες τις άλλες, και στη συνέχεια

κατατάσσουμε τις υπόλοιπες μεταβλητές ως προς την σχέση τους μ' αυτή την μεταβλητή. Έτσι έχουμε μια διάταξη των μεταβλητών ανάλογα με το πόσο εξαρτώνται ή όχι από την θεωρούμενη μεταβλητή. Η διάταξη αυτή είναι χρήσιμη όταν θελήσουμε να μην λάβουμε υπόψιν μας κάποιες μεταβλητές, οπότε μπορούμε πολύ εύκολα να τις επιλέξουμε. [Tallur, 1988].

Προτείνουμε την αναγραφή της τιμής των συντελεστών του X^2, στην εμφάνιση ενός πίνακα υπό συνθήκη σχετικών συχνοτήτων. Με τον τρόπο αυτό αναδεικνύονται εύκολα οι μεταβλητές οι οποίες επιδρούν σε μια συγκεκριμένη μεταβλητή, αλλά και το κατά πόσο επιδρούν σ'αυτή. Έτσι ο παραπάνω πίνακας 11 του παραδείγματος θα εμφανίζεται στη μορφή του πίνακα 12:

Πίνακας 12

	0-5	6-8	9-10	σύνολο
άνδρας	1 33.3% 50%	2 66.7% 50%	0 0% 0%	3 (42.9%)
γυναίκα	1 25% 50%	2 50% 50%	1 25% 100%	4 (57.1%)
σύνολο	2 (28.6%)	4 (57.1%)	1 (14.3%)	7 (100%) X^2=5.255

Βέβαια η τιμή X^2=5.255 είναι ενδεικτική και δεν έχει νόημα γιατί το πλήθος του δείγματος (7) είναι πολύ μικρό.

7. Πρόταση δημιουργίας πίνακα Burt υπό συνθήκη σχετικών συχνοτήτων

Η μελέτη και ανάλυση ενός πίνακα Burt, μας βοηθάει πολύ στην ανάλυση και περιγραφή του φαινομένου, αλλά έχει το μειονέκτημα ο πίνακας αυτός, να εξαρτάται από το πλήθος των παρατηρήσεων. Για το λόγο αυτό χρησιμοποιούμε τον πίνακα Burt σχετικών συχνοτήτων. Θέλοντας ακόμη να κερδίσουμε και τα πλεονεκτήματα της παρουσίασης των πινάκων υπό συνθήκη σχετικών συχνοτήτων, προτείνουμε την χρησιμοποίηση ενός νέου πίνακα που προκύπτει από τον πίνακα Burt, και τον οποίο ονομάζουμε πίνακα **Burt υπό συνθήκη σχετικών συχνοτήτων**.

Η κατασκευή του πίνακα αυτού, γίνεται ως εξής [Παπαδημητρίου, 1990]:

1. Τα στοιχεία κάθε στήλης του πίνακα Burt απολύτων συχνοτήτων, τα διαιρούμε με τον αριθμό που βρίσκεται στη διασταύρωση της στήλης αυτής με την κύρια διαγώνιο, και ο οποίος αντιστοιχεί στο πλήθος των παρατηρήσεων που χαρακτηρίζονται από την κατηγορία που αντιστοιχεί στη θεωρούμενη στήλη. Έτσι έχουμε τις υπό συνθήκη σχετικές συχνότητες των κλάσεων κάθε

μεταβλητής (αντίστοιχες σ' όλες τις γραμμές), σε σχέση με μία κλάση (αντίστοιχη στην στήλη).

2. Τα στοιχεία της κυρίας διαγωνίου του πίνακα Burt απολύτων συχνοτήτων, τα διαιρούμε με το πλήθος των παρατηρήσεων μας. Έτσι στη διαγώνιο, εμφανίζονται οι σχετικές συχνότητες των κλάσεων κάθε μεταβλητής.

Για παράδειγμα ο πίνακας Burt που παρουσιάστηκε στον πίνακα 7, μετά τους παραπάνω μετασχηματισμούς, γίνεται πίνακας Burt υπό συνθήκη σχετικών συχνοτήτων και παρουσιάζεται στον πίνακα 13.

Πίνακας 13

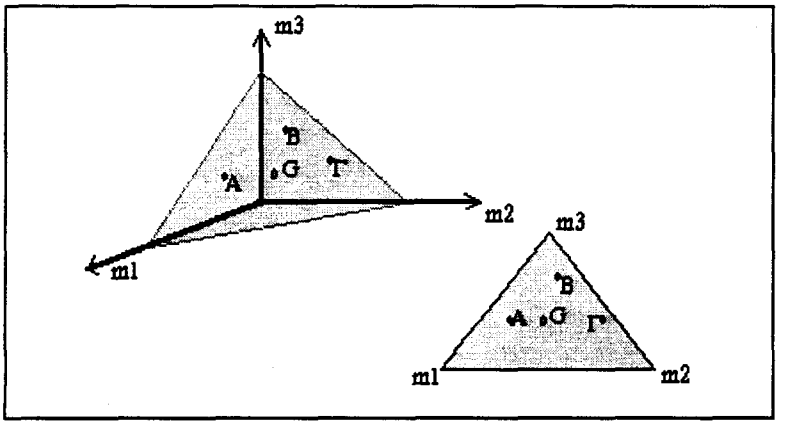
	άνδρας	γυναίκα	Μακ.	Θράκ.	Ηπειρ.	Θεσ.	‹3000000	›3000000	0-5	6-8	9-10
άνδρας	3/7	0	2/2=1	1/2=0.50	0	0	1/2=0.50	2/5=0.40	1/2=0.50	2/4=0.50	0
γυναίκα	0	4/7	0	1/2=0.50	2/2=1	1	1/2=0.50	3/5=0.60	1/2=0.50	2/4=0.50	1
Μακ.	2/3=0.67	0	2/7	0	0	0	1/2=0.50	1/5=0.20	1/2=0.50	1/4=0.25	0
Θράκ.	1/3=0.33	1/4=0.25	0	2/7	0	0	0	2/5=0.40	0	2/4=0.50	0
Ηπειρ.	0	2/4=0.50	0	0	2/7	0	1/2=0.50	1/5=0.20	0	1/4=0.25	1
Θεσ.	0	1/4=0.25	0	0	0	1/7	0	1/5=0.20	1/2=0.50	0	0
‹3000000	1/3=0.33	1/4=0.25	1/2=0.50	0	1/2=0.50	0	2/7	0	0	1/4=0.25	1
›3000000	2/3=0.67	3/4=0.75	1/2=0.50	2/2=1	1/2=0.50	1	0	5/7	2/2=1	3/4=0.75	0
0-5	1/3=0.33	1/4=0.25	1/2=0.50	0	0	1	0	2/5=0.40	2/7	0	0
6-8	2/3=0.67	2/4=0.50	1/2=0.50	2/2=1	1/2=0.50	0	1/2=0.50	3/5=0.60	0	4/7	0
9-10	0	1/4=0.25	0	0	1/2=0.50	0	1/2=0.50	0	0	0	1/7

Με τον τρόπο αυτό, θεωρώντας τις στήλες του πίνακα Burt υπό συνθήκη σχετικών συχνοτήτων, σαν διανύσματα στο χώρο των mJ (όσο το πλήθος όλων των κατηγοριών) διαστάσεων, μπορούμε να προβούμε στην σύγκριση δύο κατηγοριών, συγκρίνοντας τα αντίστοιχα διανύσματα.

Το κύριο πλεονέκτημα της πρότασης αυτής είναι η δυνατότητα που έχουμε να συγκρίνουμε δύο κατηγορίες όχι μόνο της ίδιας μεταβλητής, αλλά και διαφορετικών μεταβλητών. Αυτό γίνεται γιατί όλα τα διανύσματα του χώρου των mJ διαστάσεων, παριστάνονται σε μια κοινή βάση (όλων των κατηγοριών όλων των μεταβλητών) και έτσι έχει νόημα η μεταξύ τους σύγκριση. Η σύγκριση αυτή, μας επιτρέπει να ανακαλύψουμε τις υπάρχουσες ποιοτικές σχέσεις και συσχετίσεις μεταξύ δύο ή περισσότερων κατηγοριών.

Επειδή οι τιμές των διανυσμάτων αυτών (σχετικές συχνότητες), είναι μικρότερες από την μονάδα, η γεωμετρική τους αναπαράσταση στο χώρο των mJ διαστάσεων, θα περιορίζεται από ένα (mJ-1) υποχώρο του χώρου αυτού, που ονομάζεται "simplex". Ο simplex υποχώρος, περιορίζεται από mJ ακμές, οι οποίες ενώνουν τις κορυφές που αντιστοιχούν στα σημεία των αξόνων συντεταγμένων που αντιστοιχούν στις μονάδα. Το σύστημα συντεταγμένων μέσα στον simplex υποχώρο, ονομάζεται "βαρυκεντρικό σύστημα συντεταγμένων" αφού έχει για αρχή των αξόνων το κέντρο βάρους των mJ διανυσμάτων. [Greenacre,1993]

Για παράδειγμα αν mJ=3 (1 μεταβλητή με 3 κατηγορίες), τα διανύσματα που αντιστοιχούν στις στήλες του πίνακα Burt υπό συνθήκη σχετικών συχνοτήτων, θα αναπαριστάνονται στο χώρο των 3 διαστάσεων από τα σημεία Α, Β, Γ τα οποία βρίσκονται μέσα σε ένα τρίγωνο, όπως φαίνεται στο σχήμα 3. Αν ένα σημείο προσεγγίζει μια πλευρά του τριγώνου, σημαίνει ότι χαρακτηρίζεται κύρια από τις κατηγορίες που αντιστοιχούν στην πλευρά αυτή. Στο σχήμα 3, μπορούμε να συμπεράνουμε ότι τα σημεία Β, Γ χαρακτηρίζονται κυρίως από τις κατηγορίες m2, m3 και λιγότερο από την m1.



σχήμα 3. Γεωμετρική αναπαράσταση των διανυσμάτων-στηλών στο χώρο των 3 διαστάσεων και η αντίστοιχη παράσταση τους στο επίπεδο τρίγωνο

Άξιο ιδιαίτερης προσοχής είναι ότι με τον τρόπο αυτό, πετυχαίνουμε να αναπαραστήσουμε τις πληροφορίες τριών διαστάσεων σε ένα επίπεδο (τρίγωνο) δύο διαστάσεων. Παρόμοια για τις mJ διαστάσεις, δημιουργώντας τον πίνακα Burt υπό συνθήκη σχετικών συχνοτήτων, μειώνουμε τις διαστάσεις γεωμετρικής αναπαράστασης των σημείων σε mJ-1, χωρίς να χάσουμε καθόλου πληροφορία.

Ο πίνακας Burt υπό συνθήκη σχετικών συχνοτήτων, μπορεί να επεξεργαστεί είτε με την Παραγοντική Ανάλυση των Αντιστοιχιών, που παρουσιάσαμε στο προηγούμενο κεφάλαιο, είτε με την Ιεραρχική Ταξινόμηση όπως θα δούμε στα επόμενα κεφάλαια.

8. Γενίκευση του πίνακα Burt για ποσοτικές μεταβλητές

Όπως είδαμε και παραπάνω, μπορούμε να μετατρέψουμε μια ποσοτική μεταβλητή σε ποιοτική, χωρίζοντάς την σε κλάσεις, και στη συνέχεια να κατασκευάσουμε τον πίνακα Burt. Με τον τρόπο αυτό όμως χάνουμε ένα μέρος της πληροφορίας. Π.χ. για την κατηγορία 25-30, δεν υπάρχει διαφορά μεταξύ των τιμών 26 και 29. Αν μας ενδιαφέρουν οι διαφορές αυτές και παράλληλα επιθυμούμε να

κερδίσουμε τα πλεονεκτήματα ενός πίνακα Burt, μπορούμε να χρησιμοποιήσουμε τον γενικευμένο πίνακα Burt.

Ο πίνακας αυτός προτάθηκε από τον Ouadrani. [Ouadrani,1994]. Στο πίνακα αυτό, αντί να χωρίσουμε την ποσοτική μεταβλητή σε κάποια διαστήματα τιμών και να δημιουργήσουμε τον λογικό πίνακα με τιμές 0 ή 1, ανάλογα με το αν μια παρατήρηση ανήκει ή όχι στο αντίστοιχο διάστημα, επιλέγουμε κάποιες χαρακτηριστικές τιμές της ποσοτικής μεταβλητής. Δηλαδή τιμές που παρουσιάζονται πολύ συχνά ή σπάνια, η τιμές που θεωρούμε σημαντικές για κάποιο ειδικό λόγο. [Cibois,1983] [Tenenhaous,1994]. Κατόπιν σχηματίζουμε τόσες κατηγορίες όσες και οι τιμές αυτές. Δημιουργούμε ένα γενικευμένο λογικό πίνακα με πραγματικές τιμές στο διάστημα [0,1].

Αν για ένα αντικείμενο i η τιμή της μεταβλητής συμπίπτει με μια χαρακτηριστική τιμή, λέμε ότι το αντικείμενο βρίσκεται ακριβώς μέσα στην κλάση της j. Αν η τιμή της μεταβλητής για ένα αντικείμενο βρίσκεται ανάμεσα σε δύο χαρακτηριστικές τιμές αντιστοιχούμε στο αντικείμενο αυτό, δύο τιμές, μια για κάθε κλάση των χαρακτηριστικών τιμών. Το άθροισμα των δύο αυτών τιμών είναι 1. Αν η μία τιμή είναι κοντύτερα στο 1, αυτό σημαίνει ότι η τιμή του i αντικειμένου είναι κοντύτερα στην αντίστοιχη χαρακτηριστική τιμή. Για κάθε αντικείμενο i και για την κατηγορία της j χαρακτηριστικής τιμής, θα έχουμε την τιμή που είναι συμπληρωματική ως προς την μονάδα, της διαφοράς (σε ποσοστό) της τιμής της μεταβλητής για το i αντικείμενο, από την χαρακτηριστική τιμή αυτή. Για τιμές χαμηλότερες από την πρώτη χαρακτηριστική τιμή, θέτουμε 1 στην κλάση της πρώτης χαρακτηριστικής τιμής, ενώ για τιμές μεγαλύτερες από την τελευταία χαρακτηριστική τιμή θέτουμε 1 στην τελευταία κλάση.

8.1 Παράδειγμα

Έστω ότι έχουμε τον πίνακα δεδομένων 14 για δύο ποσοτικές μεταβλητές. Επιλέγουμε σαν χαρακτηριστικές τιμές για την μεταβλητή x_1 τις τιμές 3, 5, 10 και για την x_2 τις τιμές 22, 25 και 28.

Πίνακας 14

αντικείμενα	x_1	x_2
1	3	22
2	5	25
3	3	28
4	10	30
5	4	25
6	5	27
7	9	21
8	10	25
9	11	27
10	6	31
11	7	28
12	5	24
13	8	27
14	10	25

Τότε σχηματίζεται ο επόμενος γενικευμένος λογικός πίνακας 15, από όπου προκύπτει αντίστοιχα ο γενικευμένος πίνακας Burt 16. Στον γενικευμένο λογικό πίνακα 15, παρατηρούμε ότι δεν έχουμε σαν τιμές των κελιών μόνο 0 ή 1, αλλά και τιμές όπως 1/2, 2/5 κλπ. Η τιμή 0 σε κάποιο κελί σημαίνει ότι δεν έχει σχέση η αντίστοιχη παρατήρηση (γραμμή) με την αντίστοιχη κλάση (στήλη). Η τιμή 1, σημαίνει ότι η παρατήρηση ταυτίζεται με τη χαρακτηριστική τιμή της κλάση που αντιστοιχεί στην στήλη. Δύο τιμές μεταξύ 0 και 1 που υπάρχουν για μια παρατήρηση σε δύο διαδοχικές στήλες, σημαίνουν ότι η αντίστοιχη παρατήρηση δεν ανήκει εξ ολοκλήρου σε καμιά από τις κλάσεις των στηλών, αλλά βρίσκεται κάπου ενδιάμεσα. Το αν είναι πλησιέστερα ή όχι στην θεωρούμενη στήλη, απ' ότι στη γειτονική της, παριστάνεται από το αν η τιμή του κελιού είναι κοντύτερα στο 1 ή στο 0.

Στον πίνακα 15 και για την 7η παρατήρηση, μπορούμε να δούμε ότι οι τιμές των κελιών για τις 3 κλάσεις της μεταβλητής x1 είναι 0, 1/5, 4/5. Αυτό σημαίνει ότι η τιμή της μεταβλητής x1 που είναι 9 για την 7η παρατήρηση, δεν ανήκει στην πρώτη κλάση (η οποία έχει χαρακτηριστική τιμή 3), και βρίσκεται ανάμεσα στην δεύτερη και τρίτη κλάση (οι οποίες έχουν χαρακτηριστικές τιμές 5 και 10 αντίστοιχα). Η τιμή του κελιού 4/5 για την στήλη της τρίτης κλάσης, σημαίνει ότι βρίσκεται πολύ κοντύτερα σ' αυτή απ' ότι στην δεύτερη κλάση για την οποία η τιμή του κελιού είναι 1/5.

Ο γενικευμένος πίνακας Burt που παρουσιάζεται στον πίνακα 16, προκύπτει από τον λογικό πίνακα 15, με την διανυσματική εξίσωση B=D'D, όπως και στην περίπτωση του απλού πίνακα Burt. Η διαφορά από ένα κλασικό πίνακα Burt, είναι ότι οι πίνακες της κυρίας διαγωνίου δεν είναι διαγώνιοι αλλά τρι-διαγώνιοι

πίνακες. Δηλαδή πίνακες με μη-μηδενικά στοιχεία τόσο στην κύρια διαγώνιο, όσο και στις διαγώνιες πάνω και κάτω από αυτήν.

Πίνακας 15

	μεταβλητή x1			μεταβλητή x2		
αντικείμενα	3	5	10	22	25	28
1	1	0	0	1	0	0
2	0	1	0	0	1	0
3	1	0	0	0	0	1
4	0	0	1	0	0	1
5	1/2	1/2	0	0	1	0
6	0	1	0	0	1/3	2/3
7	0	1/5	4/5	1	0	0
8	0	0	1	0	1	0
9	0	0	1	0	1/3	2/3
10	0	4/5	1/5	0	0	1
11	0	3/5	2/5	0	0	1
12	0	1	0	1/3	2/3	0
13	1	0	0	0	1/3	2/3
14	0	0	1	0	1	0

Πίνακας 16

	3	5	10	22	25	28
3	3.25	0.25	0	1	5/6	5/3
5	0.25	4.10	0.56	8/15	2.5	31/15
10	0	0.56	4.84	4/5	2.33	34/15
22	1	8/15	4/5	2.11	0.22	0
25	5/6	2.5	2.33	0.22	4.78	0.67
28	5/3	31/15	34/15	0	0.67	5.33

Παρατηρούμε ότι ο γενικευμένος πίνακας Burt, παρά του ότι δεν είναι πίνακας σχετικών συχνοτήτων, περιέχει μη ακέραιες τιμές, επειδή προέρχεται από γινόμενα και αθροίσματα μη ακεραίων τιμών.

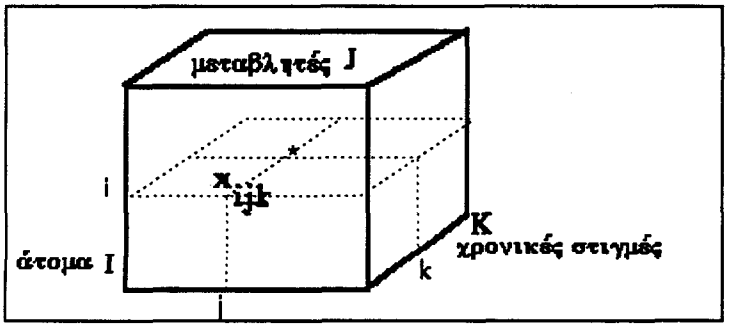
9. Τρισδιάστατοι πίνακες

Όλοι οι μέχρι τώρα πίνακες δύο διαστάσεων που παρουσιάστηκαν, επιδέχονται περαιτέρω ανάλυση με μεθόδους της Ανάλυσης Δεδομένων. Η γνώση των μεθόδων και η εμπειρία του ερευνητή, του επιτρέπει να επιλέξει την κατάλληλη μέθοδο ανάλογα με το είδος του πίνακα δεδομένων και το στόχο της έρευνας.

Πολλές φορές τα δεδομένα δεν παρουσιάζονται με την μορφή διδιάστατου πίνακα, αλλά μας παρέχονται με την μορφή πολλών διδιάστατων πινάκων. Τότε μπορούμε να οργανώσουμε τα δεδομένα μας σε μορφή τρισδιάστατου πίνακα, που είναι και η τελική μορφή πινάκων που αντιμετωπίζουμε στην παρούσα διατριβή.

Ιδιαίτερα ενδείκνυται ο πίνακας αυτός, για χρονολογικές σειρές, όπου μελετάμε αντικείμενα και μεταβλητές για διαφορετικά χρονικά διαστήματα. Στην περίπτωση αυτή η τρίτη διάσταση του τρισδιάστατου πίνακα είναι ο χρόνος. [Lavit,1988] [Escofier, 1990]

Τυπικό παράδειγμα τρισδιάστατου πίνακα, αποτελεί ο πίνακας *αντικείμενα x μεταβλητές* όταν έχουμε παρατηρήσεις για διαφορετικές χρονικές στιγμές ή σε διαφορετικούς πληθυσμούς (σχήμα 4). Στην περίπτωση αυτή, εκτός από τις σχέσεις μεταξύ των μεταβλητών ή μεταξύ των αντικειμένων, μας ενδιαφέρει και η πορεία του φαινομένου στις διαφορετικές στιγμές παρατήρησης, ή οι διαφορές και οι ομοιότητες ανάμεσα στους διαφορετικούς πληθυσμούς απ' όπου πήραμε το δείγμα μας.



σχήμα 4. Δεδομένα με την μορφή τρισδιάστατου πίνακα

Το πρόβλημα που αντιμετωπίζουμε στους τρισδιάστατους πίνακες είναι πως θα τους αναλύσουμε περαιτέρω, ώστε χωρίς καμία a priori υπόθεση να αναδεικνύονται οι χαρακτηριστικές ιδιότητες του φαινομένου που περιγράφουν.

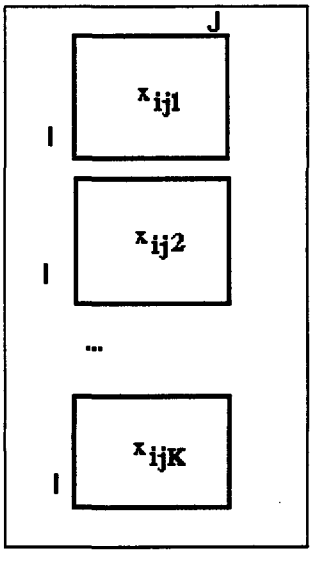
Πολύ συχνά οι ερευνητές καλούνται να αναλύσουν τρισδιάστατους πίνακες δεδομένων. Ας υποθέσουμε ότι θέλουμε να μελετήσουμε τις σχέσεις μεταξύ τριών συνόλων I, J, K ποιοτικών μεταβλητών. Τα δεδομένα των τιμών των τριών συνόλων μπορούν να συγκεντρωθούν σ' ένα τρισδιάστατο πίνακα συχνοτήτων των κλάσεων των μεταβλητών μεταξύ τους. Το στοιχείο x_{ijk} του πίνακα αυτού αντιστοιχεί στο πλήθος των παρατηρήσεων που χαρακτηρίζονται συγχρόνως από τη κλάση i μιας των μεταβλητών I, από την j μιας των μεταβλητών J και την k κλάση μιας των μεταβλητών K, στο πληθυσμό που μελετάμε. Τα δεδομένα λοιπόν, των τρισδιάστατων πινάκων, μπορούν να θεωρηθούν κλεισμένα σε ένα ορθογώνιο παραλληλεπίπεδο με ακμές I, J, K.

Στη συνέχεια θεωρούμε ότι τα τρία σύνολα που αντιστοιχούν στις 3 διαστάσεις του πίνακα, αντιπροσωπεύουν μεταβλητές. Η ανάλυση δεν διαφέρει πολύ αν κάποιο από τα σύνολα αυτά παριστάνει αντικείμενα.

9.1 Μεθοδολογία για την επεξεργασία τρισδιάστατου πίνακα

Θεωρούμε ότι τα τρία σύνολα μεταβλητών ενός τρισδιάστατου πίνακα παίζουν τον ίδιο ρόλο στο φαινόμενο που μελετάμε, ώστε από στατιστικής σκοπιάς το πρόβλημα να τίθεται συμμετρικά. Με την παραγοντική ανάλυση των αντιστοιχιών, έχουμε την δυνατότητα ανάλυσης και επεξεργασίας μόνο διδιάστατων πινάκων και προσδιορισμού των σχέσεων που συνδέουν μεταξύ τους τις κλάσεις δύο συνόλων μεταβλητών. Συγχρόνως όμως, παραθέτοντας δίπλα ή και κάτω από τον αρχικό διδιάστατο πίνακα άλλους πίνακες, έχουμε την δυνατότητα να διευρύνουμε τη σχέση των αρχικών δύο συνόλων μεταβλητών με άλλα που τίθενται ως συμπληρωματικά. Μ' αυτό τον τρόπο επιτυγχάνουμε πολλές αναλύσεις η κάθε μια των οποίων αντιστοιχεί στα κριτήρια που επιλέγουμε.

Η ανάλυση ενός τρισδιάστατου πίνακα (σχήμα 5) προτείνουμε να γίνεται παραθέτοντας κατακόρυφα ή οριζόντια τους δισδιάστατους πίνακες από τους οποίους συντίθεται, ή ακόμη παίρνοντας αθροίσματα των δισδιάστατων πινάκων και δημιουργώντας ένα νέο δισδιάστατο πίνακα με την περιγραφή του φαινομένου (σχήμα 5). [Παπαδημητρίου-Φλώρου,1992]



σχήμα 5. Κατακόρυφα απλωμένος τρισδιάστατος πίνακας

Μερικοί από τους πίνακες που μπορούμε να σχηματίσουμε και στη συνέχεια να αναλύσουμε [Παπαδημητρίου,1991] [Foucart,1984] είναι:

- Οι τρεις περιθωριακοί πίνακες IxJ, IxK, JxK με στοιχεία x_{ij}, $x_{i.k}$ $x_{.jk}$ αντίστοιχα όπου:

$$x_{ij.}=\sum_{k=1}^{K} x_{ijk} \quad x_{i.k}=\sum_{j=1}^{J} x_{ijk} \quad x_{.jk}=\sum_{i=1}^{I} x_{ijk}$$

- Οι διδιάστατοι πίνακες φέτες που ορίζονται από μια κλάση του ενός συνόλου, και από τα άλλα δύο σύνολα. Ο πίνακας που ορίζεται από την κλάση k_1 του συνόλου K είναι:

  x_{ijk1} : i=1,2,...I j=1,2,....,J

- Οι διδιάστατοι πίνακες που δημιουργούνται αν παρατεθούν οριζόντια ή κάθετα οι πίνακες φέτες του τρισδιάστατου πίνακα (σχήμα 5).
- Ο πίνακας συμπτώσεων (I∪J∪K)x(I∪J∪K) αν παραθέσουμε οριζόντια και κάθετα όλες τις μεταβλητές,

Κάθε ομάδα από τους παραπάνω πίνακες οδηγεί και σε διαφορετικά συμπεράσματα και πρέπει κάθε φορά να αναλύονται εκείνοι οι πίνακες που παρουσιάζουν ενδιαφέρον για την συγκεκριμένη έρευνα.

9.1.1 Ανάλυση των τριών περιθωριακών πινάκων

Για να σχηματισθούν οι εικόνες των σχέσεων οι οποίες συνδέουν μεταξύ τους ανά δύο τα τρία σύνολα των μεταβλητών I, J, K είναι απαραίτητη σε πρώτη φάση η ανάλυση με την AFC και των τριών περιθωριακών πινάκων IxJ, IxK, JxK. Επίσης, όταν οι διαστάσεις των τριών συνόλων είναι μεγάλες, είναι πολύ χρήσιμες οι απλές μέθοδοι της Περιγραφικής Στατιστικής για την απομάκρυνση των "ενοχλητικών" δεδομένων, όπως οι κλάσεις των μεταβλητών με πολύ υψηλή ή πολύ χαμηλή συχνότητα.

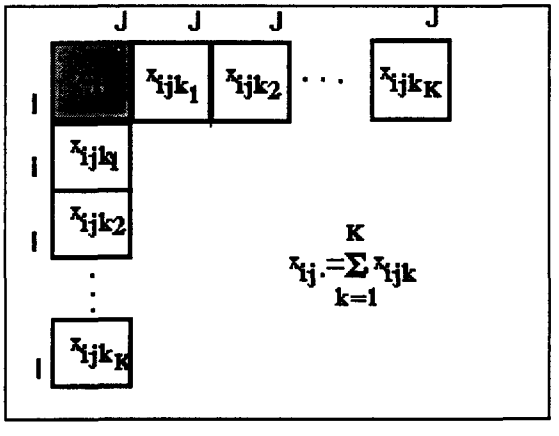
Όταν ένα από τα τρία σύνολα των μεταβλητών, έστω το K, έχει μια ιδιαίτερη βαρύτητα στο πρόβλημα που μελετάμε, όταν δηλαδή το σύνολο K μπορεί να θεωρηθεί προνομιούχο ως προς τη σημασία του, είναι δυνατό να εξαχθούν χρήσιμα συμπεράσματα από την σύγκριση των αναλύσεων AFC των Card(K) διδιάστατων πινάκων-φέτες του αρχικού τρισδιάστατου πίνακα. Αν π.χ. το προνομιούχο σύνολο K αφορά το φύλο (Card(K)=2), συγκρίνοντας τα αποτελέσματα των δύο αναλύσεων AFC των δύο διδιάστατων πινάκων-φέτες που αναφέρονται ο ένας στους άνδρες και ο άλλος στις γυναίκες, προσδιορίζονται οι διαφορές και ομοιότητες, που υπάρχουν στις σχέσεις των δύο συνόλων I και J στους άνδρες και γυναίκες.

Όταν οι κλάσεις του προνομιούχου συνόλου είναι περισσότερες από δύο, τότε με τη σύγκριση των αναλύσεων AFC των διδιάστατων πινάκων-φέτες που αντιστοιχούν σε κάθε μία, προσδιορίζεται από την μια η κοινή δομή τους, από τις ομοιότητες στις σχέσεις των συνόλων I και J, όταν αυτή υπάρχει, και από την άλλη προσδιορίζεται η εξέλιξη στις σχέσεις των συνόλων I και J ως προς το K, από τις διαφορές που αυτά παρουσιάζουν, και συγχρόνως συγκρίνοντας τους παράγοντες των αναλύσεων AFC διαπιστώνονται οι δομικές διαφοροποιήσεις τους.

Η ταξινόμηση σε κάθε περιθωριακό πίνακα, μπορεί να αναδείξει τις ομάδες μεταβλητών που σχηματίζονται βάσει των μεταξύ τους συσχετίσεων. Εφαρμόζοντας την ταξινόμηση στον πίνακα IxJ, ιεραρχούμε το σύνολο I βάσει των δεδομένων του συνόλου J. Η ταξινόμηση του πίνακα IxK, ιεραρχεί το σύνολο I βάσει των δεδομένων του συνόλου K. Αντίστοιχα μπορούμε να εφαρμόσουμε την ταξινόμηση στους πίνακες JxK, JxI, KxJ, KxI.

Ακόμη εφαρμόζοντας την ταξινόμηση στους πίνακες φέτες, μπορούμε να εξάγουμε συμπεράσματα για τις δημιουργούμενες ομάδες των μεταβλητών τις ιεραρχήσεις τους, καθώς και τις διαφοροποιήσεις των ομάδων μεταξύ των μεταβλητών του συνόλου βάσει του οποίου χωρίζονται οι πίνακες φέτες.

Τα αποτελέσματα των ταξινομήσεων, συμπληρώνουν τα αποτελέσματα των παραγοντικών αναλύσεων, αφού δεν παρουσιάζουν μόνο τις υπάρχουσες ομάδες, αλλά και τις σχέσεις ιεράρχησης μεταξύ των ομάδων αυτών, όπως θα δούμε σε επόμενο κεφάλαιο.



σχήμα 6. Ο περιθωριακός πίνακας άθροισμα IxJ με τις συμπληρωματικές γραμμές και στήλες

Ο περιθωριακός πίνακας άθροισμα IxJ είναι ο πίνακας που προκύπτει από το άθροισμα των Card(K) διδιάστατων πινάκων του αρχικού πίνακα IxJxK. Αν στο περιθωριακό πίνακα άθροισμα IxJ παραθέσουμε ως συμπληρωματικές γραμμές και στήλες, τους πίνακες φέτες $(IxJ)_{k1}$, $(IxJ)_{k2}$, ..., $(IxJ)_{kK}$ λαμβάνουμε υπόψη και την τρίτη διάσταση K. (σχήμα 6). Η ερμηνεία των αποτελεσμάτων της εφαρμογής της AFC στον πίνακα αυτό, στηρίζεται στις θέσεις που παρουσιάζουν στα παραγοντικά επίπεδα:

- τα I σημεία, που αντιστοιχούν στις κλάσεις των μεταβλητών του συνόλου I
- τα J σημεία, που αντιστοιχούν στις κλάσεις των μεταβλητών του συνόλου J

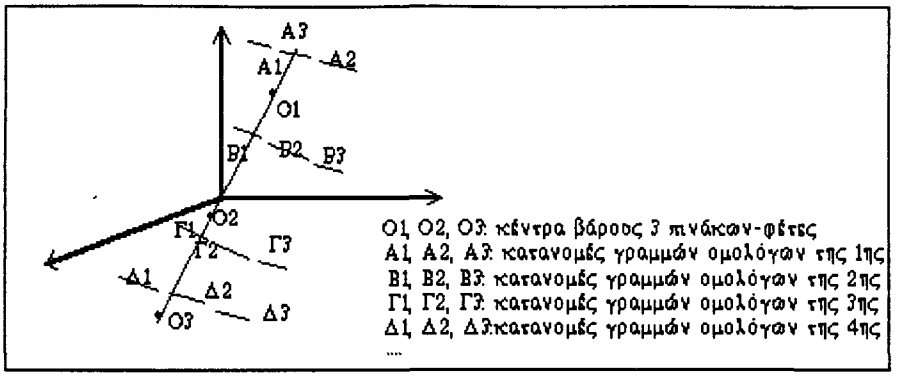
- τα IxK σημεία, που αντιστοιχούν στις κλάσεις των μεταβλητών του συνόλου I για κάθε μια από τις κλάσεις των μεταβλητών του συνόλου K (συμπληρωματικές γραμμές)
- τα JxK σημεία, που αντιστοιχούν στις κλάσεις των μεταβλητών του συνόλου J για κάθε μια από τις κλάσεις των μεταβλητών του συνόλου K (συμπληρωματικές στήλες)

Αν θεωρήσουμε το νέφος των I σημείων του πίνακα-άθροισμα IxJ στο χώρο R^J, οι προβολές τους στα παραγοντικά επίπεδα θα ορίζουν τα κέντρα βάρους των IxK σημείων που ορίζουν οι αντίστοιχες κλάσεις των μεταβλητών στους πίνακες-φέτες. Δηλαδή, η κατανομή της i γραμμής του πίνακα-άθροισμα IxJ ορίζει στα παραγοντικά επίπεδα το κέντρο βάρους (μέση κατανομή) των K γραμμών (i,k) k=1,2,..,K των K πινάκων-φέτες που παρατέθηκαν ως συμπληρωματικές γραμμές.

Με την ανάλυση AFC του πίνακα-άθροισμα IxJ έχουμε την δυνατότητα να μελετήσουμε τη συμπεριφορά των κέντρων βάρους των ομολόγων γραμμών, τη δομή που αυτές παρουσιάζουν, καθώς και να προσδιορίσουμε τις κοινές τάσεις που πιθανώς να υπάρχουν. Επίσης στην ίδια ανάλυση AFC τα σημεία που εμφανίζονται σαν συμπληρωματικά στα παραγοντικά επίπεδα και αντιστοιχούν σε γραμμές των πινάκων-φέτες, που παραθέτουμε συμπληρωματικά, προσδιορίζουν τις διαφοροποιήσεις που παρουσιάζουν οι κατανομές τους από τη μέση κατανομή τους (κέντρο βάρους ομολόγων γραμμών). Πρέπει να τονισθεί στο σημείο αυτό ότι η διαφοροποίηση αυτή αναφέρεται μόνο ως προς την κοινή τάση τους.

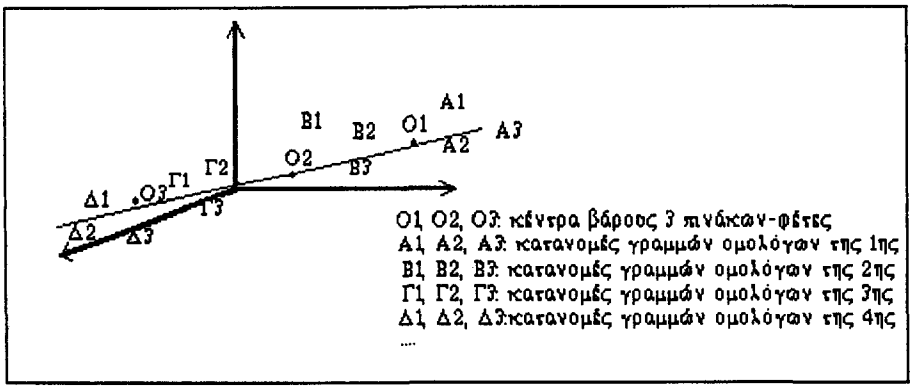
Με παρόμοιο τρόπο γίνεται η μελέτη των J σημείων (στηλών) του πίνακα-άθροισμα IxJ στον χώρο R^I.

Αν στη συνέχεια παραθέσουμε σαν συμπληρωματικές γραμμές (K γραμμές) και στήλες (K στήλες) τα αθροίσματα των στηλών και γραμμών αντίστοιχα του κάθε πίνακα-φέτα, τότε το σημείο που αντιστοιχεί σε κάθε μια απ' αυτές τις στήλες ορίζει στα παραγοντικά επίπεδα το κέντρο βάρους (μέση κατανομή) των m κατανομών ως προς το σύνολο I, και το σημείο που αντιστοιχεί σε κάθε μια απ' αυτές τις γραμμές ορίζει το κέντρο βάρους (μέση κατανομή) των n κατανομών ως προς το σύνολο J.



σχήμα 7. Η διασπορά των πινάκων-φέτες ορθογώνια ως προς την διασπορά των ομολόγων γραμμών

Θα θέλαμε να παρατηρήσουμε εδώ ότι είναι δυνατό, σε ειδικές περιπτώσεις, από τις απεικονίσεις των παραγοντικών επιπέδων να μη γίνονται εμφανείς οι διαφοροποιήσεις που πιθανόν να υπάρχουν στις κατανομές των γραμμών ή στηλών των πινάκων-φέτες. Όταν π.χ. η διεύθυνση διασποράς μεταξύ των πινάκων-φέτες είναι ορθογώνια ως προς την διεύθυνση διασποράς των ομολόγων γραμμών ή στηλών (σχήμα 7). Σ'αυτή την περίπτωση η προβολή των σημείων στους άξονες αδράνειας των κέντρων βάρους δεν μπορεί να φανερώσει τις διαφοροποιήσεις που παρουσιάζουν μεταξύ τους οι κατανομές των ομολόγων γραμμών. Επίσης όταν τα σημεία που αντιστοιχούν στις κατανομές των ομολόγων γραμμών είναι πολύ συγκεντρωμένα (σχήμα 8). Όταν δηλαδή, η εξωτερική αδράνεια μεταξύ των γραμμών είναι πολύ μεγαλύτερη από την εσωτερική.



σχήμα 8. Οι κατανομές ομολόγων γραμμών συγκεντρωμένων γύρω από την διασπορά των πινάκων-φέτες

Η ταξινόμηση του πίνακα άθροισμα IxJ, με συμπληρωματικές γραμμές ή στήλες τους πίνακες-φέτες, παρουσιάζει τις ομάδες και τις ιεραρχήσεις τους, για το σύνολο I, οι οποίες μπορούν να ερμηνευθούν (μέθοδος vacor), με τις μεταβλητές του συνόλου J. Η αντιστοίχηση των IxK συμπληρωματικών σημείων στις ομάδες αυτές, αναδεικνύει τις τυχόν υπάρχουσες διαφορές των μεταβλητών των κλάσεων

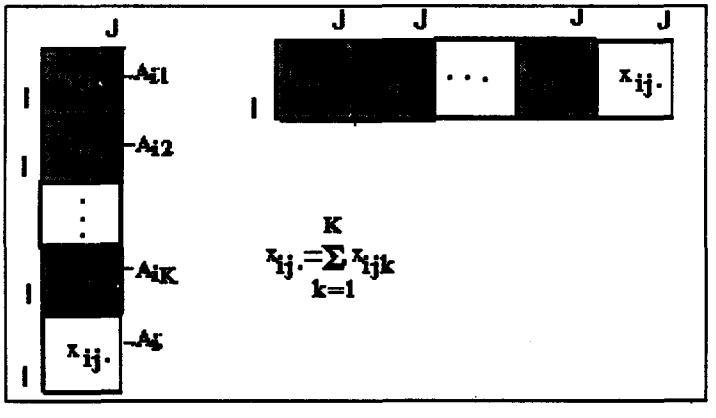
του συνόλου K, από την μέση κατανομή τους. Η μέση κατανομή τους αντιστοιχεί στην θέση της γραμμής άθροισμα του πίνακα IxJ. Με μια λοιπόν ταξινόμηση έχουμε την ταυτόχρονη ερμηνεία και των τριών συνόλων μεταβλητών.

Παρόμοια ανάλυση μπορεί να γίνει και για τους υπόλοιπους πίνακες-αθροισμάτων με συμπληρωματικούς πίνακες-φέτες.

Αν το πλήθος των μεταβλητών είναι πολύ μεγάλο, μπορούμε να μειώσουμε τον αριθμό των μεταβλητών που χρησιμοποιούνται για την ταξινόμηση, λαμβάνοντας υπόψιν, αντί για τις μεταβλητές, τους πρώτους κύριους παραγοντικούς άξονες που αποτελούν γραμμικούς συνδυασμούς των μεταβλητών. Με τον τρόπο αυτό έχουμε ερμηνεία των σπουδαιότερων μεταβλητών και έτσι χρησιμότερα συμπεράσματα.

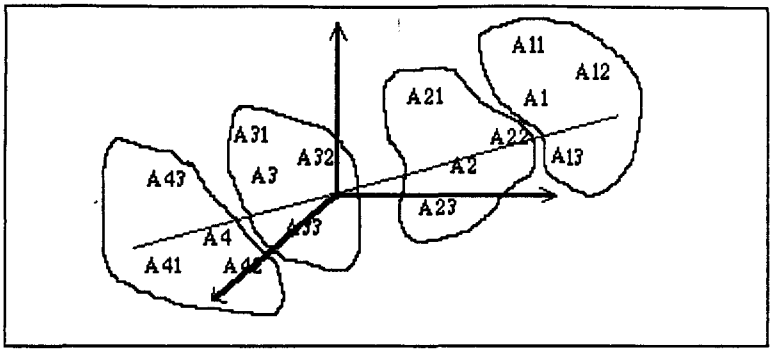
9.1.2 Ανάλυση όλων των πινάκων-φέτες

Αν θέλουμε την ανάλυση των Card(K) πινάκων παρατιθεμένων σαν γραμμές ή σαν στήλες, αναλύουμε δύο πίνακες σ' αυτή την περίπτωση: Τον απλωμένο κατακόρυφα πίνακα (I*K)xJ, όταν παραθέσουμε τους πίνακες φέτες σαν γραμμές, και τον απλωμένο οριζόντιο πίνακα Ix(J*K) όταν παραθέσουμε τους πίνακες φέτες σαν στήλες (σχήμα 9). Και στις δύο περιπτώσεις παραθέτουμε στο τέλος σαν συμπληρωματικό και τον περιθωριακό πίνακα-άθροισμα IxJ.



σχήμα 9. Ο απλωμένος κατακόρυφα και οριζόντια πίνακας με τις συμπληρωματικές γραμμές και στήλες

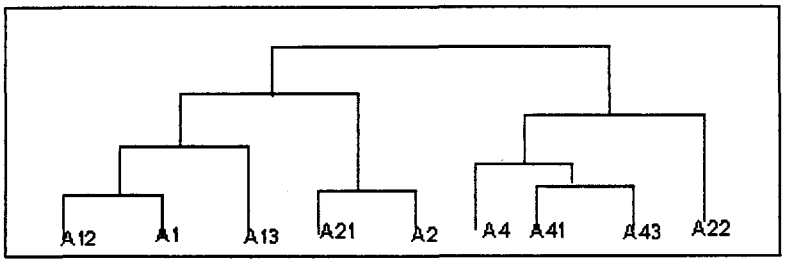
Αν θεωρήσουμε τον απλωμένο κατακόρυφα πίνακα (I*K)xJ, οι παραγοντικοί άξονες δημιουργούνται από τα διανύσματα γραμμές και προβάλλονται στα παραγοντικά επίπεδα τα σημεία A_{1k}, A_{2k}, A_{3k},...,A_{ik}, k=1,2,.., K που αντιστοιχούν στα K διανύσματα (ομόλογες γραμμές των πινάκων φέτες) για μια i γραμμή από τις Card(I) γραμμές του συνόλου I. Στα ίδια παραγοντικά επίπεδα προβάλλονται και οι γραμμές A_i του περιθωριακού πίνακα-άθροισμα IxJ που παραθέτουμε σαν συμπληρωματικό (σχήμα 10).



σχήμα 10. Το νέφος των σημείων του πίνακα (I*K)xJ

Η ολική αδράνεια των IxK σημείων του νέφους του R^m είναι στη περίπτωση αυτή ίση με το άθροισμα της εσωτερικής αδράνειας (intra) μεταξύ των ομολόγων γραμμών και της εξωτερικής (inter) μεταξύ των I διαφορετικών γραμμών.

Η ταξινόμηση του κατακόρυφα απλωμένου πίνακα όπως στο σχήμα 9, εξάγει τις ομαδοποιήσεις και ιεραρχήσεις των I*K σημείων που αντιστοιχούν στους K πίνακες φέτες. Οι ομαδοποιήσεις αυτές ερμηνεύονται από τις μεταβλητές του συνόλου J (με την μέθοδο vacor). Έτσι αν δύο ομόλογες γραμμές βρίσκονται στην ίδια ομάδα, συμπεραίνουμε ότι δεν υπάρχει διαφοροποίηση μεταξύ τους ενώ αν βρίσκονται σε αντιπαρατεθέμενες ομάδες, σημαίνει ότι υπάρχει σημαντική διαφορά μεταξύ τους. Παράδειγμα από ένα κομμάτι του δένδρου ταξινόμησης, παρουσιάζεται στο σχήμα 11.



σχήμα 11. Η ταξινόμηση των σημείων του πίνακα (I*K)xJ

9.1.3 Συμπεράσματα

Μετά από όσα αναφέραμε για την επεξεργασία των τρισδιάστατων πινάκων μπορούμε να συμπεράνουμε ότι:

1. Όταν θέλουμε να μελετήσουμε την εξωτερική δομή μεταξύ των κλάσεων των μεταβλητών του συνόλου I ή J, του πίνακα IxJxK, εφαρμόζουμε την ανάλυση AFC στον περιθωριακό πίνακα-άθροισμα IxJ. Η ερμηνεία αυτής της ανάλυσης στηρίζεται στις προβολές των σημείων, που αντιστοιχούν στα διανύσματα-κλάσεις των μεταβλητών, στους παραγοντικούς άξονες (άξονες αδράνειας) που

δημιουργούνται από τα κέντρα βάρους (αθροιστικές κατανομές) των ομολόγων γραμμών και στηλών.

2. Η ταξινόμηση σε κάθε περιθωριακό πίνακα, μπορεί να αναδείξει τις ομάδες μεταβλητών που σχηματίζονται βάσει των μεταξύ τους συσχετίσεων. Εφαρμόζοντας την ταξινόμηση στον πίνακα IxJ, ταξινομούμε ιεραρχικά το σύνολο I βάσει των δεδομένων του συνόλου J. Η ταξινόμηση του πίνακα IxK, ιεραρχεί το σύνολο I βάσει των δεδομένων του συνόλου K. Αντίστοιχα μπορούμε να εφαρμόσουμε την ταξινόμηση στους πίνακες JxK, JxI, KxJ, KxI. Ακόμη εφαρμόζοντας την ταξινόμηση στους πίνακες φέτες, μπορούμε να εξάγουμε συμπεράσματα για τις δημιουργούμενες ομάδες των μεταβλητών τις ιεραρχήσεις τους, καθώς και τις διαφοροποιήσεις των ομάδων μεταξύ των μεταβλητών του συνόλου βάσει του οποίου χωρίζονται οι πίνακες φέτες.
3. Όταν θέλουμε να μελετήσουμε την εσωτερική δομή μεταξύ των ομολόγων κλάσεων των μεταβλητών των K πινάκων του πίνακα IxJxK, εφαρμόζουμε την ανάλυση AFC στον κατακόρυφα απλωμένο πίνακα (I*K)xJ ή στον οριζόντια απλωμένο πίνακα Ix(J*K). Η ερμηνεία της ανάλυσης αυτής στηρίζεται στις προβολές των σημείων που αντιστοιχούν στα διανύσματα-ομόλογες κλάσεις των μεταβλητών, στους παραγοντικούς άξονες που δημιουργούνται από το σύνολο των γραμμών ή στηλών (σύνολο κλάσεων των μεταβλητών) και όχι στα διανύσματα που ορίζουν οι μέσες κατανομές των ομολόγων μεταβλητών. Αυτή η ανάλυση AFC μας επιτρέπει να διαπιστώσουμε τις διαφοροποιήσεις και ομοιότητες που υπάρχουν μεταξύ των ομολόγων γραμμών ή στηλών.
4. Η ταξινόμηση του πίνακα άθροισμα IxJ, με συμπληρωματικές γραμμές ή στήλες τους πίνακες-φέτες, παρουσιάζει τις ομάδες και τις ιεραρχήσεις τους, για το σύνολο I, οι οποίες μπορούν να ερμηνευθούν (μέθοδος vacor), με τις μεταβλητές του συνόλου J. Η αντιστοίχηση των IxK συμπληρωματικών σημείων στις ομάδες αυτές, αναδεικνύει τις τυχόν υπάρχουσες διαφορές των μεταβλητών των κλάσεων του συνόλου K, από την μέση κατανομή τους. Η μέση κατανομή τους αντιστοιχεί στην θέση της γραμμής άθροισμα του πίνακα IxJ. Με μια λοιπόν ταξινόμηση έχουμε την ταυτόχρονη ερμηνεία και των τριών συνόλων μεταβλητών.

Τα αποτελέσματα των ταξινομήσεων, συμπληρώνουν τα αποτελέσματα των παραγοντικών αναλύσεων, αφού δεν παρουσιάζουν μόνο τις υπάρχουσες ομάδες, αλλά και τις σχέσεις ιεράρχησης μεταξύ των ομάδων αυτών.

9.2 Εφαρμογή-Στατιστική επεξεργασία του τρισδιάστατου πίνακα δεδομένων και τρόπος ερμηνείας των αποτελεσμάτων

Στην εφαρμογή αυτή μελετούνται οι σχέσεις με τις οποίες συνδέονται μεταξύ τους οι παράμετροι, που χαρακτηρίζουν τους πρωτοετείς φοιτητές του 1991 των πέντε τμημάτων του Πανεπιστημίου Μακεδονίας Οικονομικών και Κοινωνικών Επιστημών και επιχειρείται μία σύγκριση των πέντε πινάκων διπλής εισόδου (IxI) των Card (I) παραμέτρων.

Αρχικά, γίνεται μία προσπάθεια συστηματικής ανάλυσης των αποστάσεων των γραμμών και στηλών των πέντε διαφορετικών πινάκων που αναφέρονται στις ίδιες παραμέτρους. Στη συνέχεια με τη χρήση της παραγοντικής ανάλυσης των αντιστοιχιών επιτυγχάνεται ταυτόχρονη απεικόνιση των διανυσμάτων που ορίζει το σύνολο των παραμέτρων στους πέντε πίνακες.

Τέλος, εξάγονται χρήσιμα συμπεράσματα που αφορούν στις ομοιότητες και διαφορές, που παρατηρούνται στο σύνολο των μεταβλητών στους πέντε διαφορετικούς πίνακες, καθώς και σ' αυτές που αφορούν στους πέντε πίνακες μεταξύ τους.

9.2.1 Εισαγωγή

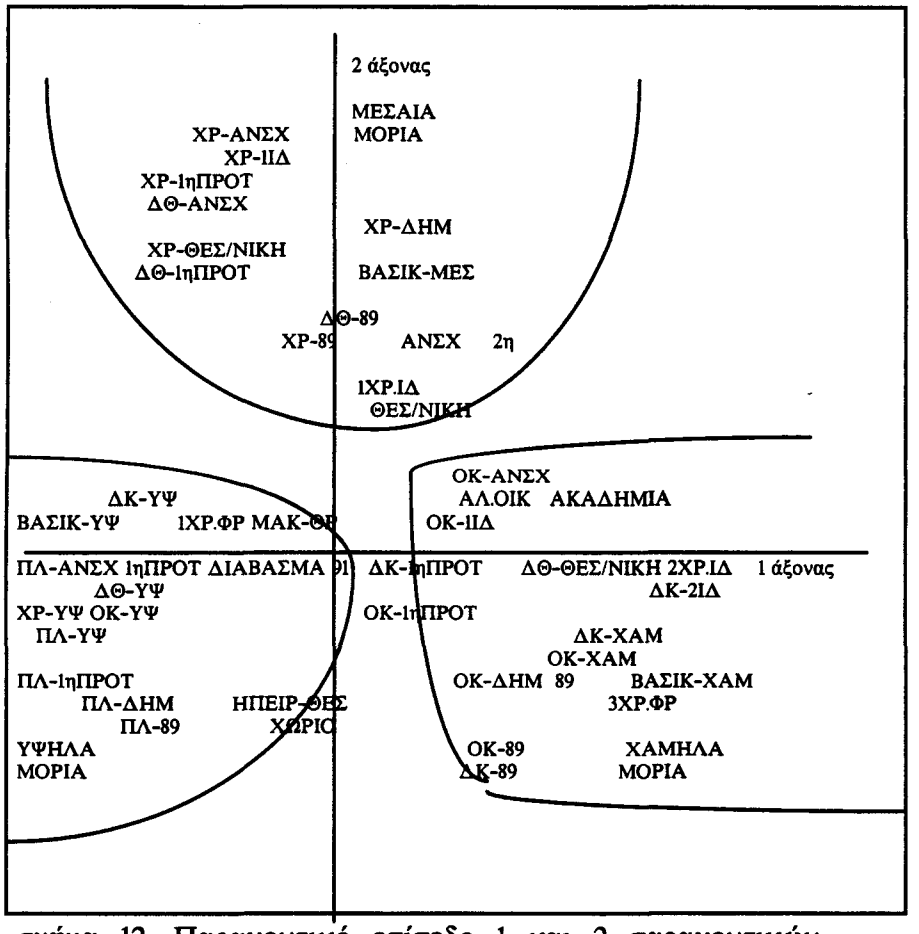
Η εφαρμογή βασίζεται στην επεξεργασία 691 ερωτηματολογίων που συμπλήρωσαν συγχρόνως με την εγγραφή τους οι πρωτοετείς των 5 τμημάτων του Πανεπιστημίου Μακεδονίας το 1991. Δημιουργήθηκαν έτσι 5 διδιάστατοι πίνακες. Η πρώτη διάσταση αντιστοιχεί στο πλήθος των πρωτοετών του κάθε τμήματος, και η δεύτερη στις 18 απαντήσεις τους, για κάθε ερώτηση.

Για τους σκοπούς της έρευνας, οι μεταβλητές που εξετάζουμε αφορούν στην προέλευση των φοιτητών, στη μόρφωση των γονιών τους, τη σχολή και το τμήμα πρώτης προτίμησης τους, τα μόρια που συγκέντρωσαν, τον τρόπο προετοιμασίας, τις φορές εισαγωγικών εξετάσεων κτλ. Οι 18 μεταβλητές χωρίζονται σε 51 συνολικά κλάσεις και δημιουργούνται 5 πίνακες με στοιχεία μόνο 0 και 1, για να έχουμε ομοιογένεια και για τα 5 τμήματα, οι πίνακες αυτοί μετασχηματίζονται σε πίνακες συμπτώσεων 51x51. Δημιουργείται έτσι ένας τρισδιάστατος πίνακας 51x51x5. Το στοιχείο ijk του πίνακα αυτού, αντιστοιχεί στον αριθμό που εκφράζει την απόλυτη συχνότητα, με την οποία η κλάση i μιας μεταβλητής εμφανίζεται συγχρόνως με την κλάση j μιας άλλης μεταβλητής για το τμήμα k.

Σε τμήματα αυτού του πίνακα εφαρμόσθηκε η παραγοντική ανάλυση των αντιστοιχιών AFC. Η ερμηνεία των αποτελεσμάτων της AFC γίνεται στα πρώτα παραγοντικά επίπεδα, αυτά δηλαδή, που δημιουργούν οι τρεις πρώτοι παραγοντικοί άξονες.

9.2.2 Ανάλυση όλων των τμημάτων

Για να μελετήσουμε την εσωτερική δομή (μεταξύ των ομολόγων κλάσεων) των 5 τμημάτων, θα εφαρμόσουμε AFC στον κατακόρυφο απλωμένο πίνακα, που δημιουργείται τοποθετώντας τους 5 πίνακες συμπτώσεων (φέτες για κάθε τμήμα) τον ένα κάτω από τον άλλο. Σχηματίζεται έτσι ο πίνακας διαστάσεων 255(5*51)x51 όπου 255 είναι οι μεταβλητές για τα πέντε τμήματα και 51 οι κλάσεις των μεταβλητών. Εφαρμόζουμε την AFC σ'αυτόν τον πίνακα.



σχήμα 12. Παραγοντικό επίπεδο 1 και 2 παραγοντικών αξόνων της AFC των 5 τμημάτων

Παρατηρώντας λοιπόν τα σημεία στο επίπεδο 1x2 (σχήμα 12), σχηματίζονται 3 ομάδες. Στην πρώτη με υψηλό σύνολο μορίων, στο κάτω αριστερό μέρος του επιπέδου είναι το τμήμα Εφαρμοσμένης Πληροφορικής. Στην ίδια ομάδα βρίσκονται και 2/212 φοιτητές του Οικονομικών Επιστημών, 3/239 φοιτητές της Διοίκησης Επιχειρήσεων, 3/78 της Χρηματοοικονομικής Λογιστικής και 15/79 του τμήματος Διεθνών-Ευρωπαϊκών Οικονομικών σπουδών που συγκέντρωσαν πολύ υψηλή βαθμολογία για τα τμήματά τους.

Στο ίδιο μέρος του επιπέδου συγκεντρώνονται αυτοί που περάσαν στο τμήμα πρώτης προτίμησης τους. Εκεί, λιγότερο ισχυρά βέβαια, βρίσκονται οι καταγόμενοι από χωριό, από τις περιοχές Ηπείρου - Θεσσαλίας, Μακεδονίας - Θράκης, που τελείωσαν το λύκειο το 1991 και εισήχθηκαν αμέσως στο Πανεπιστήμιο, προετοιμαζόμενοι με 1 χρόνο φροντιστήριο. Οι πρωτοετείς αυτής της ομάδας θεωρούν κύριο παράγοντα επιτυχίας τους το προσωπικό διάβασμα.

Στο πάνω μέρος του επιπέδου, βρίσκονται οι πρωτοετείς των 2 άλλων νέων τμημάτων Χρηματοοικονομικής Λογιστικής και Διεθνών-Ευρωπαϊκών Οικονομικών σπουδών. Αυτοί οι φοιτητές χαρακτηρίζονται από μεσαία βαθμολογία στο βασικό μάθημα και στο σύνολο μορίων τους. Επίσης, έχουν γονείς αποφοίτους ανώτερης σχολής, παρακολούθησαν 1 χρόνο ιδιαίτερα μαθήματα, κύρια είναι από Θεσσαλονίκη και πέτυχαν δίνοντας 2η χρονιά εξετάσεις.

Στο κάτω δεξιό τμήμα του επιπέδου, βρίσκεται η τρίτη ομάδα, που περιλαμβάνει τα παλιά τμήματα του Πανεπιστημίου, Οικονομικών Επιστημών και Διοίκησης Επιχειρήσεων. Οι πρωτοετείς αυτοί, συγκέντρωσαν χαμηλή βαθμολογία, τελειώσαν το λύκειο 1989, με συνέπεια να παρακολουθήσουν 3 χρόνια φροντιστήριο ή 2 χρόνια ιδιαίτερα. Άλλο χαρακτηριστικό τους είναι η προτίμηση να εισαχθούν σε παιδαγωγικά τμήματα ή σε Οικονομικό τμήμα άλλου πανεπιστημίου.

9.2.3 Ανάλυση του πίνακα άθροισμα

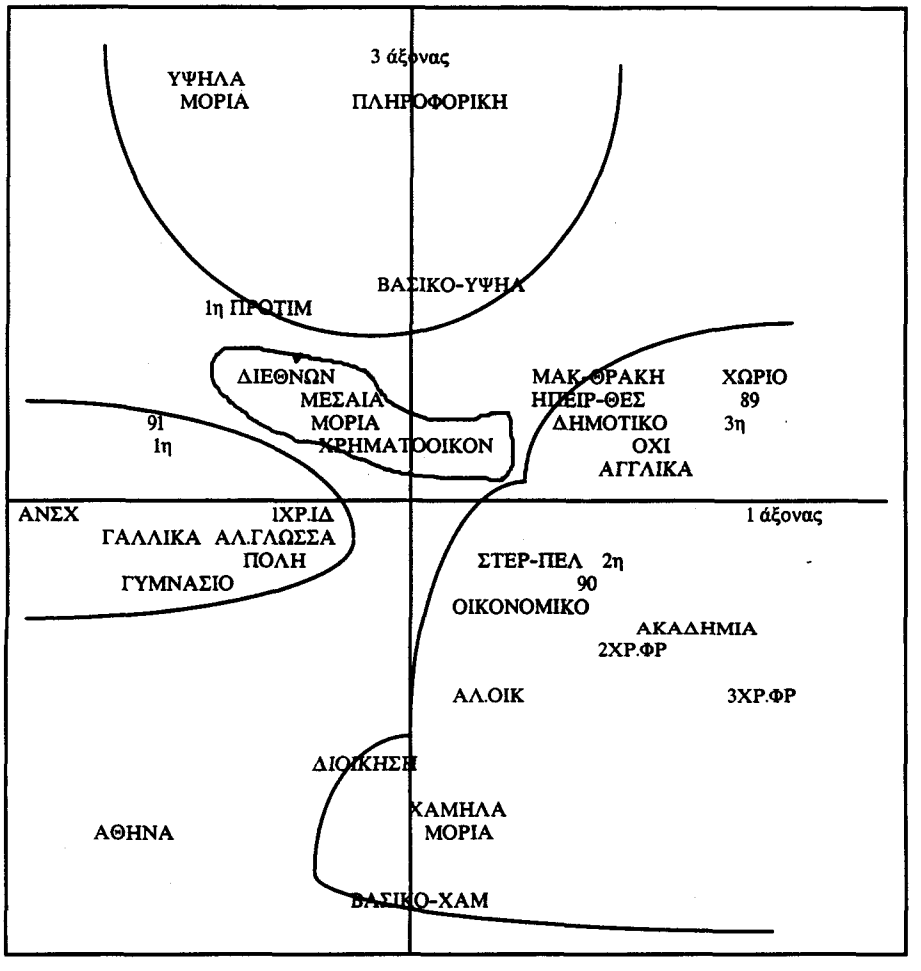
Από τους 5 πίνακες συμπτώσεων (51x51) των τμημάτων σχηματίζουμε τον πίνακα άθροισμα, που περιέχει για κάθε γραμμή I και στήλη j το άθροισμα των αντίστοιχων στοιχείων, από κάθε τμήμα.

Με την παραγοντική ανάλυση σ' αυτόν τον πίνακα έχουμε την δυνατότητα να μελετήσουμε την συμπεριφορά των κέντρων βάρους (μέση κατανομή) των μεταβλητών και για τα πέντε τμήματα. Μελετάμε δηλαδή, την εξωτερική δομή (inter=μεταξύ των κλάσεων των μεταβλητών) του πίνακα δεδομένων.

Στη συνέχεια παραθέτουμε σαν συμπληρωματικές γραμμές, 5 γραμμές, κάθε μία να αντιστοιχεί στο άθροισμα των στηλών για κάθε τμήμα. Τα σημεία στα παραγοντικά επίπεδα που αντιστοιχούν σ' αυτές τις γραμμές ορίζουν το κέντρο βάρους (μέση κατανομή) για κάθε τμήμα.

Παρατηρώντας τα κέντρα βάρους των 5 τμημάτων βλέπουμε ότι βρίσκονται στο κέντρο των 1 και 2 αξόνων, δηλαδή δεν διαφοροποιούνται ως προς αυτούς. Στο επίπεδο των 1x3 (σχήμα 13) διακρίνουμε 4 ομάδες. Πάνω το κέντρο βάρους (μέση κατανομή) των πρωτοετών φοιτητών του τμήματος Εφαρμοσμένης Πληροφορικής που τους χαρακτηρίζει κυρίως υψηλή βαθμολογία και η εισαγωγή στο τμήμα πρώτης προτίμησης τους.

Πιο κάτω στο κέντρο του επιπέδου παρατηρούμε τα κέντρα βάρους (μέσες κατανομές) των πρωτοετών των τμημάτων Διεθνών-Ευρωπαϊκών Οικονομικών Σπουδών και Χρηματοοικονομικής Λογιστικής, με κύριο χαρακτηριστικό τους τη μεσαία βαθμολογία.



σχήμα 13. Παραγοντικό επίπεδο 1 και 3 παραγοντικών αξόνων της AFC του πίνακα άθροισμα με συμπληρωματικά. όλα τα σημεία 5 τμημάτων

Κάτω δεξιά του επιπέδου παρατηρούμε τα κέντρα βάρους (μέσες κατανομές) των πρωτοετών των τμημάτων Οικονομικών Επιστημών και Διοίκησης Επιχειρήσεων. Τα χαρακτηριστικά των φοιτητών αυτών είναι η χαμηλή βαθμολογία, η εισαγωγή στο τμήμα δίνοντας 2η ή 3η χρονιά εξετάσεις, με 2 ή 3 χρόνια φροντιστήριο, χωρίς την εισαγωγή τους όμως σε Παιδαγωγικό ή σε άλλο Οικονομικό τμήμα, όπως το προτιμούσαν. Ακόμη οι φοιτητές αυτοί κατάγονται από χωριό των περιοχών Ηπείρου-Θεσσαλίας, Μακεδονίας-Θράκης, Στερεάς-Πελοποννήσου και οι γονείς τους τελείωσαν μόνο δημοτικό σχολείο.

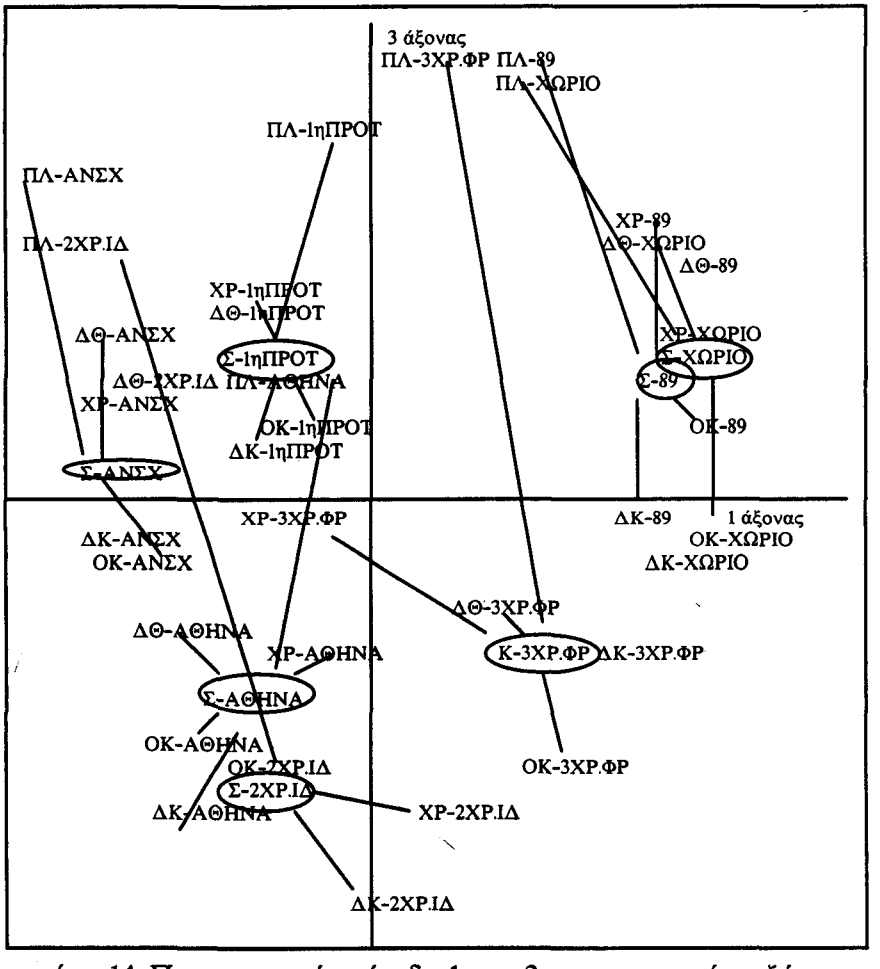
Τέλος κάτω αριστερά του επιπέδου, διακρίνουμε την ομάδα εκείνων των πρωτοετών, που είναι από Αθήνα-Θεσσαλονίκη ή άλλη πόλη, γνωρίζουν ξένες γλώσσες, έχουν γονείς μόρφωσης ανώτερης σχολής ή γυμνασίου, προετοιμάστηκαν

με ένα χρόνο ιδιαίτερα μαθήματα και πέρασαν στο Πανεπιστήμιο αμέσως μόλις τελείωσαν το λύκειο, χωρίς να συγκεντρώνουν πολύ υψηλή βαθμολογία.

9.2.4 Ανάλυση πίνακα άθροισμα με όλα τα σημεία των 5 τμημάτων

Θεωρούμε τον πίνακα άθροισμα 51x51 και κάτω απ' αυτόν παραθέτουμε τους 5 πίνακες φέτες 51x51 που αντιστοιχούν στα 5 τμήματα του πανεπιστημίου Μακεδονίας. Στην ανάλυση A.F.C. τα σημεία που αντιστοιχούν σε γραμμές των πινάκων φέτες και λαμβάνονται σαν συμπληρωματικά στα παραγοντικά επίπεδα, προσδιορίζουν τις διαφοροποιήσεις που παρουσιάζουν οι κατανομές τους, από την μέση κατανομή τους (κέντρο βάρους των ομολόγων γραμμών). Η διαφοροποίηση αυτή αναφέρεται μόνο ως προς την κοινή τάση τους.

Όταν στην ανάλυση του πίνακα άθροισμα πάρουμε σαν συμπληρωματικά σημεία και όλες τις μεταβλητές των 5 τμημάτων, καταλήγουμε στα ίδια συμπεράσματα για το επίπεδο των αξόνων 1 και 2, που καταλήξαμε αναλύοντας μόνο τον πίνακα άθροισμα.



σχήμα 14. Παραγοντικό επίπεδο 1 και 3 παραγοντικών αξόνων της AFC του πίνακα άθροισμα με συμπληρ. όλα τα σημεία 5 τμημάτων

Παρατηρώντας όμως το παραγοντικό επίπεδο των 1x3 (σχήμα 14), βλέπουμε ότι οι μεταβλητές του τμήματος Εφαρμοσμένης Πληροφορικής βρίσκονται στην κορυφή, με την υψηλή βαθμολογία, και απομακρυσμένες από την κοινή τάση όλων των μεταβλητών. Πιο κάτω και λίγο ψηλότερα από την κοινή τάση βρίσκονται των Διεθνών-Ευρωπαϊκών Οικονομικών σπουδών, στο κέντρο της Χρηματοοικονομικής Λογιστικής, ενώ κάτω και χαμηλότερα από την κοινή τάση όλων των μεταβλητών συναντάμε μαζί χωρίς να ξεχωρίζουν τις μεταβλητές των τμημάτων Διοίκησης Επιχειρήσεων και Οικονομικών Επιστημών.

Η διαφοροποίηση των τμημάτων υπάρχει μόνο ως προς τον 3ο άξονα, που ερμηνεύει από κάτω προς τα πάνω την χαμηλή βαθμολογία και την επιθυμία για εισαγωγή σε παιδαγωγικό ή σε άλλο Οικονομικό τμήμα (κάτω), ενώ προχωρώντας συναντάμε την μεσαία και υψηλή βαθμολογία καθώς και την εισαγωγή στο τμήμα επιθυμίας (πάνω).

9.2.5 Συμπεράσματα εφαρμογής

Όπως διαπιστώθηκε σε όλες τις αναλύσεις, οι πρωτοετείς φοιτητές του πανεπιστημίου Μακεδονίας, ως προς την προέλευση χωρίζονται σε δύο ομάδες.

Η πρώτη περιλαμβάνει τους φοιτητές με κύρια προέλευση από χωριό ή κωμόπολη, που έχουν γονείς επιπέδου μόρφωσης δημοτικού, εισήχθηκαν στο Πανεπιστήμιο μετά από προσπάθεια 2 ή 3 χρόνων και δεν ξέρουν ξένες γλώσσες. Κάποια διαφοροποίηση παρατηρείται μόνο στο τμήμα Εφαρμοσμένης Πληροφορικής, όπου οι πρωτοετείς από χωριό πέρασαν σ' αυτό αμέσως μετά την αποφοίτηση τους από το λύκειο.

Η δεύτερη αντίθετα περιλαμβάνει αυτούς που προέρχονται κύρια, από μεγάλα αστικά κέντρα, με γονείς μόρφωσης ανώτερης σχολής, οι οποίοι εισήχθηκαν στο πανεπιστημίο, αμέσως μόλις τελείωσαν το λύκειο, γνωρίζοντας και κάποια ξένη γλώσσα.

Οι μαθητές που κατοικούν σε πόλη έχουν την ευκαιρία και την δυνατότητα να προετοιμάζονται με φροντιστήρια και ιδιαίτερα μαθήματα και να μαθαίνουν ξένες γλώσσες. Αυτό έχει σαν συνέπεια να εισάγονται στο τμήμα πρώτης επιλογής τους χωρίς να χάνουν καμία χρονιά μετά το λύκειο. Αντίθετα οι μαθητές των χωριών, δεν έχουν τη δυνατότητα να μαθαίνουν ξένες γλώσσες, κύρια επιθυμούν να εισαχθούν σε Παιδαγωγικά τμήματα, και προετοιμάζονται παρακολουθώντας φροντιστήριο, μόνο όταν τελειώσουν το λύκειο και δεν εισαχθούν στο Πανεπιστήμιο, οπότε προσπαθούν 2η ή 3η φορά, μη βρίσκοντας άλλη διέξοδο επαγγελματικής αποκατάστασης.

Παρά του ότι τα χρόνια παρακολούθησης φροντιστηρίου, δεν συμβάλλουν σημαντικά στην τελική επίδοση, οι μαθητές που παρακολούθησαν ιδιαίτερα μαθήματα, (κυρίως αυτοί με γονείς πτυχιούχους ανώτερης σχολής), εισήχθηκαν αμέσως μετά το λύκειο, και μάλιστα σε τμήμα της πρώτης προτίμησης τους.

Η βαθμολογία στο βασικό μάθημα (Μαθηματικά) ακολουθεί το σύνολο των μορίων, δηλαδή η επίδοση των μαθητών στο βασικό μάθημα ήταν περίπου ίδια και με την επίδοση στα άλλα μαθήματα. Αυτό φανερώνει το θετικό, κατά τη γνώμη μας, της απόφασης για κατάργηση του βασικού μαθήματος, αφού δεν διαφοροποιούσε τους μαθητές.

Από την ανάλυση του πίνακα άθροισμα, παρατηρούμε ότι τα κέντρα βάρους των παλιών τμημάτων του Πανεπιστημίου Μακεδονίας, Οικονομικών Επιστημών και Διοίκησης Επιχειρήσεων, βρίσκονται στην ομάδα των πρωτοετών που είναι από χωριό, προετοιμάστηκαν με 2 ή 3 χρόνια φροντιστήριο, δίνοντας 2η ή 3η φορά εξετάσεις και οι περισσότεροι επιθυμούσαν να εισαχθούν κύρια σε Παιδαγωγικό ή άλλο Οικονομικό τμήμα.

Οι πρωτοετείς λοιπόν προτιμούσαν την εισαγωγή τους στα νέα τμήματα του Πανεπιστημίου Μακεδονίας, Εφαρμοσμένης Πληροφορικής, Χρηματοοικονομικής Λογιστικής και Διεθνών-Ευρωπαϊκών Οικονομικών σπουδών, ίσως επειδή το εντυπωσιακό του ονόματός τους, τους κάνει να πιστεύουν ότι υπόσχονται καλύτερες προοπτικές, από ότι τα παλιά τμήματα.

Επίσης από την ανάλυση του πίνακα άθροισμα με συμπληρωματικά όλα τα σημεία των 5 τμημάτων βλέπουμε ότι εκτός από την διαφοροποίηση της βαθμολογίας, τα τμήματα διαφοροποιούνται και ως προς την προτίμηση των πρωτοετών για το τμήμα εισαγωγής τους. Περισσότερο ικανοποιημένοι είναι οι πρωτοετείς της Εφαρμοσμένης Πληροφορικής αφού πέρασαν όπου επιθυμούσαν, ακολουθούν οι πρωτοετείς των τμημάτων Χρηματοοικονομικής Λογιστικής και Διεθνών-Ευρωπαϊκών Οικονομικών σπουδών, ενώ μπορούμε να πούμε ότι στην πλειοψηφία τους οι πρωτοετείς των τμημάτων Οικονομικών Επιστημών και Διοίκησης Επιχειρήσεων, επιθυμούσαν να σπουδάσουν κάτι άλλο, αλλά δεν τα κατάφεραν. Αυτό πιθανόν να επιδράσει αρνητικά στην παραπέρα φοίτηση τους.

Διαπιστώνουμε λοιπόν ότι εφαρμόζοντας την Παραγοντική Ανάλυση των Αντιστοιχιών στους τροποποιημένους πίνακες που προτείνουμε, καταφέρνουμε να αναδείξουμε χωρίς καμία a priori υπόθεση, τα βασικά χαρακτηριστικά του φαινομένου που περιγράφονται από τον τρισδιάστατο πίνακα των πρωτοετών φοιτητών του Πανεπιστημίου Μακεδονίας.

10. Συμπεράσματα

Παρουσιάστηκε στο κεφάλαιο αυτό ο τρόπος δημιουργίας των πινάκων μεταβλητών, η χρησιμότητα των πινάκων αυτών στην περιγραφή δεδομένων αλλά και στην εύρεση σχέσεων μεταξύ μεταβλητών. Χρησιμοποιώντας τους πίνακες αυτούς, μπορούμε να δούμε συνοπτικά τις τάσεις του φαινομένου που μελετάμε, καθώς και να εξάγουμε κάποια συμπεράσματα για τις σχέσεις μεταξύ των μεταβλητών.

Η συμβολή της παρούσας μελέτης έγκειται στην καλύτερη παρουσίαση των μελετώμενων πινάκων. Ένας πίνακας συμπτώσεων παρουσιάζεται με τρόπο ώστε στο ίδιο κελί να εμφανίζεται η αντίστοιχη συχνότητα, η σχετική συχνότητα ως προς την γραμμή και η σχετική συχνότητα ως προς την στήλη, και ακόμη ο συντελεστής συσχέτισης μεταξύ των δύο μεταβλητών του πίνακα συμπτώσεων. Με τον τρόπο αυτό οι πίνακες είναι έτοιμοι για την εξαγωγή σχέσεων μεταξύ μεταβλητών, αλλά και για μια παραπέρα ανάλυση με τις μεθόδους της Ανάλυσης Δεδομένων.

Η παρουσίαση του γενικευμένου πίνακα Burt, γίνεται πρώτη φορά στην ελληνική βιβλιογραφία. Ο πίνακας αυτός είναι χρήσιμος στην καλύτερη κωδικοποίηση ποσοτικών μεταβλητών, ώστε να διατηρούμε όσο το δυνατόν περισσότερη πληροφορία.

Επίσης στο κεφάλαιο αυτό προτείναμε μια μέθοδο Ανάλυσης Δεδομένων για την επεξεργασία των τρισδιάστατων πινάκων δεδομένων. Ο πίνακας που περιγράφει τα δεδομένα αυτά έχει τρεις διαστάσεις και παρουσιάσαμε τον τρόπο επεξεργασίας του, ώστε να εξαχθούν όλες οι τάσεις των δεδομένων και για τις τρεις διαστάσεις του. Νομίζουμε ότι η πρόταση μας στο μετασχηματισμό των τρισδιάστατων πινάκων συμβάλλει σημαντικά στην επεξεργασία αυτών με τις μεθόδους της Ανάλυσης Δεδομένων.

Στο παρόν κεφάλαιο προτείναμε την δημιουργία νέων πινάκων κατάλληλων να δεχθούν επεξεργασία με τις μεθόδους της Ανάλυσης Δεδομένων, ώστε να αναδειχθούν οι επικρατέστερες τάσεις που χαρακτηρίζουν τα φαινόμενα τα οποία περιγράφονται από τους πίνακες. Οι μέθοδοι της Παραγοντικής Ανάλυσης των Αντιστοιχιών και της Ανάλυσης σε Κύριες Συνιστώσες, που χρησιμοποιούνται για την επεξεργασία των πινάκων, παρουσιάστηκαν στο προηγούμενο κεφάλαιο και θα περιγράψουμε την μέθοδο της Ταξινόμηση σε επόμενο κεφάλαιο.

ΚΕΦΑΛΑΙΟ 3
ΜΕΤΡΙΚΕΣ ΚΑΙ ΔΕΙΚΤΕΣ ΑΠΟΣΤΑΣΗΣ-ΟΜΟΙΟΤΗΤΑΣ

Στην κεφάλαιο αυτό, μελετάμε τις διάφορες μετρικές αποστάσεων που έχουν προταθεί για την μέτρηση της προσέγγισης τόσο μεταξύ μεταβλητών, όσο και μεταξύ αντικειμένων. Αποστάσεις μεταξύ μεταβλητών, χρησιμοποιούνται όταν θέλουμε να ταξινομήσουμε τις μεταβλητές, βρίσκοντας τις ομάδες εκείνες που σχετίζονται μεταξύ τους, ενώ όταν αναζητούμε τις ομογενείς ομάδες των αντικειμένων, χρησιμοποιούμε μετρικές ή δείκτες ομοιότητας μεταξύ των αντικειμένων.

Η διαφορά μεταξύ "μετρικής" και "δείκτη" αφορά κυρίως στις μαθηματικές ιδιότητες των αντίστοιχων συναρτήσεων. Ο ρόλος τους στην ταξινόμηση είναι κοινός, και για το λόγο αυτό στο εξής αντιμετωπίζουμε με τον ίδιο τρόπο μετρικές και δείκτες απόστασης ή ομοιότητας.

Στο πρώτο μέρος ασχολούμαστε με την ταξινόμηση κατά αύξουσα ιεραρχία των αντικειμένων. Περιγράφονται οι αποστάσεις που χρησιμοποιούνται μεταξύ αντικειμένων, ανάλογα με το είδος των μεταβλητών που τα χαρακτηρίζουν.

Στο δεύτερο μέρος, μελετάμε τις διάφορες μετρικές αποστάσεων που έχουν προταθεί για την μέτρηση της προσέγγισης μεταξύ μεταβλητών. Εξετάζουμε ποσοτικές, ποιοτικές και δυαδικές μεταβλητές. Κατόπιν περιγράφονται οι μέθοδοι που χρησιμοποιούν τις αποστάσεις αυτές, για να την μέτρηση της ομοιότητας μεταξύ ομάδων μεταβλητών προκειμένου να πραγματοποιηθεί η ιεράρχησή τους.

Στο τέλος του μέρους αυτού γίνεται η αξιολόγηση των παραπάνω μετρικών αποστάσεων και παρουσιάζουμε τον τρόπο επιλογής της κατάλληλης κάθε φορά μετρικής, ανάλογα με τα δεδομένα και τις ανάγκες της έρευνάς μας.

1. Αποστάσεις για αντικείμενα

Συχνά χρησιμοποιούμε τις μεθόδους ταξινόμησης, για να βρούμε ομοιογενείς ομάδες και ιεραρχήσεις, μεταξύ αντικειμένων. (Λέγοντας αντικείμενα, εννοούμε γενικά τις παρατηρήσεις μας.)

Η ταξινόμηση κατά αύξουσα ιεραρχία, διενεργείται σε δύο στάδια. Στο πρώτο υπολογίζουμε την ομοιότητα μεταξύ των αντικειμένων, χρησιμοποιώντας μια μετρική απόστασης (ή κάποιο συντελεστή συσχέτισης), ενώ στο δεύτερο στάδιο η δημιουργία της ιεράρχησης συντελείται με την συνένωση ομάδων (από τις μικρότερες με ένα αντικείμενο μέχρι την ομάδα που τα περιέχει όλα). Οι μέθοδοι που χρησιμοποιούνται στο δεύτερο στάδιο είναι MIN, MAX, MOYENNE, κ.ά. Στο κεφάλαιο 4, θα ασχοληθούμε με τις μεθόδους που αφορούν το δεύτερο στάδιο.

Για το πρώτο στάδιο, χρειάζεται ένα μέτρο σύγκρισης δύο αντικειμένων. Πρέπει δηλαδή να είμαστε σε θέση να πούμε αν δύο αντικείμενα θεωρούνται "παρόμοια" ή όχι και κατά πόσο μοιάζουν. Αυτό γίνεται χρησιμοποιώντας ένα μέτρο ομοιότητας ή μια μετρική (δείκτη) απόστασης, μεταξύ δύο αντικειμένων.

Η επιλογή της απόστασης που θα χρησιμοποιηθεί για την μέτρηση της ομοιότητας μεταξύ αντικειμένων, εξαρτάται από το είδος των αντικειμένων, από το είδος των μεταβλητών που τα χαρακτηρίζουν καθώς και από τα στοιχεία εκείνα στα οποία θέλουμε να δώσουμε βαρύτητα στην ανάλυσή μας.

Γενικά τα αντικείμενα (παρατηρήσεις) μπορούν να χαρακτηρίζονται από:

- Ποσοτικές μεταβλητές
- Ποιοτικές μεταβλητές με πολλές κλάσεις
- Ποιοτικές μεταβλητές με δύο κλάσεις (παρουσία-απουσία) ή αλλιώς δυαδικές μεταβλητές

Υπάρχουν διάφορα μέτρα απόστασης ή ομοιότητας για δυαδικές, ποιοτικές και ποσοτικές μεταβλητές. Κάτω από κάποιες συνθήκες, ποσοτικά δεδομένα, είναι καλύτερο να αντιμετωπιστούν σαν ποιοτικά ή δυαδικά, όπως στην περίπτωση που διχοτομούνται ποσοτικές μεταβλητές με θόρυβο, ή όταν έχουμε ανάγκη μόνο την πληροφορία που εξαρτάται από κάποιο όριο. Για παράδειγμα, όταν ταξινομούμαι ποταμούς σύμφωνα με την ικανότητα τους για εκτροφή ψαριών, μας ενδιαφέρει αν οι τιμές περιεκτικότητας των βαρέων μετάλλων είναι κάτω ή πάνω από το επιτρεπτό επίπεδο τοξικότητας. Έτσι αν και τα δεδομένα είναι ποσοτικά τα χειριζόμαστε σαν δυαδικά.

1.1 Αντικείμενα που χαρακτηρίζονται από ποιοτικές μεταβλητές

1.1.1 Εισαγωγή

Όπως ήδη αναφέραμε, ποιοτικές είναι οι μεταβλητές που δεν μετρούνται με νούμερα, και οι τιμές των οποίων χωρίζονται σε κατηγορίες. Οι κατηγορίες αυτές μπορούν να κωδικοποιηθούν με ακεραίους αριθμούς, προκειμένου να διευκολυνθούμε στην επεξεργασία και αποθήκευση των μεταβλητών αυτών. Διακρίνονται σε διατεταγμένες όταν υπάρχει μια διάταξη μεταξύ των κατηγοριών τους, και σε μη διατεταγμένες. Για παράδειγμα, η μεταβλητή επίδοση που παίρνει τιμές μέτρια, καλά, άριστα είναι διατεταγμένη μεταβλητή ενώ η μεταβλητή χρώμα ματιών με τιμές μαύρο, καστανό γαλάζιο πράσινο είναι μη διατεταγμένη.

Υπάρχουν διαφορετικοί δείκτες απόστασης - ομοιότητας ανάλογα με το αν οι ποιοτικές μεταβλητές είναι διατεταγμένες ή όχι, αφού μια διαφορά μεταξύ δύο κατηγοριών έχει διαφορετική σπουδαιότητα αν οι κατηγορίες είναι διατεταγμένες.

Η διαφορά μεταξύ άριστα-μέτρια είναι σημαντικότερη από την διαφορά μεταξύ καλά-μέτρια.

Διατεταγμένες ποιοτικές μεταβλητές

Πολλοί αντιμετωπίζουν τις διατεταγμένες μεταβλητές, κωδικοποιώντας τες σύμφωνα με την διάταξη των κατηγοριών τους, αντιστοιχώντας ένα ακέραιο αριθμό από 1 έως το πλήθος κατηγοριών, και χρησιμοποιώντας στη συνέχεια μια μετρική απόστασης για ποσοτικές μεταβλητές, όπως την ευκλείδεια ή την Manhattan απόσταση. Έτσι όμως είναι δύσκολο να συγκρίνουμε μεταβλητές με διαφορετικό πλήθος κατηγοριών αφού η διαφορά 2(=3-1) μεταξύ πρώτης και τρίτης κατηγορίας για μια μεταβλητή με 3 κατηγορίες, είναι σημαντικότερη από την ίδια διαφορά 2, μια μεταβλητής με 10 κατηγορίες. Προκειμένου να αποφύγουμε την επίδραση του διαφορετικού πλήθους κατηγοριών μεταξύ μεταβλητών, μετατρέπουμε τους κωδικούς των κατηγοριών τους, ώστε να βρίσκονται στο διάστημα 0 έως 1, εφαρμόζοντας των μετασχηματισμό:

$$z_{if} = (r_{if}-1)/(M_f-1)$$

όπου M_f είναι η μεγαλύτερη τάξη της μεταβλητής f και r_{if} η τάξη που μετασχηματίζεται σε z_{if}. [Kaufman,1990]

Ένας άλλος τρόπος αντιμετώπισης διατεταγμένων ποιοτικών μεταβλητών είναι η μετατροπή τους σε μεταβλητές 0-1 και στη συνέχεια η χρησιμοποίηση κάποιας μετρικής ή δείκτη ομοιότητας που είναι κατάλληλη για λογικά δεδομένα. Π.χ. για μια μεταβλητή με m κατηγορίες κωδικοποιημένες από 1 έως m, μετατρέπεται σε m νέες λογικές μεταβλητές ως εξής:

$x_k=1$ αν $k \leq j$ και 0 διαφορετικά, k=1..m
όπου j η τιμή της ποιοτικής μεταβλητής.

Έτσι για μια μεταβλητή με 4 κατηγορίες θα έχουμε την κωδικοποίηση σε 4 λογικές μεταβλητές x_1, x_2, x_3, x_4

κατηγορίες	x_1	x_2	x_3	x_4
1	1	0	0	0
2	1	1	0	0
3	1	1	1	0
4	1	1	1	1

Μια απόσταση που μπορεί να χρησιμοποιηθεί στην περίπτωση αυτή, είναι ο αριθμός των διαφορετικών συνιστωσών μεταξύ δύο σημείων. [Marchetti,1989]

$\delta(x,y)=\sum \delta q(x_i,y_i)$ όπου $\delta q(x_i,y_i)=1$ αν $x_i \neq y_i$ και $\delta q(x_i,y_i)=0$ διαφορετικά

Μη διατεταγμένες ποιοτικές μεταβλητές

Ο συχνότερος τρόπος αντιμετώπισης των μη διατεταγμένων ποιοτικών μεταβλητών, είναι η μετατροπή μιας ποιοτικής μεταβλητής σε πολλές δυαδικές μεταβλητές, (παρουσία ή απουσία κάποιας κατηγορίας), και στη συνέχεια ο υπολογισμός της ομοιότητας ή διαφοράς των νέων αυτών μεταβλητών με τους αντίστοιχους δείκτες που υπάρχουν για δυαδικά δεδομένα.

Μια απλούστερη προσέγγιση είναι η χρήση του "simple matching" [Sokal,1958]. Η απόσταση μεταξύ δύο αντικειμένων i και j είναι

$d(i,j)=(p-u)/p$ (για ομοιότητα $s(i,j)=u/p$)

όπου u ο αριθμός των κοινών κατηγοριών, δηλαδή ο αριθμός των μεταβλητών για τις οποίες τα αντικείμενα i και j έχουν ακριβώς τις ίδιες τιμές, ενώ p είναι το πλήθος όλων των μεταβλητών που μελετάμε. Είναι δυνατό να δώσουμε και βάρη στους αριθμούς u και p-u όπως κάνουμε και για τις λογικές μεταβλητές. [Kaufman,1990]

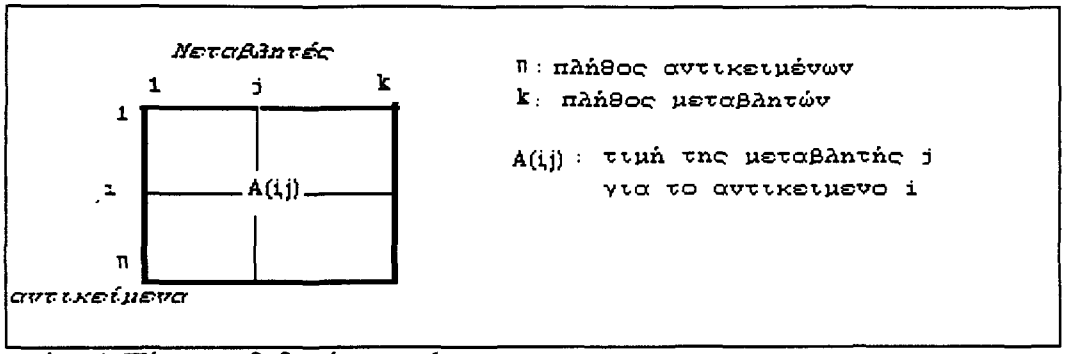
Σύγκριση των τριών μετρικών

Οι πιο γνωστές αποστάσεις που χρησιμοποιούνται για την ταξινόμηση αντικειμένων κατ' αύξουσα ιεραρχία, για δεδομένα που χαρακτηρίζονται από ποιοτικές μεταβλητές είναι του X^2, η ευκλείδεια, και ο δείκτης Jaccard.

Προσπαθούμε να μελετήσουμε την επίδραση αυτών των τριών μετρικών απόστασης, της ευκλείδειας, της X^2 και του δείκτη Jaccard, στον υπολογισμό αποστάσεων για αντικείμενα που χαρακτηρίζονται από ποιοτικές μεταβλητές ή ποσοτικές μεταβλητές όταν είναι χωρισμένες σε κλάσεις. Σκοπός μας είναι η σύγκριση των αποστάσεων και η επίδρασή τους στην δημιουργία της ιεράρχησης κατά την ταξινόμηση κατ' αύξουσα ιεραρχία.

Εξετάζονται οι περιπτώσεις αρχικά κάθε αντικείμενο να ανήκει σε μία μόνο κλάση κάθε μεταβλητής, χωρίς ή με επαναλήψεις, και στη συνέχεια ένα αντικείμενο να ανήκει σε περισσότερες κλάσεις ή να μην ανήκει σε καμία κλάση για κάποιες μεταβλητές.

Ο πίνακας δεδομένων A nxk, παριστάνει n αντικείμενα που χαρακτηρίζονται ή όχι από κάθε μία από τις k κλάσεις των μεταβλητών (σχήμα 1). Κάθε στήλη του πίνακα αντιστοιχεί σε μια κλάση, κι ένα αντικείμενο μπορεί να χαρακτηρίζεται ή όχι από την κλάση αυτή. Στη διασταύρωση, λοιπόν, της i γραμμής και j στήλης του πίνακα A(i,j), υπάρχει 1, όταν το i αντικείμενο χαρακτηρίζεται από την αντίστοιχη κλάση της j στήλης, και 0 διαφορετικά.



σχήμα 1. Πίνακας δεδομένων nxk

Όταν μελετάμε ποιοτικές μεταβλητές, αναλύουμε τον πίνακα A nxk, με στοιχεία 0 και 1, όπου το σύνολο των στηλών k αντιστοιχεί στο σύνολο των κλάσεων όλων των μεταβλητών, η απόσταση δύο αντικειμένων i και i′ είναι:

με την **ευκλείδεια μετρική**: $d_e^2(i,i') = \sum_{j=1}^{k} [A(i,j) - A(i',j)]^2$

με την **X2**: $d_{x2}^2(i,i') = \sum_{j=1}^{k} \frac{1}{A(i,j)} \left[\frac{A(i,j)}{A(i,.)} - \frac{A(i',j)}{A(i',.)} \right]^2$

και με τον **δείκτη Jaccard**: $d_j(i,i') = 1 - \frac{c}{p+q+c}$

όπου
p το πλήθος των στηλών με 1 του i αντικειμένου
q το πλήθος των στηλών με 1 του i αντικειμένου
c το πλήθος των στηλών με κοινά 1 στα δύο αντικείμενα
A(i,.) το άθροισμα της i γραμμής και A(.,j) το άθροισμα της j στήλης

Η ευκλείδεια μετρική [Benzecri,1973] βασίζεται στον υπολογισμό των διαφορών μεταξύ όλων των συντεταγμένων των αντικειμένων. Το ίδιο ισχύει για την X^2 [Benzecri,1980], μόνο που κάθε διαφορά σταθμίζεται με το άθροισμα της στήλης για την αντίστοιχη συντεταγμένη.

Ο δείκτης Jaccard [Roux,1985], υπολογίζεται μετρώντας τον αριθμό των συντεταγμένων με κοινά 1 και τον αριθμό αυτών που έχουν 1 μόνο στο ένα από τα δύο αντικείμενα. Το πλήθος των κοινών 0 δεν λαμβάνεται υπόψη.

1.1.2.Υπολογισμός αποστάσεων

α. Πλήρης πίνακας μιας μεταβλητής χωρίς επαναλήψεις (ανόμοια αντικείμενα)

Θεωρούμε κατ' αρχήν μια μόνο μεταβλητή χωρισμένη σε k κλάσεις και κάθε αντικείμενο, ανήκει οπωσδήποτε σε μία μόνο από αυτές. Εξετάζουμε όλα τα δυνατά διαφορετικά αντικείμενα, που είναι σε πλήθος k. Ο πίνακας δεδομένων

είναι διαστάσεων kxk, το άθροισμα κάθε γραμμής του είναι 1 και το άθροισμα κάθε στήλης του είναι επίσης 1.

Δεν έχει νόημα η ταξινόμηση όλων των διαφορετικών αντικειμένων όταν τα δεδομένα αφορούν μια μεταβλητή. Για καθαρά εισαγωγικούς όμως λόγους μελετάμε αρχικά την περίπτωση αυτή.

Βρίσκουμε λοιπόν τις αποστάσεις ανά δύο, όλων των αντικειμένων, που είναι διαφορετικά μεταξύ τους, με τις 3 μετρικές.

Έστω $i,i' \le k$ δύο διαφορετικές κλάσεις και δύο διαφορετικά αντικείμενα a_i και $a_{i'}$, που χαρακτηρίζονται από τις κλάσεις αυτές αντίστοιχα.

$a_i=(0,0,1...,0,...,0)$ και $a_{i'}=(0,0,...,1,...,0)$

Οι αποστάσεις τους είναι:

$$d_e(a_i,a_{i'})=\sqrt{(0-0)^2+...+(1-0)^2+...+(1-0)^2...+(0-0)^2}=\sqrt{2}$$

$$d_{x2}(a_i,a_{i'})=\sqrt{\frac{1}{1}\left(\frac{1}{1}-0\right)^2+\frac{1}{1}\left(0-\frac{1}{1}\right)^2}=\sqrt{2} \qquad d_J(a_i,a_{i'})=1-\frac{0}{2}=1$$

Η απόσταση λοιπόν, ανά δύο όλων των διαφορετικών αντικειμένων, όπως είναι φυσικό, είναι η ίδια για την κάθε μετρική, και η ταξινόμηση δεν έχει νόημα.

β. Πίνακας μιας μεταβλητής με επαναλήψεις

Υποθέτουμε ότι έχουμε επαναλήψεις αντικειμένων στον πίνακα δεδομένων (υπάρχουν κι όμοια αντικείμενα εκτός από τα k διαφορετικά).

Έστω ότι έχουμε ένα πίνακα δεδομένων nxk (n>k). Αν το a_i ($i \le k$) αντικείμενο εμφανίζεται p φορές και το $a_{i'}$ ($i' \le k$) εμφανίζεται q φορές, τότε το άθροισμα της i στήλης είναι p, της i' στήλης είναι q.

Οι αποστάσεις μεταξύ των δύο αντικειμένων a_i, $a_{i'}$ είναι:

$$d_e(a_i,a_{i'})=\sqrt{(0-0)^2+...+(1-0)^2+...+(0-1)^2+...+(0-0)^2}=\sqrt{2}$$

$$d_{x2}(a_i,a_{i'})=\sqrt{\frac{1}{p}\left(\frac{1}{1}-0\right)^2+\frac{1}{q}\left(0-\frac{1}{1}\right)^2}=\sqrt{\frac{1}{p}+\frac{1}{q}}=d_{x2}(p,q) \qquad d_J(a_i,a_{i'})=1-\frac{0}{2}=1$$

Έστω τα αντικείμενα a_i, $a_{i'}$ και $a_{i''}$, που εμφανίζονται p,q και r φορές αντίστοιχα. Αν p < q, τότε: $d_{X}2(r,p) > d_{X}2(r,q) \Rightarrow d_{X}2\,(a_{i''},a_i) > d_{X}2\,(a_{i''},a_{i'})$

δηλαδή η απόσταση με την μετρική X^2 του $a_{i''}$ από το a_i (αντικείμενο με p επαναλήψεις), είναι μεγαλύτερη από την απόστασή του με το $a_{i'}$ (αντικείμενο με q επαναλήψεις), όταν p<q. Αυτό οφείλεται στη στάθμιση που χρησιμοποιείται στην X^2, με συνέπεια τα "βαριά" (με πολλές επαναλήψεις) αντικείμενα να ενώνονται ταχύτερα για τον σχηματισμό των κόμβων της ταξινόμησης.

Η X^2, λοιπόν, οδηγεί σε λάθος υπολογισμό των αποστάσεων, όταν δεν θέλουμε να παίζει σημαντικό ρόλο στην ταξινόμηση ο αριθμός των επαναλήψεων. Στην περίπτωση αυτή ενδείκνυται η χρησιμοποίηση της ευκλείδειας απόστασης ή του δείκτη Jaccard. Αν όμως επιθυμούμε να λαμβάνεται υπόψη η επανάληψη ενός αντικειμένου στην ταξινόμηση, πρέπει να χρησιμοποιήσουμε τη μετρική X^2.

γ. Πίνακας δύο μεταβλητών, χωρίς επαναλήψεις.

Έστω ότι μελετάμε διαφορετικά αντικείμενα που χαρακτηρίζονται από δύο μεταβλητές, η πρώτη με k_1 κλάσεις κι η δεύτερη με k_2. Όλα τα δυνατά διαφορετικά αντικείμενα είναι σε πλήθος k_1k_2, με συνέπεια ο πίνακας δεδομένων να έχει k_1k_2 γραμμές και k_1+k_2 στήλες. Το άθροισμα κάθε γραμμής είναι 2 (πλήθος μεταβλητών), ενώ το άθροισμα των στηλών της πρώτης μεταβλητής είναι k_2 και της δεύτερης k_1.

Έστω a_{ij} το αντικείμενο που χαρακτηρίζεται από τη i-οστή κλάση της πρώτης μεταβλητής και την j-οστή κλάση της δεύτερης μεταβλητής ($i \le k_1$, $j \le k_2$).

Αν δύο αντικείμενα a_{ij}, $a_{i'j}$, διαφέρουν μόνο ως προς μια κλάση της πρώτης μεταβλητής οι αποστάσεις τους είναι:

$$d_e(a_{ij}, a_{i'j}) = \sqrt{1^2+1^2} = \sqrt{2} \qquad d_{x^2}(a_{ij}, a_{i'j}) = \sqrt{\frac{2}{k_2}\left(\frac{1}{2}\right)^2} = d_{x^2}(k_2)$$

$$d_J(a_{ij}, a_{i'j}) = 1 - \frac{1}{2+2-1} = \frac{2}{3}$$

Αν δύο αντικείμενα a_{ij}, $a_{ij'}$, διαφέρουν μόνο ως προς μια κλάση της δεύτερης μεταβλητής οι αποστάσεις τους είναι:

$$d_e(a_{ij}a_{ij}) = \sqrt{1^2+1^2} = \sqrt{2} \qquad d_{x^2}(a_{ij}, a_{ij'}) = \sqrt{\frac{2}{k_1}\left(\frac{1}{2}\right)^2} = d_{x^2}(k_1) \qquad d_J(a_{ij}a_{ij}) = 1\frac{1}{2+2-1} = \frac{2}{3}$$

Τέλος όταν δύο αντικείμενα a_{ij}, $a_{i'j'}$, χαρακτηρίζονται από διαφορετικές κλάσεις τόσο για την πρώτη όσο και για την δεύτερη μεταβλητή, οι αποστάσεις τους είναι:

$$d_e(a_{ij}, a_{i'j'}) = \sqrt{1^2+1^2+1^2+1^2} = \sqrt{4}$$

$$d_{x^2}(a_{ij}, a_{i'j'}) = \sqrt{\frac{2}{k_2}\left(\frac{1}{2}\right)^2 + \frac{2}{k_1}\left(\frac{1}{2}\right)^2} = d_{x^2}(k_1, k_2) \quad d_J(a_{ij}, a_{i'j'}) = 1 - \frac{0}{2+2} = 1$$

Παρατήρηση: Η απόσταση (με τη μετρική X^2) δύο αντικειμένων που διαφέρουν σε μία μόνο κλάση, εξαρτάται από την μεταβλητή στην οποία ανήκει η κλάση αυτή. **Όσο περισσότερες κλάσεις έχει μια μεταβλητή, τόσο μικρότερη είναι η απόσταση δύο σημείων που διαφέρουν ως προς μια κλάση της άλλης μεταβλητής**. Δηλαδή αν την ίδια μεταβλητή την χωρίσουμε σε διαφορετικό αριθμό κλάσεων, η νέα ταξινόμηση των αντικειμένων θα διαφέρει στους χαμηλούς κόμβους (στους πρωτοσχηματιζόμενους).

Στην περίπτωση που δεν θέλουμε ο αριθμός κλάσεων κάθε μεταβλητής να επηρεάζει την ταξινόμηση, η χρήση της μετρικής X^2, οδηγεί σε εσφαλμένα αποτελέσματα, και ενδείκνυται η χρησιμοποίηση της ευκλείδειας μετρικής ή του δείκτη Jaccard.

δ. Πίνακας m μεταβλητών, χωρίς επαναλήψεις.

Γενικεύουμε τα παραπάνω για m μεταβλητές, όπου η μεταβλητή j ($1 \le j \le m$) έχει k_j κλάσεις και όλα τα δυνατά διαφορετικά αντικείμενα είναι σε πλήθος $k_1k_2...k_m$.

Ο πίνακας δεδομένων έχει $k_1k_2...k_m$ γραμμές και $k_1+k_2+..+k_m$ στήλες. Το άθροισμα των στοιχείων κάθε γραμμής είναι m, ενώ το άθροισμα μιας στήλης για την j μεταβλητή είναι $k_1k_2...k_{j-1}k_{j+1}...k_m$.

Έστω $a_{i1i2...ij...im}$, το αντικείμενο που χαρακτηρίζεται από την i1-οστή κλάση της 1ης μεταβλητής, την i2-οστή της 2ης,..., την im-οστή κλάση της m μεταβλητής.

Οι αποστάσεις δύο αντικειμένων $a_{i1i2...ij...im}$, $a_{i1i2...ij'...im}$, που διαφέρουν μόνο ως προς μια κλάση της j μεταβλητής, είναι:

$$d_e(a_{i1i2...ij...im}, a_{i1i2...ij'...im}) = \sqrt{1^2+1^2} = \sqrt{2}$$

$$d_{x^2}(a_{i1i2...ij...im}, a_{i1i2...ij'...im}) = \sqrt{\frac{2}{k_1k_2..k_{j-1}k_{j+1}..k_m}\left(\frac{1}{m}\right)^2} = d_{x^2}(m, k_1, k_2..k_{j-1}, k_{j+1}..k_m)$$

$$d_J(a_{i1i2\ldots ij..im}, a_{i1i2\ldots ij'..im}) = 1 - \frac{m-1}{m+m-(m-1)} = \frac{2}{m+1} = d_J(m)$$

Αν $a_{i1i2\ldots ij\ldots im}$ και $a'_{i1i2\ldots ij\ldots im}$ είναι δύο αντικείμενα που διαφέρουν σε h κλάσεις, h μεταβλητών, τότε η απόστασή τους είναι:

$$d_e(a_{i1i2\ldots ij..im}, a'_{i1i2\ldots ij..im}) = \sqrt{2h} = d_e(h)$$

$$d_{x^2}(a_{i1i2\ldots ij..im}, a'_{i1i2\ldots ij..im}) = \sqrt{2\left(\frac{1}{m}\right)^2 \sum_{j=1}^{h} \frac{2}{k_1 k_2 .. k_{j-1} k_{j+1} .. k_m}} = d_{x^2}(m, h, k_1, k_2, .. k_{j-1}, k_{j+1}, .. k_m)$$

$$d_J(a_{i1i2\ldots ij..im}, a'_{i1i2\ldots ij..im}) = 1 - \frac{m-h}{m+m-(m-h)} = \frac{2}{m+h} = d_J(m, h)$$

Παρατήρηση: Όπως και προηγουμένως, παρατηρούμε ότι στην μετρική του X^2, συντελεί στον υπολογισμό των αποστάσεων και το πλήθος των κλάσεων κάθε μεταβλητής.

Όσο περισσότερες κλάσεις έχει μια μεταβλητή, τόσο μικρότερη είναι η απόσταση δύο σημείων που διαφέρουν ως προς μια κλάση κάποιας άλλης μεταβλητής. Δηλαδή αν μια μεταβλητή την χωρίσουμε σε διαφορετικό αριθμό κλάσεων, η νέα ταξινόμηση θα διαφέρει στους χαμηλούς κόμβους (στους πρωτοσχηματιζόμενους).

Στην περίπτωση που δεν θέλουμε ο αριθμός κλάσεων κάθε μεταβλητής να επηρεάζει την ταξινόμηση, η χρήση της μετρικής X^2, οδηγεί σε εσφαλμένα αποτελέσματα, και ενδείκνυται η χρησιμοποίηση της ευκλείδειας μετρικής ή του δείκτη Jaccard.

Όταν όμως έχει σημασία για την ταξινόμηση και το πλήθος κλάσεων των μεταβλητών, πρέπει να χρησιμοποιείται η μετρική του X^2.

ε. Πίνακας m μεταβλητών, με επαναλήψεις.

Αν τα αντικείμενα που χαρακτηρίζονται από m μεταβλητές επαναλαμβάνονται, η απόστασή τους με την ευκλείδεια μετρική ή με τον δείκτη Jaccard δεν αλλάζει, αλλά ο αριθμός επαναλήψεων επιδρά στον υπολογισμό των αποστάσεων τους με τη μετρική X^2.

Οι αποστάσεις δύο αντικειμένων που διαφέρουν σε h κλάσεις h μεταβλητών στην περίπτωση αυτή, δίνεται από τους τύπους:

$$d_e(a_{ili2..ij..im}, a'_{ili2..ij..im}) = \sqrt{2h} = d_e(h)$$

$$d_{x2}(a_{i1i2...ij..im}, a'_{i1i2...ij..im}) =$$

$$= \sqrt{2\left(\frac{1}{m}\right)^2 \sum_{j=1}^{h} \frac{2}{g_j k_1 k_2 .. k_{j-1} k_{j+1} .. k_m}} = d_{x2}(m,h,g,k_1,k_2..k_{j-1},k_{j+1},..k_m)$$

όπου g_j ο αριθμός επαναλήψεων της j γραμμής.

$$d_J(a_{ili2..ij..im}, a'_{ili2..ij..im}) = 1 - \frac{m-h}{m+m-(m-h)} = \frac{2}{m+h} = d_J(m,h)$$

Στην περίπτωση που δεν θέλουμε ο αριθμός των κλάσεων των μεταβλητών ή η επανάληψη ενός αντικειμένου, να συντελεί στην ταξινόμηση του, δεν πρέπει να χρησιμοποιούμε την μετρική του X^2 (που οδηγεί σε εσφαλμένη ιεράρχηση) αλλά την ευκλείδεια μετρική ή τον δείκτη Jaccard.

Όταν όμως έχει σημασία για την ταξινόμηση η επανάληψη κάποιου αντικειμένου ή το πλήθος κλάσεων των μεταβλητών, πρέπει να χρησιμοποιείται η μετρική του X^2.

1.1.3. Σύγκριση ευκλείδειας -Jaccard

Έστω ότι έχουμε πίνακα δεδομένων, όπου τα αντικείμενα χαρακτηρίζονται από ποιοτικές ή και ποσοτικές μεταβλητές χωρισμένες σε κλάσεις.

Όταν κάθε αντικείμενο ανήκει σε μία και μόνο κλάση κάθε μεταβλητής, η διάταξη των αντικειμένων είναι ίδια είτε χρησιμοποιούμε την ευκλείδεια μετρική είτε τον δείκτη Jaccard και οδηγούμαστε σε όμοια ταξινόμηση κατ' αύξουσα ιεραρχία (C.A.H.).

Περίπτωση που τα αντικείμενα μπορούν να ανήκουν σε περισσότερες από μία κλάσεις της ίδιας μεταβλητής.

Στις ειδικές περιπτώσεις που κάθε αντικείμενο είναι δυνατό να ανήκει σε περισσότερες από μία κλάσεις ή και σε καμία κλάση μιας μεταβλητής (π.χ. για την μεταβλητή ξένη γλώσσα, ένα άτομο μπορεί να μην γνωρίζει καμία ή να γνωρίζει μία και περισσότερες ξένες γλώσσες), η ευκλείδεια απόσταση μεταξύ δύο

αντικειμένων που έχουν σε μια στήλη 1, είναι ίδια με την ευκλείδεια απόσταση αν έχουν στην στήλη αυτή 0. Με τον δείκτη Jaccard όμως, που λαμβάνει υπόψη τις κοινές εμφανίσεις μιας κλάσης (κοινά 1), οι δύο αποστάσεις είναι διαφορετικές.

Έστω ο πίνακας δεδομένων A. Οι αποστάσεις μεταξύ δύο αντικειμένων a_i και a_i' όταν c είναι το πλήθος των κοινών 1 στις δύο γραμμές i και i', είναι:

$$d_e(a_i,a_{i'}) = \sqrt{[A(i,.)+A(i',.)-2c]^2} \qquad d_J(a_i,a_{i'}) = 1-\frac{c}{A(i,.)+A(i',.)-c} = \frac{A(i,.)+A(i',.)-2c}{A(i,.)+A(i',.)-c}$$

Όταν λοιπόν θέλουμε η παρουσία κάποιας κλάσης (1) να χαρακτηρίζει ένα αντικείμενο περισσότερο από την απουσία της (0), στον υπολογισμό της απόστασης, πρέπει να χρησιμοποιούμε τον δείκτη Jaccard και όχι την ευκλείδεια μετρική που εξισώνει τις κοινές παρουσίες και απουσίες.

1.1.4. Συμπεράσματα

Όταν η επανάληψη ενός αντικειμένου παίζει ή επιθυμούμε να παίζει σημαντικό ρόλο στην ταξινόμησή του, πρέπει να χρησιμοποιούμε την μετρική του X^2. Με την μετρική αυτή, αντικείμενα που επαναλαμβάνονται πολλές φορές, ταξινομούνται γρηγορότερα, και στα πρώτα επίπεδα στο δενδρόγραμμα ιεράρχησης.

Όταν δεν θέλουμε να λάβουμε υπ' όψιν την επανάληψη κάποιου αντικειμένου ή τον αριθμό κλάσεων που έχει κάθε μεταβλητή, πρέπει να χρησιμοποιήσουμε σαν μετρική απόστασης, είτε την ευκλείδεια μετρική, είτε τον δείκτη απόστασης του Jaccard, και όχι τη μετρική του X^2, γιατί όταν οι μεταβλητές έχουν διαφορετικό πλήθος κλάσεων, στον υπολογισμό των αποστάσεων με τη χρήση της μετρικής X^2, επομένως και στη δημιουργία της ιεράρχησης, επιδρά το πλήθος κλάσεων.

Η επιλογή μεταξύ ευκλείδειας απόστασης και δείκτη Jaccard, εξαρτάται από τον αριθμό των κλάσεων στις οποίες μπορεί να ανήκει κάθε αντικείμενο. Αν οποιοδήποτε αντικείμενο, ανήκει σε μόνο μία κλάση κάθε μεταβλητής, τότε η ιεράρχηση που παίρνουμε με τον δείκτη Jaccard, είναι όμοια μ' αυτή που λαμβάνουμε χρησιμοποιώντας την ευκλείδεια απόσταση. Όταν όμως ένα αντικείμενο μπορεί να χαρακτηρίζεται από περισσότερες κλάσεις της ίδιας μεταβλητής, και η παρουσία μιας κλάσης είναι σημαντικότερη (στον χαρακτηρισμό του αντικειμένου) απ' ότι η απουσία της, η χρήση της ευκλείδειας απόστασης οδηγεί σε λάθη, αφού θεωρεί τις κοινές παρουσίες και απουσίες, ισοδύναμες. Στην περίπτωση αυτή πρέπει να χρησιμοποιούμε τον δείκτη του Jaccard.

1.2 Αντικείμενα που χαρακτηρίζονται από ποσοτικές μεταβλητές

Η μέτρηση απόστασης για αντικείμενα που χαρακτηρίζονται από ποσοτικές μεταβλητές, είναι εύκολη αφού οι τιμές των μεταβλητών αυτών είναι πραγματικοί αριθμοί. Η πιο διαδεδομένη απόσταση δύο αντικειμένων i και i′ είναι:

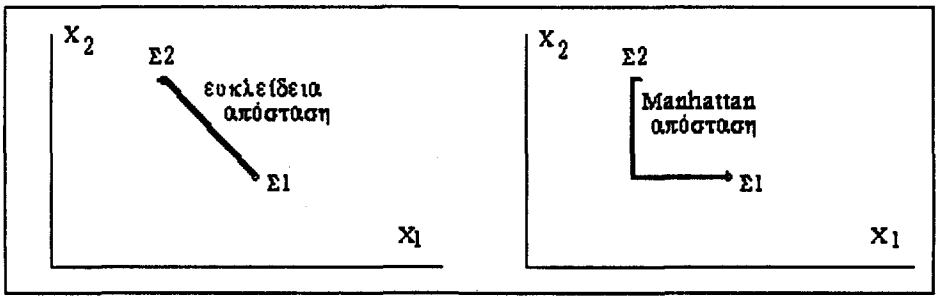
η **ευκλείδεια μετρική**: $d_e^2(i,i') = \sum_{j=1}^{p} [A(i,j) - A(i',j)]^2$

που αντιστοιχεί στην πραγματική ευκλείδεια απόσταση μεταξύ των δύο σημείων με συντεταγμένες τις τιμές των p μεταβλητών. Για να αποφύγουμε την επίδραση των μονάδων μέτρησης, κανονικοποιούμε τα δεδομένα.

Μερικές φορές μπορούμε να χρησιμοποιήσουμε την ευκλείδεια απόσταση δίνοντας και κάποια βάρη στις μεταβλητές, πράγμα που ισοδυναμεί με κανονικοποίηση των δεδομένων ώστε να απαλλαγούμε από την επίδραση των μονάδων μέτρησης. Επίσης δίνοντας βάρος 0 σε μια μεταβλητή που δεν παρέχει καμιά πληροφορία για τον υπολογισμό της απόστασης, απαλλαγόμαστε εύκολα απ' αυτή. [Kaufman,1990]

Μια άλλη γνωστή μετρική είναι η απόσταση Manhattan ή city block ή L1 που ορίζεται από τον τύπο: [Kaufman,1990]

$d(i,j)=|A(i,1)-A(j,1)|+...+|A(i,p)-A(j,p)|$



σχήμα 2. Ευκλείδεια και Manhattan απόσταση

Στο σχήμα 2 παρουσιάζουμε την ευκλείδεια και **Manhattan απόσταση** για δύο άτομα Σ1 και Σ2, που χαρακτηρίζονται από 2 ποσοτικές μεταβλητές. Η δεύτερη απόσταση ονομάζεται και city block για τον εξής λόγο: Υποθέτουμε ότι ζούμε σε μια πόλη όπου όλοι οι δρόμοι είναι από ανατολικά προς τα δυτικά ή από βορρά προς νότο και είναι κάθετοι μεταξύ τους. Έτσι αν θέλουμε να πάμε με αυτοκίνητο από την θέση i στη j, η μικρότερη διαδρομή που μπορούμε να διανύσουμε είναι $|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|$. Μόνο ένα πουλί θα μπορούσε να πάει κατευθείαν διανύοντας την απόσταση που αντιστοιχεί στην ευκλείδεια μετρική. Η Manhattan απόσταση

χρησιμοποιείται στις καταστάσεις εκείνες όπου για παράδειγμα μια διαφορά 1 στην πρώτη μεταβλητή και 3 στην δεύτερη είναι όμοια με μια διαφορά 2 στην πρώτη και 2 στην δεύτερη μεταβλητή.

Μια γενίκευση των παραπάνω δύο αποστάσεων είναι η απόσταση **Minkowski** που δίνεται από τον τύπο:

$d(i,j)=(|A(i,1)-A(j,1)|^q +... +|A(i,p)-A(j,p)|^q)^{1/q}$

όπου q πραγματικός αριθμός μεγαλύτερος ή ίσος του 1.

Καθώς $q\to\infty$, η απόσταση του Minkowski, γίνεται $d(i,j)=\max_k |x_{ik}-x_{jk}|$ που είναι η **απόσταση Chebyshev**.

Όταν τα δεδομένα που έχουμε αποτελούνται από πίνακα V διασποράς-συνδιασποράς μεταξύ των αντικειμένων, χρησιμοποιείται η **απόσταση Mahalanobis**, η οποία για δύο αντικείμενα i και j, δίνεται από τον τύπο:

$d^2(i,j)={}^t(x_i-x_j)V^{-1}(x_i-x_j)$

όπου x_i, x_j τα διανύσματα συντεταγμένων των αντικειμένων i και j αντίστοιχα.

Η απόσταση αυτή χρησιμοποιείται κυρίως στην ανάλυση discriminante. [Ngouenet,1992][Marchetti,1989][Batagelg,1990]

1.3 Απόσταση για αντικείμενα με διαφορετικού τύπου μεταβλητές

Όταν οι μεταβλητές που χαρακτηρίζουν τα αντικείμενα δεν είναι όλες του ίδιου τύπου, μπορούμε να ακολουθήσουμε δύο προσεγγίσεις.

1. Μετατρέπουμε όλες τις μεταβλητές σε ένα κοινό τύπο, ώστε να μπορούμε να χρησιμοποιήσουμε μια μετρική απόστασης. Έτσι οι ποσοτικές μεταβλητές μετατρέπονται σε ποιοτικές ή δυαδικές μεταβλητές, χωρίζοντας τες σε κλάσεις.
2. Χρησιμοποιούμε διαφορετικές μετρικές απόστασης για κάθε τύπο μεταβλητών και κατόπιν συνδέουμε τις μετρικές αυτές δίνοντάς την κατάλληλη βαρύτητα. Έτσι για δύο αντικείμενα i και j, που χαρακτηρίζονται από ποσοτικές, ποιοτικές και δυαδικές μεταβλητές θα έχουμε τον υπολογισμό της μεταξύ τους απόστασης d(i,j), με τον παρακάτω τύπο:

$$d(i,j)=\frac{w_1 d_1(i,j)+w_2 d_2(i,j)+w_3 d_3(i,j)}{w_1+w_2+w_3}$$

όπου $d_1(i,j)$, $d_2(i,j)$, $d_3(i,j)$, μετρικές απόστασης για τις ποσοτικές, ποιοτικές και δυαδικές μεταβλητές αντίστοιχα, ενώ τα βάρη w_1, w_2, w_3, επιλέγονται κατάλληλα, και ανάλογα με την σπουδαιότητα που θέλουμε να δώσουμε σε κάθε τύπο μεταβλητών. [Kaufman,1990].

2. Αποστάσεις για μεταβλητές

2.1 Εισαγωγή

Λέγοντας ταξινόμηση μεταβλητών, εννοούμε την εφαρμογή των μεθόδων ταξινόμησης (classification) που χρησιμοποιούνται για την ιεράρχηση αντικειμένων, στις μεταβλητές. Χωρίζονται τότε οι μεταβλητές σε ομοιογενείς ομάδες και οι ομάδες με την σειρά τους σε νέες ομοιογενείς ομάδες σχηματίζοντας έτσι ένα δένδρο ταξινόμησης μεταβλητών.

Η ταξινόμηση των μεταβλητών είναι χρήσιμη, όταν θέλουμε να προσδιορίσουμε σχέσεις μεταξύ των μεταβλητών ή ομάδες μεταβλητών για να προχωρήσουμε στη συνέχεια σε μια άλλη ανάλυση (παλινδρόμηση, παραγοντική ανάλυση κτλ)

Βέβαια μερικές φορές η ταξινόμηση των μεταβλητών δεν έχει νόημα, όταν αυτές είναι τελείως ανομοιογενείς (π.χ. όταν ομαδοποιούνται μαζί το φύλο και η ηλικία, δεν μπορούμε να βγάλουμε κάποιο συμπέρασμα). Αλλά έχει μεγάλη σημασία όταν αναζητούμε μεταβλητές που η συμπεριφορά της μιας επηρεάζει τις άλλες.

Για να γίνει η ταξινόμηση των μεταβλητών χρειάζεται καταρχήν να μπορούμε να βρούμε πότε δύο μεταβλητές είναι "κοντά", δηλαδή να ορίσουμε ένα μέτρο απόστασης μεταξύ τους. Αυτό μπορεί να γίνει είτε χρησιμοποιώντας τις γνωστές αποστάσεις, που χρησιμοποιούνται και για την ταξινόμηση αντικειμένων, είτε κάποιον συντελεστή συσχέτισης.

Θεωρούμε γενικά μια μεταβλητή από το σύνολο V όλων των μεταβλητών, σαν διάνυσμα με n συντεταγμένες (n είναι το πλήθος των παρατηρήσεων). Ορίζουμε μια αριθμητική συνάρτηση από το VxV στο R, όπου σε κάθε ζεύγος μεταβλητών (x,y) αντιστοιχεί ένας αριθμός που μετράει την "ομοιότητα" μεταξύ x και y. Η "ομοιότητα" αυτή έχει διαφορετική έννοια ανάλογα με το είδος των μεταβλητών. Διακρίνουμε γενικά δύο μεγάλες κατηγορίες. Ποσοτικές και ποιοτικές μεταβλητές.

2.2 Ποσοτικές μεταβλητές

Είναι οι μεταβλητές εκείνες που σε κάθε παρατήρηση αντιστοιχεί ένας πραγματικός αριθμός. π.χ. βάρος, εισόδημα κτλ. Μέτρα ομοιότητας για ποσοτικές μεταβλητές είναι:

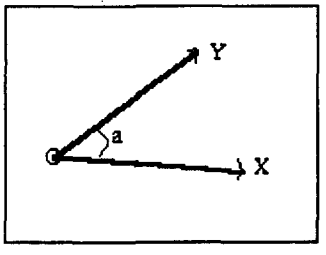
1. Γωνία των δύο διανυσμάτων που αντιστοιχούν στις μεταβλητές

Το εσωτερικό γινόμενο δύο διανυσμάτων δίνεται από τη σχέση:

$\langle XY\rangle=\Sigma x_i y_i=|X||Y|\cos a$

οπότε: $\cos a = \dfrac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$

Σαν μέτρο για την ομοιότητα των διανυσμάτων αυτών, χρησιμοποιούμε το συνημίτονο της γωνίας a που σχηματίζουν (σχήμα 3). Όσο τείνουν να συμπέσουν τα δύο διανύσματα, τόσο το συνημίτονο είναι μεγαλύτερο.



σχήμα 3. Γωνία a μεταξύ δύο διανυσμάτων

Σημαντική ιδιότητα του συνημιτόνου, είναι ότι δεν εξαρτάται από το μέτρο των δύο διανυσμάτων. (Δεν αλλάζει αν οι μεταβλητές πολλαπλασιαστούν με κάποιον αριθμό).

2. Συντελεστής συσχέτισης

Όταν η μια μεταβλητή φαίνεται να επιδρά στην άλλη (π.χ. απασχόληση και εισόδημα), σαν μέτρο απόστασης των δύο μεταβλητών μπορούμε να λάβουμε τον συντελεστή συσχέτισής τους.

Ο συντελεστής συσχέτισης δύο ποσοτικών μεταβλητών, δίνεται από τη σχέση:

$$r(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i x_i' y_i'}{\sqrt{\sum_i x_i'^2 \sum_i y_i'^2}} = \cos a'$$

όπου $x_i' = x_i - \bar{x}$ και $y_i' = y_i - \bar{y}$

Παρατηρούμε ότι είναι ίσος με το συνημίτονο της γωνίας μεταξύ των κεντραρισμένων διανυσμάτων των X και Y. (Δηλαδή των αντίστοιχων διανυσμάτων που έχει αφαιρεθεί η μέση τιμή). Η τιμή του συντελεστή συσχέτισης δεν αλλάζει σε οποιοδήποτε γραμμικό μετασχηματισμό των διανυσμάτων. (Δεν αλλάζει αν αντί για X, θέσουμε αX+β). Μειώνεται όμως έτσι σε κάποιες

περιπτώσεις, η δύναμη διακριτοποίησης των μεταβλητών, αφού όλες οι μεταβλητές κX+λ θεωρούνται ίδιες.

Η βασική διαφορά μεταξύ συνημιτόνου και συντελεστή συσχέτισης για την μέτρηση της απόστασης είναι ότι το συνημίτονο βασίζεται στις αρχικές μετρήσεις, (απόκλιση από την αρχή των αξόνων), ενώ ο συντελεστής συσχέτισης σε κεντραρισμένες τιμές (απόκλιση από το μέσο όρο). Αν η αρχή των αξόνων έχει έννοια και είναι καλά θεμελιωμένη, τότε οι αρχικές τιμές έχουν σημασία με απόλυτη έννοια και το συνημίτονο είναι κατάλληλο μέτρο απόστασης. Αν η αρχή είναι τυχαία επιλεγμένη ή για ευκολία στους υπολογισμούς μας, τότε οι αρχικές τιμές έχουν έννοια σε σχέση η μία με την άλλη και όχι σε σχέση με την αρχή. Η διάκριση λοιπόν είναι όμοια με την διάκριση μεταξύ λόγου και διαστήματος. Έτσι το συνημίτονο χρησιμοποιεί κλίμακα λόγου ενώ ο συντελεστής συσχέτισης χρησιμοποιεί μόνο κλίμακα διαστήματος. [Anderberg,1973]

3. Ευκλείδεια απόσταση

Μπορούμε να ομογενοποιήσουμε τις μεταβλητές, (αφαιρούμε την μέση τιμή και διαιρούμε με την τυπική τους απόκλιση) και στη συνέχεια να υπολογίσουμε την ευκλείδεια απόσταση, όπως κάνουμε και για την ταξινόμηση των αντικειμένων. Αυτό γίνεται όταν οι μεταβλητές είναι παρόμοιες π.χ. απασχόληση στην γεωργία και απασχόληση στη βιομηχανία.

Η αντίστοιχη σχέση είναι: $d_e^2(X,Y) = \sum_{i=1}^{n} (u_i - v_i)^2$

όπου u_i, v_i είναι οι συντεταγμένες των δύο μεταβλητών μετά την ομογενοποίηση.

4. X^2 απόσταση

Μερικές φορές οι ποσοτικές μεταβλητές που μελετάμε δεν αφορούν παρά συχνότητα εμφάνισης κάποιας ποιοτικής μεταβλητής. Για παράδειγμα μπορεί να μας ενδιαφέρει το πλήθος των ατόμων που ασχολούνται με κάποια είδη εργασίας σε κάθε νομό της Ελλάδας. Τότε αν και οι μεταβλητές (πλήθος εργατών, πλήθος βιοτεχνών, πλήθος υπαλλήλων) είναι ποσοτικές, εκφράζουν συχνότητες εμφάνισης κάποιων ποιοτικών μεταβλητών (είδος εργασίας). Τέτοιου είδους δεδομένα περιγράφονται με πίνακες συχνοτήτων. Για το παράδειγμα μας θα έχουμε τον πίνακα με γραμμές τους νομούς της Ελλάδας και με στήλες τις αντίστοιχες μεταβλητές απασχόλησης που εκφράζουν συχνότητα.

Όταν τα δεδομένα μας είναι σε μορφή πίνακα συχνοτήτων B, χρησιμοποιούμε την απόσταση X^2 [Benzecri,1973], η οποία για την μέτρηση της

απόστασης μεταξύ δύο κατηγοριών x και y (στήλες του πίνακα δεδομένων) δίνεται από τη σχέση:

$$d_x^2(x,y) = \sum_{i=1}^{n} \frac{1}{B(i,.)} \left(\frac{B(i,x)}{B(.,x)} - \frac{B(i,y)}{B(.,y)} \right)^2$$

όπου:

B(i,j) η συχνότητα της i γραμμής και j στήλης

B(i,.) το άθροισμα της i γραμμής

B(.,j) το άθροισμα της j στήλης

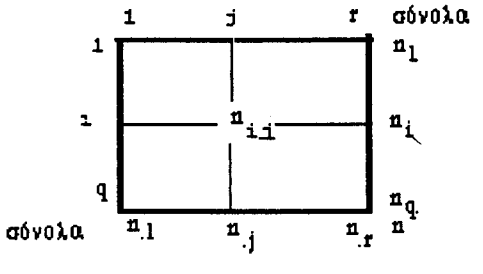
n το πλήθος των αντικειμένων (γραμμών του πίνακα)

Η απόσταση αυτή, ικανοποιεί την αρχή της "κατανεμημένης ισοδυναμίας". Δηλαδή αν ομαδοποιήσουμε δύο ή περισσότερες ανάλογες γραμμές, δεν τροποποιούνται οι αποστάσεις μεταξύ στηλών. Το γεγονός αυτό εξασφαλίζει σταθερότητα στα αποτελέσματα.

2.3 Ποιοτικές μεταβλητές

Ποιοτικές είναι οι μεταβλητές των οποίων τις τιμές δεν μπορούμε να μετρήσουμε με έναν αριθμό. Στις ποιοτικές μεταβλητές τα αντικείμενα δεν παίρνουν αριθμητική τιμή για κάθε μεταβλητή, αλλά χαρακτηρίζονται από κάποια κατηγορία της. Για παράδειγμα όταν μελετάμε τη μεταβλητή φύλο, τα αντικείμενα θα είναι άνδρας ή γυναίκα.

Σε ποιοτικές μεταβλητές μετράμε πόσες φορές παρουσιάστηκε μια κατηγορία κάποιας μεταβλητής και τον αριθμό αυτό ονομάζουμε συχνότητα αυτής της κατηγορίας. Με τις συχνότητες όλων των κατηγοριών για όλες τις μεταβλητές, σχηματίζουμε ένα πίνακα, τον οποίο ονομάζουμε πίνακα συμπτώσεων, όπως αναφέραμε στο κεφάλαιο 2. Ένας τρόπος παρουσίασης ποιοτικών μεταβλητών, είναι με ένα πίνακα συμπτώσεων που παρουσιάζει την κατανομή των παρατηρήσεων μεταξύ δύο μεταβλητών χωρισμένων σε κατηγορίες (κλάσεις). Παράδειγμα τέτοιου πίνακα είναι ο πίνακας στο σχήμα 4.



σχήμα 4. Πίνακας συμπτώσεων για δύο μεταβλητές

Έστω ότι έχουμε δύο ποιοτικές μεταβλητές και των πίνακα συμπτώσεων τους. Η πρώτη έχει q κατηγορίες και η δεύτερη r.

Παρακάτω δίνονται μερικά παλαιότερα μέτρα προσέγγισης δύο μεταβλητών, βασισμένα στο στατιστικό x^2 που χρησιμοποιείται στον έλεγχο ανεξαρτησίας δύο μεταβλητών. Μετράμε την σχέση των 2 μεταβλητών, ποσοκοποιώντας την απομάκρυνση από τον θεωρητικό πίνακα συμπτώσεων. (Δηλαδή από τον πίνακα που προκύπτει υποθέτοντας ανεξαρτησία των 2 μεταβλητών). Σαν απόσταση των δύο μεταβλητών μπορούμε να χρησιμοποιήσουμε:

1) συντελεστή $x^2 = n\sum_{i=1}^{q}\sum_{j=1}^{r}\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2 \frac{1}{n_{i.}n_{.j}}$

Η τιμή του μέτρου αυτού αυξάνει όσο αυξάνει το n. [Pearson,1904]

Αν θέλουμε η απόσταση να κυμαίνεται μεταξύ 0 και 1 τότε χρησιμοποιούμε:

2) T του Tschuprow

$$T=\frac{\phi^2}{\sqrt{(r-1)(q-1)}} \qquad \text{όπου } \varphi^2=x^2/n$$

3) C του Cramer που τροποποίησε το T, θέτοντας:

$$C=\sqrt{\frac{\phi^2}{\min\{(r-1),(q-1)\}}} \quad \text{όπου } \varphi^2=x^2/n$$

[Παπαϊωάννου,1994] [Goodman,1954]

4) Pearson πρότεινε ένα μέτρο γνωστό σαν συντελεστής συμπτώσεων βασισμένο στο φ^2

$$P=\sqrt{\frac{\phi^2}{(1+\phi^2)}}=\sqrt{\frac{x^2}{(n+x^2)}}$$

[Παπαϊωάννου,1994][Anderberg,1973]

Τα μέτρα x^2 και όσα παράγονται απ' αυτά είναι χρήσιμα στον έλεγχο υποθέσεων αλλά όχι σαν μέτρο συσχετισμού. Είναι δύσκολο να συγκρίνουμε τιμές μεταξύ δύο πινάκων διαφορετικών μεταβλητών. Δεν μπορούμε λοιπόν να χρησιμοποιήσουμε αυτά στην ταξινόμηση μεταβλητών.

Έτσι προτάθηκαν νεώτερα μέτρα για πίνακες συμπτώσεων. Αναφέρουμε:

5) Συντελεστής συσχέτισης για πίνακα συμπτώσεων [Tallur,1986]

Η προσέγγιση δύο μεταβλητών j και h (ή δύο κατηγοριών μεταβλητών), μετριέται με τον συντελεστή συσχέτισης τους, όπως προσδιορίζεται από τον πίνακα συμπτώσεων.

Ο τύπος είναι: $$s(j,h) = \frac{\sum_i \left(\frac{f_{ij} f_{ih}}{p_{i.}} \right) - p_{.j} p_{.h}}{\sqrt{\left[\sum_i \left(\frac{f_{ij}^2}{p_{i.}} \right) - p_{.j}^2 \right] \left[\sum_i \left(\frac{f_{ih}^2}{p_{i.}} \right) - p_{.h}^2 \right]}}$$

$f_{ij}=n_{ij}/n \quad p_{i.}=n_{i.}/n \quad p_{.j}=n_{.j}/n$

6) Γενικευμένος συντελεστής συσχέτισης για τις μεταβλητές X και Y, που είναι χωρισμένες σε κατηγορίες. Προτάθηκε από τον Guy Der Megreditchian [Megreditchian,1988][Pottier,1994]. Τα δεδομένα περιγράφονται από πίνακα συμπτώσεων.

$$Q[X,Y]=\sqrt{\det(K_{X|Y}\ K_{Y|X})}$$

όπου $K_{X|Y}$ ο πίνακας με στοιχείο στην i γραμμή και j στήλη το $n_{ij}/n_{.j}$, και

$K_{Y|X}$ ο πίνακας με στοιχείο στην i γραμμή και j στήλη το $n_{ij}/n_{i.}$

Ο συντελεστής αυτός είναι εύκολος στον υπολογισμό του και η συμπεριφορά του τείνει προς την συμπεριφορά του συντελεστή x^2 του Pearson. Έτσι, όταν οι δύο μεταβλητές είναι ανεξάρτητες Q[X,Y]=0, ενώ όταν η μεταβλητή X εξαρτάται από την Y, $Q^2[X,Y]=1$

7) Δείκτης πιθανότητας σχέσης.

Πρόκειται για ένα δείκτη που προτάθηκε από τον Lerman [Lerman,1981] και βασίζεται στην χρήση της θεωρίας πιθανοτήτων για την πιθανότητα σχέσης μεταξύ δύο μεταβλητών. Ο υπολογισμός του δείκτη αυτού είναι αρκετά πολύπλοκος και δεν θα ασχοληθούμε περισσότερο.

Πίνακες συμπτώσεων πολλών μεταβλητών (Burt)

Όταν ο πίνακας συμπτώσεων δεν αφορά μόνο δύο μεταβλητές αλλά περισσότερες, λέμε ότι έχουμε ένα πίνακα συμπτώσεων τύπου Burt (κεφάλαιο 2). Από τον υποπίνακα για δύο μεταβλητές του πίνακα Burt, βρίσκουμε την απόσταση αυτών των μεταβλητών.

Αν θέλουμε την εύρεση σχέσης μεταξύ ενός ζεύγους (v1,v2) και μιας τρίτης μεταβλητής v3 χρησιμοποιώ τον υποπίνακα v3x(v1,v2).

Οι διαγώνιοι υποπίνακες που αφορούν μόνο μία μεταβλητή, δεν χρησιμοποιούνται.

Μπορούμε ακόμη να μελετήσουμε τις κατηγορίες ποιοτικών μεταβλητών σαν ξεχωριστές μεταβλητές, οπότε έχουμε και περισσότερο ερμηνεύσιμα συμπεράσματα αφού διακρίνονται οι σχέσεις μεταξύ κατηγοριών.

2.4 Δυαδικές μεταβλητές

Δυαδικές ονομάζονται οι ποιοτικές μεταβλητές που χαρακτηρίζουν τα αντικείμενα μόνο με την παρουσία ή απουσία τους. Έτσι έχουν μόνο δύο τιμές που αντιστοιχούν σε 0 (απουσία) ή 1 (παρουσία).

Ο πίνακας συμπτώσεων δύο δυαδικών μεταβλητών X και Y έχει την παρακάτω μορφή:

	1	0
1	a	b
0	c	d

Τα προηγούμενα μέτρα μπορούν να ισχύουν και για δυαδικές μεταβλητές.

1. Γωνία δύο διανυσμάτων

Από τον πίνακα συμπτώσεων, προκύπτουν εύκολα οι σχέσεις:

$\Sigma x_i y_i = a$ $\quad \Sigma x_i^2 = a+b$ $\quad \Sigma y_i^2 = a+c$. Τότε η γωνία των X και Y είναι:

$$\cos(X,Y) = \frac{a}{\sqrt{(a+b)(a+c)}} = \sqrt{\frac{a}{a+b}\frac{a}{a+c}}$$

Το κυριότερο μειονέκτημα του μέτρου αυτού είναι το ότι εξαρτάται από το ποια από τις δύο κλάσεις (παρουσία ή απουσία) θα θεωρείται με τιμή 0 και ποια με τιμή 1.

2. Συντελεστής συσχέτισης

$$r(X,Y) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

και παρατηρούμε ότι ισχύει $r^2 = \varphi^2 = x^2/n = T^2 = C^2$

Η διόρθωση του Yate στο x^2 δίνει

$$X_c^2 = \frac{n(|ad - bc| - 0.5n)^2}{(a+b)(c+d)(a+c)(b+d)}$$

3. Μέτρα μόνο για δυαδικές μεταβλητές.

Κατά καιρούς προτάθηκαν διάφοροι δείκτες ομοιότητας ή απόστασης, που ισχύουν ειδικά και μόνο για δυαδικές μεταβλητές. [Anderberg,1973]. Τέτοιοι δείκτες είναι:

Russell-Rao $d(X,Y) = \frac{c}{a+b+c+d}$

Jaccard: $d_j(X,Y) = 1 - \frac{c}{2a+b}$

Dice $d(X,Y) = \frac{2a}{2a+b+c}$

Ο δείκτης των Russell-Rao, λαμβάνει υπόψιν του στον παρανομαστή, τις κοινές απουσίες (0-0) των δύο μεταβλητών, κάτι που δεν συμβαίνει στους άλλους δείκτες απόστασης. Η διαφορά του δείκτη Dice από Jaccard, είναι ότι ο πρώτος δίνει διπλή βαρύτητα στις κοινές παρουσίες (1-1).

2.5 Διαφορετικού τύπου μεταβλητές

Όταν οι μεταβλητές δεν είναι όλες του ίδιου τύπου, μετατρέπουμε κάποιες απ' αυτές σε κατάλληλο τύπο, ώστε να μπορούμε να υπολογίσουμε αποστάσεις μεταξύ τους.

Δηλαδή αν έχουμε ποσοτικές και ποιοτικές μεταβλητές, μετατρέπουμε τις ποσοτικές σε ποιοτικές, χωρίζοντάς τες σε κλάσεις. Κατόπιν χρησιμοποιούμε μέτρα μεταξύ ποιοτικών μεταβλητών.

2.6 Πορεία μετά τον υπολογισμό των αποστάσεων (τρόπος υπολογισμού αποστάσεων μεταξύ ομάδων)

Αφού υπολογίσουμε τις αποστάσεις ανά δύο όλων των μεταβλητών, προχωρούμε στην συνένωση ομάδων, ομαδοποιώντας τις μεταβλητές που είναι "κοντύτερα". Υπάρχουν διάφορες μέθοδοι για την συνένωση ομάδων. Οι κυριότερες είναι:

MIN

MAX

MOYENNE

Κριτήριο ελάχιστης αύξησης της αδράνειας (κριτήριο Ward) [Benzecri,1980]

Κριτήριο Denimal [Denimal,1994]

Οι παραπάνω μέθοδοι απαιτούν τον υπολογισμό αποστάσεων, όχι πια μεταξύ μεταβλητών, αλλά μεταξύ ομάδων μεταβλητών.

Η απόσταση μεταξύ ομάδων για MIN,MAX,MOYENNE είναι αντίστοιχα η ελάχιστη, μέγιστη και μέση απόσταση, μεταξύ όλων των μεταβλητών που αποτελούν τις δύο ομάδες.

Με τη μέθοδο αύξησης της αδράνειας ενώνονται οι δύο εκείνες ομάδες, που με την συνένωσή τους προκαλούν την μικρότερη αύξηση της αδράνειας μέσα στις κλάσεις. Δηλαδή οι ομάδες εκείνες που ελαχιστοποιούν το

$$\frac{m_s\, m_s'}{(m_s+m_s')} d^2(s,s')$$

όπου s και s' οι δύο κλάσεις που ενώνονται με αντίστοιχα βάρη m_s m_s'.

Όταν στη μέτρηση της απόστασης χρησιμοποιείται συντελεστής συσχέτισης, για τον υπολογισμό αποστάσεων μεταξύ ομάδων, αθροίζουμε τις αντίστοιχες στήλες των μεταβλητών που ενώνονται, σχηματίζοντας έτσι μια νέα μεταβλητή, και στη συνέχεια υπολογίζουμε τον συντελεστή συσχέτισής της με τις υπόλοιπες. [Diday,1982]

Μπορεί επίσης να χρησιμοποιηθεί και η μέθοδος MAX για την εύρεση των δύο ατόμων με την μεγαλύτερη συσχέτιση. [Tenenbaus,1994]

Με την μέθοδο του κριτηρίου Denimal, ξεκινάμε από ένα a priori διαμελισμό P1, και σε κάθε στάδιο λαμβάνουμε το διαμελισμό εκείνο, για τον οποίο το παρακάτω κριτήριο ελαχιστοποιείται.

$$\delta(r,s)= \frac{4a_k}{m^2}\left[\sum_{l=1}^{k-2} a_l\delta_l(r,s) + a_k - 1\ \mathrm{Ward}(r,s)\right]$$

όπου:

r, s οι δύο κλάσεις που ενώνονται

$a_1,a_2, \ldots ,a_k$ τα βάρη που αντιστοιχούν σε μια διαδοχική σειρά διαμελισμών $P_1,P_2, ,P_k$ και έχουν άθροισμα m

$$\delta l(r,s)= \frac{f_r f_s}{f_r+f_s}\left[\left\|(\overline{f_{Jr}}^{l} - f_{Jr}) - (\overline{f_{Js}}^{l} - f_{sr})\right\|^2 - \left\|(\overline{f_{Jr}}^{l} - \overline{f_{Js}}^{l})\right\|^2\right]$$

$$\mathrm{Ward}(r,s)= \frac{f_r f_s}{f_r+f_s}\left\|f_{Jr} - f_{sr})\right\|^2$$

f_r, f_s είναι τα βάρη των κλάσεων r και s αντίστοιχα

f_{Jr}, f_{Js} τα αντίστοιχα profil. ($f_{Ji}=n_{ij}/n_{i.}$)

$\overline{f_{Jr}}^{l}$,$\overline{f_{Js}}^{l}$ παριστάνουν την ανακατασκευή των κλάσεων r,s μέσω του διαμελισμού P_l.

Δηλαδή από την μία ελαχιστοποιούμε το κριτήριο του Ward (μείωση αδράνειας μέσα στις ομάδες), και από την άλλη η ελαχιστοποίηση του υπόλοιπου όρου σημαίνει ότι ο νέος διαμελισμός οφείλει να είναι όσο γίνεται πιο μακριά από την διαμέριση P1, η οποία αντιπροσωπεύει μια ποιοτική μεταβλητή και θέλουμε να ελαχιστοποιήσουμε την επίδρασή της.

2.7 Συμπεράσματα

Το ερώτημα που τίθεται είναι ποια απ' όλες τις παραπάνω αποστάσεις θα χρησιμοποιήσουμε; Η απάντηση εξαρτάται από το είδος των μεταβλητών που θέλουμε να ταξινομήσουμε, από τη φύση των δεδομένων που έχουμε, καθώς και από το τι είδους σχέσεις μεταξύ των μεταβλητών θέλουμε να ερευνήσουμε.

Αν μελετάμε συσχετίσεις μεταβλητών ή την επίδραση μιας μεταβλητής σε κάποια άλλη, πρέπει να χρησιμοποιούμε σαν απόσταση ένα συντελεστή συσχέτισης. Για ποσοτικές μεταβλητές τον κλασσικό συντελεστή συσχέτισης ή το συνημίτονο της γωνίας των διανυσμάτων που σχηματίζουν. Για ποιοτικές μεταβλητές, τον συντελεστή συσχέτισης του Tallur ή τον γενικευμένο συντελεστή συσχέτισης Pottier.

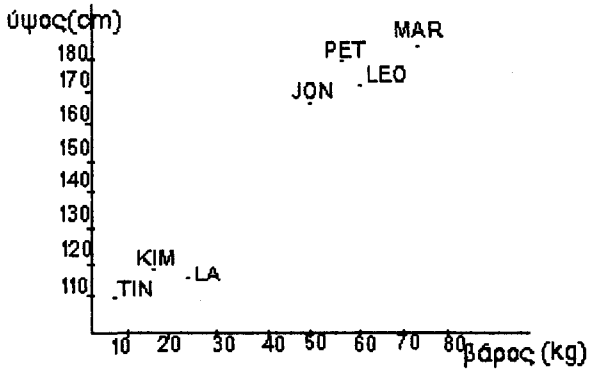
Όταν οι μεταβλητές μετρώνται με ίδιες ή ανάλογες μονάδες μέτρησης και μας ενδιαφέρει πόσο κοντά ή μακρυά βρίσκονται δύο μεταβλητές, μπορούμε να χρησιμοποιήσουμε μια μετρική απόστασης. Την ευκλείδεια μετρική για ποσοτικές μεταβλητές, την μετρική του X^2 όταν έχουμε πίνακα συχνοτήτων και την Jaccard ή άλλη παρόμοια όταν οι μεταβλητές παίρνουν μόνο δύο τιμές.

ΚΕΦΑΛΑΙΟ 4
ΤΑΞΙΝΟΜΗΣΗ ΑΡΧΕΣ-ΜΕΘΟΔΟΙ

1. Εισαγωγή-Βασικές αρχές

Η ταξινόμηση των "ομοίων" αντικειμένων σε ομάδες, είναι μια σημαντική ανθρώπινη διαδικασία. Ένα παιδί μαθαίνει να διακρίνει τραπέζια από καρέκλες, γάτες και σκύλους, άνδρες και γυναίκες κτλ. [Kaufman,1990] Όταν έχουμε μία ή δύο μεταβλητές οι ομάδες διακρίνονται πολύ εύκολα ενώ όταν το πλήθος των μεταβλητών αυξάνει, χρειάζεται να προβούμε σε μια μαθηματική διαδικασία (υπολογισμός αποστάσεων-ομοιοτήτων μεταξύ αντικειμένων και ομάδων).

Κατά την διαδικασία της ταξινόμησης, σκοπός μας είναι να ανακαλύψουμε τις υπάρχουσες ομάδες στα δεδομένα. Βλέποντας στο σχήμα 1, την γραφική αναπαράσταση 7 ατόμων ως προς δύο μεταβλητές (βάρος και ύψος), διακρίνουμε αμέσως 2 ομάδες ατόμων. Η πρώτη περιλαμβάνει τους {TIN,KIM,LA} και η δεύτερη τους {MAR,PET,LEO,JON}. Τα αντικείμενα μέσα στις ομάδες είναι όσο γίνεται παρόμοια, ενώ αυτά που ανήκουν σε διαφορετικές ομάδες διαφέρουν μεταξύ τους, ως προς τις μεταβλητές που τα περιγράφουν.



σχήμα 1. Γραφική αναπαράσταση 7 ατόμων

Η μεγάλη και γρήγορη ανάπτυξη των ηλεκτρονικών εργαλείων υπολογισμού των τελευταίων δεκαετιών, έδωσε μεγάλη ώθηση στην δημιουργία τεχνικών ταξινόμησης. Ωστόσο η ταξινόμηση υπήρχε από πολύ παλιά. Ήδη τον 5ο αιώνα π.Χ. ο Πλάτων θεωρεί την ύπαρξη "όλων" σε ένα ιδανικό κόσμο, τα οποία αναπαριστούν αντικείμενα. Ο μαθητής του Αριστοτέλης, εισάγει την έννοια της "Επιστήμης αντικειμένων" και περιγράφει αντικείμενα με βάσει τις ιδιότητες τους, τονίζοντας την δυικότητα μεταξύ αντικειμένων και των περιγραφών τους.

[Louis,1956] Εξακόσια χρόνια μετά τον Αριστοτέλη, ο Πορφύρος [Bochenski,1970], εισάγει διαδοχικές διαιρέσεις των κλάσεων αντικειμένων χρησιμοποιώντας "γένη" και "είδη". [Diday,1994] Στον 18ο αιώνα ο Carlous Linnaaeus (σουηδός βοτανολόγος) ασχολήθηκε με το πρόβλημα της κατάταξης φυτών και ζώων στο βιβλίο του "Genera Plantarum" (δημοσιεύτηκε το 1737). Το 1756 ο Adanson, εργαζόμενος στην Αφρική έγραφε ότι οι κλάσεις οργανισμών ορίζονται από ένα υψηλό βαθμό ομοιότητας μεταξύ των ατόμων που ανήκουν σ' αυτές. Οι ομοιότητες αυτές υπολογίζονται με βάση τα χαρακτηριστικά των ατόμων. [Adanson,1757].

Έτσι τέθηκαν οι βάσεις για την Αριθμητική Ταξινόμηση. [Sneath,1973] Από τότε εμφανίστηκαν διάφορες μέθοδοι ταξινόμησης. Παρουσιάσεις των μεθόδων έχουν γίνει από τους Good(1965), Ball(1971), Benzecri(1973), Caillez και Pages (1976), Lerman (1970,1981), Diday (1980) κτλ. Μια πλήρη ανασκόπηση των μεθόδων της ταξινόμησης μπορεί κάποιος να βρεί στο άρθρο του Bock [Bock,1994]. Σε πολλές επιστήμες χρησιμοποιείται τόσο η κατάταξη όσο και η ταξινόμηση. Λέγοντας κατάταξη εννοούμε την τοποθέτηση παρατηρήσεων σε κάποιες ομάδες οι οποίες όμως είναι προσδιορισμένες εκ των προτέρων με βάσει κάποια επιθυμητά χαρακτηριστικά τους, ενώ με την ταξινόμηση οι ομάδες δεν προκαθορίζονται αλλά δημιουργούνται από τις ιδιότητες των δεδομένων.

Είναι δύσκολο να διακρίνουμε με σαφή τρόπο μια "μέθοδο" από ένα "αλγόριθμο" ταξινόμησης, αφού η βιβλιογραφία δεν είναι ξεκάθαρη σ'αυτό το θέμα. Στη συνέχεια αναφερόμαστε σε μέθοδο ταξινόμησης και θεωρούμε τον αλγόριθμο για την κατασκευή της καθώς και τον τρόπο υπολογισμού της προσέγγισης μεταξύ δύο ομάδων. Τα χαρακτηριστικά γνωρίσματα για την κατάταξη των μεθόδων είναι:

1. Το αν ταξινομούμε τα αντικείμενα ή τις μεταβλητές που τα χαρακτηρίζουν.
2. Η φορά ταξινόμησης
3. Ο ιεραρχικός ή μη ιεραρχικός χαρακτήρας των αναζητούμενων διαμελισμών σε κλάσεις αντικειμένων.

Τα ερωτήματα όμως που τίθενται είναι:

1. Η χρησιμοποιούμενη μέθοδος δίνει ικανοποιητικά αποτελέσματα; Ποιό κριτήριο είναι εκείνο που θα μας βοηθήσει να αποφασίσουμε;
2. Σε ποιό σημείο της ιεράρχησης πρέπει να σταματήσουμε παίρνοντας τον ιδανικό διαμελισμό των δεδομένων, έτσι ώστε αφενός μεν τα αποτελέσματα να ανταποκρίνονται όσο το δυνατό περισσότερο στην πραγματικότητα, και αφετέρου οι ομάδες του διαμελισμού να μην είναι πάρα πολλές;

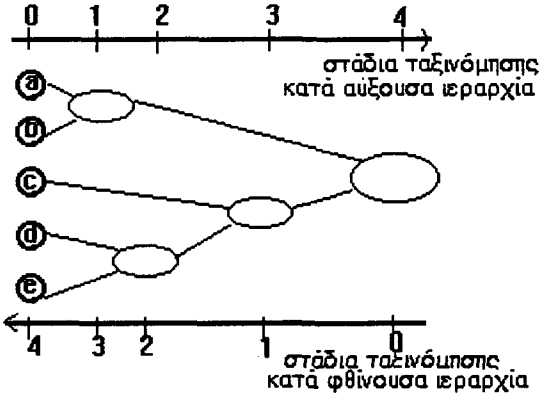
Η καλύτερη μέθοδος, θα ήταν να πάρουμε όλες τις δυνατές ιεραρχήσεις των δεδομένων και να τις αξιολογήσουμε με κάποιο κριτήριο. Δυστυχώς όμως ο αριθμός των ιεραρχήσεων θα ήταν πολύ μεγάλος, ιδιαίτερα όταν αυξάνει το

μέγεθος των παρατηρήσεων. Γι' αυτό καταφεύγουμε σε ευρετικούς τοπικούς αλγορίθμους. Οι "τοπικοί" αλγόριθμοι, εξετάζουν ένα μικρότερο αριθμό ενδεχομένων απ' ότι οι "ολικοί".

Διακρίνουμε 3 βασικές κατηγορίες τέτοιων αλγορίθμων:

1. Ταξινόμηση κατά αύξουσα ιεραρχία (προς τα πάνω). Ενώνουμε πρώτα-πρώτα, τα δύο πιο ομογενή στοιχεία, και στη συνέχεια κάνουμε το ίδιο για τις ομάδες. Όσο προχωρούμε, προκύπτουν διαμελισμοί με μικρότερο πλήθος κλάσεων, μέχρι να φθάσουμε σε μία κλάση που περιέχει όλες τις παρατηρήσεις μας. [Roux, 1985]
2. Ταξινόμηση κατά φθίνουσα ιεραρχία (προς τα κάτω). Ξεκινάμε χωρίζοντας το σύνολο των παρατηρήσεων σε δύο ομάδες, όσο το δυνατό πιο ομογενείς. Στη συνέχεια κάθε μια από τις ομάδες αυτές, τη διασπάμε σε δύο νέες ομάδες. Συνεχίζουμε έτσι, μέχρι όλες οι ομάδες να περιέχουν μόνο μία παρατήρηση. Δηλαδή μέχρι να φθάσουμε στο σύνολο των αρχικών παρατηρήσεων. Η ομάδα που κάθε φορά διαλέγουμε να διασπαστεί, είναι αυτή που έχει μέγιστη αδράνεια (διασπορά). [Roux, 1985] [Benzecri,1973]

Μια γραφική αναπαράσταση των βημάτων ταξινόμησης κατά αύξουσα και φθίνουσα ιεραρχία, παρουσιάζεται στο σχήμα 2.



σχήμα 2. Διαφορά μεταξύ ταξινόμησης κατά αύξουσα και φθίνουσα ιεραρχία

3. Μέθοδοι που περιορίζονται μόνο στη κατασκευή ενός διαμελισμού. Με τις μεθόδους αυτές αναζητείται μόνο ο κατάλληλος διαμελισμός σε συγκεκριμμένο πλήθος κλάσεων. Τέτοιες μέθοδοι είναι k-means [Ball και Hall,1965] formes fortes, voisins reciproques ή nuees dynamiques [Diday,1972]. Επιλέγεται με αυτές τις μεθόδους ένα σύνολο αρχικών κέντρων και στη συνέχεια εντοπίζεται ο διαμελισμός με ομάδες γύρω από τα κέντρα αυτά. Η διαδικασία επαναλαμβάνεται έχοντας τώρα σαν σύνολο αρχικών κέντρων νέα κέντρα (π.χ. τα κέντρα βάρους των δημιουργημένων ομάδων). Η επανάληψη συνεχίζεται μέχρι να σταθεροποιηθούν οι ομάδες.

Όπως ήδη αναφέραμε, σκοπός της ταξινόμησης είναι να επιτευχθεί η βέλτιστη "οργάνωση" των δεδομένων σε κλάσεις και υποκλάσεις ώστε να είναι όσο γίνεται κοντύτερα τόσο στα δεδομένα όσο και στις μεταβλητές που τα χαρακτηρίζουν [Tallur,1988]. Σε κάθε στάδιο της ταξινόμησης, αναζητούμε έναν διαμελισμό του συνόλου των παρατηρήσεων σε ομογενείς ομάδες, ως προς το σύνολο των μεταβλητών, ώστε κάθε ομάδα να διαφέρει σημαντικά από τις υπόλοιπες. Μ' αυτό τον τρόπο επιτυγχάνεται μια ιεράρχηση των παρατηρήσεων, δηλαδή, μια σειρά διαμελισμών "ο ένας μέσα στον άλλο", που όσο προχωρά, τόσο πιο λεπτομερής γίνεται. Βέβαια υπάρχουν και μέθοδοι που περιορίζονται μόνο στην εύρεση ενός βέλτιστου διαμελισμού των δεδομένων, για δοσμένο αριθμό κλάσεων.

Στην παρούσα διατριβή, δεν θα ασχοληθούμε με τέτοιες μεθόδους, αλλά θα χρησιμοποιήσουμε τους διαμελισμούς που προκύπτουν από μια ταξινόμηση καθώς και τις χαρακτηριστικές ιδιότητες τους. Ένας τρόπος για να πάρουμε ένα διαμελισμό, είναι να κόψουμε το δένδρο ιεράρχησης σε κάποιο επίπεδο. (Στη συνέχεια ορίζουμε το δένδο ιεράρχησης και τον τρόπο επιλογής διαμελισμού). Αν έχουμε πολλά δεδομένα, εφαρμόζουμε τη μέθοδο nuees dynamique για να μειώσουμε τον αριθμό κλάσεων, και μετά την ταξινόμηση κατά αύξουσα ιεραρχία. Έχοντας επιλέξει ένα διαμελισμό που προκύπτει απ' αυτή, ξαναεφαρμόζουμε nuees dynamique για να βελτιστοποιήσουμε τα αποτελέσματα. [Diday,1992] [Molliere,1985] [Marchetti,1989].

Ένας διαμελισμός που προκύπτει αυτόνομα, μερικές φορές μπορεί να είναι καλύτερος από κάποιο διαμελισμό που προκύπτουν από μια ταξινόμηση, γιατί στην δεύτερη περίπτωση δεν μπορούμε να ανατρέψουμε κάτι που έγινε σε προηγούμενο στάδιο. Αν δύο αντικείμενα ομαδοποιήθηκαν προηγουμένως, δεν μπορούμε πλέον να τα χωρίσουμε. Στην ιδιότητα αυτή βέβαια οφείλεται και η επιτυχία των μεθόδων ταξινόμησης (μικρός υπολογιστικός χρόνος), αλλά και η αδυναμία τους να διορθώσουν μια λάθος απόφαση αν αυτή ληφθεί νωρίς. [Kaufman,1990].

Στο εξής θα ασχοληθούμε μόνο με την ταξινόμηση κατ' αύξουσα ιεραρχία (στη συνέχεια αναφέρεται σαν C.A.H.). Θεωρούμε ότι η μέθοδος αυτή είναι κοντύτερα στην διαδικασία ανάλυσης, σύνθεσης και ταξινόμησης που εφαρμόζει ο ανθρώπινος εγκέφαλος. Για την δημιουργία των κλάσεων ομαδοποιούμε τα κοντυνότερα σημεία (ανάλυση) ενώ για την ερμηνεία και διαχωρισμό των ομάδων βρίσκουμε τις διαφοροποιήσεις τους (σύνθεση).

Με την ιεραρχική ταξινόμηση προσπαθούμε να διαμελίσουμε ένα σύνολο δεδομένων σε όσο το δυνατό ομογενείς ομάδες, ώστε κάθε ομάδα να είναι διαφορετική από τις υπόλοιπες. Αναλύοντας τα δεδομένα μιας ταξινόμησης, παρατηρούμε ότι το σύνολο των δεδομένων, χωρίζεται σε δύο ομάδες με όσο

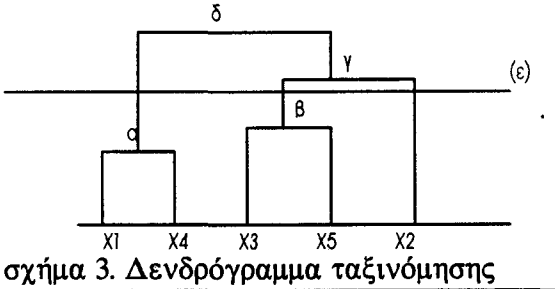
γίνεται περισσότερο ομογενή στοιχεία η κάθε μια. Στη συνέχεια οι ομάδες αυτές διασπώνται η κάθε μία σε δύο νέες κατά το δυνατό ομογενείς ομάδες, δημιουργώντας μια νέα ομοδοποίηση των δεδομένων. Η διαδικασία συνεχίζεται, μέχρι να καταλήξουμε στις αρχικές μας παρατηρήσεις.

Δημιουργούμε δηλαδή, μια ιεραρχία από διαδοχικούς διαμελισμούς των παρατηρήσεων, όλο και καλύτερα "προσαρμοζόμενους" στο αρχικό σύνολο παρατηρήσεων.

2. Σχηματική απεικόνιση της ταξινόμησης

Σχηματική απεικόνιση της ιεραρχίας αυτής, επιτυγχάνουμε με ένα "δενδρόγραμμα" (σχήμα 3) του οποίου κάθε κόμβος παριστάνει μια ομάδα και χωρίζεται σε δύο υποομάδες (τους κλάδους του κόμβου). Τερματικοί κόμβοι είναι οι αρχικές μας παρατηρήσεις που εμφανίζονται μόνο μία φορά.

Στο σχήμα 3 τα X1, X2, X3, X4, X5, είναι οι αρχικές παρατηρήσεις. Ενώνονται τα X1 και X4, σχηματίζοντας την ομάδα α. Κατόπιν τα X3 και X5, οπότε προκύπτει η ομάδα β. Στη συνέχεια η ομάδα β ενώνεται με το X2, σχηματίζοντας την ομάδα γ. Αυτή ενώνεται με την α, και προκύπτει έτσι η ομάδα δ που περιέχει όλες τις παρατηρήσεις.



σχήμα 3. Δενδρόγραμμα ταξινόμησης

Αν δεν ενδιαφερόμαστε για μια συνολική ιεράρχηση των παρατηρήσεων, αλλά θέλουμε απλώς να έχουμε ένα διαμελισμό σε συγκεκριμμένο αριθμό κλάσεων, δεν έχουμε παρά να διαλέξουμε ένα διαμελισμό, "κόβοντας" το δενδρόγραμμα με μια ευθεία γραμμή. Οι κλάδοι που απομένουν παριστάνουν τις ομάδες του διαμελισμού.

Στο σχήμα 3, αν θέλουμε διαμελισμό σε 3 κλάσεις δεν έχουμε παρά να κόψουμε το δένδρο με την ευθεία ε, οπότε θα προκύψει ο διαμελισμός με τις κλάσεις: {X1,X4} {X2} {X3,X5}.

3. Στάδια ταξινόμησης

Η ταξινόμηση χωρίζεται σε δύο βασικά στάδια. Στο πρώτο υπολογίζουμε τις ομοιότητες μεταξύ αντικειμένων, ενώ στο δεύτερο σχηματίζουμε τις ομάδες αντικειμένων, που περιλαμβάνουν τα πιο "όμοια" αντικείμενα.

Για τον υπολογισμό των ομοίων αντικειμένων, χρησιμοποιείται είτε μια μετρική απόστασης, είτε κάποιος συντελεστής συσχέτισης. Το ποια μέθοδος θα χρησιμοποιηθεί, εξαρτάται από το είδος και την φύση των δεδομένων. Πλήρη αναφορά των μετρικών και δεικτών αποστάσεων κάναμε στο προηγούμενο κεφάλαιο. Από τις πιο διαδεδομένες μετρικές απόστασης μεταξύ δύο αντικειμένων, είναι:

- Η ευκλείδεια απόσταση, για αντικείμενα που χαρακτηρίζονται από ποσοτικές μεταβλητές.
- Η X^2 απόσταση όταν μελετάμε ποιοτικές μεταβλητές, και ιδιαίτερα όταν αναλύουμε πίνακες συχνοτήτων.

Τέλος όταν αναλύουμε πίνακες ποιοτικών μεταβλητών, με "λογικά δεδομένα" (1 ή 0 δηλαδή παρουσία ή απουσία κάποιου χαρακτηριστικού), χρησιμοποιούνται διάφοροι άλλοι δείκτες απόστασης που μετράνε τις κοινές ή μη παρουσίες. Ένας τέτοιος δείκτης είναι ο δείκτης Jaccard. Υπάρχουν και πολλοί ακόμη παρόμοιοι δείκτες, για δεδομένα 0 και 1, οι οποίοι είναι στατιστικά ισοδύναμοι [Lerman,1980].

Αν θέλουμε να μετρήσουμε την απόσταση μεταξύ μεταβλητών, εκτός από τις προηγούμενες αποστάσεις, μπορούμε να χρησιμοποιήσουμε και ένα συντελεστή συσχέτισης: Τον γνωστό συντελεστή για ποσοτικές μεταβλητές, ενώ για ποιοτικές μεταβλητές το συντελεστή Cramer, Tschuprow, Shannon. [Diday,1992]. (βλέπε προηγούμενο κεφάλαιο)

Στο δεύτερο στάδιο της ταξινόμησης, για να υπολογίσουμε το πόσο "κοντά" βρίσκονται δύο ομάδες, χρησιμοποιούμε διαφορετικές μεθόδους όπως MIN, MAX, MOYENNE, απόσταση των κέντρων βάρους, αύξηση της αδράνειας κ.α. Κάθε μία από τις μεθόδους αυτές καταλήγει σε διαφορετική ιεράρχηση των παρατηρήσεων.

4. Επίδραση των δεδομένων στην επιλογή μετρικής

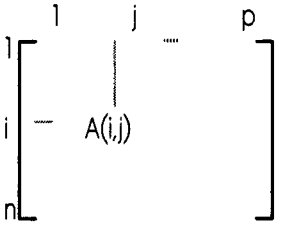
Με την C.A.H. προσπαθούμε να βρούμε τις όσο γίνεται πιο ομογενείς ομάδες των δεδομένων. Με ποιο κριτήριο όμως θα αποφασίσουμε αν μια ομάδα είναι ομογενής ή όχι; Ποιο είναι το μέτρο απόστασης που θα μας βοηθήσει να ελέξουμε αν δύο παρατηρήσεις ή αντικείμενα ανήκουν στην ίδια ομάδα ή όχι;

Η απάντηση στα ερωτήματα αυτά, εξαρτάται από την φύση των δεδομένων που έχουμε να μελετήσουμε. Διακρίνουμε δύο μεγάλες κατηγορίες:

. Δεδομένα που περιγράφονται από ποσοτικές μεταβλητές

. Δεδομένα που περιγράφονται από ποιοτικές μεταβλητές

Γενικά ο πίνακας δεδομένων Α, είναι ένας ορθογώνιος πίνακας nxp, όπου n το πλήθος των παρατηρούμενων ατόμων και p ο αριθμός των μεταβλητών που χαρακτηρίζει κάθε άτομο. Στη διασταύρωση της i γραμμής του πίνακα και της j στήλης, έχουμε την τιμή της j μεταβλητής που χαρακτηρίζει το i άτομο. Όταν η τιμή αυτή είναι είναι αριθμητική έχουμε ποσοτικά δεδομένα ενώ στην αντίθετη περίπτωση μιλάμε για δεδομένα ποιοτικά.



Κάποιες φορές αντί για πίνακα δεδομένων *αντικείμενα x μεταβλητές*, έχουμε πίνακα nxn αποστάσεων ή ομοιοτήτων μεταξύ των αντικειμένων μας. [Kaufman,1990] Τέτοια είναι και η περίπτωση ενός πίνακα Burt σχετικών συχνοτήτων (κεφάλαιο 2), όταν πρόκειται να προβούμε στην ταξινόμηση μεταβλητών. [Tallur,1988]

Από τον πίνακα των δεδομένων Α διαστάσεων nxp, υπολογίζουμε αρχικά έναν άλλο συμμετρικό πίνακα B(n,n) που περιέχει όλες τις αποστάσεις μεταξύ των παρατηρήσεων ανά δύο, με κάποια από τις μετρικές που αναφέραμε στο προηγούμενο κεφάλαιο. (κεφάλαιο 3).

5. Μετρικές για την ένωση δύο ομάδων

Αρχικά βρίσκουμε τα δύο "πλησιέστερα" στοιχεία (αυτά που έχουν τη μικρότερη απόσταση μεταξύ τους), και στη θέση τους θέτουμε ένα νέο, με συντεταγμένες (τιμές για τις p μεταβλητές), τις συντεταγμένες του κέντρου βάρους των στοιχείων που αντικαθίστανται. Αν θέλουμε να βρούμε την απόσταση ενός στοιχείου k, από μια ομάδα, ή την απόσταση μεταξύ δύο ομάδων, υπολογίζουμε τις αποστάσεις μεταξύ του στοιχείου αυτού και των υπολοίπων, με μία από τις ακόλουθες 3 μεθόδους: MIN, MAX, MOYENNE.

Αν i και i' είναι τα δύο συνενούμενα στοιχεία σε μια κλάση, που συμβολίζεται iUi', και k ένα άλλο στοιχείο, η απόσταση του k από την κλάση αυτή είναι:

MIN $\quad dm(iUi',k)=\min \{d(i,k),d(i',k)\}$

Ένα αντικείμενο εντάσσεται σε μια ομάδα, αν είναι κοντινότερα σε ΕΝΑ αντικείμενο της ομάδας αυτής, απ' ότι σ' ένα αντικείμενο μιας άλλης ομάδας.

MAX $dM(iUi',k)=\max\{d(i,k),d(i',k)\}$

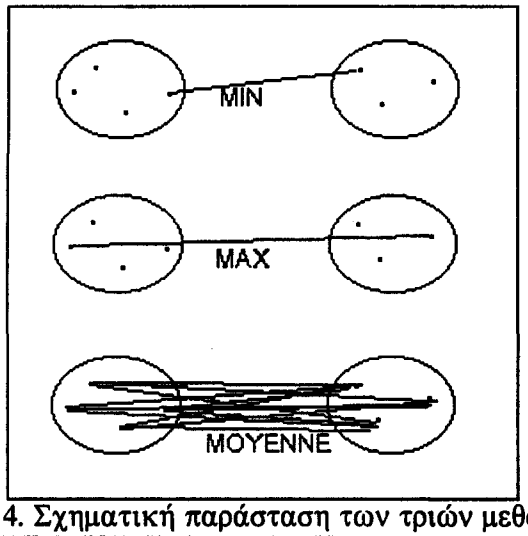
Ένα αντικείμενο εντάσσεται σε μια ομάδα, αν είναι κοντινότερα σε ΟΛΑ τα αντικείμενα της ομάδας αυτής, απ' ότι σ' όλα τα αντικείμενα μιας άλλης ομάδας.

MOYENNE $d\mu(iUi',k)=\dfrac{p(i)\,d(i,k)+p(i')\,d(i',k)}{p(i)+p(i')}$

(όπου p(x) είναι το πλήθος στοιχείων της κλάσης x)

Ένα αντικείμενο εντάσσεται σε μια ομάδα, αν είναι κοντινότερα κατά μέσο όρο στα αντικείμενα της ομάδας αυτής, απ' ότι στα αντικείμενα μιας άλλης ομάδας.

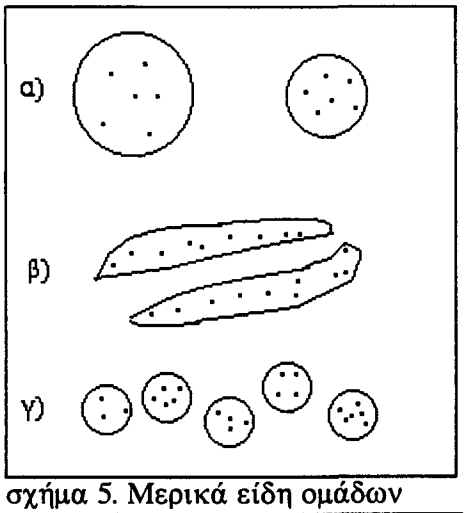
Εποπτικά οι τρεις μέθοδοι παριστάνονται στο σχήμα 4.



σχήμα 4. Σχηματική παράσταση των τριών μεθόδων

Το ποια από τις 3 παραπάνω μεθόδους θα χρησιμοποιήσουμε, εξαρτάται από την φύση των δεδομένων και το είδος των ομάδων που επιθυμούμε. Αν θέλουμε ομάδες όπως αυτές στο σχήμα 5α, χρησιμοποιούμε την απόσταση MOYENNE. Στην περίπτωση αυτή, η MIN δεν είναι κατάλληλη, γιατί αν δύο μόνο σημεία των ομάδων βρεθούν κοντά, τότε ενώνονται οι δύο ομάδες και δεν χωρίζονται στα επόμενα στάδια. Το αποτέλεσμα αυτό ονομάζεται επίδραση αλυσίδας. Αντικείμενα μπορούν να βρίσκονται στην ίδια κλάση αν και απομακρυσμένα μεταξύ τους, επειδή υπάρχουν αντικείμενα ανάμεσά τους, που τα συνδέουν (αλυσίδα). Κάποιες φορές βέβαια αυτό είναι πλεονέκτημα. Αν τα δεδομένα μας σχηματίζουν ομάδες σε σχήμα σαν κι αυτό στο σχήμα 5β, κατάλληλη απόσταση για την εύρεσή τους είναι η MIN, γιατί οι MAX και MOYENNE θα δημιουργούσαν μόνο μία ομάδα. Τέλος αν οι ομάδες είναι συμπαγείς αλλά όχι πολύ απομακρυσμένες μεταξύ τους (σχήμα

5γ), η χρήση της MIN θα τις συνέδεε, η χρήση της MOYENNE θα δημιουργούσε δύο ή τρείς κλάσεις, ενώ εκείνη που μπορεί να τις αναδείξει, είναι η μέθοδος MAX. [Kaufman,1990] [Roux,1985]



σχήμα 5. Μερικά είδη ομάδων

Άλλη μέθοδος για τον υπολογισμό της απόστασης μεταξύ ομάδων, είναι η απόσταση μεταξύ δύο χαρακτηριστικών σημείων ("noyaux"). Αν θέλουμε να ενώσουμε τις ομάδες A_1 και A_2, παίρνουμε δύο υποσύνολά τους N_1 και N_2.
$D(A_1,A_2)=\Sigma\, d(w_i,w_j)$ όπου w_i ανήκει N_1 και w_j στο N_2.

Επίσης μπορούμε να χρησιμοποιήσουμε την απόσταση d, των κέντρων βάρους g_1, g_2 των δύο ομάδων (κριτήριο Ward). [Tenenbaus,1994]

$$D(A_1,A_2)=\frac{\mu(A_1)\,\mu(A_2)}{\mu(A_1)+\mu(A_2)}d^2(g_1,g_2)$$

όπου μ(A) το βάρος της ομάδας A. Η απόσταση αυτή αντιπροσωπεύει το χάσιμο αδράνειας με την ένωση των A_1 και A_2. Αποδεικνύεται:

$D(A_1,A_2)= I(A_1\cup A_2) -I(A_1) -I(A_2)$

Είναι $I(A_1\cup A_2) =I(A_1) +I(A_2)+ \frac{\mu(A_1)\,\mu(A_2)}{\mu(A_1)+\mu(A_2)}d^2(g_1,g_2)$

Μια άλλη μέθοδος για την ένωση δύο ομάδων είναι η "Moment d' ordre deux". [Roux,1985] [Benzecri,1973] Με την μέθοδο αυτή διαλέγουμε ποιες ομάδες θα ενωθούν, όχι βασιζόμενοι στην μεταξύ τους απόσταση, αλλά στην αύξηση της αδράνειας (διασποράς) που προκύπτει με την ένωση των δύο ομάδων σε μία. Σε κάθε βήμα, ενώνονται οι ομάδες που προκαλούν την μικρότερη αύξηση της αδράνειας (διασποράς).

Δεν μπορούμε, γενικά, να συγκρίνουμε μεταξύ τους τις ιεραρχήσεις των μεθόδων αυτών, αφού μια μέθοδος που είναι ιδανική για ένα σύνολο δεδομένων, δεν είναι τόσο ικανοποιητική για κάποιο άλλο.

Η σύγκριση λοιπόν των μεθόδων, μπορεί να πραγματοποιηθεί μόνο εμπειρικά όταν εφαρμόζονται σ' ένα συγκεκριμένο σύνολο παρατηρήσεων. Για να επιλέξουμε την καταλληλότερη από τις τρεις μεθόδους (MIN, MAX, MOYENNE) για το συγκεκριμένο αυτό σύνολο, συγκρίνουμε την αδράνεια (διασπορά) των παρατηρήσεων μετά την ιεράρχηση, με την αρχική ολική αδράνεια (διασπορά). Η μέθοδος κατά την οποία οι δύο αυτές τιμές βρίσκονται πλησιέστερα, είναι η καταλληλότερη.

Πολλές φορές, δεν ενδιαφερόμαστε για την συνολική ιεράρχηση των παρατηρήσεων, αλλά επιθυμούμε απλώς να προσδιορίσουμε τον διαμελισμό τους στο ιδανικό πλήθος κλάσεων ή σε συγκεκριμένο πλήθος κλάσεων. Για κάθε μέθοδο, λοιπόν, και για κάθε διαφορετικό πλήθος κλάσεων υπολογίζουμε το ποσοστό της αρχικής ολικής αδράνειας που ερμηνεύει κάθε διαμελισμός, και με την σύγκριση αυτών μεταξύ τους, επιλέγουμε τον ιδανικότερο.

Στη συνέχεια της διατριβής ασχολούμαστε μόνο με τις μεθόδους MIN, MAX, MOYENNE.

6. Διαδικασία δημιουργίας της ιεραρχίας

Ξεκινώντας από τα δεδομένα, βρίσκουμε τον πίνακα αποστάσεων όλων των αντικειμένων ανά δύο, και κατόπιν εφαρμόζουμε τις μεθόδους της ταξινόμησης, στον πίνακα αυτό.

Ενώνουμε καταρχήν τα δύο πλησιέστερα αντικείμενα σε μια ομάδα και συνεχίζουμε έτσι, ενώνοντας διαδοχικά τις δύο πλησιέστερες κλάσεις, ξαναυπολογίζουμε τις αποστάσεις αντικειμένων και κλάσεων, μέχρι να καταλήξουμε σε δύο μόνο κλάσεις, τις οποίες ενώνουμε σε μία τελική κλάση, που περιέχει όλα τα δεδομένα.

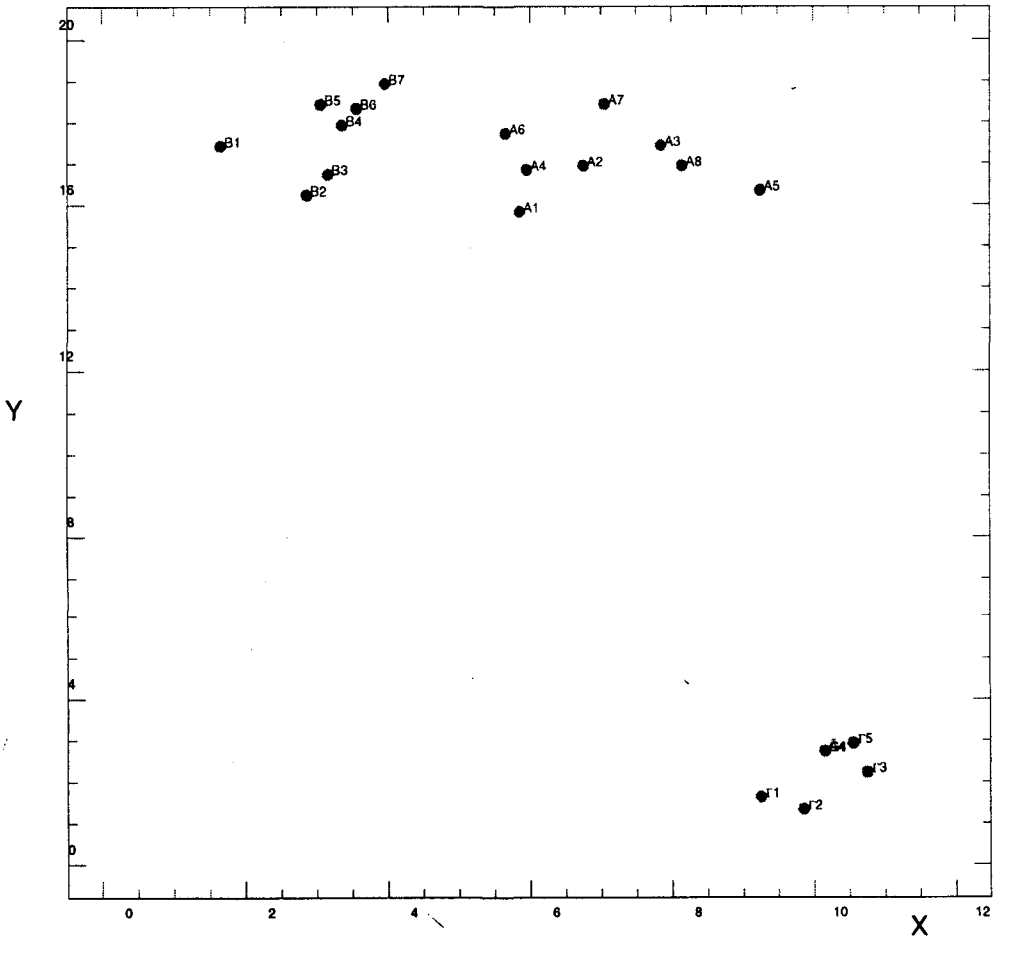
6.1Παράδειγμα

Θα περιγράψουμε την δημιουργία της ταξινόμησης κατ' αύξουσα ιεραρχία, με ένα παράδειγμα, ώστε να γίνει εύκολα κατανοητή.

Έστω ότι έχουμε τα δεδομένα 20 παρατηρήσεων και 2 μεταβλητών, όπως φαίνονται στον πίνακα 1.

Πίνακας 1
Οι τιμές των δύο μεταβλητών

ΟΝΟΜΑ ΑΝΤΙΚΕΙΜΕΝΟΥ	ΤΙΜΕΣ Χ	ΤΙΜΕΣ Υ
Α1	5.5	15.4
Α2	6.4	16.5
Α3	7.5	17.0
Α4	5.6	16.4
Α5	8.9	15.9
Α6	5.3	17.3
Α7	6.7	18.0
Α8	7.8	16.5
Β1	1.3	17.0
Β2	2.5	15.8
Β3	2.8	16.3
Β4	3.0	17.5
Β5	2.7	18.0
Β6	3.2	18.5
Γ1	8.9	1.2
Γ2	9.5	0.9
Γ3	10.4	1.8
Γ4	9.8	2.3
Γ5	10.2	2.5



σχήμα 6. Γραφική αναπαράσταση των 20 αντικειμένων

Δημιουργούμε τον συμμετρικό πίνακα αποστάσεων ανά δύο (πίνακας 2), χρησιμοποιώντας την ευκλείδεια απόσταση.

Πίνακας 2. Ο πρώτος πίνακας αποστάσεων μεταξύ των 20 στοιχείων

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	B4	B5	B6	B7	Γ1	Γ2	Γ3	Γ4
A2	1.42127																		
A3	2.56125	1.20830																	
A4	1 499	0.80623	1.99249																
A5	3.43657	2.57099	1.78045	3.33766															
A6	1.91050	1.36015	2.22036	0.94868	3.86264														
A7	2.86356	1.52971	1.28062	1.94165	3.04138	1.56525													
A8	2.54951	1.40000	0.58310	2.20227	1.25300	2.62488	1.86011												
B1	4.49444	5.12445	6.20000	4.34166	7.67919	4.01123	5.49181	6.51920											
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706										
B3	2.84605	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5 400	1.65529	0.58310									
B4	3.26497	3.54401	4.52769	2.82312	6.11310	2.30868	3.73363	4.90306	1.77200	1.77200	1.21655								
B5	3.82099	3.99249	4.90306	3.31210	6.54599	2.69258	4.000	5.31601	1.72047	2.20907	1.70294	0.58310							
B6	3.39706	3.49285	4.39318	2.83019	6.04070	2.18403	3.50143	4.80833	2.10238	2.21359	1.64924	0.44721	0.50990						
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	2.34094	1.16619	1.02956	0.72111					
Γ1	14.60137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.33494	17.90754	17.64596	18.09365				
Γ2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.69235	18.06793	16.46238	16.79434	17.82723	18.40245	18.12981	18.56260	0.67082			
Γ3	14.45579	15.23450	15.47417	15.36880	14.17956	16.31748	16.61716	14.92816	17.71581	16.07514	16.37101	17.35655	17.93683	17.63661	18.03136	1.61555	1.27279		
Γ4	13.78768	14.60137	14.87884	14.71224	13.62975	15.66046	16 312	14.34015	16.98058	15.34731	15.65248	16.65173	17.23079	16.93871	17.34589	1.42127	1.43178	0.78102	
Γ5	13.72953	14.50655	14.74924	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	15.65886	16.63851	17.21918	16.91626	17.30780	1.83848	1.74642	0.72801	0.44721

Πίνακας 3. Δεύτερος πίνακας αποστάσεων με τη μέθοδο MOYENNE

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	21	B5	Γ5	B7	Γ1	Γ2	Γ3
A2	1.42127																	
A3	2.56125	1.20830																
A4	1 499	0.80623	1.99249															
A5	3.43657	2.57099	1.78045	3.33766														
A6	1.91050	1.36015	2.22036	0.94868	3.86264													
A7	2.86356	1.52971	1.28062	1.94165	3.04138	1.56525												
A8	2.54951	1.40000	0.58310	2.20227	1.25300	2.62488	1.86011											
B1	4.49444	5.12445	6.20000	4.34166	7.67919	4.01123	5.49181	6.51920										
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706									
B3	2.84605	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5 400	1.65529	0.58310								
21	3.33101	3.51843	4.46043	2.82666	6.07690	2.24636	3.61753	4.85569	1.93719	1.99280	1.43290							
B5	3.82099	3.99249	4.90306	3.31210	6.54599	2.69258	4 000	5.31601	1.72047	2.20907	1.70294	0.54650						
Γ5	13.72953	14.50655	14.74924	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	15.65886	16.77739	17.21918					
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	2.34094	0.94365	1.02956	17.3078				
Γ1	14.60137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.49045	17.90754	1.83848	18.09365			
Γ2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.69235	18.06793	16.46238	16.79434	17.97852	18.40245	1.74642	18.56260	0.67082		
Γ3	14.45579	15.23450	15.47417	15.36880	14.17956	16.31748	16.61716	14.92816	17.71581	16.07514	16.37101	17.49658	17.93683	0.72801	18.03136	1.61555	1.27279	
Γ4	13.78768	14.60137	14.87884	14.71224	13.62975	15.66046	16 312	14.34015	16.98058	15.34731	15.65248	16.79522	17.23079	**0.44721**	17.34589	1.42127	1.43178	0.78102

Πίνακας 4. Δεύτερος πίνακας αποστάσεων με τη μέθοδο MIN

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	21	B5	Γ5	B7	Γ1	Γ2	Γ3
A2	1.42127																	
A3	2.56125	1.20830																
A4	1 499	0.80623	1.99249															
A5	3.43657	2.57099	1.78045	3.33766														
A6	1.91050	1.36015	2.22036	0.94868	3.86264													
A7	2.86356	1.52971	1.28062	1.94165	3.04138	1.56525												
A8	2.54951	1.40000	0.58310	2.20227	1.25300	2.62488	1.86011											
B1	4.49444	5.12445	6.20000	4.34166	7.67919	4.01123	5.49181	6.51920										
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706									
B3	2.84605	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5 400	1.65529	0.58310								
21	3.26497	3.49285	4.39318	2.82312	6.04070	2.18403	3.50143	4.80833	1.77200	1.77200	1.21655							
B5	3.82099	3.99249	4.90306	3.31210	6.54599	2.69258	4 000	5.31601	1.72047	2.20907	1.70294	0.50990						
Γ5	13.72953	14.50655	14.74924	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	15.65886	16.63851	17.21918					
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	2.34094	0.72111	1.02956	17.3078				
Γ1	14.60137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.33494	17.90754	1.83848	18.09365			
Γ2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.69235	18.06793	16.46238	16.79434	17.82723	18.40245	1.74642	18.56260	0.67082		
Γ3	14.45579	15.23450	15.47417	15.36880	14.17956	16.31748	16.61716	14.92816	17.71581	16.07514	16.37101	17.35655	17.93683	0.72801	18.03136	1.61555	1.27279	
Γ4	13.78768	14.60137	14.87884	14.71224	13.62975	15.66046	16 312	14.34015	16.98058	15.34731	15.65248	16.65173	17.23079	**0.44721**	17.34589	1.42127	1.43178	0.78102

Πίνακας 5. Δεύτερος πίνακας αποστάσεων με τη μέθοδο MAX

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	21	B5	Γ5	B7	Γ1	Γ2	Γ3
A2	1.42127																	
A3	2.56125	1.20830																
A4	1 499	0.80623	1.99249															
A5	3.43657	2.57099	1.78045	3.33766														
A6	1.91050	1.36015	2.22036	0.94868	3.86264													
A7	2.86356	1.52971	1.28062	1.94165	3.04138	1.56525												
A8	2.54951	1.40000	0.58310	2.20227	1.25300	2.62488	1.86011											
B1	4.49444	5.12445	6.20000	4.34166	7.67919	4.01123	5.49181	6.51920										
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706									
B3	2.84605	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5 400	1.65529	0.58310								
21	3.39706	3.54401	4.52769	2.83019	6.11310	2.30868	3.73363	4.90306	2.10238	2.21359	1.64924							
B5	3.82099	3.99249	4.90306	3.31210	6.54599	2.69258	4 000	5.31601	1.72047	2.20907	1.70294	0.58310						
Γ5	13.72953	14.50655	14.74924	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	15.65886	16.91626	17.21918					
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	2.34094	1.16619	1.02956	17.3078				
Γ1	14.60137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.64596	17.90754	1.83848	18.09365			
Γ2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.69235	18.06793	16.46238	16.79434	18.12981	18.40245	1.74642	18.56260	0.67082		
Γ3	14.45579	15.23450	15.47417	15.36880	14.17956	16.31748	16.61716	14.92816	17.71581	16.07514	16.37101	17.63661	17.93683	0.72801	18.03136	1.61555	1.27279	
Γ4	13.78768	14.60137	14.87884	14.71224	13.62975	15.66046	16 312	14.34015	16.98058	15.34731	15.65248	16.93871	17.23079	**0.44721**	17.34589	1.42127	1.43178	0.78102

Τα δεδομένα επιλέκτηκαν έτσι ώστε να σχηματίζουν 3 ομάδες στο επίπεδο των 2 διαστάσεων (σχήμα 6). Σε κάθε βήμα, και για τις 3 μεθόδους, με την προτεινόμενη διαδικασία, τυπώνεται ο πίνακας αποστάσεων ώστε να φαίνεται ποια είναι τα "κοντινότερα" στοιχεία, τα οποία ενώνονται για να σχηματίσουν τον νέο κόμβο.

Ο πρώτος πίνακας αποστάσεων (πίνακας 2) είναι ίδιος και για τις 3 μεθόδους, και η ελάχιστη απόσταση είναι 0.44721 για τα στοιχεία Β6 και Β4. Στο επόμενο βήμα τα στοιχεία αυτά ενώνονται σε ένα που συμβολίζεται με τον κόμβο νούμερο 21.

Ο δεύτερος πίνακας αποστάσεων (πίνακας 3, 4 και 5 αντίστοιχα για τις τρεις μεθόδους) διαφέρει στις 3 μεθόδους ως προς την γραμμή και την στήλη που βρίσκεται ο κόμβος 21. Η απόσταση ενός στοιχείου από την ομάδα, εξαρτάται από τη μέθοδο που ακολουθούμε. Για τη μέθοδο MIN, η απόσταση του Β3 από την ομάδα 21 είναι 1.21655 (δηλαδή η απόσταση του Β3-Β4), για τη μέθοδο MAX είναι 1.64924 (δηλαδή η απόσταση του Β3-Β6), ενώ για τη μέθοδο MOYENNE είναι 1.43290 (ο μέσος όρος των δύο προηγουμένων αποστάσεων).

Η ελάχιστη απόσταση στον δεύτερο πίνακα αποστάσεων (πίνακας 3,4,5) είναι και για τις 3 μεθόδους, 0.44721, μεταξύ των στοιχείων Γ4-Γ5, . Στη συνέχεια ενώνονται τα στοιχεία αυτά και σχηματίχουν τον κόμβο 22. Ο επόμενος πίνακας αποστάσεων, ο τρίτος για κάθε μέθοδο, διαφέρει από τον αντίστοιχο δεύτερο στην γραμμή και στήλη του κόμβου 22.

Πίνακας 6
Κόμβοι της ταξινόμησης με τη μέθοδο MOYENNE

κόμβος	αριστερό	δεξιό	βάρος	απόσταση
21	Β6	Β4	2	0.447214
22	Γ4	Γ5	2	0.447214
23	Β5	21	3	0.546499
24	Α8	Α3	2	0.583095
25	Β3	Β2	2	0.583095
26	Γ1	Γ2	2	0.670820
27	22	Γ3	3	0.754518
28	Α4	Α2	2	0.806226
29	23	Β7	4	0.972288
30	Α6	28	3	1.154415
31	30	Α1	4	1.445584
32	Α5	24	3	1.516723
33	26	27	5	1.554382
34	Β1	25	3	1.676175
35	Α7	31	5	1.975042
36	29	34	7	2.030048
37	32	35	8	2.409936
38	36	37	15	4.236877
39	33	38	20	16.112806

Πίνακας 7
Κόμβοι της ταξινόμησης με τη μέθοδο MIN

κόμβος	αριστερό	δεξιό	βάρος	απόσταση
21	B6	B4	2	0.447214
22	Γ4	Γ5	2	0.447214
23	B5	21	3	0.509902
24	A8	A3	2	0.583095
25	B3	B2	2	0.583095
26	Γ1	Γ2	2	0.670820
27	23	B7	4	0.721110
28	Γ3	22	3	0.728011
29	A4	A2	2	0.806226
30	A6	29	3	0.948683
31	30	A1	4	1 4988
32	24	31	6	1.208305
33	27	25	6	1.216553
34	A5	32	7	1.252996
35	26	28	5	1.272792
36	A7	34	8	1.280625
37	B1	33	7	1.655295
38	37	36	15	2.080865
39	35	38	20	13.462912

Πίνακας 8
Κόμβοι της ταξινόμησης με τη μέθοδο MAX

κόμβος	αριστερό	δεξιό	βάρος	απόσταση
21	B6	B4	2	0.447214
22	Γ4	Γ5	2	0.447214
23	A8	A3	2	0.583095
24	B3	B2	2	0.583095
25	B5	21	3	0.583095
26	Γ1	Γ2	2	0.670820
27	22	Γ3	3	0.781025
28	A4	A2	2	0.806226
29	25	B7	4	1.166190
30	A6	28	3	1.360147
31	24	B1	3	1.697056
32	A5	23	3	1.780449
33	27	26	5	1.838478
34	30	A1	4	1.910497
35	A7	34	5	2.863564
36	31	29	7	2.915476
37	32	35	8	3.862642
38	36	37	15	7.679193
39	33	38	20	18.562597

Στον τρίτο πίνακα αποστάσεων, διαπιστώνουμε ότι: Με την μέθοδο MAX η ελάχιστη απόσταση 0.58310 είναι των τριών ζευγών B5-21 , B3-B2, και των A3-A8. Ενώνονται λοιπόν τα A3 και A8 σχηματίζοντας τον κόμβο 23, στο επόμενο βήμα τα B3-B2, σχηματίζοντας τον κόμβο 24 και στο μεθεπόμενο τα B5 και 21 που σχηματίζουν τον κόμβο 25.

Με τη μέθοδο MIN η ελάχιστη απόσταση είναι 0.50990 μεταξύ B5 και 21, τα οποία ενώνονται στον κόμβο 23. Στο επόμενο βήμα ενώνονται και εδώ τα A3-A8 με απόσταση 0.58310 που σχηματίζουν τον κόμβο 24, ενώ στο μεθεπόμενο τα B3-B2 με την ίδια απόσταση σχηματίζουν τον κόμβο 25.

Με τη μέθοδο MOYENNE, η ελάχιστη απόσταση είναι 0.54650 για τα B5 και 21 τα οποία σχηματίζουν τον κόμβο 23. Στη συνέχεια ενώνονται τα A3-A8 με απόσταση 0.58310, δημιουργώντας τον κόμβο 24. Τέλος στο επόμενο βήμα ενώνονται τα B3-B2 με απόσταση πάλι 0.58310 και δίνουν τον κόμβο 25.

Στο 6ο βήμα και με τις 3 μεθόδους, ενώνονται τα Γ1-Γ2 με απόσταση 0.67082 και δίνουν τον κόμβο 26.

Στο 7ο βήμα για την μέθοδο MAX, ενώνεται το Γ3 με τον κόμβο (Γ4-Γ5) 22, (d(Γ3,Γ4)=0.78102) και δίνουν τον κόμβο 27. Με τη μέθοδο MIN, ενώνεται το B7 με τον κόμβο (B5-21) 23 (d(B7,21)=0.72111=d(B7,B6)) ενώ στο 8ο βήμα ενώνεται το Γ3 με (Γ4-Γ5) 22 (d(Γ3,Γ5)=0.72801). Με τη μέθοδο MOYENNE, το Γ3 με τον (Γ4-Γ5) 22 ενώνονται στο 7ο βήμα (0.75452=μέσος όρος των d(Γ3,Γ4) και d(Γ3,Γ5)).

Στο 8ο βήμα με τη μέθοδο MAX και MOYENNE, ενώνονται τα A2-A4 (d(A2,A4)=0.80623) δημιουργώντας τον κόμβο 28. Το ίδιο συμβαίνει με τη μέθοδο MIN αλλά στο βήμα 9.

Στο 9ο βήμα με τη μέθοδο MAX, συνδέεται ο κόμβος (B5-21) 25 με το B7 (d(B7,21)=1.16619=d(B7,B4)), ενώ με τη μέθοδο MOYENNE, ο κόμβος (B5-21) 23 με το B7 (απόσταση=0.97229).

Στο 10ο βήμα, με τη μέθοδο MAX, ενώνεται το A6 με τον κόμβο (A2-A4) 28 (d(A6,A2)=1.36015). Με τη μέθοδο MIN ενώνεται το A6 με τον κόμβο (A2-A4) 29 (d(A6,A4)=0.94868). Τέλος με τη μέθοδο MOYENNE, ενώνεται το A6 με τον κόμβο (A2-A4) 28 (μέσος όρος (των d(A6,A2), d(A6,A4))=1.15442).

Στο 11ο βήμα, με τη μέθοδο MAX, ενώνονται τα B1 και (B2-B3) 24 (d(B1,B2)=1.69706). Με τις άλλες 2 μεθόδους στο 11ο βήμα, ενώνεται το A1 με τον κόμβο 30 (με αποστάσεις 1 499, 1.44558 αντίστοιχα).

Στο βήμα 12 με τη μέθοδο MAX, ενώνονται τα στοιχεία A5 και (A8-A3) 23 (d(A5,A3)=1.78045). Με τη μέθοδο MIN οι κόμβοι (A8-A3) 24 και (A1-30) 31, (απόσταση 1.20830) ενώ με τη μέθοδο MOYENNE ενώνονται τα A5 και (A8-A3) 24 (απόσταση 1.51672).

Στο 13ο βήμα, με τη μέθοδο MAX, ενώνονται οι κόμβοι (Γ1-Γ2) 26 και (Γ3-22) 27 (d(Γ1,22)=1.83848), με τη μέθοδο MIN οι κόμβοι (B2-B3) 25 και (B7-23) 27 (d(B3,23)=1.21655) ενώ με τη μέθοδο MOYENNE οι κόμβοι (Γ1-Γ2) 26 και (Γ3-22) 27 (απόσταση 1.55438).

Στο 14ο βήμα, με τη μέθοδο MAX το στοιχείο A1 ενώνεται με τον κόμβο (A6-28) 30 (d(A1,A6)=1.91050). Με τη μέθοδο MIN ενώνεται το A5 με τον κόμβο 32 (απόσταση 1.253). Τέλος με τη μέθοδο MOYENNE, ενώνεται το B1 με το (B3-B2) 25 (απόσταση 1.67618).

Στο βήμα 15 και με τη μέθοδο MAX και με τη MOYENNE, ενώνεται το A7 με τον κόμβο (30-A1) (αντίστοιχοι κόμβοι 34 31), αλλά με διαφορετική τιμή απόστασης. Για τη μέθοδο MAX d(A7,A1)=2.86356 ενώ για τη MOYENNE (μέσος (των d(A7,A1) και d(A7,30))=1.97504). Στο ίδιο βήμα με τη μέθοδο MIN ενώνονται οι κόμβοι 26-28, ενώ με τη μέθοδο αυτή η ένωση του A7 με τον κόμβο (A5-32) 34 γίνεται στο επόμενο βήμα. Επίσης στο 16ο βήμα, με τη μέθοδο MAX ενώνονται οι κόμβοι 29 και 31, με τη MOYENNE οι 29 και 34.

Στο βήμα 18 ενώ με τις άλλες μεθόδους ενώνονται ανώτεροι κόμβοι, με την MIN ενώνεται το αρχικό στοιχείο B1 με τον κόμβο 33.

Τέλος στο 19ο βήμα και με τις 3 μεθόδους ενώνονται οι κόμβοι 36 και 37 δίνοντας τον κόμβο 38 ο οποίος στο επόμενο βήμα ενώνεται με τον κόμβο 33 δίνοντας ένα μόνο κόμβο που περιέχει και τις 20 αρχικές παρατηρήσεις.

Με τη μέθοδο MIN μπορούμε να έχουμε στην ίδια κλάση απομακρυσμένα στοιχεία. Η μέθοδος MAX υποχρεώνει πιο απομακρυσμένα σημεία να μην

θεωρηθούν κοντινά. Τέλος η μέθοδος MOYENNE προκύπτει από ένα συμβιβασμό των δύο παραπάνω μεθόδων.

Προχωράμε έτσι υπολογίζοντας ξανά τον πίνακα αποστάσεων και παριστώντας τους διαμελισμούς που σχηματίζονται στο επίπεδο, ώστε να φαίνονται εύκολα οι σχηματιζόμενες ομάδες.

Στο συγκεκριμμένο παράδειγμα, οι μέθοδοι MAX και MOYENNE δίνουν περίπου τις ίδιες ομάδες με κάποιες μικρές διαφοροποιήσεις, ένω η μέθοδος MIN ξεχωρίζει γρήγορα κάποιο στοιχείο που είναι απομακρυσμένο από τα υπόλοιπα.

7. Δενδρόγραμμα ιεραρχίας

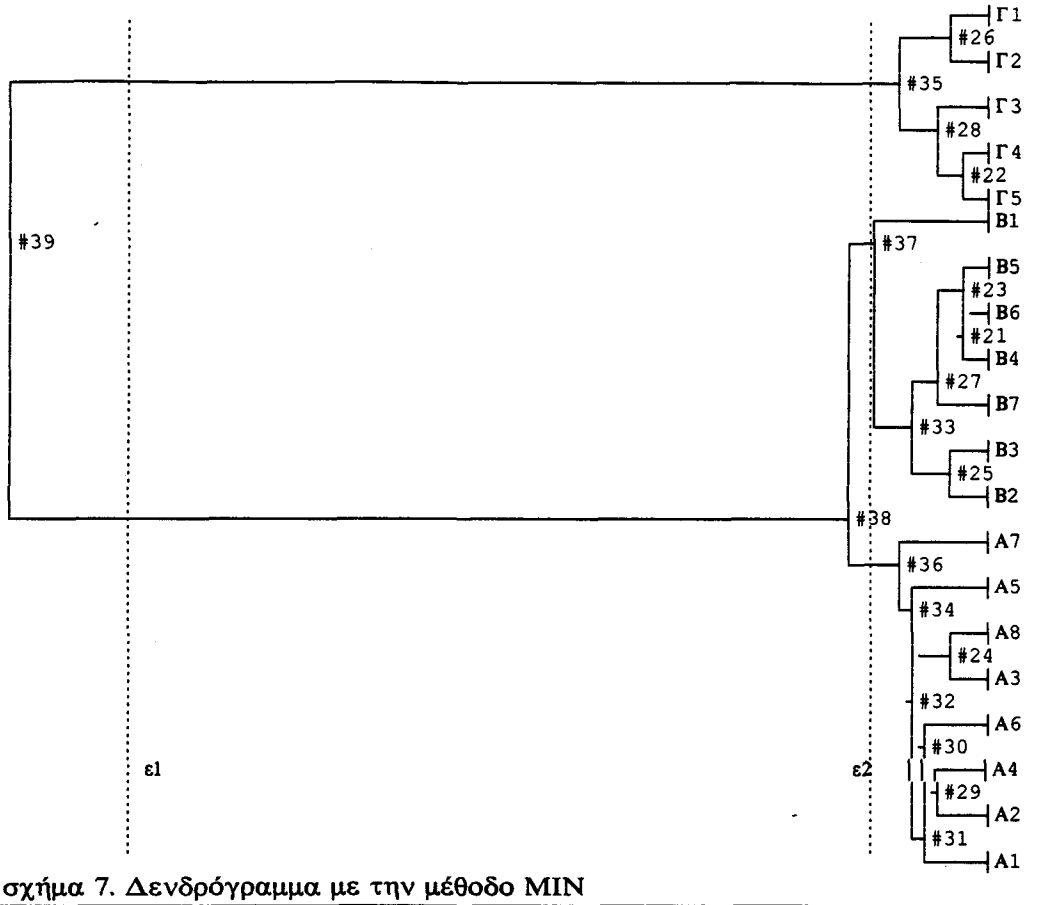
Η ταξινόμηση κατά αύξουσα ιεραρχία (C.A.H), όπως προαναφέραμε, σχηματικά παριστάνεται με ένα δενδρόγραμμα, όπου κάθε κόμβος αντιστοιχεί σε μία κλάση (ομάδα παρατηρήσεων) και στο επόμενο επίπεδο, χωρίζεται σε δύο υποκλάσεις (υποομάδες) (στις κλάσεις που ενώθηκαν για να σχηματίσουν τον κόμβο). Η κάθε μία από τις υποκλάσεις χωρίζεται με τη σειρά της σε δύο υποκλάσεις μέχρι να καταλήξουμε στους τερματικούς κόμβους οι οποίοι αντιστοιχούν στις αρχικές παρατηρήσεις και εμφανίζονται μόνο μία φορά.

Ένα δενδρόγραμμα ιεραρχίας φαίνεται στο σχήμα 3, και παρουσιάζει την ιεραρχική ταξινόμηση κατ' αύξουσα ιεραρχία για 5 αντικείμενα.

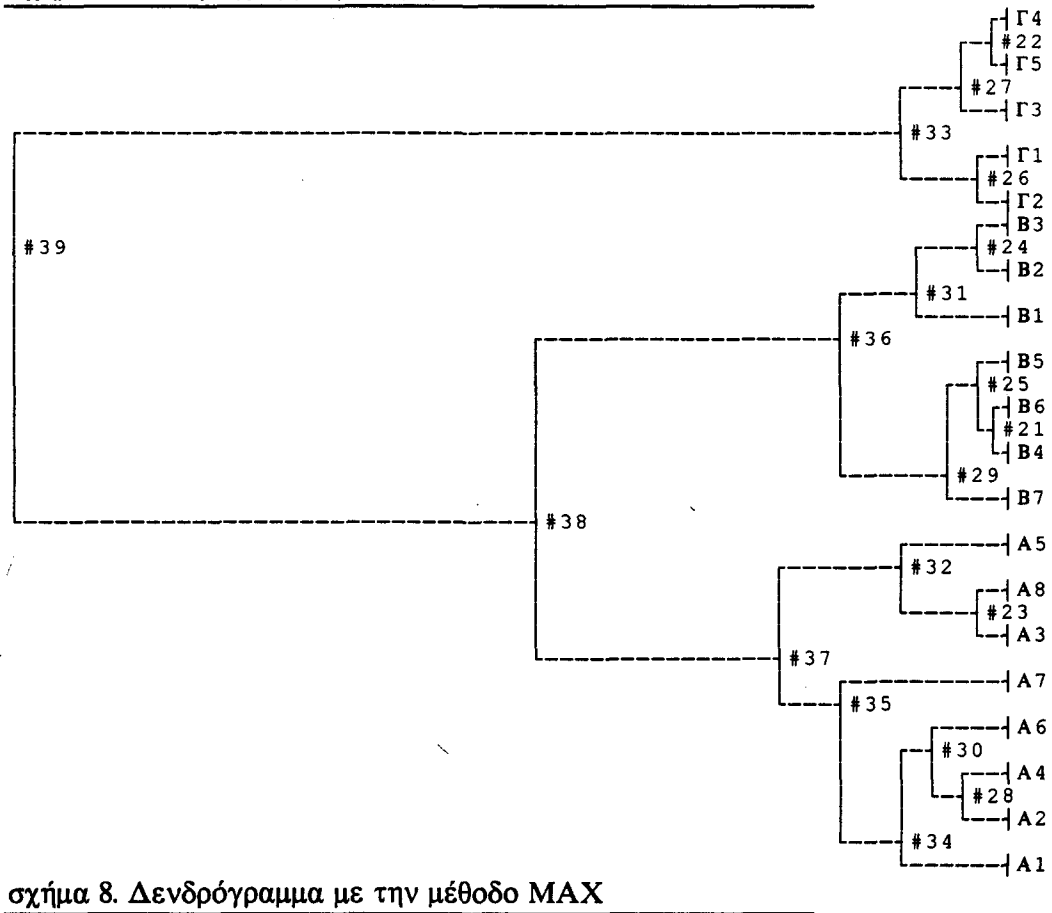
Το ύψος του κόμβου από το επίπεδο των τελικών κόμβων είναι ανάλογο με την απόσταση των δύο παρατηρήσεων που ενώθηκαν. Σε κάθε επίπεδο, ενώνουμε δύο κόμβους σχηματίζοντας ένα νέο, με απόσταση από το αρχικό επίπεδο, ανάλογη με την απόσταση των δύο ενωμένων κόμβων. Καταλήγουμε έτσι σε δύο μόνο κόμβους οι οποίοι ενώνονται και σχηματίζουν τη ρίζα του δένδρου. Η ρίζα αντιστοιχεί σε μια μόνο κλάση που περιλαμβάνει όλες τις παρατηρήσεις.

Τα δενδρογράμματα του παραδείγματος που αντιστοιχούν στις μεθόδους MIN, MAX και MOYENNE, φαίνονται στο σχήμα 7, 8 και 9. Παρατηρώντας το ύψος του κάθε κόμβου, και ειδικότερα το ύψος του κόμβου #38 που σχηματίζεται από τις δύο ομάδες των (A1,A2,A3,A4,A5,A6,A7,A8) και (B1,B2,B3,B4,B5,B6,B7), βλέπουμε την διαφοροποίηση μεταξύ των τριών μεθόδων. Με την MIN (σχήμα 7) οι δύο ομάδες είναι πολύ κοντά. Στο σχήμα 8 της μεθόδου MAX απέχουν περισσότερο ενώ για την μέθοδο MOYENNE (σχήμα 9) έχουν την μέση απόσταση. Ανάλογη παρατήρηση μπορούμε να κάνουμε και για κατώτερους κόμβους.

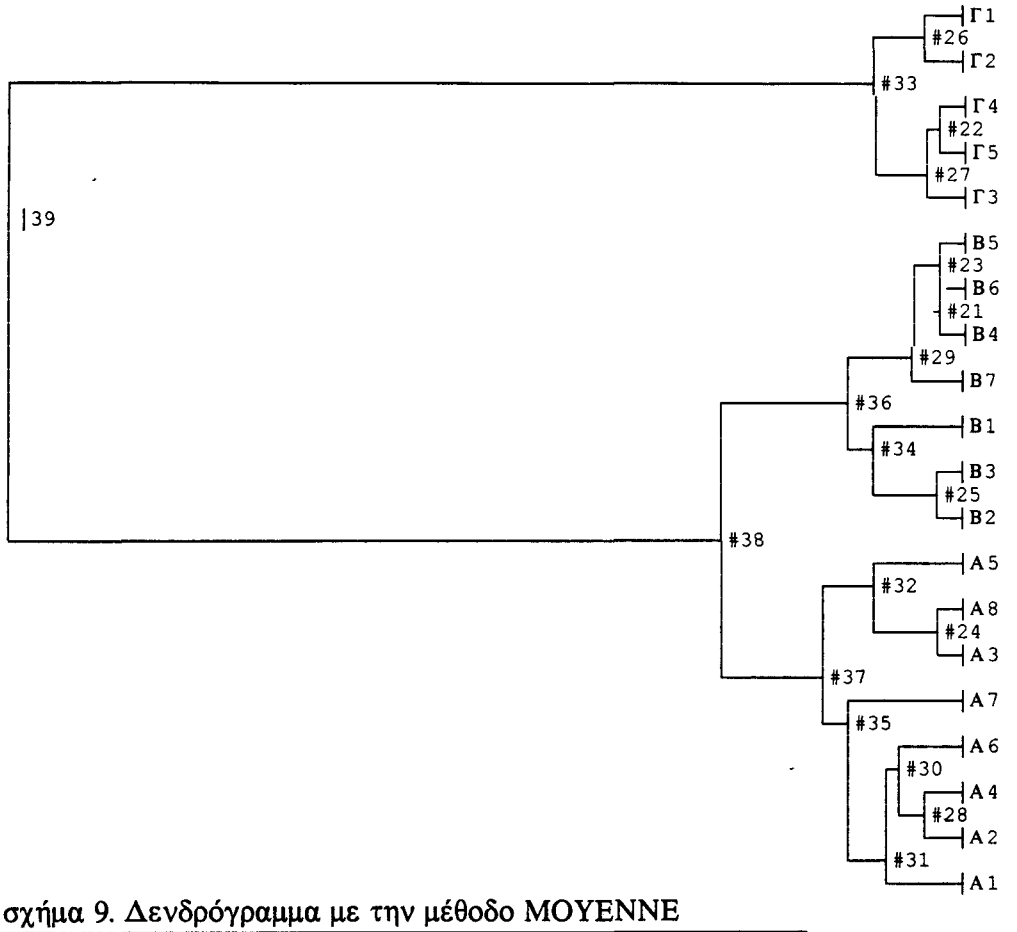
Κάθε φορά που "κόβουμε" το δενδρόγραμμα με μια ευθεία γραμμή σε κάποιο επίπεδο, οι κόμβοι που απομένουν σχηματίζουν ένα διαμελισμό του συνόλου των αρχικών παρατηρήσεων.



σχήμα 7. Δενδρόγραμμα με την μέθοδο MIN



σχήμα 8. Δενδρόγραμμα με την μέθοδο MAX



σχήμα 9. Δενδρόγραμμα με την μέθοδο MOYENNE

Προχωρώντας από πάνω προς τα κάτω, κάθε διαμελισμός προκύπτει από τον προηγούμενό του, διασπώντας μια κλάση του σε δύο.

"Κόβουμε" λοιπόν διαδοχικά το δενδρόγραμμα (σχήμα 7) με την ευθεία ε1 σχηματίζοντας στην αρχή διαμελισμό σε 2 κλάσεις, κατόπιν με την ευθεία ε2 σε 3 κλάσεις κτλ, μέχρι να φθάσουμε στον διαμελισμό σε n κλάσεις που είναι τα αρχικά μας δεδομένα. Οι αποστάσεις των ευθειών που ορίζουν οι κόμβοι, είναι ανάλογες με τις πραγματικές αποστάσεις τους. Π.χ. η απόσταση μεταξύ των κόμβων 38 και 35 είναι ανάλογη με την απόσταση μεταξύ των ευθειών ε1 και ε2 (σχήμα 7). Δηλαδή είναι πολύ μεγάλη σε σχέση με τις αποστάσεις μεταξύ των υπολοίπων κόμβων. Αυτό συμβαίνει γιατί οι κόμβοι αυτοί ορίζουν τις δύο περισσότερο ανομοιογενείς ομάδες.

8. Οι σχηματιζόμενοι διαμελισμοί

Οι διαδοχικοί διαμελισμοί που σχηματίζονται και με τις 3 μεθόδους για το προηγούμενο παράδειγμα, φαίνονται στα αντίστοιχα δενδρογράμματα (σχήμα 7,8 και 9). Όσο προχωράμε από τις αρχικές παρατηρήσεις προς την κορυφή, παρατηρούμε ότι η απόσταση των στοιχείων που ενώνονται αυξάνει, και επομένως μειώνεται η ομοιογένεια των ομάδων (πίνακες 9,10,11).

Πίνακας 9

	Διαμελισμοί με τη MOYENNE	
κλάσεις	αδράνεια διαμελισμού	ποσοστό αρχικής αδράνειας
2	47.7961	0.9148
3	50.8403	0.9731
4	51.2865	0.9816
5	51.5815	0.9873
6	51.7239	0.9900
7	51.8148	0.9917
8	51.9464	0.9943
9	52.0225	0.9957
10	52.0919	0.9971
11	52.1323	0.9978
12	52.1657	0.9985
13	52.1820	0.9988
14	52.1993	0.9991
15	52.2105	0.9993
16	52.2190	0.9995
17	52.2275	0.9996
18	52.2359	0.9998
19	52.2409	0.9999
20	52.2459	1.0000

Πίνακας 10

Διαμελισμοί με τη MIN	
αδράνεια διαμελισμού	ποσοστό αρχικής αδράνειας
47.7961	0.9148
50.8403	0.9731
50.9642	0.9755
51.0722	0.9775
51.2038	0.9801
51.4988	0.9857
51.7609	0.9907
52.0225	0.9957
52.0919	0.9971
52.1323	0.9978
52.1485	0.9981
52.1659	0.9985
52.1993	0.9991
52.2105	0.9993
52.2190	0.9995
52.2275	0.9996
52.2359	0.9998
52.2409	0.9999
52.2459	1.0000

Πίνακας 11

Διαμελισμοί με τη MAX	
αδράνεια διαμελισμού	ποσοστό αρχικής αδράνειας
47.7961	0.9148
50.8403	0.9731
51.2865	0.9816
51.5815	0.9873
51.7239	0.9900
51.7933	0.9913
51.9249	0.9939
52.0010	0.9953
52.0919	0.9971
52.1323	0.9978
52.1657	0.9985
52.1820	0.9988
52.1993	0.9991
52.2105	0.9993
52.2189	0.9995
52.2274	0.9996
52.2359	0.9998
52.2409	0.9999
52.2459	1.0000

Εξετάζουμε τώρα τα 3 δενδρογράμματα, αρχίζοντας την ανάλυσή τους από την κορυφή προς τα κάτω.

Για τον διαμελισμό σε 3 κλάσεις και με τις 3 μεθόδους, οδηγούμαστε σε 3 ομάδες. Η πρώτη περιέχει τα στοιχεία A1,A2,A3,A4,A5,A6,A7,A8 η δεύτερη τα B1,B2,B3,B4,B5,B6,B7 και η τρίτη τα Γ1,Γ2,Γ3,Γ4,Γ5.

Ο χωρισμός σε 4 κλάσεις είναι διαφορετικός για κάθε μέθοδο. Ενώ με την MOYENNE και τη MAX, η ομάδα Α χωρίζεται σε δύο υποομάδες (A1-A3-A8) και (A2-A4-A5-A6-A7) με τη μέθοδο MIN αποχωρίζεται το B1 από την υπόλοιπη ομάδα Β.

Στις 5 κλάσεις, με τις μεθόδους MAX και MOYENNE διασπάται η ομάδα Β σε δύο υποομάδες (B1-B2-B3) και (B4-B5-B6-B7), ενώ με τη μέθοδο MIN αποχωρίζεται το Α7 από την ομάδα Α.

Στις 6 κλάσεις, με τις μεθόδους MAX και MOYENNE αποχωρίζεται το Α7 από την υποομάδα της Α. ενώ με τη μέθοδο MIN διασπάται η ομάδα Γ, σε δύο υποομάδες: την (Γ1-Γ2) και (Γ3-Γ4-Γ5).

Σε 7 κλάσεις η εικόνα είναι διαφορετική και με τις 3 μεθόδους. Με τη μέθοδο MOYENNE χωρίζεται το B1 από τα B2-B3, με τη μέθοδο MAX το A1 από το A2-A4-A6 ενώ με τη μέθοδο MIN το Α5 από την υπόλοιπη ομάδα Α.

Στις 8 κλάσεις με τις μεθόδους MAX και MOYENNE διασπάται η ομάδα Γ σε δύο υποομάδες, την (Γ1-Γ2) και (Γ3-Γ4-Γ5), ενώ με τη μέθοδο MIN αποχωρίζεται το Α5 από την υπόλοιπη ομάδα Α.

Ο χωρισμός σε εννιά κλάσεις, με τις μεθόδους MOYENNE και MAX, γίνεται διαχωρίζοντας το Α5 από την υποομάδα των Α3-Α8, ενώ με την μέθοδο MIN, διασπώντας την υποομάδα της Α, σε δύο νέες υποομάδες, τις Α3-Α8 και Α1-Α2-Α4-Α6.

Στις 10 κλάσεις με τη μέθοδο MOYENNE και MIN διαχωρίζεται το Α1 ενώ με τη μέθοδο MAX το Β1. Ετσι προκύπτει ίδια ομαδοποίηση σε 10 ομάδες, και με τις τρεις μεθόδους.

Στις 11 κλάσεις και με τις 3 μεθόδους διαχωρίζεται το Α6 από την ομάδα που ανήκε.

Στις 12 κλάσεις με τις μεθόδους MAX και MOYENNE διαχωρίζεται το Β7 ενώ με τη μέθοδο MIN διαχωρίζονται τα Α2 και Α4.

Ο διαχωρισμός των Α2 και Α4 γίνεται με τις μεθόδους MAX και MOYENNE στις 13 κλάσεις, ενώ με τη μέθοδο MIN αποχωρίζεται το Γ3 από την υποομάδα Γ4-Γ5.

Ο χωρισμός του Γ3 γίνεται με τις μεθόδους MAX και MOYENNE στις 14 κλάσεις, ενώ με τη μέθοδο MIN αποχωρίζεται το Β7.

Ο διαμελισμός σε 14 κλάσεις είναι ίδιος τώρα και στις 3 μεθόδους και ο χωρισμός σε 15 κλάσεις γίνεται σ' όλες διαχωρίζοντας τα Γ1 και Γ2.

Αν θελήσουμε να παρατηρήσουμε τα αποτελέσματα της ιεράρχησης και με τις 3 μεθόδους, στο επίπεδο με άξονες τις δύο μεταβλητές, διαπιστώνουμε ότι για το συγκεκριμένο παράδειγμα οι μέθοδοι MAX και MOYENNE δίνουν περίπου τις ίδιες ομάδες με κάποιες μικρές διαφοροποιήσεις, ενώ η μέθοδος MIN διαχωρίζει από την αρχή κάποιο στοιχείο που είναι απομακρυσμένο από τα υπόλοιπα.(Π.χ. το Β1 από την υπόλοιπη ομάδα Β)

Πάντως ο διαμελισμός σε 3 κλάσεις είναι ίδιος και στις 3 μεθόδους και συμπίπτει με τις 3 ομάδες των αρχικών δεδομένων (σχήμα 6).

Οι μέθοδοι MAX και MOYENNE δίνουν τα ίδια αποτελέσματα για 4, 5 ή 6 κλάσεις και σαφώς καλύτερα από τη μέθοδο MIN (πίνακες 4,5,6). Καλύτερος διαμελισμός σε 7 κλάσεις είναι αυτός που παίρνουμε με τη μέθοδο MOYENNE, ακολουθεί της μεθόδου MAX με παρόμοια αποτελέσματα και τέλος ο διαμελισμός της μεθόδου MIN.

Ένας διαμελισμός θεωρείται τόσο καλύτερος, όσο η αδράνεια του, πλησιάζει στην αρχική ολική αδράνεια όπως εξηγείται στην επόμενη παράγραφο.

9. Τομή του δενδρογράμματος

Με μια ευθεία γραμμή (παράλληλη στις αρχικές παρατηρήσεις), κόβουμε το δενδρόγραμμα σε κάποιο σημείο. Οι κλάδοι που απομένουν κάτω από την ευθεία, αντιστοιχούν σε ομάδες ενός διαμελισμού με πλήθος κλάσεων, όσος και ο αριθμός των κλάδων.

Αν κόψουμε το δένδρο ψηλότερα έχουμε διαμελισμό με λιγότερο πλήθος κλάσεων, ενώ αν το κόψουμε χαμηλώτερα προκύπτει διαμελισμός με μεγαλύτερο πλήθος κλάσεων. Γενικά κόβοντας το δένδρο στον κόμβο k προκύπτει διαμελισμός με 2n-k κλάσεις. Μπορούμε λοιπόν να εκλέξουμε διαμελισμό με όποιο πλήθος κλάσεων επιθυμούμε, κόβοντας σε κατάλληλο σημείο το δενδρόγραμμα.

Το πρόβλημα που έχουμε είναι: Ποιό πλήθος είναι καταλληλότερο για ένα διαμελισμό των δεδομένων, ώστε ούτε οι κλάσεις να είναι πάρα πολλές αλλά και οι δημιουργούμενες ομάδες να ανταποκρίνονται στην πραγματικότητα;

Ένα δεύτερο πρόβλημα είναι εφόσον με κάθε μέθοδο παίρνουμε συνήθως διαφορετικούς διαμελισμούς, ποια μέθοδος είναι η καταλληλότερη;

Δυστυχώς δεν μπορούμε γενικά να συγκρίνουμε τις 3 μεθόδους, αφού μια μέθοδος που είναι ιδανική για κάποιο σύνολο δεδομένων δεν είναι τόσο καλή για κάποιο άλλο. Η σύγκριση των μεθόδων μπορεί να γίνει μόνο για κάποιο συγκεκριμμένο σύνολο δεδομένων. Ευτυχώς με την μεγάλη ανάπτυξη των υπολογιστών, είναι εύκολο και γρήγορο να υπολογίσουμε την ιεραρχία και με τις 3 μεθόδους και κατόπιν να εκλέξουμε την καταλληλότερη.

10. Κριτήριο για κάθε διαμελισμό σε r κλάσεις

Για το σύνολο των n παρατηρήσεων, στο χώρο των p διαστάσεων, υπολογίζουμε την αδράνεια ως προς το κέντρο βάρους τους g, με ένα από τους παρακάτω τύπους ανάλογα αν τα δεδομένα είναι ποσοτικά ή ποιοτικά.

Ευκλείδεια μετρική: $E_r = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[A(i,j) - g(j)\right]^2$

Μετρική του X^2: $E_r = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\left[A(i,j) - A(i,.)A(.,j)\right]^2}{A(i,.)A(.,j)}$

όπου A(i,.) και A(.,j) παριστάνουν τα αθροίσματα της i γραμμής και j στήλης αντίστοιχα.

Κάθε διαμελισμός ενός συνόλου παρατηρήσεων, δίνει διαφορετικό πλήθος κλάσεων, και τα στοιχεία των κλάσεων διαφέρουν από μέθοδο σε μέθοδο. Προσπαθούμε, λοιπόν, να βρούμε το κριτήριο εκείνο, που θα μας βοηθήσει να

ελέγξουμε αν ένας διαμελισμός ομαδοποιεί ικανοποιητικά τα αρχικά δεδομένα, και αν είναι καλύτερος ή όχι από κάποιον άλλο.

Για κάθε διαμελισμό που χωρίζει τα δεδομένα σε r πλήθος κλάσεις, υπολογίζουμε την αδράνεια (διασπορά) των r κλάσεων (E_e ή E_x), ως προς το κέντρο βάρους των αρχικών παρατηρήσεων, με τις παρακάτω σχέσεις όταν χρησιμοποιούμε την ευκλείδεια μετρική:

$$E_{re} = \frac{1}{n}\sum_{i=1}^{r} p(k)\sum_{j=1}^{p}\left[g_k(i,j) - g(j)\right]^2 \quad \text{ή} \quad E_{re} = \frac{1}{n}\sum_{k=1}^{r}\sum_{j=1}^{p} p(k) g_k{}^2(j) - n\sum_{j=1}^{p} g^2(j)$$

και όταν χρησιμοποιούμε τη μετρική του X^2 τη σχέση:

$$E_{rx} = \frac{1}{n}\sum_{k=1}^{r} p(k)\sum_{j=1}^{p} \frac{g_k(j) - \dfrac{A(.,j)}{n}\displaystyle\sum_{j=1}^{p}\frac{g_k(j)}{p(k)}}{\dfrac{A(.,j)}{n}\displaystyle\sum_{j=1}^{p}\frac{g_k(j)}{p(k)}}$$

όπου g_k είναι το διάνυσμα του κέντρου βάρους της k κλάσης, p(k) το πλήθος στοιχείων της και g το διάνυσμα του κέντρου βάρους όλων των δεδομένων.

Θεωρούμε ότι κάθε κλάση αποτελείται από ένα μόνο σημείο (κέντρο βάρους) και το σημείο αυτό σταθμίζεται με ένα αριθμό p(k) που αντιστοιχεί στο πλήθος των στοιχείων (βάρος) της κλάσης. (Θεωρούμε, δηλαδή, ότι όλα τα στοιχεία της κλάσης, βρίσκονται στο κέντρο βάρους της). Την αδράνεια που προαναφέραμε την συγκρίνουμε με την αρχική αδράνεια των παρατηρήσεων (πριν γίνει οποιαδήποτε ομαδοποίηση). Όσο περισσότερο πλησιάζουν οι δύο αυτές τιμές (ο λόγος των αδρανειών πλησιάζει στο 1), τόσο περισσότερο η δημιουργούμενη ομαδοποίηση προσαρμόζεται καλύτερα στα αρχικά δεδομένα.

11. Κριτήριο συνολικής ιεραρχικής ταξινόμησης

Εφαρμόζοντας τη μέθοδο C.A.H, σ' ένα πίνακα δεδομένων, προκύπτει η ολική ιεραρχική ταξινόμηση των παρατηρήσεων που παρουσιάζεται εποπτικά με το δενδρόγραμμα της ιεράρχησης. Για κάθε διαμελισμό, με αριθμό κλάσεων r, εφαρμόζουμε το κριτήριο της προηγούμενης παραγράφου, ξεκινώντας από τις δύο τελικές κλάσεις και φθάνοντας στις n αρχικές κλάσεις.

Στη συνέχεια, για κάθε διαμελισμό, υπολογίζουμε το λόγο

$$\lambda_r = \frac{\text{αδράνεια διαμελισμού}}{\text{αρχική αδράνεια}} \quad \text{(το r από 2 μέχρι n)}$$

για την ευκλείδεια μετρική: $$\lambda_r = \frac{\frac{1}{n}\sum_{k=1}^{r} p(k)\sum_{j=1}^{p}[g_k(j)-g(j)]^2}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}[A(i,j)-g(j)]^2}$$

και για την μετρική του X^2: $$\lambda_r = \frac{\frac{1}{n}\sum_{k=1}^{r} p(k)\sum_{j=1}^{p}\frac{g_k(j)-\frac{A(.,j)}{n}\sum_{j=1}^{p}\frac{g_k(j)}{p(k)}}{\frac{A(.,j)}{n}\sum_{j=1}^{p}\frac{g_k(j)}{p(k)}}}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{[A(i,j)-A(i,.)A(.,j)]^2}{A(i,.)A(.,j)}}$$

όπου, g_κ είναι το διάνυσμα του κέντρου βάρους της k κλάσης με πλήθος στοιχείων p(k) και g είναι το διάνυσμα του κέντρου βάρους όλων των δεδομένων.

Ο λόγος αυτός εκφράζει το ποσοστό της ολικής αδράνειας που ερμηνεύεται από τον συγκεκριμένο διαμελισμό.

Με τις τιμές αυτού του λόγου, σχηματίζουμε ένα διάγραμμα (διάγραμμα 1), όπου εποπτικά παρουσιάζεται η μεταβολή του λόγου αδράνειας του διαμελισμού ως προς την αρχική ολική αδράνεια, σαν συνάρτηση του πλήθους κλάσεων κάθε διαμελισμού. Παρατηρώντας το διάγραμμα που εκφράζει το ποσοστό της ερμηνευόμενης αδράνειας, ως προς το πλήθος κλάσεων του διαμελισμού, από δεξιά προς τα αριστερά, δηλαδή από τον διαμελισμό σε n κλάσεις προς την μία κλάση (που περιλαμβάνει όλες τις παρατηρήσεις), διαπιστώνεται ότι, ο λόγος λ_r, μειώνεται, όσο ελαττώνεται ο αριθμός των κλάσεων κάθε διαμελισμού.

Η μεταβολή του λόγου αυτού μεταξύ δύο διαδοχικών διαμελισμών, προκύπτει από τον παρακάτω τύπο:

d_r= (λόγος διαμέλ. r+1 κλάσεις)-(λόγος διαμέλ. r κλάσεων)

Ειδικά για την ευκλείδεια μετρική είναι

$$d_r = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{p}[A(i,j)-g(j)]^2}\left[\left[\frac{1}{n}\sum_{k=1}^{r+1}\sum_{j=1}^{p}p(k)g_k{}^2(j)-n\sum_{j=1}^{p}g^2(j)\right]-\left[\frac{1}{n}\sum_{k=1}^{r}\sum_{j=1}^{p}p'(k)g'^2_k(j)-n\sum_{j=1}^{p}g^2(j)\right]\right]$$

Επειδή οι ομάδες του διαμελισμού σε r κλάσεις είναι ίδιες μ' αυτές σε r+1 κλάσεις, εκτός από μία ομάδα (έστω την x) η οποία διασπάται σε δύο υποομάδες x′

και x″ στο διαμελισμό σε r+1 κλάσεις, ο τύπος της μεταβολής του λόγου d_r λαμβάνει την μορφή:

$$\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{p}[A(i,j)-g(j)]^2}\left[p(x')\sum_{j=1}^{p}g_x{}^2(j)+p(x'')\sum_{j=1}^{p}g_x{}^2(j)-p'(x)\sum_{j=1}^{p}g^2(j)\right]$$

και επειδή p′(x)=p(x′)+p(x″) ο τύπος λαμβάνει την τελική μορφή:

$$d_r=\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{p}[A(i,j)-g(j)]^2}\left[p(x')\left[\sum_{j=1}^{p}g_x{}^2(j)-\sum_{j=1}^{p}g^2(j)\right]+p(x'')\left[\sum_{j=1}^{p}g_x{}^2(j)-\sum_{j=1}^{p}g^2(j)\right]\right]$$

Παρατηρούμε λοιπόν, ότι η μεταβολή του ερμηνευομένου ποσοστού αδράνειας μεταξύ δύο διαδοχικών διαμελισμών, εξαρτάται μόνο από την διασπώμενη κλάση x καθώς και από τις δύο νέες υποκλάσεις στις οποίες διασπάται.

Όταν μεταβαίνοντας από το διαμελισμό σε r+1 κλάσεις στο διαμελισμό σε r κλάσεις, ο λόγος λ_r μεταβάλλεται απότομα, σημαίνει ότι επιλέγοντας το διαμελισμό με μία κλάση λιγότερη, χάνουμε σημαντικό ποσοστό πληροφορίας. Στο σημείο αυτό, δηλαδή στις r+1 κλάσεις έχουμε, χωρίς καμία a priori υπόθεση, τον πιο ικανοποιητικό διαμελισμό.

Επειδή δεν είναι εμφανής ο εντοπισμός των απότομων μεταβολών του ποσοστού της ερμηνευόμενης αδράνειας, από τη γραφική παράσταση του λόγου λ_r σε σχέση με τον αριθμό κλάσεων (διάγραμμα 1), προτείνουμε τη μελέτη της γραφικής παράστασης των διαφορών d_r μεταξύ των ποσοστών ερμηνευόμενης αδράνειας όλων των διαδοχικών διαμελισμών όπου εμφανίζονται πολύ πιο έντονα οι διακυμάνσεις του (διάγραμμα 2).

Από το ποιας μεθόδου, μεταξύ των MIN, MAX, MOYENNE, η μεταβολή του λόγου λr είναι πιο ομαλή, συμπεραίνουμε ποια απ' αυτές δίνει την καλύτερα προσαρμοσμένη στα δεδομένα συνολική ιεράρχηση. Καλύτερη είναι η μέθοδος εκείνη στην οποία ο λόγος έχει τιμή κοντά στο 1 για μεγαλύτερο αριθμό διαμελισμών (από δεξιά προς τα αριστερά στο διάγραμμα), και οι μεταβολές του λόγου από διαμελισμό σε διαμελισμό, είναι μικρές.

Επίσης προτείνουμε για κάθε μια από τις τρεις μεθόδους, τον προσδιορισμό του ιδανικότερου διαμελισμού των δεδομένων, χωρίς να είναι γνωστός εκ των προτέρων ο αριθμός των κλάσεων. Ο ιδανικότερος διαμελισμός είναι εκείνος στον οποίο παρατηρείται απότομη μεταβολή του λόγου αδράνειας του διαμελισμού προς

την αρχική ολική αδράνεια. Και αυτό γιατί αν δεχθούμε τον διαμελισμό με μια κλάση λιγότερη, χάνουμε σημαντικό ποσοστό ερμηνευόμενης αδράνειας, ενώ προχωρώντας στο διαμελισμό με μια κλάση περισσότερη διαχωρίζουμε τα δεδομένα σε επιπλέον κλάσεις χωρίς να κερδίζουμε μεγάλο ποσοστό ερμηνευόμενης αδράνειας.

12. Επιλογή διαμελισμού για συγκεκριμένο πλήθος κλάσεων

Πολλές φορές για διάφορους λόγους (είδος δεδομένων, βιβλιογραφία κλπ.), είναι σκόπιμο να διαμελισθούν τα δεδομένα σε συγκεκριμένο αριθμό, έστω x κλάσεων. Σ' αυτή την περίπτωση, για να βρούμε ποια από τις 3 μεθόδους δίνει τον καλύτερο διαμελισμό, ελέγχουμε το λόγο αδράνειας του διαμελισμού σε x κλάσεις προς την αρχική ολική αδράνεια, και με τις 3 μεθόδους. Ο λόγος αυτός υπολογίζεται από τον τύπο: $\lambda_x = \dfrac{\frac{1}{n}\sum_{k=1}^{x} p(k)\sum_{j=1}^{p}[g_k(j)-g(j)]^2}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}[A(i,j)-g(j)]^2}$

όταν για την απόσταση χρησιμοποιούμε την ευκλείδεια μετρική,

και $\lambda_x = \dfrac{\frac{1}{n}\sum_{k=1}^{x} p(k)\sum_{j=1}^{p}\dfrac{g_k(j)-\frac{A(.,j)}{n}\sum_{j=1}^{p}\frac{g_k(j)}{p(k)}}{\frac{A(.,j)}{n}\sum_{j=1}^{p}\frac{g_k(j)}{p(k)}}}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\dfrac{[A(i,j)-A(i,.)A(.,j)]^2}{A(i,.)A(.,j)}}$ για την απόσταση με τη X^2 μετρική

Η μέθοδος που παρουσιάζει την μεγαλύτερη τιμή του λόγου αυτού, δίνει τον ιδανικότερο διαμελισμό σε x κλάσεις.

Για να διαπιστωθεί αυτό πιο παραστατικά, απεικονίζεται γραφικά ο λόγος λ_x ως προς τον αριθμό κλάσεων, στο ίδιο διάγραμμα (διάγραμμα 1) και για τις 3 μεθόδους. Η μέθοδος της οποίας η γραφική παράσταση για το συγκεκριμένο αριθμό κλάσεων βρίσκεται ψηλότερα από τις υπόλοιπες, δίνει τον ιδανικότερο διαμελισμό.

Είναι δυνατό μια μέθοδος να θεωρηθεί συνολικά καλλίτερη για τα δεδομένα μας ενώ μια άλλη να ερμηνεύει μεγαλύτερο ποσοστό αρχικής ολικής αδράνειας για το συγκεκριμμένο πλήθος κλάσεων.

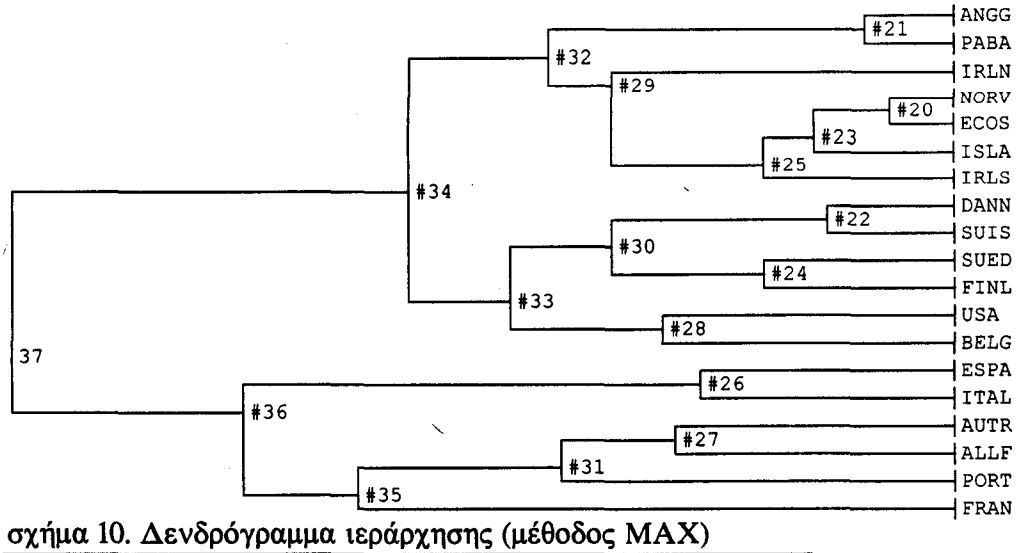
13. Παράδειγμα: Ταξινόμηση 19 χωρών ως προς τις αιτίες θανάτου (ευκλείδεια μετρική)

Εφαρμόσαμε τις 3 μεθόδους, σε ένα πίνακα διαστάσεων 19x6 (πίνακας 12), χρησιμοποιώντας την ευκλείδεια μετρική. Κάθε γραμμή αντιστοιχεί σε μία χώρα, και κάθε στήλη σε μια αιτία θανάτου. Στη διασταύρωση της γραμμής i και της στήλης j, αντιστοιχεί το πλήθος ατόμων της χώρας i που πέθαναν λόγω της j αιτίας. (βιβλίο Todd "Le fou et le proletaire" 1979 annexe 2, σελ 283) [Roux,1985].

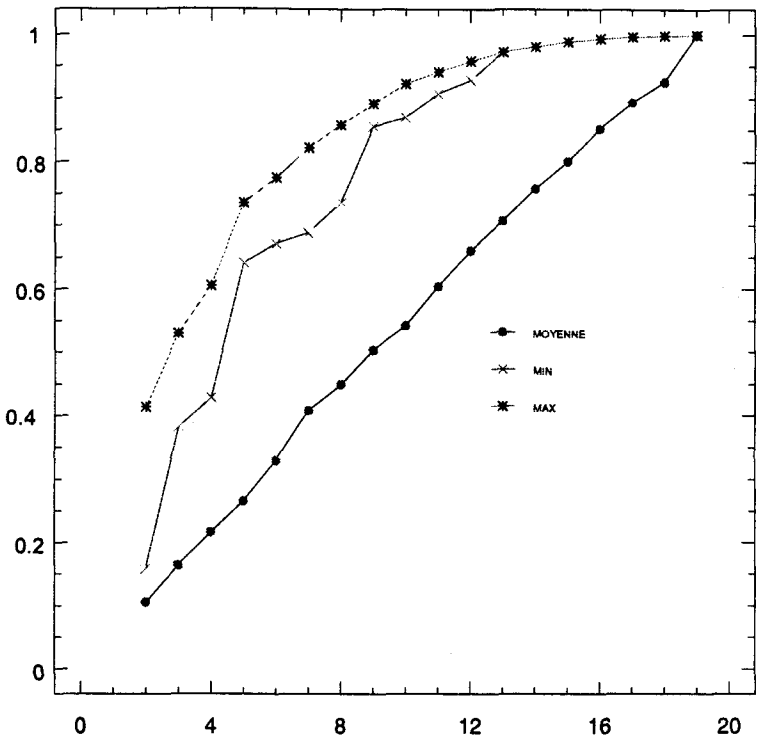
Πίνακας 12
Οι 6 μεταβλητές για τις 19 χώρες

AUTRICHE	241	16	330	43	363	325
FRANCE	156	9	225	10	535	328
PORTUGAL	85	19	349	7	281	345
ALLFEDER	210	12	230	21	298	269
BELGIQUE	156	10	260	13	367	144
FINLANDE	251	26	180	29	387	55
SUEDE	194	11	151	13	384	122
SUISSE	225	9	195	26	276	128
ITALIE	54	11	219	19	224	319
IRLNORD	40	136	215	18	320	43
DANNEMAR	241	6	168	11	230	107
ISLANDE	101	5	179	23	380	9
ECOSSE	82	15	155	18	342	59
ESPAGNE	40	4	136	17	237	225
NORVEGE	104	6	138	22	346	41
IRLSUD	38	7	182	32	314	37
PABAS	89	7	169	10	218	47
ANGGALLE	79	10	130	14	203	36
USA	121	102	220	26	273	158

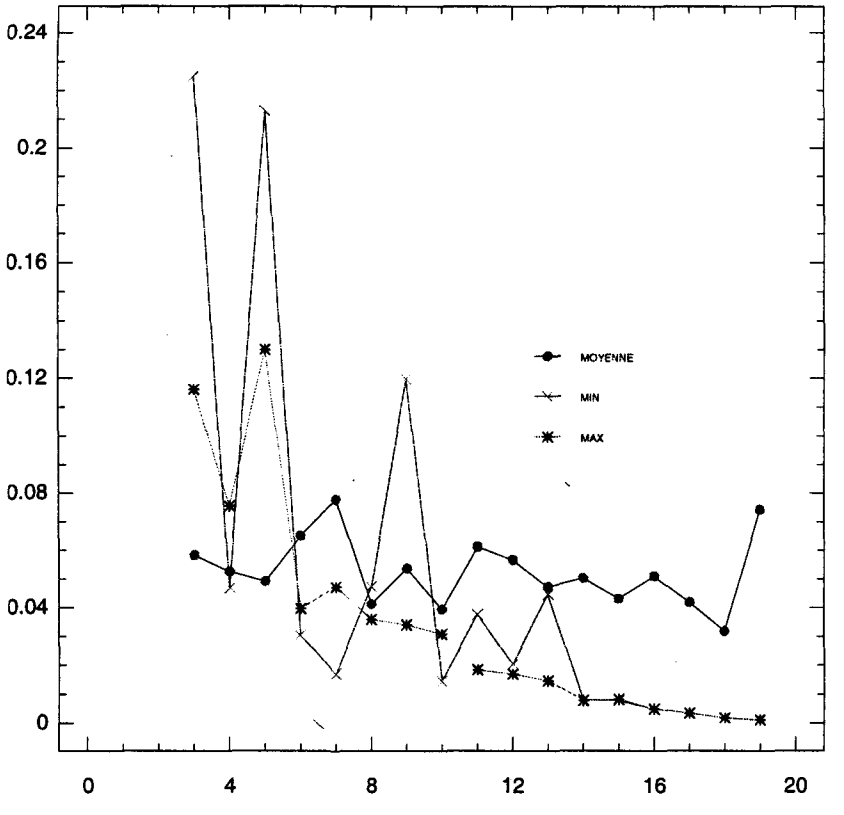
Το δένδρο ιεράρχησης (με τη μέθοδο MAX) για τα δεδομένα του παραπάνω πίνακα, φαίνεται στο σχήμα 10.



σχήμα 10. Δενδρόγραμμα ιεράρχησης (μέθοδος MAX)



διάγραμμα 1. Ερμηνευόμενο ποσοστό αδράνειας



διάγραμμα 2. Μεταβολές του ερμηνευόμενου ποσοσστού αδράνειας

Στο διάγραμμα 1 παριστάνεται το ερμηνευόμενο ποσοστό αδράνειας, σε σχέση με το πλήθος κλάσεων κάθε διαμελισμού και για τις 3 μεθόδους. Παρατηρούμε ότι η μέθοδος MAX παρουσιάζει μεγαλύτερες τιμές του ερμηνευόμενου ποσοστού αδράνειας, για κάθε πλήθος κλάσεων. Συμπεραίνουμε, λοιπόν, ότι για τα συγκεκριμένα δεδομένα συνολικά καλύτερη θεωρείται η ιεράρχηση με τη μέθοδο MAX.

Στη συνέχεια στο διάγραμμα 2 παριστάνονται γραφικά οι μεταβολές (διαφορές) του ποσοστού ερμηνευόμενης αδράνειας μεταξύ δύο διαδοχικών διαμελισμών, ως προς το πλήθος κλάσεων τους.

Στο πρώτο διάγραμμα (διάγραμμα 1) οι μεταβολές των ποσοστών είναι πολύ μικρές σχεδόν ανεπαίσθητες, ενώ στο δεύτερο (διάγραμμα 2) παρατηρούμε εύκολα ότι:

Με τη μέθοδο MOYENNE και για διαμελισμό σε 5 ή περισσότερες κλάσεις, η καμπύλη των διαφορών είναι αρκετά ομαλή και οι μεταβολές παρουσιάζουν ελάχιστες διαφοροποιήσεις. Όσο όμως ελλατώνεται το πλήθος των κλάσεων, οι διαφοροποιήσεις γίνονται πιο σημαντικές. Αυτό δείχνει πιο φανερά ότι χρησιμοποιώντας την μέθοδο MOYENNE, ο ιδανικότερος διαμελισμός των παρατηρήσεων είναι σε 5 κλάσεις όπου ερμηνεύεται ποσοστό 72.92% (πίνακας 13) της αρχικής ολικής αδράνειας των παρατηρήσεων.

Με τη μέθοδο MAX η καμπύλη των διαφορών είναι πιο ομαλή και οι διαφοροποιήσεις είναι μικρότερες. Ο ιδανικότερος διαμελισμός, ως προς αυτή τη μέθοδο, είναι σε 5 κλάσεις, όπου ερμηνεύεται ποσοστό 73.55% (πίνακας 14) της αρχικής ολικής αδράνειας των παρατηρήσεων.

Τέλος για τη μέθοδο MIN, παρατηρώντας το διάγραμμα 2 βλέπουμε την μη ομαλότητα της καμπύλης. Αυτό κάνει ακόμη πιο έντονη την μη καταλληλότητα της μεθόδου για τα συγκεκριμένα δεδομένα.

Πίνακας 13

	Διαμελισμοί με τη MOYENNE	
κλάσεις	αδράνεια διαμελισμού	ποσοστό αρχικής αδράνειας
2	73550.9132	0.1052
3	114311.6061	0.1635
4	151058.6433	0.2161
5	185506.3934	0.2653
6	231015.2484	0.3304
7	285300.5980	0.4081
8	314137.1390	0.4493
9	351774.2501	0.5031
10	379301.5425	0.5425
11	422151.0803	0.6038
12	461777.1895	0.6605
13	494853.8992	0.7078
14	530106.4059	0.7582
15	560260.0538	0.8013
16	595908.1679	0.8523
17	625094.7858	0.8941
18	647336.3935	0.9259
19	699160.2187	1.0000

Πίνακας 14

	Διαμελισμοί με τη MIN	
	αδράνεια διαμελισμού	ποσοστό αρχικής αδράνειας
	4590.9207	0.1569
	11173.9845	0.3818
	12542.0108	0.4286
	18769.0966	0.6414
	19656.8827	0.6717
	20152.1985	0.6886
	21540.7400	0.7361
	25042.7599	0.8557
	25467.5757	0.8703
	26569.3301	0.9079
	27162.7073	0.9282
	28467.3871	0.9728
	28706.0450	0.9809
	28935.8652	0.9888
	29081.0055	0.9937
	29180.3740	0.9971
	29232.7950	0.9989
	29264.6371	1.0000

Πίνακας 15

	Διαμελισμοί με τη MAX	
	αδράνεια διαμελισμού	ποσοστό αρχικής αδράνειας
	12115.9651	0.4140
	15513.3993	0.5301
	17719.7808	0.6055
	21522.8251	0.7355
	22683.6277	0.7751
	24063.4503	0.8223
	25112.7661	0.8581
	26107.0161	0.8921
	27008.0450	0.9229
	27547.2555	0.9413
	28042.5713	0.9582
	28467.3871	0.9728
	28697.2073	0.9806
	28935.8652	0.9888
	29081.0055	0.9937
	29180.3740	0.9971
	29232.7950	0.9989
	29264.6371	1.0000

14. Παράδειγμα: Ταξινόμηση των 10 περιοχών της Ελλάδας ως προς την εκπαίδευση (X^2 μετρική)

Εφαρμόζουμε πάλι τις 3 μεθόδους σε ένα πίνακα δεδομένων διαστάσεων 10x10 (πίνακας 16), χρησιμοποιώντας τώρα την μετρική X^2 αφού τα δεδομένα που έχουμε είναι συχνότητες. Κάθε γραμμή αντιστοιχεί σε ένα γεωγραφικό διαμέρισμα της Ελλάδας και κάθε στήλη στα είδη μόρφωσης των κατοίκων άνω των 10 ετών (πτυχιούχοι, φοιτητές, απόφοιτοι μόνο δημοτικού, αγράμματοι κτλ.). Τα στοιχεία αφορούν στην απογραφή του 1991 και πάρθηκαν από την Στατιστική επετηρίδα της Ελλάδας 1994-95. Στη διασταύρωση της γραμμής i και στήλης j, βρίσκεται το πλήθος ατόμων του i διαμερίσματος για την j μεταβλητή (συχνότητα).

Πίνακας 16
Οι συχνότητες των 10 μεταβλητών για τα 10 γεωγραφικά διαμερίσματα

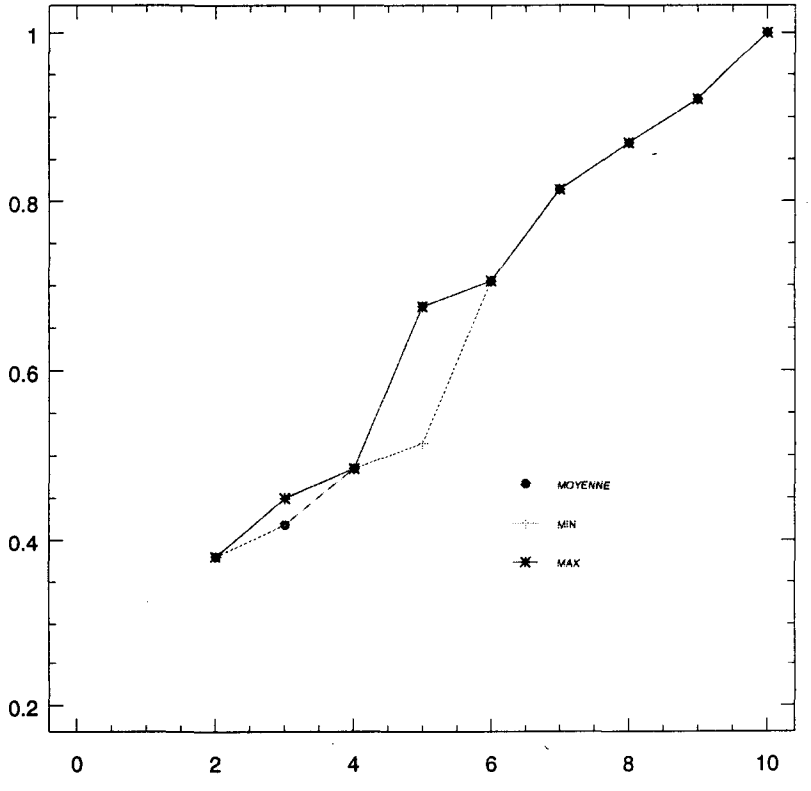
περιοχή	μεταπτυχιακοί	πτυχίο Πανεπ.	πτυχίο ΤΕΙ	πτυχίο σχολής	φοιτητές	μέση εκπαίδ.	Γ τάξη	απόφοιτοι δημοτικού	δεν τελείωσαν δημοτικό	αγράμματοι
ATTIKH	22248	259672	46881	31137	107384	803021	308025	857211	201531	95265
STEREA	2507	46806	11983	6657	19879	196314	124513	467980	124740	102286
PELOPON	1920	43012	9587	4561	19630	165926	109289	426179	106993	67151
IONIO	288	6669	1495	890	1455	25170	19101	78118	19330	18679
HPEIROS	594	14509	3240	1162	7911	37496	29123	135688	37346	32931
THESSALIA	902	31653	8061	2771	11626	96324	63897	275805	83433	69034
MAKEDONIA	6101	111059	28356	7693	63768	353908	205536	823462	244000	127252
THRAKI	431	11411	3103	909	5555	38075	22821	126513	42433	44799
AIGAIO	794	16725	3873	3456	4120	69945	41941	181007	49441	28458
KRHTH	1080	22495	6334	2929	10422	80315	49372	211334	51473	31791

Το ποσοστό ερμηνευόμενης αδράνειας σε συνάρτηση με το πλήθος κλάσεων φαίνεται στον πίνακα 17, για τις 3 μεθόδους. Από τον πίνακα αυτό, παρατηρούμε ότι η μόνη σημαντική διαφορά μεταξύ των 3 μεθόδων βρίσκεται στον διαμελισμό

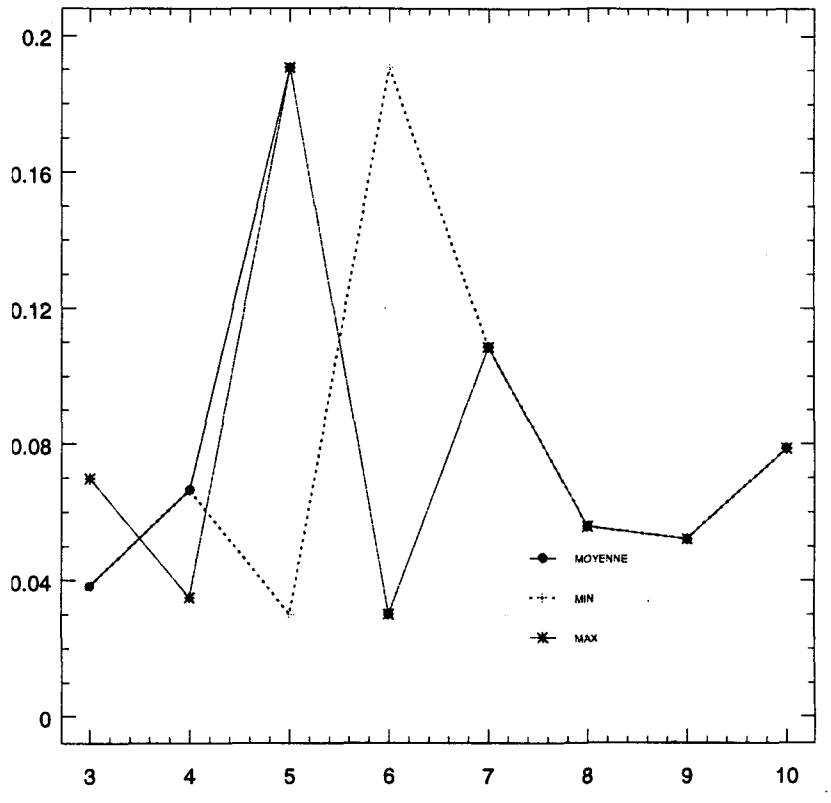
σε 5 κλάσεις όπου οι μέθοδοι MOYENNE και MAX ερμηνεύουν μεγαλύτερο ποσοστό (67.51%) αρχικής αδράνειας απ' ότι η μέθοδος MIN (51.45%).

Πίνακας 17
Ποσοστό αδράνειας για τις 3 μεθόδους

πλήθος κλάσεων	μέθοδος MOYENNE	μέθοδος MIN	μέθοδος MAX
2	0.3798	0.3798	0.3798
3	0.4181	0.4181	0.4496
4	0.4845	0.4845	0.4845
5	0.6751	0.5145	0.6751
6	0.7051	0.7051	0.7051
7	0.8135	0.8135	0.8135
8	0.8692	0.8692	0.8692
9	0.9214	0.9214	0.9214
10	1.000	1.000	1.000



διάγραμμα 3. Ποσοστό ερμηνευόμενης αδράνειας



διάγραμμα 4. Μεταβολή του ερμηνευόμενου ποσοστού αδράνειας

Στο διάγραμμα 3 παριστάνεται το ερμηνευόμενο ποσοστό αδράνειας, σε σχέση με το πλήθος κλάσεων κάθε διαμελισμού, και για τις 3 μεθόδους και για πλήθος κλάσεων από 2 μέχρι 10. Παρατηρούμε ότι η μέθοδος MIN παρουσιάζει κάποια διαφοροποίηση για 5 κλάσεις και ότι τα αποτελέσματα των μεθόδων MOYENNE και MAX συμπτίπτουν για πλήθος κλάσεων από 10 έως 4. Ο χωρισμός σε δύο, τρεις ή τέσσερεις κλάσεις, είναι ανεπαρκής για τα δεδομένα, γιατί το ποσοστό ερμηνευόμενης αδράνειας είναι πολύ μικρό.

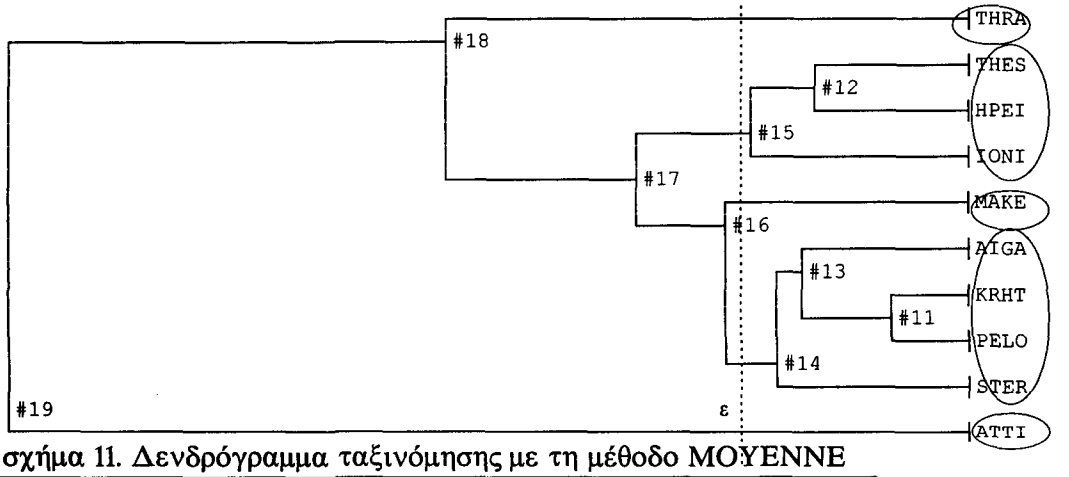
Συνολικά καλύτερη μέθοδος για τα συγκεκριμμένα δεδομένα, για κάθε πλήθος κλάσεων θεωρείται η ταξινόμηση με την μέθοδο MAX, αφού ο λόγος ερμηνευόμενης αδράνειας είναι ομαλώτερος για την μέθοδο αυτή.

Στο διάγραμμα 4 απεικονίζουμε τις μεταβολές του ερμηνευόμενου ποσοστού αδράνειας μεταξύ δύο διαδοχικών διαμελισμών, ως προς το πλήθος κλάσεων τους, και με τις 3 μεθόδους. Σ' αυτό διακρίνουμε εύκολα τις απότομες μεταβολές του ποσοστού ερμηνευόμενης αδράνειας, που παριστάνονται με τις μεγάλες μεταβολές της γραφικής παράστασης.

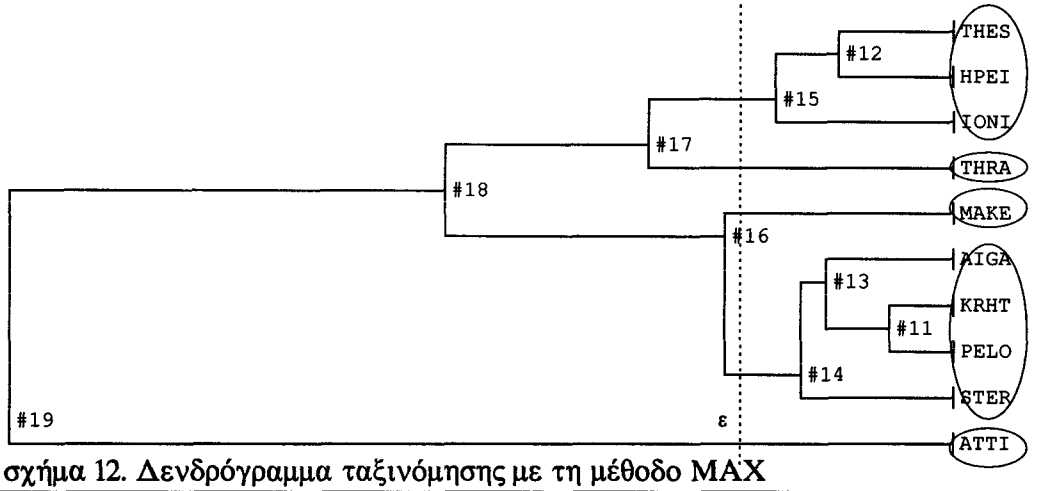
Με τη μέθοδο MOYENNE και MAX ο ιδανικότερος διαμελισμός των συγκεκριμένων δεδομένων επιτυγχάνεται χωρίζοντάς τα σε 5 κλάσεις με τις οποίες ερμηνεύεται ποσοστό 67.51% της αρχικής ολικής αδράνειας των δεδομένων.

Με τη μέθοδο MIN ο ιδανικότερος διαμελισμός είναι εκείνος που παίρνουμε χωρίζοντας τα δεδομένα σε 6 κλάσεις με ποσοστό ερμηνευόμενης αδράνειας 70.51%.

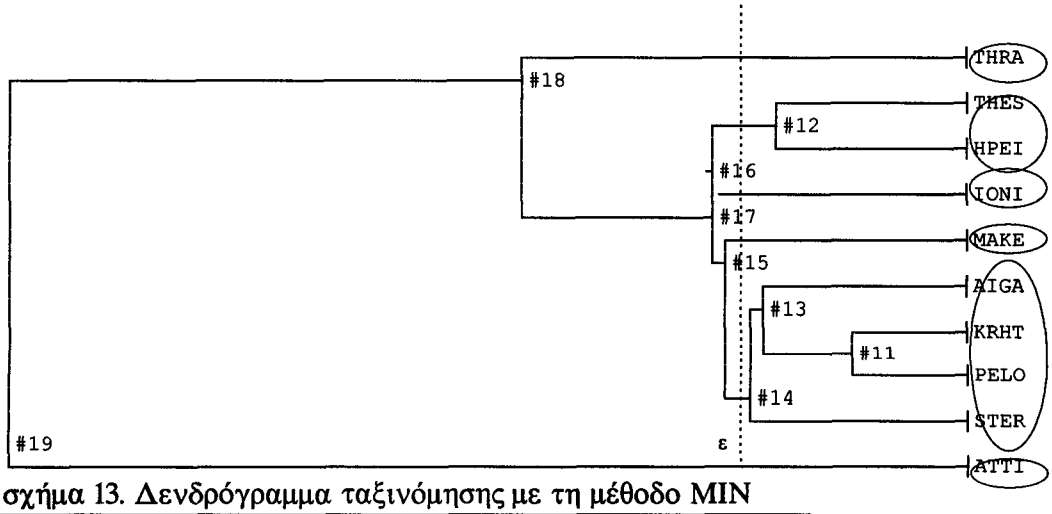
Τα δενδρογράμματα ταξινόμησης για τις τρεις μεθόδους, φαίνονται αντίστοιχα στο σχήμα 11, 12 και 13. Στα σχήματα αυτό φαίνεται και η ιδανικότερη τομή του κάθε δενδρογράμματος με την ευθεία ε. Για τις μεθόδους MOYENNE και MAX κόβουμε το δενδρόγραμμα στον κόμβο 16 (5 κλάσεις), και σχηματίζονται οι εξής ομαδοποιήσεις: {Αττική} {Θράκη} {Θεσσαλία, Ήπειρος, Ιόνιο} {Μακεδονία} {Αιγαίο, Κρήτη, Πελοπόνησος, Στερεά}. Για την μέθοδο MIN η ιδανικότερη τομή γίνεται στον κόμβο 15 (6 κλάσεις) και παίρνουμε τις ομαδοποιήσεις: {Αττική} {Θράκη} {Θεσσαλία, Ήπειρος} {Ιόνιο} {Μακεδονία} {Αιγαίο, Κρήτη, Πελοπόνησος, Στερεά}. Η διαφορά δηλαδή βρίσκεται στον διαχωρισμό του Ιονίου από την κλάση των περιοχών {Θεσσαλία, Ήπειρος}. Και αυτό συμβαίνει γιατί με την μέθοδο MIN πρώτα δημιουργείται η κλάση που ενώνει την Μακεδονία με την κλάση {Αιγαίο, Κρήτη, Πελοπόνησος, Στερεά} και κατόπιν η κλάση {Θεσσαλία, Ήπειρος, Ιόνιο} ενώ με τις άλλες δύο μεθόδους η πορεία είναι ανάστροφη.



σχήμα 11. Δενδρόγραμμα ταξινόμησης με τη μέθοδο MOYENNE



σχήμα 12. Δενδρόγραμμα ταξινόμησης με τη μέθοδο MAX



σχήμα 13. Δενδρόγραμμα ταξινόμησης με τη μέθοδο MIN

15. Ταξινόμηση μεταβλητών

Έχοντας ένα πίνακα συμπτώσεων μεταξύ κατηγοριών των μεταβλητών, μπορούμε να προβούμε σε μια ιεραρχική ταξινόμηση μεταβλητών, και να συνάγουμε συμπεράσματα που αφορούν στις σχέσεις των κατηγοριών των μεταβλητών. [Tallur,1988] [Daude,1992]. Θα δούμε την διαδικασία αυτή, με ένα παράδειγμα.

Έχουμε τον επόμενο πίνακα 18, συμπτώσεων Burt, για τα δεδομένα 1056 πολυπληθών καθολικών οικογενειών στο Quebec. Μελετήσαμε τον αριθμό παιδιών των οικογενειών σε σχέση με την ηλικία γάμου της μητέρας (δύο κλάσεις), σε σχέση με τα χρόνια εκπαίδευσης τους, με το εισόδημα της οικογένειας (κλάσεις χαμηλό και υψηλό) και επίσης σε σχέση με την απόσταση κατοικίας από την πλησιέστερη πόλη (αν είναι μακρυα ή κοντά). Τα δεδομένα πάρθηκαν από το βιβλίο του Andrews [Andrews,1985], και παρουσιάζονται στον πίνακα μετά από μια προσαρμογή για να δημιουργηθεί ο πίνακας Burt.

Πίνακας 18

	gam15-19	gam20-24	sx0-6	sx7	eislow	eishigh	aposm	aposk	paid7-9	paid9-11	paid11-14
gam15-19	399	0	230	169	163	236	248	151	54	181	164
gam20-24	0	657	297	360	363	294	312	345	316	301	40
sx0-6	230	297	527	0	248	279	321	206	192	281	54
sx7	169	360	0	529	278	251	239	290	178	201	150
eislow	163	363	248	278	526	0	273	253	243	239	44
eishigh	236	294	279	251	0	530	287	243	127	243	160
aposm	248	312	321	239	273	287	560	0	73	322	165
aposk	151	345	206	290	253	243	0	496	297	160	39
paid7-9	54	316	192	178	243	127	73	297	370	0	0
paid9-11	181	301	281	201	239	243	322	160	0	482	0
paid11-14	164	40	54	150	44	160	165	39	0	0	204

Κατόπιν εφαρμόσαμε την μέθοδο της ταξινόμησης κατ' αύξουσα ιεραρχία, στις γραμμές αυτού του πίνακα, χρησιμοποιώντας την ευκλείδεια απόσταση και την μέθοδο MOYENNE, και πήραμε το δενδρόγραμμα που φαίνεται στο σχήμα 14. Χωρίζοντας τα δεδομένα σε 2 κλάσεις ερμηνεύεται ποσοστό 34.18% της ολικής αδράνειας, ενώ χωρίζοντάς τα σε 3 κλάσεις έχουμε ερμηνεία της ολικής αδράνειας κατά 51.48%, όπως φαίνεται στο πίνακα 19.

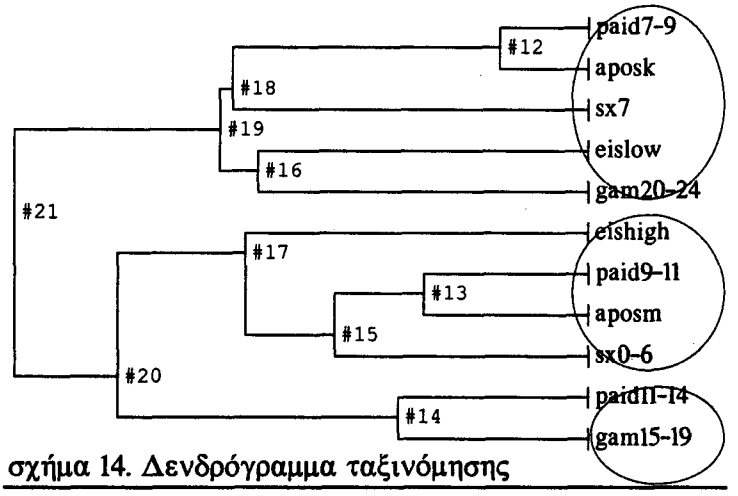
Πίνακας 19
Ποσοστό αδράνειας για τις 3 μεθόδους

πλήθος κλάσεων	μέθοδος MOYENNE	μέθοδος MIN	μέθοδος MAX
2	0.3418	0.3418	0.3418
3	0.5148	0.4266	0.5148
4	0.6071	0.5996	0.6168
5	0.6904	0.6773	0.6904
6	0.7681	0.7495	0.7681
7	0.8328	0.8328	0.8328
8	0.8906	0.8906	0.8906
9	0.9325	0.9325	0.9325
10	0.9710	0.9710	0.9710
11	1.0000	1.0000	1.0000

Παρατηρούμε ότι αρχικά οι οικογένειες χωρίζονται σε δύο μεγάλες κατηγορίες. Η πρώτη περιλαμβάνει τις οικογένεις που έχουν από 9 έως 14 παιδιά, μένουν μακρυά από πόλη και η ηλικία γάμου της μητέρας είναι από 15 έως 19 χρόνων. Η δεύτερη κατηγορία περιλαμβάνει τις οικογένειες με αριθμό παιδιών από 7 έως 9, οι οποίες διαμένουν κοντά σε πόλη και οι μητέρες έχουν παντρευτεί μεταξύ 20 και 24 χρόνων. Η πρώτη κατηγορία χωρίζεται περαιτέρω σε άλλες δύο υποομάδες. Η μία για τις οικογένειες που έχουν 11-14 παιδιά και ηλικία γάμου μητέρας από 15 έως 19 ετών, και η άλλη περιλαμβάνει τις οικογένειες με 9-11 παιδιά υψηλού εισοδήματος και οι οποίες κατοικούν μακρυά από πόλη.

Στο δενδρόγραμμα παρατηρούμε ακόμη ότι στην πρώτη κατηγορία ανήκουν οικογένειες με υψηλό εισόδημα και 0 έως 6 χρόνια εκπαίδευσης, ενώ στην δεύτερη οικογένειες με επίπεδο εκπαίδευσης άνω των 7 χρόνων και χαμηλό εισόδημα.

Βέβαια οι μεταβλητές αυτές χαρακτηρίζουν την κάθε κατηγορία, αλλά δεν είναι αυτές που κυρίως συμβάλλουν στον διαχωρισμό τους. Παρατηρώντας τον πίνακα με τα βοηθήματα ερμηνείας VACOR, όπως θα περιγράψουμε στο επόμενο κεφάλαιο, συμπεραίνουμε ότι οι οικογένειες διαχωρίζονται κυρίως ως προς τον αριθμό παιδιών, την ηλικία γάμου της μητέρας και την απόσταση κατοικίας από πόλη.



σχήμα 14. Δενδρόγραμμα ταξινόμησης

16. Συμπεράσματα

Για κάθε συγκεκριμένο σύνολο δεδομένων ποσοτικών ή ποιοτικών μεταβλητών, προτείνουμε ένα τρόπο προσδιορισμού ποια από τις τρεις μεθόδους ταξινόμησης (MIN, MAX, MOYENNE), οδηγεί στην ιεράρχηση που είναι καλύτερα προσαρμοζόμενη σ' αυτά τα δεδομένα. Αυτό επιτυγχάνεται με τον υπολογισμό του λόγου της αδράνειας του κάθε διαμελισμού προς την αρχική ολική αδράνεια, σε συνδυασμό με τη μελέτη δύο γραφικών παραστάσεων. Στην πρώτη απεικονίζεται γραφικά ο λόγος λ_r, που εκφράζει το ποσοστό της ολικής αδράνειας που ερμηνεύεται από τον κάθε διαμελισμό, ως προς τον αριθμό κλάσεων και για τις 3 μεθόδους στο ίδιο διάγραμμα, ενώ στη δεύτερη οι μεταβολές d_r του λόγου αυτού μεταξύ διαδοχικών διαμελισμών, πάλι σε κοινό διάγραμμα και για τις τρεις μεθόδους.

Καλύτερη μέθοδος για τα δεδομένα θεωρείται εκείνη, στην οποία ο λόγος λ_r έχει υψηλότερες τιμές για μεγαλύτερο σύνολο κλάσεων και ακόμη εκείνη για την οποία οι μεταβολές d_r του λόγου αυτού μεταξύ διαδοχικών διαμελισμών είναι ομαλώτερες.

Επίσης προτείνουμε για κάθε μια από τις τρεις μεθόδους, τον προσδιορισμό του ιδανικότερου διαμελισμού των δεδομένων, χωρίς να είναι γνωστός εκ των προτέρων ο αριθμός των κλάσεων. Ο ιδανικότερος διαμελισμός είναι εκείνος στον οποίο παρατηρείται απότομη μεταβολή του λόγου αδράνειας του διαμελισμού προς

την αρχική ολική αδράνεια. Και αυτό γιατί αν δεχθούμε τον διαμελισμό με μια κλάση λιγότερη, χάνουμε σημαντικό ποσοστό ερμηνευόμενης αδράνειας, ενώ προχωρώντας στο διαμελισμό με μια κλάση περισσότερη διαχωρίζουμε τα δεδομένα σε επιπλέον κλάσεις χωρίς να κερδίζουμε μεγάλο ποσοστό ερμηνευόμενης αδράνειας.

Τέλος επιθυμώντας τον διαχωρισμό των δεδομένων σε συγκεκριμένο πλήθος κλάσεων, προσδιορίζουμε ποια μέθοδος δίνει διαμελισμό με τον οποίο ερμηνεύεται το μεγαλύτερο ποσοστό της αρχικής ολικής αδράνειας. Αυτό επιτυγχάνεται συγκρίνοντας τους λόγους αδράνειας του διαμελισμού προς την αρχική ολική αδράνεια μεταξύ των 3 μεθόδων, για το συγκεκριμένο πλήθος κλάσεων. Η μέθοδος που παρουσιάζει μεγαλύτερη τιμή του λόγου αυτού, δίνει τον καλύτερο διαμελισμό σ' αυτό το πλήθος κλάσεων.

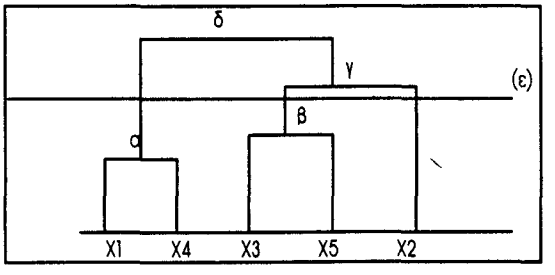
ΚΕΦΑΛΑΙΟ 5
ΕΡΜΗΝΕΙΑ ΤΟΥ ΔΕΝΔΡΟΓΡΑΜΜΑΤΟΣ ΤΗΣ ΤΑΞΙΝΟΜΗΣΗΣ (ΜΕΘΟΔΟΣ VACOR)

1. Εισαγωγή

Στο κεφάλαιο αυτό ασχολούμαστε με τους τρόπους ερμηνείας των αποτελεσμάτων μιας ταξινόμησης. Η ερμηνεία των αποτελεσμάτων είναι απαραίτητη για την εξαγωγή συμπερασμάτων, αλλά και για να καθοριστούν με ακρίβεια οι σημαντικότερες ομάδες των δεδομένων και των μεταβλητών που συμβάλλουν στην δημιουργία του δενδρογράμματος και οι οποίες χαρακτηρίζουν τις ομάδες των δεδομένων. Αν τα δεδομένα μας είναι λίγα σε αριθμό μπορούμε πολύ εύκολα να ερμηνεύσουμε το δενδρόγραμμα που προκύπτει από μία ταξινόμηση. Όταν όμως το πλήθος των παρατηρήσεων είναι μεγάλο, η ερμηνεία του δενδρογράμματος, απαιτεί περισσότερους υπολογισμούς, και τον υπολογισμό μερικών δεικτών, που ονομάζονται "δείκτες βοήθειας" για την ερμηνεία της ταξινόμησης.

2. Ερμηνεία δενδρογράμματος ταξινόμησης

Όπως είδαμε και στο κεφάλαιο 4, κατά την διαδικασία μιας ταξινόμησης, δημιουργούνται διαδοχικοί διαμελισμοί των δεδομένων, ο ένας μέσα στον άλλο, που παριστάνονται με ένα δενδρόγραμμα. Διαμελισμός ενός συνόλου δεδομένων ονομάζεται η δομή εκείνη όπου το σύνολο δεδομένων χωρίζεται σε ομάδες (ή κλάσεις) και κάθε αντικείμενο, ανήκει σε μία και μόνο ομάδα. Οι διαδοχικοί διαμελισμοί, που δημιουργούνται με την ιεραρχική ταξινόμηση κατ' αύξουσα ιεραρχία, περιγράφονται από ένα δενδρόγραμμα όπως αυτό στο σχήμα 1. Με X1,X2,X3,X4,X5 παριστάνονται τα αντικείμενα, ενώ τα α,β,γ,δ παριστάνουν τις διαδοχικά σχηματιζόμενες ομάδες, που αντιστοιχούν στους κόμβους του δένδρου.



σχήμα 1. Δενδρόγραμμα ταξινόμησης

Κόβοντας το δενδρόγραμμα με μια ευθεία γραμμή, παίρνουμε ένα διαμελισμό των δεδομένων με το επιθυμητό πλήθος κλάσεων. Π.χ. στο σχήμα 1, η ευθεία (ε) κόβει το δενδρόγραμμα σχηματίζοντας 3 κλάσεις αντικειμένων. Τις {X1,X4}, {X3,X5}, {X2}. Όσο χαμηλότερα κόβουμε το δενδρόγραμμα, τόσο μεγαλύτερο είναι το πλήθος των κλάσεων του σχηματιζόμενου διαμελισμού. Επίσης, όσο μεγαλύτερο είναι το πλήθος των κλάσεων, τόσο αυξάνει η ομοιογένεια μέσα στις κλάσεις. Μας ενδιαφέρει ένας διαμελισμός χωρίς πολύ μεγάλο πλήθος κλάσεων, αλλά με αρκετά ομοιογενείς κλάσεις. (Για τον τρόπο εύρεσης του βέλτιστου πλήθους κλάσεων, αναφερθήκαμε αναλυτικά στο προηγούμενο κεφάλαιο 4). Αυτό που μας ενδιαφέρει στο παρόν κεφάλαιο, είναι ο καθορισμός των χαρακτηριστικών των κλάσεων που επιλέγουμε. Δηλαδή ο καθορισμός των μεταβλητών εκείνων που χαρακτηρίζουν κάθε κλάση, αλλά και η εύρεση των μεταβλητών που είναι υπεύθυνες για τη διάσπαση κάθε κλάσης.

Με την ερμηνεία του δενδρογράμματος μιας ιεραρχικής ταξινόμησης, προσπαθούμε να αποκωδικοποιήσουμε τους παράγοντες (μεταβλητές), που επιδρούν στη διαμόρφωση των ομάδων των διαδοχικών διαμελισμών, της ταξινόμησης.

Η ερμηνεία του δενδρογράμματος γίνεται προχωρώντας από την κορυφή του δενδρογράμματος προς τη βάση του. Κάθε κόμβος του δενδρογράμματος αντιστοιχεί σε μια κλάση διαμελισμού, και διαχωρίζεται σε δύο άλλους κόμβους. Προσπαθούμε να εντοπίσουμε τις μεταβλητές εκείνες που συντελούν περισσότερο στη διάσπαση αυτή, καθώς και τις μεταβλητές εκείνες που χαρακτηρίζουν κάθε κόμβο του δενδρογράμματος.

3. Βασικές αρχές

Για κάθε κόμβο λοιπόν έχουμε δύο στάδια ερμηνείας. Προσπαθούμε να διακρίνουμε τις:

- Μεταβλητές που κυρίως χαρακτηρίζουν τον κόμβο αυτό με την παρουσία ή την απουσία τους.
- Μεταβλητές που είναι υπεύθυνες για την διάσπαση του κόμβου αυτού σε δύο νέους κόμβους.

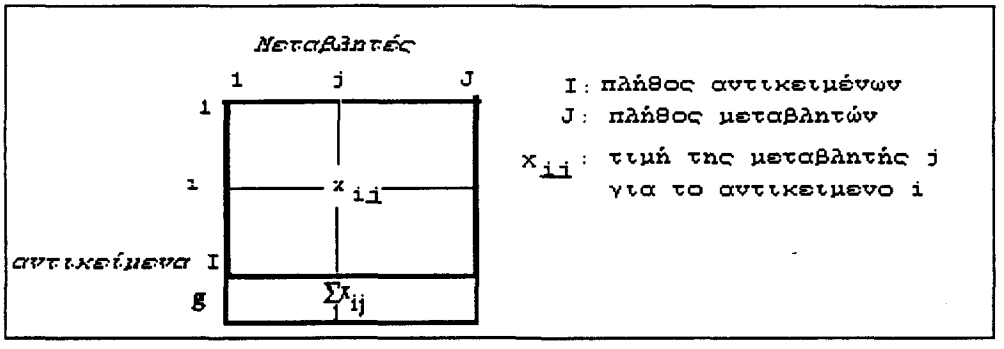
Στο πρώτο στάδιο, για κάθε κόμβο, συγκρίνονται οι μέσες ποσοστιαίες τιμές κάθε μεταβλητής με τις αντίστοιχες τιμές τους στην κορυφή του δενδρογράμματος, που αντιστοιχούν και στο κέντρο βάρους όλων των αντικειμένων. Δηλαδή βρίσκουμε την διαφορά της κλάσης που αντιστοιχεί στον κόμβο, από την μέση τιμή όλων των αντικειμένων.[Benzecri J.,1980]

Στο δεύτερο στάδιο, κάθε κόμβος αντιμετωπίζεται σαν ένα "δίπολο". Προσπαθούμε να εντοπίσουμε κατά πόσο διαφέρουν μεταξύ τους, οι δύο επόμενοί

του κόμβοι. Δηλαδή αναλύουμε τους δύο κόμβους στους οποίους χωρίζεται, αναζητώντας τις διαφορές ανάμεσα τους, που προκαλούν τον διαχωρισμό.

Η βασική αρχή ερμηνείας των αποτελεσμάτων (μέθοδος Vacor) είναι:

Για κάθε κόμβο, υπολογίζουμε τις μέσες ποσοστιαίες τιμές του για κάθε μεταβλητή. Κατόπιν εντοπίζουμε ανάμεσα στα ποσοστά αυτά, εκείνα που με το απροσδόκητα μεγάλο ή απροσδόκητα μικρό μέγεθός τους, χαρακτηρίζουν την κλάση του αντίστοιχου κόμβου. Κάθε κόμβος δηλαδή χαρακτηρίζεται τόσο από την έντονη παρουσία κάποιων μεταβλητών, όσο και από την έντονη απουσία κάποιων άλλων.



σχήμα 2. Ο πίνακας δεδομένων

Το νέφος των σημείων, παριστάνεται από I αντικείμενα και J μεταβλητές. Στον χώρο των J μεταβλητών, έχουμε ένα σύστημα συντεταγμένων, όπου κάθε άξονας συντεταγμένων αντιστοιχεί σε μια μεταβλητή και η συντεταγμένη κάθε αντικειμένου στον άξονα αυτό δεν είναι παρά η τιμή του για την αντίστοιχη μεταβλητή. Η αρχή των αξόνων δεν είναι το κέντρο βάρους του νέφους, αλλά το σημείο με τιμή 0 για όλες τις μεταβλητές. Αν διαιρέσουμε κάθε μεταβλητή με το άθροισμα $\sum_j x_{ij}$ των τιμών της, το σύστημα συντεταγμένων γίνεται ορθοκανονικό.

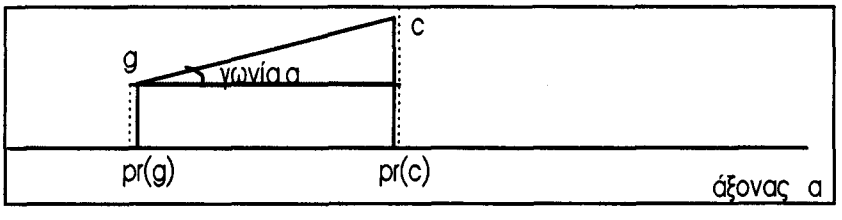
Το κέντρο βάρους g, του νέφους βρίσκεται στο σημείο που έχει για συντεταγμένες τιμές για κάθε μεταβλητή, την τιμή που αντιστοιχεί στο άθροισμα τιμών όλων των αντικειμένων για την μεταβλητή αυτή (τελευταία γραμμή για το σύνολο, στον πίνακα δεδομένων στο σχήμα 2). Στο εξής θα συμβολίζουμε με g το κέντρο βάρους του νέφους, ενώ με c το κέντρο μιας κλάσης ή τον κόμβο που αντιστοιχεί στην κλάση αυτή. Οι συντεταγμένες ενός κέντρου c είναι το άθροισμα των τιμών όλων των αντικειμένων που ανήκουν στην κλάση c, για κάθε μεταβλητή.

Έστω το παρακάτω σχήμα 3 όπου με g παριστάνεται το κέντρο βάρους του νέφους και με c το κέντρο μιας κλάσης. Αντίστοιχα με Pr(g), Pr(c) παριστάνονται

οι αντίστοιχες προβολές σε ένα άξονα a, που αντιστοιχεί σε μία από τις J μεταβλητές. Η γωνία α είναι η γωνία που σχηματίζει το ευθύγραμμο τμήμα που ενώνει τα g και c, με τον άξονα a. Γνωρίζοντας την γωνία αυτή, μπορούμε να συμπαιράνουμε αν οι προβολές των g και c στον άξονα a, βρίσκονται κοντά ή όχι. Όσο κοντύτερα βρίσκονται οι προβολές αυτές τόσο πλησιάζουν τα δύο κέντρα ως προς τον άξονα αυτό. Δηλαδή η συμπεριφορά της μεταβλητής που αντιστοιχεί στον άξονα a, είναι παρόμοια για τα δύο κέντρα. Όσο το συνημίτονο της γωνίας α, είναι μεγάλο τόσο τα δύο κέντρα απομακρύνονται, δηλαδή η διαφοροποίησή τους για την αντίστοιχη μεταβλητή είναι μεγάλη.

Για να δούμε σε τι διαφοροποιείται μια κλάση c από το κέντρο βάρους g, συγκρίνουμε τις συντεταγμένες της με τις αντίστοιχες συντεταγμένες του κέντρου βάρους. Μας ενδιαφέρει τόσο το πρόσημο όσο και το μέγεθος των διαφορών για κάθε μεταβλητή. Βέβαια ορισμένες μόνο από τις διαφορές των μεταβλητών είναι χαρακτηριστικές για κάθε κλάση. Τις σημαντικές μεταβλητές για κάθε κλάση, τις εντοπίζουμε από την υψηλή τιμή του αντίστοιχου συνημιτόνου τους, όπως εξηγούμε στην επόμενη παράγραφο.

Επομένως αναζητούμε τους άξονες μεταβλητών οι οποίοι παρουσιάζουν υψηλές τιμές του συνημιτόνου και άρα η συμβολή των μεταβλητών αυτών είναι σημαντική στην διαφοροποίηση της κλάσης από το κέντρο βάρους.



σχήμα 3. Η γωνία μεταξύ ενός κόμβου και του κέντρου βάρους

4. Υπολογισμοί δεικτών (Βοηθήματα ερμηνείας του δενδρογράμματος)

Για να είναι αποτελεσματική και χρήσιμη η ερμηνεία του δενδρογράμματος μιας ταξινόμησης, δεν αρκεί μόνο να προσδιορίσουμε τις κλάσεις των αντικειμένων, πρέπει και να χαρακτηρίσουμε τις κλάσεις αυτές με τις σημαντικότερες μεταβλητές. Για να γίνει αυτό, χρειάζεται να υπολογίσουμε τις τιμές διαφόρων δεικτών, τους οποίους ονομάζουμε βοηθήματα ερμηνείας.

Ο αρχικός κόμβος του δενδρογράμματος, αντιστοιχεί στο διάνυσμα του R^J που έχει για συντεταγμένες τις μέσες τιμές κάθε μεταβλητής. Αντιπροσωπεύει δηλαδή το μέσο αντικείμενο. Οι τιμές των συντεταγμένων του είναι:

$$g_j = \sum_{i=1}^{I} x_{ij} \quad j = 1,2,..,J$$

Κάθε κόμβος c, αντιπροσωπεύεται από ένα διάνυσμα του R^J που έχει για συντεταγμένες τις μέσες τιμές κάθε μεταβλητής για τα αντικείμενα που ανήκουν στην κλάση του κόμβου. Οι τιμές των συντεταγμένων του είναι:

$$c_j = \sum_{i=1}^{I_c} x_{ij} \quad j = 1,2,..,J$$

Για να είναι δυνατή η σύγκριση μεταξύ των κόμβων, διαιρούμε τις συντεταγμένες με το άθροισμά τους, ώστε να έχουμε ποσοστιαίες τιμές για κάθε μεταβλητή.

Για κάθε κόμβο c του δενδρογράμματος, και για κάθε μεταβλητή a που χαρακτηρίζει τα δεδομένα, υπολογίζουμε τους επόμενους δείκτες:

1) <u>COR ή COS²</u> είναι το τετράγωνο του συνημιτόνου της γωνίας που σχηματίζει η γραμμή που ενώνει το κέντρο βάρους g του νέφους με το κέντρο της κλάσης c, με τον αντίστοιχο άξονα a, κάθε μεταβλητής. Παριστάνει την συσχέτιση του κέντρου c βάρους της κλάσης, με την μεταβλητή a.

Ο τύπος είναι: $COR_a(c) = \dfrac{\|pr_a(g,c)\|^2}{\|g-c\|^2}$

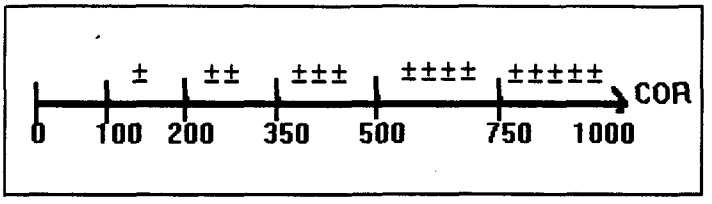
όπου $\|pr_a(g,c)\|^2 = \dfrac{(f^c{}_a - f^g{}_a)^2}{f^g{}_a}$ σύμφωνα με την απόσταση X^2, για να είναι ορθοκανονικό το σύστημα, και $\|g-c\|^2 = \sum\{\dfrac{(f^c{}_a - f^g{}_a)^2}{f^g{}_a} \mid a \in J\}$.

Επομένως είναι: $$\boxed{COR_a(c) = \dfrac{\dfrac{(f^c{}_a - f^g{}_a)^2}{f^g{}_a}}{\sum_{a \in J} \dfrac{(f^c{}_a - f^g{}_a)^2}{f^g{}_a}}}$$

Όσο η ευθεία (g,c) πλησιάζει την διεύθυνση του άξονα a, τόσο μεγαλύτερο είναι το μέρος αυτού του άξονα στην ερμηνεία της απόκλισης του κέντρου της κλάσης c από το κέντρο g.

Η τιμή του $COR_a(c)$ εκφράζει το ποσοστό ερμηνείας της κλάσης από τον άξονα a. Προσδιορίζει το μέγεθος συμβολής της κάθε μεταβλητής (άξονας a) στο χαρακτηρισμό του κόμβου c. Η συμβολή αυτή δεν είναι τίποτα άλλο παρά η ένδειξη της απόκλισης κάθε κλάσης c από το κέντρο βάρους g για την αντίστοιχη μεταβλητή.

Για τον εντοπισμό του μεγέθους απόκλισης δημιουργούμε την παρακάτω κλίμακα για τις αντίστοιχες αποκλίσεις: [Benzecri J.,1985]



σχήμα 4. Κλίμακα έντασης της τιμής του COR

Το πλήθος των + προσδιορίζει το μέγεθος της υπεροχής που παρουσιάζει για κάθε μεταβλητή η μέση τιμή της αντίστοιχης κλάσης έναντι του κέντρο βάρους, ενώ το πλήθος των - το μέγεθος της υστέρησης.

Το άθροισμα των COR για όλες τις μεταβλητές και για έναν κόμβο, δίνει την ποιότητα της αναπαράστασης (QLT) του κόμβου αυτού στο σύστημα αξόνων των μεταβλητών. Στην περίπτωση που εξετάζουμε όλες τις μεταβλητές του φαινομένου, η ποιότητα αναπαράστασης είναι 100%. Αν έχουμε μόνο ορισμένες από τις μεταβλητές, η ποιότητα αναπαράστασης είναι μικρότερη.

2) CTR είναι το ποσοστό συμμετοχής μιας μεταβλητής στον χαρακτηρισμό κάποιας κλάσης σε σχέση με τη συμμετοχή της μεταβλητής αυτής στο σύνολο των αντικειμένων. Προβάλλουμε στον θεωρούμενο άξονα a της μεταβλητής, όχι μόνο τα κέντρα βάρους, αλλά ολόκληρο το νέφος (σύνολο) σημείων και όλα τα σημεία (νέφος) κάθε κλάσης. Για την θεωρούμενη μεταβλητή a και την θεωρούμενη κλάση c, υπολογίζουμε τις προβολές $pr_a(g)$ και $pr_a(c)$ που αντιστοιχούν στα κέντρα βάρους των προβολών, του νέφους και της κλάσης c αντίστοιχα. Συμβολίζουμε με $CTR_a(c)$ την αδράνεια του κέντρου βάρους της προβαλλόμενης κλάσης $pr_a(c)$, ως προς το κέντρο βάρους του προβαλλόμενου νέφους $pr_a(g)$, διαιρεμένη με την ολική αδράνεια της προβολής του νέφους στον άξονα a.

Ο τύπος είναι: $CTR_a(c) = \frac{In_a(c)}{In_a(N)} = \frac{p_c \|pr_a(c) - pr_a(g)\|^2}{In_a(N)}$

όπου $In_a(c)$ συμβολίζει την αδράνεια του κέντρου βάρους της προβαλλόμενης κλάσης $pr_a(c)$, ως προς το κέντρο βάρους του προβαλλόμενου νέφους $pr_a(g)$, p_c είναι το βάρος (πλήθος στοιχείων) της c κλάσης, p_i είναι το βάρος του i σημείου του νέφους N όλων των σημείων (συνήθως $p_i=1$), και $In_a(N)=\Sigma\{p_i \|pr_a(i)-pr_a(g)\|^2 \mid i \in N\}$ παριστάνει την ολική αδράνεια της προβολής του νέφους στον άξονα a.

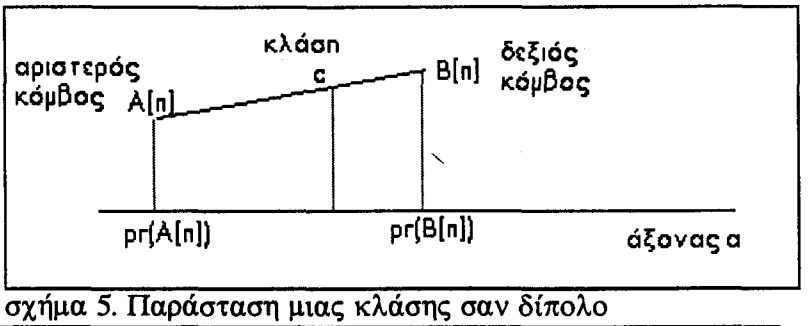
Δηλαδή η τιμή του $CTR_a(c)$ εκφράζει το ποσοστό αδράνειας μιας κλάσης για τον άξονα a σε σχέση με την αδράνεια όλου του νέφους στον άξονα αυτό. Θεωρούμε ότι τα αντικείμενα περιγράφονται μόνο από την μεταβλητή a, και επιθυμούμε να δούμε κατά πόσο η κλάση c χαρακτηρίζεται από την μεταβλητή αυτή, σε ποσοστό βέβαια του χαρακτηρισμού του νέφους από την μεταβλητή a. Π.χ. αν η τιμή του $CTR_a(c)$ είναι 0.750, σημαίνει ότι η αδράνεια της κλάσης c για τον άξονα της μεταβλητής a, είναι το 75% της ολικής αδράνειας όλου του νέφους ως προς τον αντίστοιχο άξονα.

Ο δείκτης CTR βοηθά στην αναζήτηση των κλάσεων που έχουν μεγάλη σχετική συμμετοχή στην αδράνεια των θεωρούμενων μεταβλητών. Ο δείκτης COR βοηθά στον εντοπισμό των μεταβλητών που συμβάλλουν στην δημιουργία των κλάσεων με την έννοια ότι είναι ισχυρά συσχετισμένες με τις κλάσεις αυτές. [Benzecri-Lebreaux,1980] Στις περισσότερες περιπτώσεις, σημεία με υψηλή τιμή CTR έχουν και υψηλή τιμή COR. Μπορεί όμως να συμβεί να υπάρχουν σημεία με χαμηλή τιμή CTR και υψηλή τιμή COR για κάποια μεταβλητή. Ακόμη μπορεί να υπάρχουν σημεία με υψηλή τιμή CTR και χαμηλή τιμή COR για κάποια μεταβλητή.

5. Τα δίπολα

Κάθε κλάση έχει ένα κέντρο βάρους c. Το κέντρο βάρους του νέφους των ατόμων, που αντιστοιχεί στο σύνολο όλων των ατόμων, σημειώνεται με g. Κάθε κόμβος n (αντίστοιχος της κλάσης c), ορίζεται από το κέντρο βάρους c της κλάσης και από την μάζα του m_n. Ο κόμβος n, διασπάται σε δύο άλλες κλάσεις A[n] B[n] με αντίστοιχα κέντρα βάρους c_A και c_B και μάζες m_A και m_B. Ο κόμβος με την δομή αυτή, λέμε ότι σχηματίζει ένα "δίπολο".

Σκοπός των υπολογισμών για την βοήθεια στην ερμηνεία είναι να εκτιμήσουμε την κατάσταση στον χώρο, από τη μια μεριά των ακτίνων που ενώνουν το κέντρο g του νέφους με τα πιο κάτω κέντρα της ιεραρχίας, και από την άλλη τους τομείς που ενώνουν τα A[n] B[n] μέσα στα δίπολα.



σχήμα 5. Παράσταση μιας κλάσης σαν δίπολο

Για κάθε κόμβο n και για κάθε άξονα δίνεται η διαφορά στον άξονα αυτό των συντεταγμένων μεταξύ A[n] και B[n]. Σκοπός είναι να δειχθεί καθαρά σε τι διαφέρουν οι κλάσεις A[n] και B[n] στις οποίες διαχωρίζεται ο κόμβος n. Συμαντικότερες είναι οι διαφορές εκείνες που αντιστοιχούν σε συντεταγμένες με υψηλό δείκτη COD.

Ο δείκτης COD μετράει το τετράγωνο του συνημιτόνου της γωνίας που σχηματίζει το ευθύγραμμο τμήμα A[n] B[n], με τον θεωρούμενο άξονα, ή αλλιώς τη συνεισφορά της μεταβλητής που αντιστοιχεί στον άξονα, στην ερμηνεία της διάσπασης σε A[n] και B[n] και δίνεται από τον τύπο:

$$COD_a(n) = \frac{\left\| pr_a(A[n]) - pr_a(B[n]) \right\|^2}{\left\| A[n] - B[n] \right\|^2}$$

Ο δείκτης CTD δίνει το ποσοστό της αδράνειας του διπόλου ως προς την ολική αδράνεια του νέφους, προβάλλοντάς το στον θεωρούμενο άξονα και δίνεται από τον τύπο:

$$CTD_a(n) = \frac{p_{A[n]} p_{B[n]}}{p_{[n]}} \frac{\left\| pr_a(A[n] - B[n]) \right\|^2}{In_a(N)}$$

όπου $p_{[n]}$, $p_{A[n]}$, $p_{B[n]}$ είναι τα βάρη (πλήθος στοιχείων) των κλάσεων, n, A[n] και B[n] αντίστοιχα, και $In_a(N)$ η ολική αδράνεια του νέφους N προβαλλόμενου στον άξονα της μεταβλητής a όπως περιγράψαμε προηγουμένως.

Ο δείκτης CTD παριστάνει την σχετική συνεισφορά του διπόλου στην θεωρούμενη μεταβλητή (δηλαδή κατά πόσο η αδράνεια του διπόλου ερμηνεύει την αδράνεια ως προς την θεωρούμενη μεταβλητή), ενώ ο δείκτης COD μετράει τη σχετική συνεισφορά της μεταβλητής στο θεωρούμενο δίπολο (κατά πόσο η μεταβλητή εξηγεί την συμπεριφορά του διπόλου).

6. Τρόπος παρουσίασης των αποτελεσμάτων του Vacor

Για κάθε κόμβο, προτείνουμε να εμφανίζουμε τις μεταβλητές που τον χαρακτηρίζουν, αναφέροντας τες με φθίνουσα σειρά του συνημιτόνου τους, δηλαδή ξεκινώντας από τις μεταβλητές με υψηλό συνημίτονο και προχωρώντας προς αυτές που έχουν χαμηλή τιμή. Έτσι με μια μόνο ματιά μπορούμε να βρούμε αμέσως τις υπεύθυνες μεταβλητές για τον χαρακτηρισμό και την διάσπαση κάθε κόμβου. Επίσης δημιουργούμε μια ακόμη στήλη όπου υπάρχουν σε φθίνουσα σειρά οι μεταβλητές που χαρακτηρίζουν τον κόμβο με τις αντίστοιχες συνεισφορές τους. Συγκρίνοντας τις δύο στήλες για κάθε κόμβο, μπορούμε να συμπεράνουμε εύκολα, αν υπάρχουν μεταβλητές που χαρακτηρίζουν κάποιο κόμβο αλλά δεν συντελούν

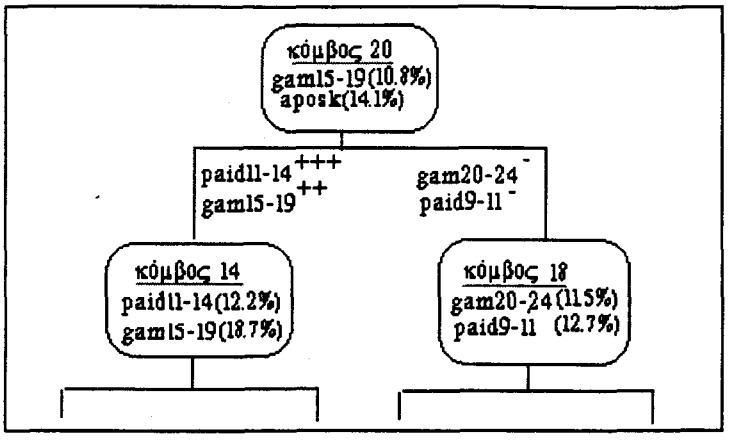
στην διάσπασή του, ή αν υπάρχουν μεταβλητές που συντελούν στην διάσπαση του κόμβου, χωρίς να τον χαρακτηρίζουν σημαντικά.

Ένα παράδειγμα αποτελεσμάτων, φαίνεται στον παρακάτω πίνακα 1. Στον πίνακα αυτό παρατηρούμε ότι η μεταβλητή που κυρίως χαρακτηρίζει (κατά ποσοστό 18.4%) τον κόμβο 20, είναι η μεταβλητή paid7-9, ενώ αυτή που κυρίως (ποσοστό 38.5%) συντελεί στην διάσπαση του κόμβου αυτού σε δύο άλλους κόμβους, είναι η μεταβλητή paid11-14 (αριθμός παιδιών 11 έως 14).

Πίνακας 1

κόμβος 20			δίπολο 20		
μεταβλητή	COS	CTR	μεταβλητή	COD	CTD
paid7-9	184	272	paid11-14	385	379
gam15-19	130	222	gam15-19	274	378
aposm	121	240	gam20-24	167	378
aposk	107	240	paid9-11	98	182
gam20-24	79	222	sxol0-6	32	65
eisl	74	180	sxol7	32	65
eish	74	180	paid7-9	11	13
paid11-14	67	82	aposm	1	1
sxol0-6	61	151	aposk	0	1
sxol7	61	151	eisl	0	1
paid9-11	43	99	eish	0	1
	1001			1000	

Για να γίνουν ευκολότερα ορατά τα συμπεράσματα αυτά, σε κάθε δίπολο του δενδρογράμματος, προτείνουμε να γράφονται οι σημαντικότερες μεταβλητές που συντελούν στην διάσπασή του. Παράδειγμα παρουσίασης ενός διπόλου στο δενδρόγραμμα ιεράρχησης, φαίνεται στο σχήμα 6. Οι μεταβλητές που βρίσκονται ακριβώς κάτω από το νούμερο του κόμβου, αντιστοιχούν στις μεταβλητές που έχουν υψηλό συνημίτονο για τον κόμβο αυτό. Το αντίστοιχο ποσοστό δείχνει πόσο συμμετέχει η θεωρούμενη μεταβλητή, στη διαμόρφωση του κόμβου. Οι μεταβλητές που βρίσκονται κάτω από την γραμμή του διπόλου, παριστάνουν τις μεταβλητές που συντελούν στην διάσπαση του κόμβου στους δύο υποκόμβους του. Το πλήθος των + ή - αντιστοιχεί στην ένταση του συνημιτόνου για την διάσπαση.



σχήμα 6. Το δίπολο του κόμβου 20 και οι σημαντικότερες μεταβλητές για την διάσπασή του

7. Στάδια της ερμηνείας του δενδρογράμματος.

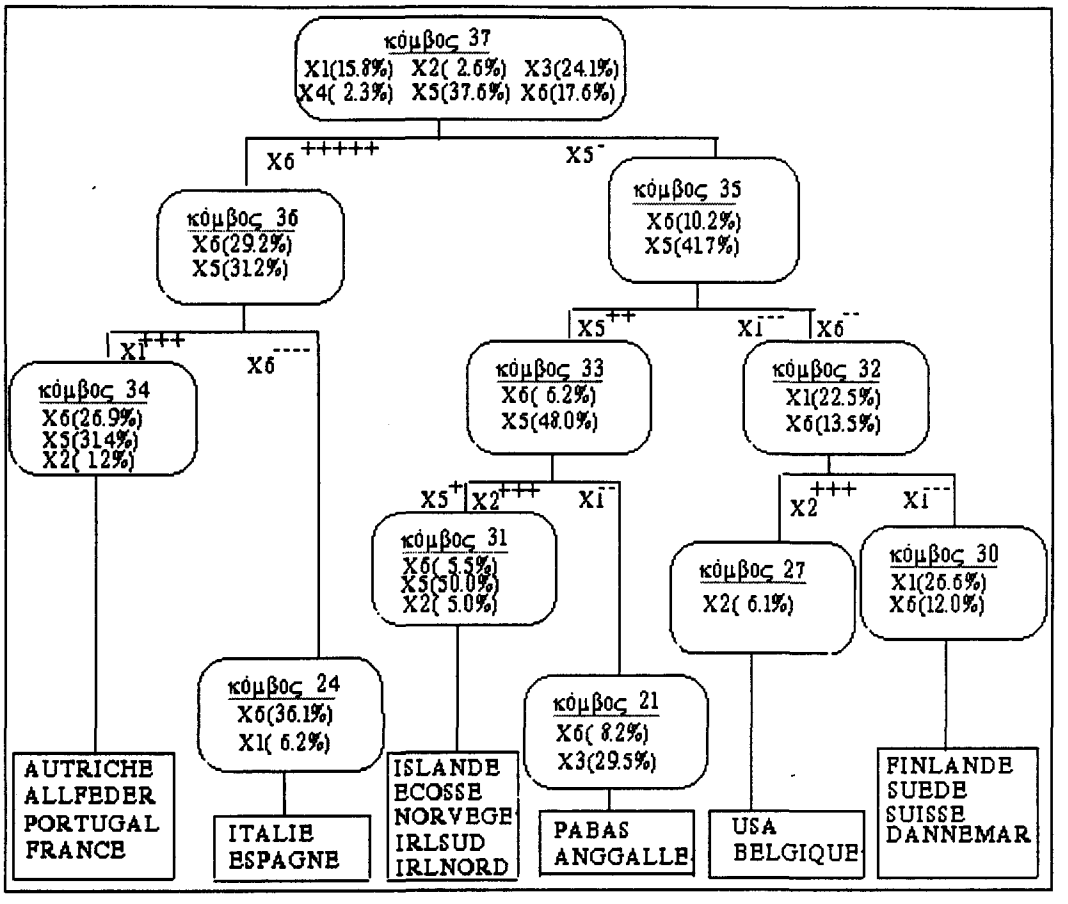
Όπως αναφέραμε και προηγουμένως η ερμηνεία του δενδρογράμματος γίνεται από πάνω προς τα κάτω. Για κάθε κόμβο ενδιαφερόμαστε καταρχήν να εξηγήσουμε τους λόγους που οδηγούν στη διάσπασή του. (Εύρεση των μεταβλητών των σημαντικότερων για την διάσπαση του κόμβου στους δύο επόμενους).

Εφαρμόζοντας τα παραπάνω για ένα αριθμό (συνήθως) μικρό ανώτερων κόμβων, καταλήγουμε σε ένα αριθμό κλάσεων οι οποίες σχηματίζουν ένα διαμελισμό του συνόλου των αντικειμένων. Η ερμηνεία των κλάσεων του διαμελισμού αυτού γίνεται εν μέρει αφού έχουμε καθορίσει τις υπεύθυνες μεταβλητές που οδηγούν στον αντίστοιχο διαχωρισμό των αντικειμένων, αλλά μπορεί να συμπληρωθεί, εξετάζοντας τον πίνακα με τις συνεισφορές των κόμβων στις μεταβλητές (δείκτης CTR).

7.1 Παράδειγμα ερμηνείας του δενδρογράμματος

Στο σχήμα 7, φαίνεται η πρότασή μας για τον τρόπο παρουσίασης ενός δενδρογράμματος. Πρόκειται για το δενδρόγραμμα που αντιστοιχεί στο παράδειγμα των 19 χωρών του προηγούμενου κεφαλαίου.

Ο ανώτερος κόμβος 37, αντιστοιχεί στην μέση τιμή των μεταβλητών για όλες τις χώρες. Κατόπιν χωρίζεται σε δύο υποκλάσεις τις 36 και 35, κυρίως εξαιτίας των μεταβλητών X6 και X5.



σχήμα 7. Το δενδρόγραμμα για το παράδειγμα των 19 χωρών με τις σημαντικότερες μεταβλητές από την μέθοδο VACOR

Ο κόμβος 36 κυρίως χαρακτηρίζεται από τις μεταβλητές X6 και X5. Το ίδιο ισχύει και για τον κόμβο 35. Η διαφορά τους, βρίσκεται στο ότι ο κόμβος 36 χαρακτηρίζεται από την υπεροχή της X6 (τιμή 29.2%) έναντι της μέσης τιμής της για όλες τις μεταβλητές (17.6%). Ο κόμβος 35 υστερεί ως προς την μεταβλητή X6 (τιμή 10.2%) έναντι της μέσης τιμής της για όλες τις μεταβλητές (17.6%). Επίσης ο κόμβος 36 χαρακτηρίζεται από την υστέρηση του ως προς την μεταβλητή X5 (31.2% έναντι 37.6% της μέσης τιμής), ενώ ο κόμβος 35 υπερέχει ως προς αυτή την μεταβλητή (41.7% έναντι 37.6% της μέσης τιμής).

Σε ανάλογα συμπεράσματα μπορούμε να καταλήξουμε μελετώντας το εμπλουτισμένο δενδρόγραμμα και για τους υπόλοιπους κόμβους.

8. Ομαδοποιημένες μεταβλητές

Μερικές φορές είναι δύσκολο να βρεθούν ισχυρά συνημίτονα (ιδίως όταν οι μεταβλητές είναι πολλές). Για παράδειγμα, έστω ότι ενδιαφερόμαστε για την κατανομή των εσόδων μιας οικονομίας που είναι περισσότερα στον πρωτογενή τομέα, απ' ότι στους άλλους τομείς. Αν έχουμε μεταβλητές όπως παραγωγή

σιταριού, βάμβακος, ψαριών, κάρβουνου κλπ., πουθενά δεν εντοπίζουμε μεγάλη σημαντικότητα.

Στην περίπτωση αυτή καλό θα ήταν να αθροίσουμε κάποιες κατηγορίες ή ομάδες μεταβλητών και να τις αντιμετωπίσουμε σαν μια μεταβλητή, στην ερμηνεία του δενδρογράμματος. [Moussaoui,1987] [Benzecri,1985].

9. Συμπεράσματα

Παρουσιάσαμε στο κεφάλαιο αυτό, τους τρόπους ερμηνείας των αποτελεσμάτων μιας ταξινόμησης. Επιμείναμε ιδιαίτερα στην εξήγηση του δενδρογράμματος μιας ταξινόμησης και προτείναμε ένα τρόπο παρουσίασης των δεικτών ερμηνείας. Οι δείκτες αυτοί παρουσιάζονται πάνω στο δενδρόγραμμα με την σειρά σπουδαιότητας τους, και γίνεται έτσι πολύ εύκολη η ερμηνεία των σημαντικότερων μεταβλητών για κάθε κόμβο.

ΚΕΦΑΛΑΙΟ 6
ΣΥΝΕΠΑΓΩΓΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΜΕΤΑΞΥ ΔΥΑΔΙΚΩΝ Η ΠΟΙΟΤΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ

1. Εισαγωγή-Ανάγκη χρήσης Συνεπαγωγικής Ταξινόμησης

Ερμηνεύοντας το δενδρόγραμμα μιας ταξινόμησης όπως είδαμε στο κεφάλαιο 5, βρίσκουμε τις ομάδες των μεταβλητών που είναι υπεύθυνες για τον χωρισμό των αντικειμένων σε κλάσεις. **Όλες όμως οι μεταβλητές έχουν την ίδια σπουδαιότητα μέσα σε κάθε κλάση, ή υπάρχουν μεταβλητές που η υπαρξή τους στον χαρακτηρισμό μιας κλάσης συνεπάγεται και την ύπαρξη κάποιων άλλων μεταβλητών;**

Στο ερώτημα αυτό έρχεται να απαντήσει η Συνεπαγωγική Ταξινόμηση, με την οποία ταξινομούνται οι μεταβλητές και οι σχέσεις τους παριστάνονται με ένα προσανατολισμένο δένδρο. Στη συνέχεια λοιπόν, αφού αρχικά μελετάμε θεωρητικά την συνεπαγωγική σχέση μεταξύ δυαδικών και αριθμητικών μεταβλητών, καταλήγουμε σε λογισμικό που μετρά και προσδιορίζει την συνεπαγωγή μεταξύ των μεταβλητών αυτών. Αφού αρχικά, εξετάζουμε θεωρητικά τις σχέσεις συνοχής και συνεπαγωγής μεταξύ ομάδων μεταβλητών, στη συνέχεια δημιουργήσαμε το απαραίτητο λογισμικό για τη μέτρηση της συνοχής και της συνεπαγωγής αυτής.

Αυτό παρουσιάζει ιδιαίτερο ενδιαφέρον και μπορεί να εφαρμοσθεί μετά την χρήση του συνδυασμού ιεραρχικής ταξινόμησης και της μεθόδου Vacor. Η εφαρμογή της μεθόδου Vacor στην ιεραρχική ταξινόμηση ενός πίνακα δεδομένων, παρέχει μεν εκείνες τις μεταβλητές που συμβάλλουν σημαντικότερα στην δημιουργία της ιεράρχησης (διάσπαση κόμβων και χαρακτηρισμός κλάσεων), δεν δίνει όμως την δυνατότητα να προσδιορισθεί η σχέση επίδρασης αυτών, μεταξύ τους. Αυτή την σημαντική έλλειψη για την πλήρη ερμηνεία του δενδρογράμματος της ιεράρχησης, έρχεται να συμπληρώσει η παρακάτω προτεινόμενη μεθοδολογία για την συνεπαγωγή των ομάδων μεταβλητών.

Στο δενδρόγραμμα της Συνεπαγωγικής Ταξινόμησης, κάθε κόμβος αντιστοιχεί σε μια μεταβλητή ή ομάδα μεταβλητών και ο επόμενος του, αντιστοιχεί στις μεταβλητές που προκύπτουν συνεπαγωγικά από την μεταβλητή του κόμβου.

Κατά την διαδικασία της Συνεπαγωγικής Ταξινόμησης, δεν αρκούμαστε μόνο στην σύγκριση των συχνοτήτων ή των συντελεστών συσχέτισης, αφού λαμβάνουμε υπόψιν ταυτόχρονα τη θετική συσχέτιση, την ομοιότητα, τη στατιστική εξάρτηση προσανατολισμένη περισσότερο θετικά από την απλή ένωση που προκύπτει από test x^2. Αναζητούμε τις σχέσεις όλων των μεταβλητών σαν

"σύνολο" όχι μεμονωμένα. Η μέθοδος περιλαμβάνεται στις μεθόδους της Πολυδιάστατης Στατιστικής Ανάλυσης, αφού ασχολείται με την μελέτη όλων των μεταβλητών και των σχέσεων τους. Ειδικότερα, μας ενδιαφέρει η μέτρηση της σχέσης συνεπαγωγής μεταξύ δύο μεταβλητών. Δηλαδή κατά πόσο, η παρουσία της μιας μεταβλητής συνεπάγεται και την παρουσία κάποιας άλλης; Η δυνατότητα μέτρησης της συνεπαγωγής αυτής, είναι χρήσιμη σε διάφορα πεδία εφαρμογών, και κυρίως στη διδακτική, όπου και πρωτοεμφανίστηκε.

Δύο είναι οι κύριοι ερευνητές που ασχολήθηκαν με την μέτρηση της συνεπαγωγής μεταξύ δυαδικών μεταβλητών, από διαφορετική οπτική γωνία ο καθένας.

- Ο Lerman [Lerman,1981] μετράει τον αριθμό των αντικειμένων (παρατηρήσεων) που παρουσιάζουν ταυτόχρονα τις δύο μεταβλητές, σε σχέση με το πλήθος αυτών που κατέχουν την κάθε μία. Πρόκειται μάλλον για ένα μέτρο ομοιότητας μεταξύ των μεταβλητών παρά για μέτρο μέτρησης της σχέσης συνεπαγωγής τους.
- Ο Gras [Gras,1992]κάνει μια στατιστική προσέγγιση, χρησιμοποιώντας πιθανότητες για τα γεγονότα που αντιτίθενται στην μετρούμενη συνεπαγωγή.

Η κύρια διαφορά είναι ότι ενώ η προσέγγιση του Lerman είναι συμμετρική για τις δύο μελετώμενες μεταβλητές, η προσέγγιση του Gras δεν είναι συμμετρική για δύο μεταβλητές. Υπολογίζεται δηλαδή τόσο η συνεπαγωγή η πρώτη μεταβλητή να επηρεάζει την δεύτερη, όσο και η συνεπαγωγή η δεύτερη μεταβλητή να επηρεάζει την πρώτη. Οι δύο αυτές συνεπαγωγές δεν έχουν πάντα την ίδια ισχύ.

Η Συνεπαγωγική Ταξινόμηση διαφέρει από την Κλασική Ταξινόμηση, στο ότι το δένδρο ιεράρχησης που δημιουργείται, έχει κόμβους προσανατολισμένους. Μπορεί να χρησιμοποιηθεί μετά από μια την Κλασική Ταξινόμηση, για την εύρεση σχέσεων και συνεπαγωγών μεταξύ των μεταβλητών που προέκυψε να είναι σημαντικές για την δημιουργία των κόμβων του δενδρογράμματος. [Καραπιστόλης,1996]

2. Μέτρηση της συνεπαγωγής μεταξύ δύο δυαδικών μεταβλητών

Στη παράγραφο αυτή ορίζουμε τις βασικές αρχές για την μέτρηση της συνεπαγωγής μεταξύ δύο δυαδικών μεταβλητών. Θεωρούμε ένα σύνολο Ε των ατόμων ή αντικειμένων, και ένα σύνολο V των μελετώμενων μεταβλητών στο σύνολο Ε. Έστω ακόμη ότι για δύο μεταβλητές a και b του συνόλου V, είναι A και B τα αντίστοιχα υποσύνολα του Ε που περιλαμβάνουν τα άτομα εκείνα με την παρουσία της μεταβλητής a και b αντίστοιχα.

Παράδειγμα:

Έστω Ε είναι ένα σύνολο 10 φοιτητών και έστω a η μεταβλητή πέρασε το μάθημα των Μαθηματικών και b η μεταβλητή πέρασε το μάθημα της Στατιστικής. Τα δεδομένα των φοιτητών για τις δύο αυτές μεταβλητές φαίνονται στον πίνακα 1.

Πίνακας 1

φοιτητές	Μαθηματικά	Στατιστική
1	0	0
2	0	1
3	1	1
4	1	0
5	0	0
6	1	1
7	1	1
8	0	0
9	0	1
10	0	0
Σύνολο	4	5

Το σύνολο Α περιλαμβάνει τους φοιτητές {3,4,6,7} και το Β τους {2,3,6,7,9}.

Η παρουσίαση των δεδομένων αυτών μπορεί να γίνει και με ένα πίνακα συμπτώσεων των δύο μεταβλητών ως εξής:

Πίνακας 2

Στατιστική / Μαθηματικά	επιτυχία (1)	αποτυχία (0)	Σύνολο
επιτυχία (1)	3	1	4
αποτυχία (0)	2	4	6
Σύνολο	5	5	10

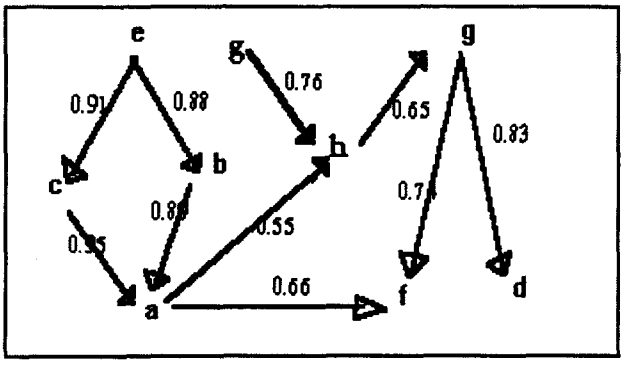
Μας ενδιαφέρει να εξετάσουμε την ισχύ της σχέσης:
"Αν ένας φοιτητής πέρασε το μάθημα των Μαθηματικών τότε πέρασε και το μάθημα της Στατιστικής" δηλαδή την συνεπαγωγή a⇒b.

Εμπνευσμένος από την μέθοδο του Lerman [Lerman,1981], για να ορίσει την "ομοιότητα" δύο μεταβλητών ο Gras [Gras,1993] θεμελιώνει την έννοια της Στατιστικής Συνεπαγωγής, η οποία αναφέρεται στο σύνολο των θεωρούμενων μεταβλητών. Η κύρια ιδέα συνίσταται στο να δώσουμε ένα κριτήριο που επιτρέπει τον αριθμητικό υπολογισμό των γεγονότων που αντιτίθενται στην υπόθεση της ύπαρξης συνεπαγωγής μεταξύ όλων των ζευγών των μεταβλητών.

Η μέτρηση της συνεπαγωγής αυτής, συνίσταται στην εύρεση ενός κριτηρίου που επιτρέπει την αριθμητική μέτρηση της απομάκρυνσης από την συνεπαγωγή. Δηλαδή τον αριθμητικό υπολογισμό των γεγονότων που αντιτίθενται στην συγκεκριμένη συνεπαγωγή. Στο παραπάνω παράδειγμα η συνεπαγωγή a⇒b τίθεται

σε αμφισβήτηση σε μια μόνο περίπτωση (Στον πίνακα 1, παρατηρούμε ότι μόνο ο 4ος φοιτητής πέρασε τα Μαθηματικά και απέτυχε στην Στατιστική).

Κατόπιν οργανώνουμε τις μεταβλητές σε δίκτυο με ένα μη-συμμετρικό μεταβατικό γράφημα, με βάρη στις ακμές και δίνουμε έτσι μια εικόνα δομής στο σύνολο των μεταβλητών. Ένα παράδειγμα τέτοιου γραφήματος για 9 μεταβλητές παρουσιάζεται στο σχήμα 1.



σχήμα 1. Δίκτυο μεταβλητών με ένα γράφημα

3. Ορισμοί - Μαθηματική Ανάλυση

Στη συνέχεια, με card(X) συμβολίζουμε τον πληθικό αριθμό του συνόλου X, ενώ με $\overline{Y}$ παριστάνουμε το συμπληρωματικό υποσύνολο ενός οποιουδήποτε τυχαίου υποσυνόλου Y του συνόλου E.

Αξίωμα 1

Θα λέμε ότι η συνεπαγωγή a⇒b είναι παραδεκτή σε επίπεδο εμπιστοσύνης 1-α αν και μόνο αν η πιθανότητα να ισχύει $\text{card}(X\cap\overline{Y})\leq\text{card}(A\cap\overline{B})$ είναι μικρότερη ή ίση με α. Δηλαδή:

$P[\text{card}(X\cap\overline{Y})\leq\text{card}(A\cap\overline{B})]\leq\alpha$

όπου X και Y είναι δύο οποιαδήποτε τυχαία υποσύνολα του E, ανεξάρτητα επιλεγμένα και έχοντας τον ίδιο πληθικό αριθμό (card) με τα A και B αντίστοιχα.

Αυτό σημαίνει ότι η συνεπαγωγή a⇒b είναι παραδεκτή εμπειρικά αν ο αριθμός των ατόμων του E που αντιτίθεται σ' αυτή είναι πολύ μικρός, σε σχέση με τον αριθμό ατόμων που αναμένεται κάτω από την υπόθεση της απουσίας σύνδεσης των δύο μεταβλητών a και b.

Για παράδειγμα αν cardE=100, cardA=36 και cardB=50, ενώ card(A∩$\bar{B}$)=3, ο αριθμός 3 είναι απίθανα μικρός, για να υποστηρίξουμε την απουσία σχέσης μεταξύ των μεταβλητών a και b. Πράγματι παρατηρούμε ότι σχεδόν όλο το A περιέχεται στο B, ενώ αν δεχόμαστε την απουσία σχέσης θα περιμέναμε περίπου το μισό των στοιχείων του A μέσα στο B.

Επανερχόμαστε τώρα στην μελέτη της τυχαίας μεταβλητής card(X∩$\bar{Y}$). Έστω n_a, n_b,$n_{\bar{b}}$,$n_{a\cap\bar{b}}$, $n_{a\cap\bar{b}}$ οι πληθάριθμοι των συνόλων A, B, $\bar{B}$, A∩$\bar{B}$, A∪$\bar{B}$ αντίστοιχα.
Η διαδικασία αποτελείται από τυχαία επιλογή, σε 3 βήματα [Gras,1992]:

1. Θεωρούμε το σύνολο αναφοράς E σαν την πραγματοποίηση ενός τυχαίου συνόλου αναφοράς &, του οποίου ο πληθάριθμος N θα είναι μια τυχαία μεταβλητή που ακολουθεί την κατανομή Poisson, με παράμετρο τον πληθάριθμο n του παρατηρούμενου συνόλου E. Η υπόθεση αυτή είναι συμβατή στις πιο συχνά παρατηρούμενες περιπτώσεις. Ο τύπος πιθανότητας είναι:

$$Pr[N = m] = \frac{n^m}{m!} e^{-n}$$

2. Η τυχαία επιλογή ενός οποιουδήποτε υποσυνόλου (π.χ. X), με τυχαίο πληθάριθμο K για μια ομοιόμορφη κατανομή πιθανοτήτων στα στοιχεία του E και ίση με d/m (στην περίπτωση του X, d=n_a), είναι δυωνυμική με τύπο:

$$Pr[K = k / N = m] = \binom{m}{k}\left(\frac{d}{m}\right)^k\left(1-\frac{d}{m}\right)^{m-k} \qquad k \le m$$

3. X και $\bar{Y}$ τα οποία είναι 2 οποιαδήποτε υποσύνολα επιλεγμένα με ανεξάρτητο τρόπο μεταξύ όλων των υποσυνόλων, έχουν αντίστοιχα πιθανότητες να επιλεγούν p(a)=n_a/n και p(b)=n_b/n.
Η πιθανότητα ενός στοιχείου του & να ανήκει στο X∩$\bar{Y}$ είναι: p(a)p(b)
Αποδεικνύεται ότι, η τυχαία μεταβλητή card(X∩$\bar{Y}$) ακολουθεί την κατανομή Poisson με παράμετρο np(a)p(b).

Έτσι, αν κανονικοποιήσουμε την παραπάνω τυχαία μεταβλητή card(X∩$\bar{Y}$), παίρνουμε την μεταβλητή Q(a,$\bar{b}$), η οποία ακολουθεί την τυπική κανονική

κατανομή αν $n_a n_{\bar{b}}/n \geq 3$. Στην πράξη, η τιμή της $Q(a, \bar{b})$ υπολογίζεται με βάση τους πληθάριθμους των συνόλων $n_a, n_{\bar{b}}, n_{a \wedge b}$, n, και είναι η:

$$q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$$

που ονομάζεται **δείκτης συνεπαγωγής**

Χρήσιμη για τον έλεγχο των υποθέσεων μας, είναι η **ένταση συνεπαγωγής** $\varphi(a, \bar{b})$ της a στην b, η οποία ορίζεται από τον τύπο:

$$\varphi(a,\bar{b}) = 1 - P[Q(a,\bar{b}) \leq q(a,\bar{b})] = \frac{1}{\sqrt{2\pi}} \int_{q(a,b)}^{\infty} e^{-\frac{t^2}{2}} dt$$

Θα λέμε ότι η συνεπαγωγή a⇒b είναι παραδεκτή σε επίπεδο εμπιστοσύνης 1-α αν και μόνο αν ισχύει η $\varphi(a, \bar{b}) = 1 - P[Q(a, \bar{b}) \leq q(a, \bar{b})] \geq 1-\alpha$

Ισχύει η ισοδυναμία $0.5 \leq \varphi(a, \bar{b}) \leq 1 \Leftrightarrow q(a, \bar{b}) \leq 0$. Καθώς είναι προφανές, σημαντικές συνεπαγωγές απαιτούν τιμές της $\varphi(a, \bar{b})$ κοντά στο 1, δηλαδή οι τιμές του $q(a, \bar{b})$ οφείλουν να είναι αρνητικές. Σημειώνουμε ότι χρήσιμη τιμή, στην περίπτωση της κανονικής προσέγγισης είναι αυτή που αντιστοιχεί στο κλασικό επίπεδο σημαντικότητας 5%:

$\varphi(a, \bar{b}) \geq 0.95 \Leftrightarrow q(a, \bar{b}) \leq -1.65$

Παράδειγμα: Για τα δεδομένα του πίνακα 2 έχουμε:

$n=10$, $n_a=4$, $n_b=5$, $n_{\bar{b}}=5$, $n_{a\cap\bar{b}}=1$

οπότε ο δείκτης συνεπαγωγής για τη σχέση a⇒b είναι $q(a, \bar{b}) = -0.707$

και η ένταση συνεπαγωγής $\varphi(a, \bar{b}) = 0.7611$

Αυτό σημαίνει ότι με πιθανότητα 76.11% όποιος πέτυχε στο μάθημα των Μαθηματικών, πέτυχε και στο μάθημα της Στατιστικής

3.1 Ιδιότητες

- Οι δείκτες q και φ είναι συμμετρικοί με την έννοια ότι:

$q(a,\bar{b}) = q(\bar{b},a)$ και $\varphi(a, \bar{b}) = \varphi(\bar{b},a)$

Η ιδιότητα αυτή μεταφράζεται στην μαθηματική λογική από την ισοδυναμία

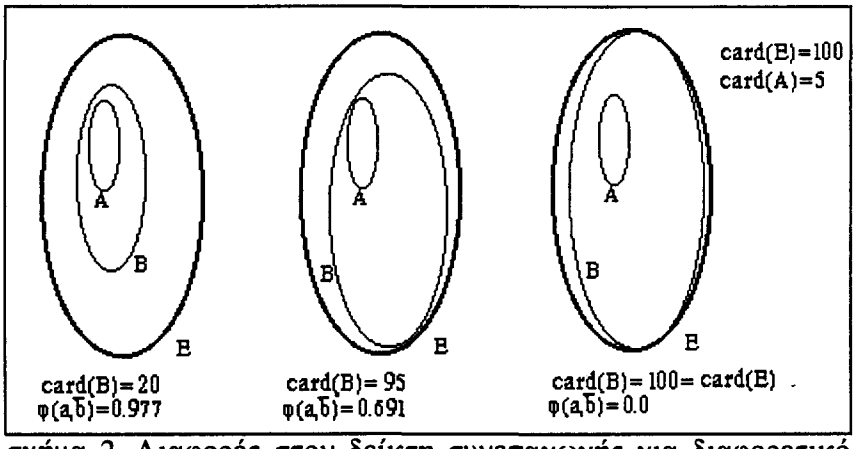
[a⇒b] ⇔ [οχι b⇒οχι a]

- Αν $0<n_a<n_b<n$ και $n_a n_b \neq n n_{a\wedge \bar{b}}$ τότε $q(a,\bar{b})/q(\bar{a},b) > 1$

Δηλαδή αν $n_a<n_b$ η πιο δυνατή στατιστικά συνεπαγωγή είναι αυτή του a⇒b. Αλλά αυτό δεν αποκλείει την δυνατότητα να οριστεί η ποιότητα της συνεπαγωγής b⇒a.

- Αν για n_a σταθερό, $n_b \to n$ (B τείνει στο E), τότε $\varphi(a,\bar{b})$ τείνει στο 0.

Αν $n_a=n$ (B=E) τότε $\varphi(a,\bar{b})=0$.



σχήμα 2. Διαφορές στον δείκτη συνεπαγωγής για διαφορετικό μέγεθος του συνόλου B

Στο σχήμα 2, βλέπουμε την μεταβολή της έντασης συνεπαγωγής a⇒b, για σταθερό μέγεθος του συνόλου A και μέγεθος του συνόλου B να μεταβάλλεται. Όσο πιο μικρό είναι το μέγεθος του συνόλου B, τόσο ισχυρότερη είναι η ένταση της συνεπαγωγής, αν και η δεσμευμένη πιθανότητα στις 3 περιπτώσεις του σχήματος, δεν αλλάζει.

Η διαφορά του δείκτη συνεπαγωγής από τα μέτρα απόστασης και ομοιότητας που παρουσιάσαμε στο 3° κεφάλαιο, έγκειται στο ότι τα μέτρα εκείνα είναι συμμετρικά για δύο μεταβλητές. Ο δείκτης συννεπαγωγής, είναι μη-συμμετρικό μέτρο πιθανότητας και βασίζεται στη μέτρηση των "περιπτώσεων" που αντιτίθενται στην θεωρούμενη συνεπαγωγή.

3.2 Σχέση της έντασης συνεπαγωγής με άλλα μέτρα συσχέτισης

Οι σχέσεις του δείκτη έντασης συνεπαγωγής, με τα παρακάτω μέτρα συσχέτισης, είναι:

- με συντελεστή συσχέτισης ρ: $\frac{\rho(a,b)}{q(a,\bar{b})} = \sqrt{\frac{n}{n_b n_{\bar{a}}}}$ Κατευθύνονται γενικά προς την "ίδια φορά", αλλά μπορεί να υπάρξει αύξηση του $q(a,\bar{b})$ και ταυτόχρονα μείωση του ρ(a,b).
- με x^2: Αφού ισχύει η σχέση $x^2 = n\rho^2$, είναι: $\frac{x^2(a,b)}{q^2(a,\bar{b})} = \frac{n^2}{n_b n_{\bar{a}}}$
- με δείκτη ομοιότητας Lerman: $\frac{q(a,\bar{b})}{s(a,b)} = -\sqrt{\frac{n_b}{n_{\bar{b}}}}$ Ίδια κατεύθυνση αλλά δεν ταυτίζονται.

4. Αριθμητικές Μεταβλητές

Εκτός από τις δυαδικές μεταβλητές με τιμές 0 και 1, μπορούμε να ορίσουμε τον δείκτη συνεπαγωγής και για αριθμητικές μεταβλητές με περισσότερες τιμές. [Larher,1991] Αυτές διακρίνονται σε δύο κατηγορίες: Σε διαβαθμισμένες (modales) μεταβλητές οι οποίες παίρνουν τιμές στο διάστημα [0,1], και σε μεταβλητές συχνοτήτων (quntitatives discretes) για τις οποίες μετράμε τον αριθμό φορών που μια μεταβλητή παρουσιάζεται σε ένα αντικείμενο. Οι διαβαθμισμένες μεταβλητές αντιστοιχούν εκτός από τις τιμές αληθές, ψευδές και στις τιμές 'ίσως", "συχνά", "μερικές φορές" κτλ. Η τιμή λοιπόν του αντικειμένου x για την a μεταβλητή, είναι a(x) και ανήκει στο διάστημα [0,1].

4.1 Παράδειγμα

Έστω ότι ρωτήσαμε 5 άτομα για τη γνώμη τους στις παρακάτω δύο ερωτήσεις:

α) Είναι αλήθεια ότι σε 50 χρόνια θα κατοικούν άνθρωποι στην σελήνη;

β) Νομίζετε ότι το 2000 θα εξαρτώνται τα πάντα από Η/Υ;

Οι δυνατές απαντήσεις τους ήταν: "αλήθεια", "ίσως", "ψέμματα"

Τις κωδικοποιούμε με τους αριθμούς 1, 0.5, 0 αντίστοιχα. Ο πίνακας με τα δεδομένα μας είναι ο επόμενος πίνακας 3.

Πίνακας 3

άτομα	ερώτηση α	ερώτηση β	$a \wedge \bar{b}$
1	0	1	0
2	0.5	1	0
3	1	1	0
4	0.5	0	0.5
5	0	0	0
σύνολο	2	3	0.5

Έχουμε $n_a=2$, $n_b=3$, $n_{\bar{b}}=5-3=2$, $n_{a\wedge\bar{b}}=0.5$
Οπότε:

$$q(a,\bar{b}) = \frac{0.5-\frac{2\cdot 2}{5}}{\sqrt{\frac{2\cdot 2}{5}}} = -0.3354$$

και

$\phi(a,\bar{b}) = 1 - P[Q(a,\bar{b}) \le q(a,\bar{b})] = 0.629$

(Η πιθανότητα υπολογίζεται από τον πίνακα της κανονικής κατανομής προσθέτουμε την τιμή που αντιστοιχεί στο 0.33 με την τιμή 0.5).

Η τιμή 0.629 σημαίνει ότι με πιθανότητα 62.9% η μεταβλητή a συνεπάγεται την b.

Γενικά αν έχουμε k δυνατές απαντήσεις, κωδικοποιούμε με τους αριθμούς

$$0,\frac{1}{k-1},\frac{2}{k-1},\ldots,\frac{k-1}{k-1}=1$$

Τα παραπάνω μας οδηγούν σε μια παρόμοια αντιμετώπιση και για τις διακριτές ποσοτικές μεταβλητές (μεταβλητές συχνοτήτων). Οι μεταβλητές αυτές ορίζονται από τον αριθμό των φορών που μια μεταβλητή a παρουσιάζεται στο αντικείμενο x. Διαιρώντας την τιμή αυτή με την μέγιστη τιμή της μεταβλητής για τα αντικείμενα, προκύπτουν τιμές στο διάστημα [0,1] για όλα τα αντικείμενα. Άρα μπορούμε να χρησιμοποιήσουμε τους παραπάνω τύπους.

Έστω α και β δύο τέτοιες μεταβλητές με μέγιστες τιμές α_m και β_m αντίστοιχα. Τότε η τιμή α_i και β_i κάθε αντικειμένου I γίνεται :

$\alpha'_i = \alpha_i / \alpha_m \qquad \beta'_i = \beta_i / \beta_m$

και έχουμε

$$v_{\acute{a}\alpha} = \sum_i a_i \qquad v_{\beta} = \sum_i \beta_i \qquad v_{\alpha\wedge\bar{\beta}} = \sum_i a_i(1-\beta_i) \qquad v_{\bar{\beta}} = n - v_{\beta}$$

οπότε ο τύπος θα είναι:

$$q(a,\bar{b}) = \frac{v_{\alpha\wedge\bar{\beta}} - \frac{v_\alpha v_{\bar{\beta}}}{n}}{\sqrt{\frac{v_\alpha v_{\bar{\beta}}}{n}}}$$

Οι παραπάνω τύποι ισχύουν ακόμη και για την περίπτωση δυαδικών μεταβλητών. Παρατηρούμε ότι οι διαβαθμισμένες μεταβλητές, από την στιγμή που ποσοκοποιούνται (όταν είναι ποιοτικές και διατεταγμένες) υπόκεινται σε στατιστική ανάλυση ανάλογη με τις μεταβλητές συχνοτήτων.

5. Αναπαράσταση με γράφημα

Προκειμένου να διευκολύνουμε την ανάλυση και ερμηνεία των σχέσεων συνεπαγωγής μεταξύ ενός συνόλου V μεταβλητών, παριστάνουμε με ένα γράφημα την δομή του V με την ένταση συνεπαγωγής φ.

Έστω η σχέση $\Re$ στο σύνολο V που ορίζεται ως εξής:

$$(\alpha,\beta)\in V^2 \quad \alpha \,\Re\, \beta \text{ αν } \varphi(\alpha,\bar{\beta})\geq 0.95 \text{ και } 0 \leq n_\alpha \leq n_\beta$$

Η σχέση αυτή είναι ανακλαστική αλλά δεν είναι συμμετρική, ούτε μεταβατική ούτε αντισυμμετρική.

Για να την παραστήσουμε με ένα γράφημα μεταβατικό και χωρίς κύκλους, παίρνουμε τον περιορισμό της $\Re'$ στις μεταβλητές που ικανοποιούν την συνθήκη:

$\alpha \,\Re'\, \beta$ και $\beta \,\Re'\, \gamma$ τότε το τόξο (α,γ) ανήκει στο γράφημα αν $\varphi(\alpha,\bar{\gamma})\geq 0.5$

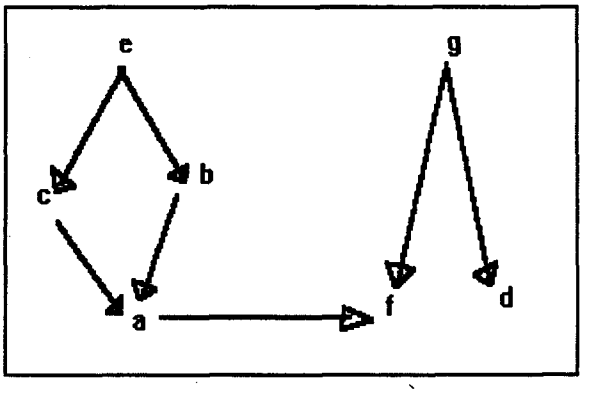
δηλαδή αν η συνεπαγωγική σχέση $\alpha \Rightarrow \gamma$ είναι "ισχυρότερη από την ουδετερότητα"

Παράδειγμα

Έστω για τις 7 μεταβλητές a,b,c,d,e,f και g έχουμε τις εξής σημαντικές συνεπαγωγές (δείκτης μεγαλύτερος 0.5):

$e \Rightarrow c,a,f,b \quad c \Rightarrow a,f \quad b \Rightarrow a,f \quad g \Rightarrow d,f \quad a \Rightarrow f$

Μπορούμε να παραστήσουμε τις σχέσεις αυτές με το επόμενο γράφημα:



6. Ανάλυση και ερμηνεία Συνεπαγωγικού Γραφήματος

Όπως αναφέραμε προηγουμένως, το γράφημα είναι μεταβατικό, και σε κάθε ακμή, υπάρχει ένα βάρος μεταξύ 0.5 και 1. Θέλοντας να ερμηνεύσουμε την σπουδαιότητα κάθε μονοπατιού στο γράφημα, ορίζουμε το **σημαντικό μονοπάτι** του γραφήματος και την **ισχύ** του μονοπατιού

Ένα μονοπάτι θα λέγεται φ_o-σημαντικό αν όλες οι ακμές έχουν βαρύτητα (δείκτη συνεπαγωγής) τουλάχιστον ίσο με φ_o. Το διάνυσμα ισχύος ενός μονοπατιού με μήκος r (άρα με r+1 κορυφές) είναι το διάνυσμα $[0,1]^r$, του οποίου οι

συντεταγμένες είναι ίσες με τους διαδοχικούς δείκτες συνεπαγωγής των κορυφών του μονοπατιού.

Τα σημαντικά μονοπάτια για μια τιμή φ_0 υψηλή (π.χ. 0.95), πιστοποιούν την συνεπαγωγική ανάλυση. Προκειμένου να μετρήσουμε την συνοχή (rigidite) ενός μονοπατιού σ' όλες τις ακμές του, αντιστοιχούμε σ'αυτό τον γεωμετρικό μέσο των εντάσεων συνεπαγωγής των ακμών του. Αν ο γεωμετικός μέσος παρουσιάζει υψηλές τιμές, τότε κατά μήκος του μονοπατιού η μείωση των εντάσεων συνεπαγωγής, είναι περιορισμένη. Περισσότερα θα αναφέρουμε στην επόμενη παράγραφο.

7. Ομάδες μεταβλητών

Το ενδιαφέρον για την έρευνα συνεπαγωγικών σχέσεων μεταξύ ομάδων μεταβλητών, για να μελετήσουμε κινήσεις μη συμμετρικές γενικότερες και όχι μόνο για απομονωμένες μεταβλητές, οδήγησε στην εισαγωγή της έννοιας του δείκτη συνεπαγωγής, μεταξύ κλάσεων μεταβλητών. Η προσέγγιση είναι διαφορετική από την προσέγγιση της κλασικής Στατιστικής, με την οποία οργανώνονται συμμετρικά κλάσεις μεταβλητών, βασισμένες σε κριτήρια αποστάσεων, συσχετίσεων ή δεικτών ομοιότητας (κεφάλαιο 3).

Για να έχουμε συνεπαγωγή μεταξύ δύο ομάδων μεταβλητών Α και Β, απαραίτητη προυπόθεση είναι να υπάρχει κάποια συνοχή των μεταβλητών μέσα στις ομάδες, έτσι ώστε η συνεπαγωγική ροή της ομάδας Α προς την ομάδα Β, να προκύπτει από μια εσωτερική ροή μέσα στην Α που θα τροφοδοτεί την εσωτερική ροή της Β. Με ποιο όμως τρόπο θα μετρήσουμε την συνοχή των μεταβλητών σε μια ομάδα;

Θα μπορούσαμε ίσως να ελέξουμε την ομοιογένεια των μεταβλητών της ομάδας, αλλά αυτή είναι συμμετρική κάτι που δεν επιθυμούμε για την έννοια της συνεπαγωγής που όπως είδαμε μπορεί να είναι ασύμμετρη.

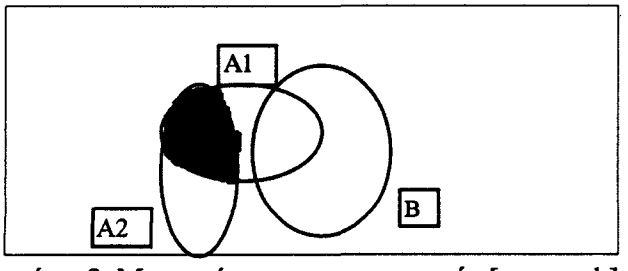
Έστω ότι έχουμε μια ομάδα δύο μεταβλητών α_1 και α_2, και b μια άλλη μεταβλητή. Η ερώτηση που αντιμετωπίζουμε είναι:
Αν ένα άτομο κατέχει τις ιδιότητες α_1 και α_2 μπορούμε να πούμε ότι κατέχει και την b;
Η απάντηση στο ερώτημα αυτό, μπορεί να δοθεί αν μετρήσουμε την συνεπαγωγή $\alpha_1 \wedge \alpha_2 \Rightarrow b$ η οποία δεν είναι ισοδύναμη με την συνεπαγωγή $[\alpha_1 \Rightarrow b] \wedge [\alpha_2 \Rightarrow b]$, αλλά είναι ισοδύναμη με την συνεπαγωγή $[\alpha_1 \Rightarrow [\alpha_2 \Rightarrow b]]$ ή με την $[\alpha_2 \Rightarrow [\alpha_1 \Rightarrow b]]$. Άρα η συνεπαγωγή της ομάδας των α_1 και α_2 προς την μεταβλητή b, θα οριστεί με βάση την συνεπαγωγική σχέση μεταξύ των α_1 και α_2 δηλαδή με βάση την συνοχή της ομάδας. Ο δείκτης συνεπαγωγή στην περίπτωση αυτή είναι:

$$q(a_1 \wedge a_2, \bar{b}) = \frac{n_{(a_1 \wedge a_2) \wedge \bar{b}} - \frac{n_{a_1 \wedge a_2} n_{\bar{b}}}{n}}{\sqrt{\frac{n_{a_1 \wedge a_2} n_{\bar{b}}}{n}}}$$

και η ένταση συνεπαγωγής δίνεται από την σχέση:

$$\phi(a_1 \wedge a_2, b) = \frac{1}{\sqrt{2\pi}} \int_{q(a_1 \wedge a_2, \bar{b})}^{\infty} e^{-t^2/2} \, dt$$

Συνήθως αν μια από τις συνεπαγωγές [$\alpha_1 \Rightarrow$b] και [$\alpha_2 \Rightarrow$b] έχει έννοια, τότε και η [$\alpha_1 \wedge a_2 \Rightarrow$b] έχει έννοια. Όμως αν παρατηρήσουμε το επόμενο σχήμα 3, βλέπουμε πως αν και [$\alpha_1 \Rightarrow$b] είναι σημαντική, η [$\alpha_2 \Rightarrow$b] δεν ισχύει και η [$\alpha_1 \wedge a_2 \Rightarrow$b] δεν ισχύει. Για το λόγο αυτό, πρέπει πριν εξετάσουμε την συνεπαγωγή μιας ομάδας μεταβλητών προς μια άλλη ομάδα, πρέπει να ελέξουμε την συνοχή των ομάδων.



σχήμα 3. Μη ισχύς της συνεπαγωγής [$\alpha_1 \wedge \alpha_2 \Rightarrow$b]

7.1 Μέτρηση της συνοχής

Έστω μια ομάδα μεταβλητών με 3 στοιχεία, τις μεταβλητές a,b,c. Αν είναι $\phi(a,\bar{b}) = 0.97 \quad \phi(b,\bar{c}) = 0.95 \quad \phi(a,\bar{c}) = 0.92$, μπορούμε να πούμε ότι η ομάδα προσανατολίζεται από την a προς την c με μια καλή συνοχή. Αυτό δεν θα συνέβαινε αν $\phi(a,\bar{b}) = 0.82 \quad \phi(b,\bar{c}) = 0.38 < \phi(c,\bar{b}) = 0.70 \quad \phi(a,\bar{c}) = 0.48 > \phi(c,\bar{a})$.

Ορίζουμε καταρχήν την συνοχή c(α_1,α_2) για μια ομάδα δύο μεταβλητών. Αυτή θα εξαρτάται από το ποια από τις δύο συνεπαγωγές [$\alpha_1 \Rightarrow \alpha_2$] ή [$\alpha_2 \Rightarrow \alpha_1$] είναι η ισχυρότερη.

$p = \max[\phi(a,\bar{b}), \phi(b,\bar{a})]$

Αν $p \geq 0.5 \quad c(a_1, a_2) = \sqrt{1 - \mathrm{E}^2}$ όπου $E = -p \log_2 p - (1-p) \log_2 (1-p)$

Αν $p = 1$ τότε $c(a_1, a_2) = 1$

Αν $p \leq 0.5$ τότε $c(a_1, a_2) = 0$

Όταν η ομάδα μεταβλητων έχει 3 στοιχεία a,b,c, μπορούμε να μετρήσουμε 6 τιμές της έντασης συνεπαγωγής $\phi(a,\bar{b}), \phi(a,\bar{c}), \phi(b,\bar{a}), \phi(b,\bar{c}), \phi(c,\bar{a})$ και $\phi(c,\bar{b})$.

Οστόσο, υποθέτοντας ότι a και b έχουν ενωθεί σε μια κλάση (a,b) εξαιτίας του γεγονότος ότι $n_a \leq n_b$, σχηματίζονται μόνο 2 ζεύγη: ((a,b),c) και (c,(a,b)). Για το πρώτο ζεύγος υπολογίζουμε τις εντάσεις συνεπαγωγής $\phi(a,\bar{b}), \phi(a,\bar{c}), \phi(b,\bar{c})$ ενώ για το δεύτερο τις $\phi(c,\bar{a})$, $\phi(c,\bar{b})$, $\phi(a,\bar{b})$.

Επιθυμούμε ο δείκτης συνεπαγωγικής συνοχής, να περιέχει την πληροφορία για τις δυαδικές συνεπαγωγικές σχέσεις μεταξύ όλων των μεταβλητών, σεβόμενος την διάταξη των όρων κάθε ζεύγους. Ταυτόχρονα, για να διατηρήσουμε την αντισυμμετρική δυναμική της συνεπαγωγής, μόνο οι πιο ισχυρές μεταξύ των τριών μεταβλητών είναι ενδιαφέρουσες για τον σκοπό μας (σύσταση νέας ομάδας). Για το λόγο αυτό, η συνοχή καθενός από τα δύο παραπάνω ζεύγη ((a,b),c) και (c,(a,b)), ορίζεται με τον γεωμετρικό μέσο των τριών συνοχών που αντιστοιχούν στο κάθε ζεύγος. Έτσι αν μια απ'αυτές είναι 0, η συνοχή του ζεύγους είναι 0.

Το ζεύγος A=((a,b),c) θα ονομάζεται κλάση και η συνεπαγωγική συνοχή c(A)του θα ορίζεται από τον γεωμετρικό μέσο των συνοχών των κλάσεων με 2 στοιχεία:

$$c(A) = \left[c(a,b) \times c(b,c) \times c(a,c)\right]^{1/3}$$

Η επιλογή του γεωμετρικού μέσου, αντί του αριθμητικού έγινε για δύο λόγους. Από τη μια για να έχουμε μηδενική συνοχή αν η συνοχή ενός τουλάχιστον ζεύγους είναι 0, και από την άλλη για να "οδηγήσουμε" την c(A) κοντά στο 1, όταν οι συνοχές όλων των ζευγών είναι αρκετά ισχυρές.

Αναγωγικά προχωρώντας, μια ομάδα A από r μεταβλητές, αποτελείται από δύο υποομάδες $A_1=\{ \alpha_1,..., \alpha_i\}$ και $A_2=\{ \alpha_{i+1},..., \alpha_{r-i}\}$. Η συνοχή της ορίζεται από τη σχέση:

$$C(A) = \left[\prod_{\substack{j>i \\ i\in\{1,...,r-1\} \\ j\in\{2,...,r\}}} c(a_i,a_j) \right]^{2/r(r-1)}$$

7.2 Συνεπαγωγή μεταξύ δύο ομάδων μεταβλητών

Η συνεπαγωγή δύο ομάδων μεταβλητών υπολογίζεται σε συνάρτηση με:

- τη συνοχή κάθε ομάδας
- τη μέγιστη ένταση συνεπαγωγής των στοιχείων της μιας ομάδας προς τα στοιχεία της άλλης
- το μέγεθος των δύο ομάδων

Ο δείκτης συνεπαγωγής ψ(A,B) της ομάδας μεταβλητών A προς την ομάδα B, με r και s στοιχεία αντίστοιχα, ορίζεται από τη σχέση:

$$\psi(A,B) = \{\sup_{\substack{i\in\{1,..r\}\\ j\in\{1,..s\}}} \phi(a_i,\overline{b_j})\}^{rs} \times [C(A)\times C(B)]^{1/2}$$

Ο δείκτης συνεπαγωγής μεταξύ των ομάδων αυξάνει με την αύξηση της συνοχή κάθε ομάδας, ενώ μηδενίζεται αν η μια έχει συνοχή 0. Επίσης αυξάνει καθώς αυξάνει η μέγιστη σύνδεση των δύο ομάδων. Τέλος μειώνεται όταν οι ομάδες έχουν μεγάλο πληθικό αριθμό. Ο εκθέτης rs για το μέγιστο των συνεπαγωγικών σχέσεων (αριθμοί μικρότεροι ή ίσοι με την μονάδα), επιτρέπει τη μείωση της επίδρασης των ομάδων με μεγάλο πληθάριθμο, οι οποίες έτσι μπορεί να παρουσιάσουν μέγιστη ένταση συνεπαγωγής.

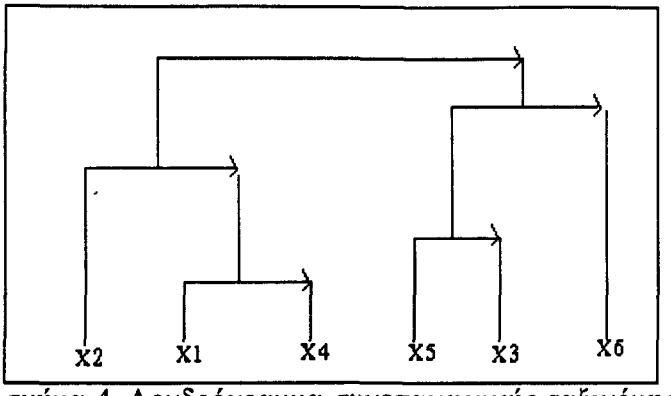
8. Αλγόριθμος ιεραρχικής ταξινόμησης βασισμένος στην συνοχή

Ακολουθώντας την μεθοδολογία των ιεραρχικών μεθόδων (κεφάλαιο 4), θα ορίσουμε έναν αλγόριθμο διαδοχικών ομαδοποιήσεων μεταβλητών και κλάσεων τους, για τη δημιουργία της Συνεπαγωγικής Ταξινόμησης. Το κριτήριο στο οποίο βασίζεται είναι η συνοχή.

Θεωρούμε όλα τα ζεύγη μεταβλητών, υπολογίζουμε τις εντάσεις συνεπαγωγής τους και κατόπιν τις συνοχές τους. Επιλέγουμε το ζεύγος (a_k,a_l) που παρουσιάζει τη μέγιστη συνοχή ή αν υπάρχουν πολλά, αυτό που παρουσιάζει μικρότερη συχνότητα εμφάνισης για την πρώτη μεταβλητή του ζεύγους. Στην περίπτωση και νέων πολλαπλών λύσεων, διαλεγουμε το ζεύγος με την μικρότερη συχνότητα και για τη δεύτερη μεταβλητή του ζεύγους. Σχηματίζουμε μια κλάση με τις μεταβλητές a_k, a_l, προσανατολισμένη από την a_k προς την a_l.

Εφαρμόζοντας διαδοχικά τον αλγόριθμο, ενώνουμε κάθε φορά δύο κλάσεις μεταβλητών είτε δύο μεταβλητές, είτε μια κλάση με μια μεταβλητή. Η διαδικασία συνεχίζεται μέχρι ότου η συνοχή κάθε νέας κλάσης που προκύπτει είναι 0 ή μέχρι να δημιουργηθεί μια κλάση που περιλαμβάνει όλες τις μεταβλητές.

Με τον τρόπο αυτό δημιουργείται ένα δένδρο ιεραρχικής Συνεπαγωγικής Ταξινόμησης, όπως αυτό που φαίνεται στο σχήμα 4.



σχήμα 4. Δενδρόγραμμα συνεπαγωγικής ταξινόμησης

8.1 Παράδειγμα

Έστω το σύνολο δεδομένων 10 παρατηρήσεων για 5 δυαδικές μεταβλητές, όπως παριστάνονται στον πίνακα 4. Από τον πίνακα αυτό, δημιουργούμε τον πίνακα Burt (πίνακας 5), όπου υπάρχει ο αριθμός συμπτώσεων μεταξύ όλων των ζευγών μεταβλητών.

Πίνακας 4

παρατήρηση	μεταβλητή X1	μεταβλητή X2	μεταβλητή X3	μεταβλητή X4	μεταβλητή X5
1	1	0	1	0	0
2	1	0	1	0	1
3	0	1	0	1	1
4	1	1	0	1	0
5	0	1	1	0	0
6	0	1	0	1	1
7	1	0	1	1	1
8	0	0	0	0	0
9	0	1	1	0	1
10	1	0	1	0	0
Σύνολο	5	5	6	4	5

Πίνακας 5
Burt-Συμπτώσεις για τις επιτυχίες-αποτυχίες των μεταβλητών

		X1		X2		X3		X4		X5	
		1	0	1	0	1	0	1	0	1	0
X1	1	5	0	1	4	4	1	2	3	2	3
	0	0	5	4	1	2	3	2	3	3	2
X2	1	1	4	5	0	2	3	3	2	3	2
	0	4	1	0	5	4	1	1	4	2	3
X3	1	4	2	2	4	6	0	1	5	3	3
	0	1	3	3	1	0	4	3	1	2	2
X4	1	2	2	3	1	1	3	4	0	3	1
	0	3	3	2	4	5	1	0	6	2	4
X5	1	2	3	3	2	3	2	3	2	5	0
	0	3	2	2	3	3	2	1	4	0	5

Χρησιμοποιώντας τον πίνακα αυτό υπολογίζουμε όλους τους δείκτες συνεπαγωγής μεταξύ όλων των ζευγών των μεταβλητών. Τα αποτελέσματα

φαίνονται στον πίνακα 6. Κατόπιν υπολογίζουμε τις αντίστοιχες εντάσεις συνεπαγωγής, τις οποίες παρουσιάζουμε στον πίνακα 7, και έπειτα υπολογίζουμε την συνοχή κάθε ζεύγους μεταβλητών. (πίνακας 8).

Πίνακας 6
Δείκτες συνεπαγωγής

⇒	X1	X2	X3	X4	X5
X1	1	0.95	-0.71	0	0.32
X2	0.95	1	0.71	-0.58	-0.32
X3	-0.58	0.58	1	0.74	0
X4	0	-0.71	1.11	1	-0.71
X5	0.32	-0.32	0	-0.58	1

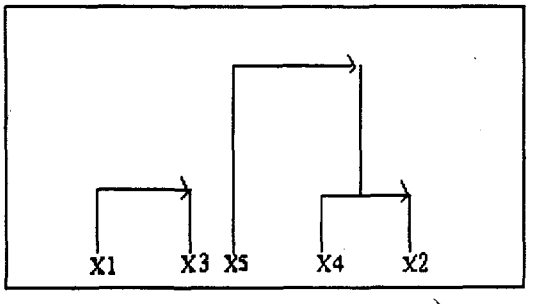
Πίνακας 7
Εντάσεις συνεπαγωγής

⇒	X1	X2	X3	X4	X5
X1	1	0.172	0.761	0.5	0.375
X2	0.172	1	0.239	0.719	0.625
X3	0.719	0.281	1	0.230	0.5
X4	0.5	0.761	0.030	1	0.761
X5	0.375	0.625	0.5	0.719	1

Πίνακας 8
Συνοχή μεταξύ ζευγών μεταβλητών

⇒	X1	X2	X3	X4	X5
X1	1	0	0.606	0	0
X2	0	1	0	0	0
X3	0	0	1	0	0
X4	0	0.606	0	1	0.606
X5	0	0.297	0	0	1

Το δενδρόγραμμα Συνεπαγωγικής Ταξινόμησης για τις 5 αυτές μεταβλητές φαίνεται στο σχήμα 5.



σχήμα 5. Δενδρόγραμμα συνεπαγωγικής ταξινόμησης του παραδείγματος

Για τα δεδομένα του παραπάνω παραδείγματος, δεν έχει νόημα η παραπέρα ένωση των κόμβων του δενδρογράμματος, αφού η συνοχή των νέων ομάδων είναι 0.

9. Προσδιορισμός των σημαντικών κόμβων

Σκοπός μας στην παράγραφο αυτή είναι να εντοπίσουμε τα επίπεδα και τους κόμβους εκείνους του δενδρογράμματος με την μεγαλύτερη σημασία.

Σε κάθε στάδιο της ιεραρχικής ταξινόμησης, δημιουργείται μια νέα ομάδα μεταβλητών. Από ένα σημείο όμως και πέρα, η νέα ομάδα δεν έχει μεγάλη συνοχή γιατί οι μεταβλητές που την αποτελούν δεν σχετίζονται πολύ μεταξύ τους. Σκοπός μας είναι να προσδιορίσουμε το επίπεδο εκείνο της ιεραρχίας που είναι σημαντικό, με την έννοια ότι μας παρέχει ένα καλό διαμελισμό των μεταβλητών, σε ομάδες που έχουν την καλύτερη δυνατή συνοχή. Η διαδικασία είναι ανάλογη με την μέθοδο Vacor, που παρουσιάσαμε στο κεφάλαιο 5. Για το λόγο αυτό ορίζουμε τον δείκτη $s(\Omega,k)$ για κάθε επίπεδο k, και τον ονομάζουμε δείκτη σημαντικότητας.

Έστω ένα σύνολο V, m μεταβλητών, k ένα επίπεδο της ιεραρχίας και Π_k ο διαμελισμός του συνόλου των μεταβλητών που ορίζεται από το επίπεδο αυτό. Κατά τον σχηματισμό του επιπέδου k, δημιουργήθηκε μια νέα κλάση με m_i μεταβλητές ($2 \le m_i \le m$). Η συνοχή της κλάσης αυτής είναι μικρότερη από τις συνοχές των κλάσεων των προηγούμενων επιπέδων και μεγαλύτερη από τις συνοχές των κλασεων των επιπέδων που ακολουθούν.

Στο σύνολο V όλων των μεταβλητών μπορούμε να αντιστοιχήσουμε το καρτεσιανό γινόμενο VxV, που αποτελείται από m(m-1) ζεύγη μεταβλητών, διαφορετικών μεταξύ τους. Αν διατάξουμε κατ' αύξουσα σειρά την συνοχή των ζευγών αυτών, σχηματίζεται μια προδιάταξη Ω, του συνόλου VxV. Συμβολίζουμε με $G(\Omega)$, το γράφημά της, δηλαδή τα ζεύγη ζευγών ((a,b), (c,d)), για τα οποία έχουμε $c(a,b) \le c(c,d)$.

Σε κάθε κόμβο υπολογίζουμε ένα δείκτη $s(\Omega,k)$ που εξαρτάται τόσο από το πλήθος των ζευγών που ακολουθούν την διάταξη, όσο και από το πλήθος αυτών που δεν την ακολουθούν. Το επίπεδο στο οποίο παρουσιάζεται ένα τοπικό μέγιστο για τον δείκτη $s(\Omega,k)$, το ονομάζουμε σημαντικό επίπεδο και τους κόμβους του, σημαντικούς κόμβους.

10. Συμβολή των αντικειμένων στους κόμβους

Για κάθε κόμβο στο δενδρόγραμμα της Συνεπαγωγικής ταξινόμησης, εντοπίζουμε τις δύο κυριώτερες μεταβλητές, που έχουν τον μεγαλύτερο δείκτη

συνεπαγωγής και είναι οι υπεύθυνες για την δημιουργία του κόμβου. Η ίδια πορεία ακολουθείται και για τους προηγούμενους κόμβους του θεωρούμενου κόμβου.

10.1 Διάνυσμα έντασης συνεπαγωγής

Με τον τρόπο αυτό, για κάθε κόμβο του επιπέδου k, ορίζεται μια ακολουθία από k εντάσεις συνεπαγωγής (φ_1,φ_2, ... φ_k) που αντιστοιχούν στην μεγαλύτερη ένταση συνεπαγωγής για όλους τους προηγούμενους κόμβους του c κόμβου. Το διάνυσμα της ακολουθίας αυτής, ονομάζεται διάνυσμα συνεπαγωγικής ισχύος του c κόμβου.

Για κάθε αντικείμενο x, υπολογίζουμε επίσης ένα διάνυσμα συνεπαγωγικής ισχύος αντικειμένου ($\varphi_{1x},\varphi_{2x}$, ... φ_{kx}), για κάθε κόμβο. Για τον κόμβο i και για τις δύο σπουδαιότερες μεταβλητές του, είναι $\varphi_{ix}=1$ αν ισχύει η σχέση συνεπαγωγής για το x αντικείμενο, $\varphi_{ix}=0$ αν δεν ισχύει, και $\varphi_{ix}=0.5$ σε κάθε άλλη περίπτωση.

Στη συνέχεια συγκρίνουμε τα δύο διανύσματα συνεπαγωγικής ισχύος του κόμβου και ενός αντικειμένου, με την σχέση:

$$d(x,c)=\left[\frac{1}{n}\sum_{i=1}^{k}\frac{(\varphi_i-\varphi_{ix})^2}{1-\varphi_i}\right]^{1/2}$$

Η απόσταση αυτή, δεν είναι παρά η απόσταση x^2, μεταξύ των κατανομών $(1-\varphi_i)_i$, $(1-\varphi_{ix})_i$. Εκφράζει την απομάκρυνση μεταξύ της προκύπτουσας εμπειρικής συνεπαγωγής και αυστηρής συνεπαγωγής.

10.2 Συνεισφορά αντικειμένου σε κόμβο

Θεωρητικά, μια κλάση C αντιπροσωπεύεται από ένα "ιδανικό" αντικείμενο x_t το οποίο έχει για εντάσεις συνεπαγωγής στους κόμβους που προηγούνται της C κλάσης, τις τιμές φ_1,φ_2, ... φ_k. Για κάθε αντικείμενο x, υπολογίζουμε την απόστασή του από το ιδανικό αντικείμενο και έχουμε έτσι ένα μέτρο σύγκρισης αντικειμένου και κλάσης. Το μέτρο αυτό δίνεται από την σχέση $\gamma(x, C)=\frac{d(x_t,C)}{d(x,C)}$ και το ονομάζουμε συνεισφορά του αντικειμένου x στην κλάση C.

Αν τώρα ενδιαφερόμαστε για την συνεισφορά μιας ομάδας αντικειμένων G στην κλάση C, υπολογίζουμε το κέντρο βάρους των συνεισφορών των αντικειμένων της ομάδας, από τη σχέση: $\gamma(G, C)=\frac{1}{Card(G)}\sum_{x\in G}\gamma(x, C)$.

11. Συμπεράσματα

Παρουσιάσαμε στο κεφάλαιο αυτό, τη μέθοδο της συνεπαγωγικής ταξινόμησης μεταξύ μεταβλητών. Η μέθοδος είναι πολύ χρήσιμη στην αναζήτηση των σχέσεων συνεπαγωγής μεταξύ των μεταβλητών, της προσανατολισμένης εξάρτησής τους, και στον καλύτερο χαρακτηρισμό των ομάδων μεταβλητών. Βασίζεται στην εισαγωγή της έννοιας έντασης συνεπαγωγής που είναι ένα μέτρο της μη-συμμετρικής σχέσης μεταξύ δύο μεταβλητών. Μας επιτρέπει να διακρίνουμε τις χαρακτηριστικές σχέσεις προσανατολισμένης διάταξης μεταξύ μεταβλητών και ομάδων τους, και ακόμη να προσδιορίσουμε τις σημαντικότερες μεταβλητές από ένα μεγάλο πλήθος μεταβλητών.

Η Συνεπαγωγική Ταξινόμηση, είναι πολύ χρήσιμη σε κάθε ερευνητή, που ενδιαφέρεται να ανακαλύψει όχι απλώς τις συμμετρικές σχέσεις μεταξύ μεταβλητών, αλλά τις χαρακτηριστικές δομές του συνόλου δεδομένων, οι οποίες συνεπάγονται κάποιες άλλες δομές.

ΚΕΦΑΛΑΙΟ 7
ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΤΗΝ ΣΥΜΒΟΛΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (symbolic data analysis)

1. Γενικά

Οι μέχρι τώρα μέθοδοι ταξινόμησης που παρουσιάσαμε και αναλύσαμε στα προηγούμενα κεφάλαια, αφορούν αντικείμενα που περιγράφονται από "κλασικές" μεταβλητές, δηλαδή ποιοτικές, λογικές ή ποσοτικές μεταβλητές. Η Συμβολική ανάλυση μας δίνει τη δυνατότητα να μελετήσουμε στατιστικά αντικείμενα που περιγράφονται από διαστήματα τιμών για κάθε ποσοτική μεταβλητή, ή από σύνολα τιμών για ποιοτικές μεταβλητές ή ακόμη από συνδυασμούς τιμών-ιδιοτήτων των μεταβλητών καθώς και από λογικές σχέσεις-ιδιότητες μεταξύ τιμών μεταβλητών. Τα αντικείμενα που περιγράφονται με τον τρόπο αυτό, ονομάζονται συμβολικά αντικείμενα.

Στο κεφάλαιο αυτό, θα προσπαθήσουμε να θεμελιώσουμε τις αρχές που διέπουν μια ιεραρχική ταξινόμηση συμβολικών αντικειμένων καθώς και τους συλλογισμούς που μας οδήγησαν στην κατασκευή λογισμικού, το οποίο παρέχει αποτελέσματα ικανά να οδηγήσουν στην εύκολη ερμηνεία τους.

Στην αρχή, περιγράφεται ο τρόπος αναπαράστασης παρατηρήσεων με πολύπλοκες μεταβλητές, ορίζοντας τα συμβολικά αντικείμενα. Περιγράφεται επίσης η μέθοδος ιεραρχικής ταξινόμησης και ο τρόπος επιλογής του βέλτιστου αριθμού κλάσεων.

Η εφαρμογή της Ανάλυσης Δεδομένων με συμβολικά αντικείμενα (symbolic data analysis), μας παρέχει το πλεονέκτημα της αντιπροσώπευσης κάθε κλάσης (ομάδας αντικειμένων στην ταξινόμηση), όχι απλά μόνο από την μέση τιμή και την διασπορά των μεταβλητών, αλλά κι από το διάστημα τιμών τους ή ακόμη κι από την κατανομή συχνοτήτων των τιμών τους.

Το κύριο πρόβλημα που παρουσιάζεται κατά την πορεία της ταξινόμησης συμβολικών αντικειμένων (όπως και απλών αντικειμένων) είναι το πώς θα αποφασίσουμε αν και κατά πόσο δύο συμβολικά αντικείμενα, θεωρούνται παρόμοια ή όχι. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί χρησιμοποιώντας είτε μια μετρική απόστασης είτε ένα δείκτη ομοιότητας μεταξύ δύο συμβολικών αντικειμένων.

Στη συνέχεια, αναφέρουμε τις χρησιμοποιούμενες μετρικές απόστασης μεταξύ συμβολικών αντικειμένων, και προτείνουμε ένα νέο δείκτη ομοιότητας για συμβολικά αντικείμενα που παριστάνονται από αριθμητικές μεταβλητές,

παίρνοντας υπόψιν και την ταυτόχρονη παρουσία των μεταβλητών για τον υπολογισμό της ομοιότητας.

Τέλος παρουσιάζονται αποτελέσματα εφαρμογής της προτεινόμενης μεθόδου χρησιμοποιώντας δεδομένα από πραγματικές φωτογραφίες.

2. Εισαγωγή

Έστω ότι σε μια έρευνα αγοράς ζητείται να περιγραφεί η εβδομαδιαία κατανάλωση κάποιων προϊόντων. Έστω επίσης ότι μια οικογένεια καταναλώνει από 300 έως 400 γραμ. ζάχαρη όταν δεν υπάρχουν παιδιά, και 600-1000 γραμ, αν έχει μέχρι 3 παιδιά ή αν δεν υπάρχουν παιδιά αλλά καταναλώνει και 200 έως 300 γραμ. καφέ.

Είναι φανερό ότι δεν είναι εύκολο να τοποθετηθούν οι πληροφορίες αυτές σε ένα κλασσικό πίνακα δεδομένων, όπου οι γραμμές αντιπροσωπεύουν τις οικογένειες και οι στήλες τις μελετώμενες μεταβλητές. Αυτό γιατί δεν υπάρχει μόνο μια τιμή σε κάθε κελί (π.χ. για την κατανάλωση ζάχαρης), και επίσης γιατί δεν είναι εύκολο να αναπαρασταθούν οι κανόνες του τύπου "**αν.., τότε...**" στον πίνακα αυτό. Όμως η παρουσίαση τέτοιου είδους δεδομένων, μπορεί να γίνει εύκολα με μια λογική έκφραση όπως:

[κατανάλωση ζάχαρης=[300,400]] ∧ [αριθμός παιδιών=0] ή
[κατανάλωση ζάχαρης=[600,1000]] ∧ [αριθμός παιδιών∈ {1,2,3}] ή
[κατανάλωση ζάχαρης=[600,1000]] ∧ [αριθμός παιδιών=0] ∧ [κατανάλωση καφέ=[200,300]]

Η χρήση της Συμβολικής Ανάλυσης Δεδομένων, προέκυψε από την ανάγκη περιγραφής δεδομένων που ήταν δύσκολο, αν όχι αδύνατο, να παρασταθούν από τους κλασικούς πίνακες δεδομένων.

Μερικές φορές μια λογική δυαδική αναπαράσταση της γνώσης [αληθές ή ψευδές] είναι αρκετή για να πάρουμε την κυρίως πληροφορία, αλλά σε κάποιες περιπτώσεις ο πραγματικός κόσμος είναι πιο αβέβαιος και χρειάζεται να περιγράψουμε την αβεβαιότητα [*συχνά, μερικές φορές, σπάνια*]. Αυτό μπορεί να γίνει χρησιμοποιώντας πιθανότητες στην παρουσίαση συμβολικών αντικειμένων. Στη συνέχεια, ασχολούμαστε μόνο με απλά συμβολικά αντικείμενα.

3. Ορισμός συμβολικών αντικειμένων

Στην Ανάλυση Δεδομένων, απλά (κλασικά) αντικείμενα είναι αυτά που περιγράφονται από έναν αριθμό μεταβλητών και τα οποία για κάθε μεταβλητή

παίρνουν μόνο μία τιμή. Η ανάγκη να ασχοληθούμε με πιο γενικές πληροφορίες οι οποίες δεν προσαρμόζονται στο κλασικό μοντέλο περιγραφής με πίνακες, οδήγησε στην ανάπτυξη της συμβολικής προσέγγισης στην Ανάλυση Δεδομένων.

Ορισμός και περιγραφές των συμβολικών αντικειμένων, έχουν δοθεί από τον Diday [Diday,1993], και βασίζονται στην έννοια του γεγονότος.

Γεγονός είναι ένα ζεύγος μεταβλητή-τιμή που συνδέει χαρακτηριστικές μεταβλητές και χαρακτηριστικές τιμές των αντικειμένων.

Για παράδειγμα:

e1=[ύψος=[1.5-2.0]]

e2=[χρώμα={άσπρο,πράσινο}]

Εδώ e1 είναι ένα γεγονός που δείχνει ότι η μεταβλητή ύψος παίρνει τιμές μεταξύ 1.5 και 2.0, και e2 είναι το γεγονός που δείχνει ότι η μεταβλητή χρώμα παίρνει είτε την τιμή άσπρο είτε την τιμή πράσινο.

Ένα **συμβολικό αντικείμενο**, περιγράφεται σαν σύζευξη γεγονότων. Ονομάζουμε επαλήθευση (assertion) τη σύζευξη αυτή.

Το συμβολικό αντικείμενο που περιγράφεται από τη σχέση:

α=[χρώμα ={πράσινο, μπλε}]∧[μέγεθος=32]∧[[τιμή=[100-150]]

δηλώνει ότι το α έχει τις επόμενες ιδιότητες:

- χρώμα είτε πράσινο είτε μπλε
- μέγεθος 32
- τιμή που κυμαίνεται μεταξύ 100 και 150.

Τα συμβολικά αντικείμενα είναι μια γενίκευση των απλών αντικειμένων και περιγράφονται από μεταβλητές (όχι απαραίτητα κοινές για όλα τα αντικείμενα), οι οποίες μπορούν να πάρουν καμία ή πολλές τιμές στο ίδιο αντικείμενο. Ακόμη μπορούμε να ορίσουμε λογικές συζεύξεις γεγονότων που συνδέουν μεταβλητές και τιμές.

Τα συμβολικά αντικείμενα είναι πολυπλοκώτερα από τα απλά αντικείμενα διότι [Gowda,1992]:

1. Όλα τα αντικείμενα δεν ορίζονται στις ίδιες μεταβλητές.
2. Κάθε μεταβλητή μπορεί να πάρει καμία, μία ή περισσότερες τιμές ή ακόμη ένα διάστημα τιμών για κάθε αντικείμενο.
3. Η τιμή των μεταβλητών μπορεί να παριστάνει συχνότητα, σχετική πιθανότητα εμφάνισης, επίπεδο σημαντικότητας μιας τιμής κτλ.
4. Η περιγραφή ενός συμβολικού αντικειμένου μπορεί να εξαρτάται από σχέσεις μεταξύ άλλων αντικειμένων

Για την περιγραφή ενός μοντέλου φορέματος που παράγεται σε διαφορετικά μεγέθη, χρώματα και χρησιμοποιώντας διαφορετικό είδος υφάσματος, μπορούμε να έχουμε:

Μοντέλο χ=[μέγεθος=[44-50]] ∧ [χρώμα={μαύρο,κόκκινο}] ή [μέγεθος=[46,48]] ∧ [χρώμα=πράσινο] ή [μέγεθος=52] ∧ [χρώμα=μαύρο-κόκκινο(δίχρωμο)] και [πουλιέται μαζί με το σακάκι μοντέλο y]

Στο παράδειγμα αυτό για το αντικείμενο μοντέλο χ, οι μεταβλητές χρώμα και μέγεθος μπορούν να πάρουν πολλές τιμές, δημιουργούνται σχέσεις μεταξύ μεταβλητών (κάποια χρώματα είναι δυνατά μόνο για ορισμένα μεγέθη) και τέλος υπάρχουν και σχέσεις μεταξύ άλλων αντικειμένων (με το μοντέλο y).

Η λογική περιγραφή ενός συμβολικού αντικειμένου ονομάζεται πρόθεση-σκοπός (intension). Επομένως ένα συμβολικό αντικείμενο, μπορεί να γραφεί ως εξής (intension):

$$s=[y_1=V_1]\wedge..\wedge[y_p=V_p] \quad \text{ή} \quad s=\wedge[y_i=V_i]$$

όπου V_i είναι διαστήματα ή σύνολα δυνατών τιμών, για την μεταβλητή y_i.
Το αντικείμενο s, εκφράζει την συνθήκη " η μεταβλητή y_1 παίρνει τιμές μέσα στο V_1, ..., και η μεταβλητή y_p παίρνει τιμές μέσα στο V_p".

Ορίζουμε ως επέκταση (extension) του αντικειμένου s, το σύνολο των παρατηρήσεων για τις οποίες οι τιμές των αντίστοιχων μεταβλητών περιέχονται στο σύνολο $V_1xV_2x...xV_p$. Η επέκταση του, είναι το σύνολο των παρατηρήσεων που είναι τέτοιες ώστε η λογική περιγραφή τους από ένα συμβολικό αντικείμενο, να είναι αληθής. Δηλαδή η επέκταση είναι η απαρίθμηση όλων των παρατηρήσεων που ικανοποιούν τις ιδιότητες ενός συμβολικού αντικειμένου.

4. Δεδομένα φωτογραφιών

Η Συμβολική Ανάλυση Δεδομένων (symbolic data analysis) μπορεί να χρησιμοποιηθεί στη διαδικασία επεξεργασίας εγχρώμων φωτογραφιών. Για λόγους κατανόησης και απλότητας τα αντικείμενα είναι απλά συμβολικά αντικείμενα (σύζευξη απλών γεγονότων), δηλαδή χωρίς περιγραφή σχέσεων μεταξύ αντικειμένων.

Στην επεξεργασία φωτογραφιών μέσω υπολογιστή, σκοπός είναι η άμεση και αυτόματη εύρεση της πληροφορίας της φωτογραφίας. Αλλά ποια πληροφορία είναι πράγματι απαραίτητη; Στην ανακατασκευή του τρι-διάστατου χώρου ή στον καθορισμό της κίνησης στον χρόνο και χώρο, χρειάζεται να βρεθούν οι ομογενείς περιοχές στην φωτογραφία, ανάλογα με το τι παριστάνουν, το χρώμα και τη θέση τους. Ακολούθως οι περιοχές αυτές, σε πολλές εικόνες, μπορούν να συγκριθούν προκειμένου να βρούμε τις διαφορές μεταξύ τους.

Στη συνέχεια εφαρμόζεται μια μέθοδος ταξινόμησης για την εύρεση της ιεράρχησης αυτών των συμβολικών αντικειμένων η οποία παριστάνεται με το δένδρο ιεράρχησης. Έτσι είναι δυνατό να επιλεγούν πολλοί διαμελισμοί με διαφορετικό αριθμό κλάσεων που ο καθένας περιλαμβάνεται στον προηγούμενό του. Χρησιμοποιώντας το κατάλληλο κριτήριο (κεφάλαιο 4), επιλέγεται ο διαμελισμός με βέλτιστο αριθμό κλάσεων [Παπαδημητρίου-Φλώρου,1993]. Με τον τρόπο αυτό, καθορίζονται οι πιο σημαντικές και ομογενείς περιοχές στην εικόνα, καθεμιά από τις οποίες αντιπροσωπεύεται από ένα συμβολικό αντικείμενο.

Κατόπιν μπορούν να βρεθούν οι επεκτάσεις αυτών των συμβολικών αντικειμένων σε άλλες εικόνες. Αν οι επεκτάσεις αυτές βρίσκονται στις ίδιες περιοχές (έχουν δηλαδή τις ίδιες σχετικές συντεταγμένες) στις άλλες εικόνες, αυτό σημαίνει ότι η περιοχή αυτή δεν άλλαξε. Αν όχι υπάρχει μετακίνηση ή μεταβολή της περιοχής αυτής.

Κάθε εικόνα (φωτογραφία) αποθηκευμένη στον υπολογιστή, παριστάνεται από ένα αριθμό pixels (συχνά 512x512). Για κάθε pixel, είναι γνωστές οι συντεταγμένες του (x,y) μέσα στο πλαίσιο της εικόνας, και επίσης το χρώμα του. Για έγχρωμες φωτογραφίες, το χρώμα είναι μια τριάδα ακεραίων από 0 έως 255 για τα τρία βασικά χρώματα κόκκινο, πράσινο και μπλε.

Ένα σημαντικό πρόβλημα στη παραπάνω διαδικασίας είναι ο υπολογισμός της ομοιότητας (ή απόστασης) μεταξύ pixels, μεταξύ pixel και κλάσης ή μεταξύ κλάσεων. Προτείνουμε μια νέα μετρική ή δείκτη ομοιότητας στην παράγραφο 6. Στην παράγραφο 7 παρουσιάζονται τα αποτελέσματα αυτής της μετρικής χρησιμοποιώντας πραγματικές φωτογραφίες.

Επειδή ο αριθμός των pixels είναι πολύ μεγάλος για μια περαιτέρω ανάλυση, επιλέγεται καταρχήν ένας διαμελισμός τους σε μικρό αριθμό k ομογενών κλάσεων (π.χ. 100 κλάσεις), διατηρώντας την μέγιστη δυνατή πληροφορία. Χρησιμοποιώντας την κλασική Στατιστική κάθε κλάση του διαμελισμού μπορεί να χαρακτηριστεί από την μέση τιμή και την τυπική απόκλιση των μεταβλητών που την χαρακτηρίζουν. Αλλά μ' αυτό τον τρόπο η εικόνα απλοποιείται πολύ και χάνεται σημαντική πληροφορία.

Χρησιμοποιώντας την Συμβολική Ανάλυση Δεδομένων (symbolic data analysis), κάθε κλάση του διαμελισμού μπορεί να παρασταθεί εκτός από την μέση τιμή και την διασπορά των μεταβλητών της, και με το διάστημα τιμών τους, ή με την κατανομή συχνοτήτων των τιμών αυτών, ή ακόμη με σχέσεις μεταξύ κλάσεων (π.χ. γειτνίαση κλπ).

Έστω S μια κλάση του αρχικού διαμελισμού της φωτογραφίας. Ο συμβολικός τύπος για την κλάση αυτή είναι:

$S=[\text{κόκκινο}=[r_l,r_u]]\wedge[\text{πράσινο}=[g_l,g_u]]\wedge[\text{μπλε}=[b_l,b_u]]\wedge[x=[x_l,x_u]]\wedge[y=[y_l,y_u]]$ (1)

όπου

r_l,r_u,g_l,g_u,b_l,b_u είναι οι αντίστοιχες μικρότερες και μεγαλύτερες τιμές για τα χρώματα κόκκινο, πράσινο και μπλε στην κλάση S.

x_l,x_u,y_l,y_u είναι οι αντίστοιχες μικρότερες και μεγαλύτερες συντεταγμένες για τον x και y άξονα συντεταγμένων.

5. Αρχικός διαμελισμός

Ξεκινώντας την επεξεργασία μιας έγχρωμης φωτογραφίας, υπολογίζεται ένας διαμελισμός όλων των pixels σε k κλάσεις. Χρησιμοποιείται γι' αυτό ο αλγόριθμος "nuees dynamiques", ο οποίος είναι ένας επαναληπτικός αλγόριθμος, με τα επόμενα βήματα [Diday,1971][Celeux,1989]:

Βήμα 0:

Έστω $B_0(0),B_1(0),...,B_k(0)$, k τυχαία επιλεγμένα αντικείμενα μεταξύ N αρχικών αντικειμένων.

Βήμα i:

Υπολογίζεται η ομοιότητα μεταξύ όλων των αρχικών αντικειμένων και των B_j (j=1,..,k). Κατασκευάζονται k νέες κλάσεις ως εξής: Ένα αντικείμενο A, αντιστοιχεί στην κλάση εκείνη με την οποία έχει την μεγαλύτερη ομοιότητα. Επιλέγονται νέα κέντρα $B_0(i),B_1(i),...,B_k(i)$, για κάθε κλάση B_j, αντιπροσωπευτικά της κλάσης τους. (π.χ. $B_j(i)$ είναι η μέση τιμή, η επικρατούσα τιμή ή η κατανομή χρωμάτων για την κλάση j).

Τερματισμός:

Ο αλγόριθμος σταματά όταν $B_j(i)=B_j(i-1)$ για κάθε j.

Σε κάθε κλάση, αντιστοιχεί ένα νέο συμβολικό αντικείμενο του τύπου (1).

6. Ιεραρχική ταξινόμηση

Κατόπιν οι κλάσεις αυτές (ομογενείς περιοχές στην εικόνα) ταξινομούνται, προκειμένου να έχουμε μια ιεράρχησή τους, και να μπορεί να επιλέγει οσοδήποτε μικρός αριθμός ομογενών κλάσεων.

Ο αλγόριθμος που χρησιμοποιείται είναι αυτός της "προς τα άνω ταξινόμησης" (κεφάλαιο 4) και αποτελείται από τα επόμενα βήματα [Benzecri,1973]:

Βήμα 0:

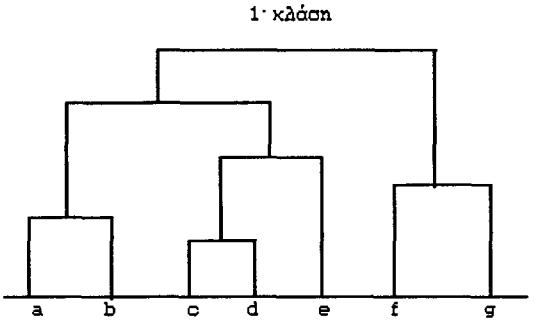
Έστω $B_0(0),B_1(0),...,B_k(0)$, k κλάσεις του διαμελισμού.

Βήμα i:

Υπολογίζεται η ομοιότητα μεταξύ όλων των παρόντων κλάσεων ανά δύο. Μετά ενώνονται οι δύο κλάσεις με την μεγαλύτερη ομοιότητα. Έτσι υπάρχουν τώρα k-1 κλάσεις

Τερματισμός:

Ο αλγόριθμος σταματά όταν k=1.



σχήμα 1. Δενδρόγραμμα ιεράρχησης

Τα αποτελέσματα του παραπάνω αλγορίθμου, παριστάνονται με ένα δενδρόγραμμα ιεράρχησης, όπως αυτό στο σχήμα 1. Η κορυφή του δένδρου αντιστοιχεί στην μεγαλύτερη κλάση που περιέχει όλες τις παρατηρήσεις. Διαιρείται σε δύο υποκλάσεις, η καθεμιά στη συνέχεια με την σειρά της διαιρείται σε δύο άλλες υποκλάσεις κοκ.

Στην παραπάνω διαδικασία χρειάζεται ένα μέτρο ομοιότητας μεταξύ δύο αντικειμένων (pixels), μεταξύ ενός αντικειμένου και μιας κλάσης (στην διαδικασία εύρεσης του διαμελισμού προκειμένου να αντιστοιχηθεί ένα pixel σε μια κλάση) και ακόμη μεταξύ δύο κλάσεων (στην διαδικασία της ταξινόμησης για την σύγκριση δύο κλάσεων). Από την στιγμή που τόσο οι κλάσεις όσο και τα pixels παριστάνονται με το ίδιο τρόπο σαν συμβολικά αντικείμενα, χρειάζεται να ορισθεί μόνο ένα μέτρο ομοιότητας μεταξύ συμβολικών αντικειμένων.

7. Μετρικές - Μέτρα ή δείκτες ομοιότητας

Για να βρεθούν σχέσεις μεταξύ συμβολικών αντικειμένων, χρειάζεται να ορισθεί πότε δύο αντικείμενα θεωρούνται "όμοια" και πότε όχι.

Ένα μέτρο ομοιότητας S μεταξύ δύο αντικειμένων $s_1, s_2 \in E$, (κεφάλαιο 3), είναι μια αντιστοιχία από το ExE στο R_+, με τις παρακάτω ιδιότητες:

$S(s_1, s_1)=S(s_2, s_2)\geq S(s_1, s_2)$ (ανακλαστική)

$S(s_1, s_2)=S(s_2, s_1)$ (συμμετρική)

Έστω δύο συμβολικά αντικείμενα $s_1 =\wedge[y_i=A_i]$ και $s_2 =\wedge[y_i=B_i]$ όπου

A_i ,B_i είναι διαστήματα πραγματικών τιμών.

Στην περίπτωση αυτή οι Gowda-Diday πρότειναν ένα μέτρο ομοιότητας που ορίζεται από 3 συντεταγμένες [Gowda-Diday,1992]:

Θέση $S_p(A_i , B_i) = 1 - |a_1 - b_1| / |U_i|$

Εύρος $S_s(A_i , B_i) = |l_a + l_b| / 2l_s$

Περιεχόμενο $S_c(A_i , B_i) = inters / l_s$

όπου:

A_i =διάστημα $[a_1 , a_u]$

B_i =διάστημα $[b_1 , b_u]$

l_a= μήκος του διαστήματος A_i

l_b= μήκος του διαστήματος B_i

l_s= εύρος των A_i και B_i $= |\max(a_u , b_u) - \min(a_1 , b_1)|$

$|U_i|$= το μεγαλύτερο από τα μήκη των A_i και B_i

inters= μήκος της τομής των A_i και B_i

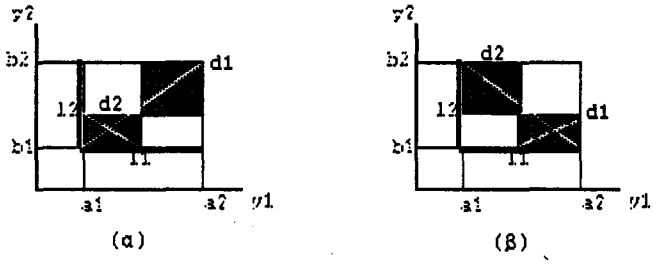
Τότε το μέτρο ομοιότητας του Gowda-Diday μεταξύ των A_i και B_i είναι:

$S(A_i , B_i) = S_p(A_i , B_i) + S_s(A_i , B_i) + S_c(A_i , B_i)$

και επομένως μεταξύ των s_1, s_2 είναι:

$S_{g.d}(A_i , B_i) = \Sigma S(A_i , B_i)$ (2)

7.1 Νέο μέτρο ομοιότητας



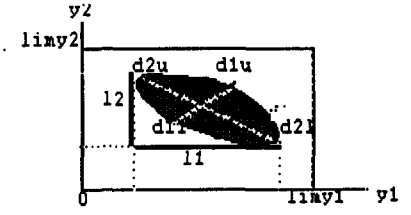
σχήμα 2. Δύο διαφορετικές κλάσεις για τα ίδια διαστήματα τιμών των y_1 και y_2

Το πρόβλημα με το μέτρο ομοιότητας των Gowda-Diday είναι ότι δεν παίρνει υπόψιν την ταυτόχρονη παρουσία πολλών μεταβλητών. Έστω για παράδειγμα, ότι έχουμε τα αντικείμενα για τα οποία οι τιμές των μεταβλητών y_1 και y_2 μεταβάλλονται στα διαστήματα $[a_1 , a_2]$ και $[b_1 , b_2]$ αντίστοιχα. Στο σχήμα 2 έχουμε δύο παραδείγματα κλάσεων. Τα αντικείμενα αυτά, παριστάνονται από τα γραμμοσκιασμένα τμήματα των σχημάτων 2α και 2β. Παρατηρούμε ότι οι ομάδες αντικειμένων των δύο σχημάτων, (αριστερά και δεξιά), έχουν ίδιες μέσες τιμές ως

προς τις μεταβλητές y_1 και y_2, και ακόμη τα ίδια διαστήματα τιμών για αυτές, αλλά είναι τελείως διαφορετικές. Με την χρήση του μέτρου ομοιότητας του Gowda (τύπος (2)) οι δύο κλάσεις θεωρούνται ακριβώς ίδιες, αφού η θέση, το εύρος και το περιεχόμενο της κάθε μεταβλητής ξεχωριστά είναι ακριβώς τα ίδια.

Γι' αυτό το είδος των προβλημάτων χρειάζεται είτε να υπάρχει ένας καλύτερος διαμελισμός (με περισσότερο ομογενείς κλάσεις), είτε να υπολογιστεί η ομοιότητα χρησιμοποιώντας τις ταυτόχρονες τιμές των δύο μεταβλητών.

Στα ήδη υπάρχοντα μέτρα ομοιότητας, χρησιμοποιείται μόνο το εύρος τιμών των μεταβλητών (δηλαδή τα μήκη a_1-a_2 και b_1-b_2 στο σχήμα 2). Εμείς προτείνουμε τη χρησιμοποίηση και του εύρους των ταυτόχρονων τιμών των μεταβλητών (δηλαδή των διαγωνίων d_1 και d_2), για την καλύτερη σύγκριση των δύο ομάδων.



σχήμα 3. Υπολογισμός των d_1 και d_2 για τις μεταβλητές y_1 και y_2

Χωρίς περιορισμό της γενικότητας, υποθέτουμε ότι η μεταβλητή y_1 παίρνει τιμές στο διάστημα $[0,limy_1]$ και η y_2 στο $[0,limy_2]$. Αν όχι μπορούμε να κάνουμε μια μετατόπιση των τιμών.

Στο σχήμα 3, έχουμε ένα παράδειγμα μιας κλάσης, με σημεία τα οποία περικλείονται από την καμπύλη. Τότε για δύο μεταβλητές έχουμε τα παρακάτω άκρα των διαγωνίων:

d_{1l}= σημείο με $\min(y_1^2+y_2^2)$

d_{1u}= σημείο με $\min((limy_1-y_1)^2+(limy_2-y_2)^2)$

d_{2l}= σημείο με $\min(y_1^2+(limy_2-y_2)^2)$

d_{2u}= σημείο με $\min((limy_1-y_1)^2+y_2^2)$

Στο σχήμα 2 έχουμε σχεδιάσει τις διαγωνίους d_1 και d_2 που υπολογίστηκαν με τον παραπάνω τρόπο. Φαίνεται καθαρά ότι d_1 στο σχήμα 2α είναι μεγαλύτερη από d_1 στο σχήμα 2β ενώ η d_2 του 2α είναι μικρότερη από την d_2 του 2β. Δεν μπορούν λοιπόν οι δύο ομάδες αντικειμένων, να θεωρηθούν όμοιες.

Γενικότερα για p μεταβλητές, υπολογίζουμε παρόμοια, τα d_{1l},d_{1u}, d_{2l},d_{2u},..., d_{ql},d_{qu} όπου $q=2^{p-1}$.

Η νέα περιγραφή του συμβολικού αντικειμένου, θα είναι:

$S=[y_1=[y_{1l},y_{1u}]]\wedge...\wedge[y_p=[y_{pl},y_{pu}]] \wedge [d_1=[d_{1l},d_{1u}]]\wedge...\wedge[d_q=[d_{ql},d_{qu}]]$ (3)

Προτείνουμε για την μέτρηση της ομοιότητας μεταξύ δύο αντικειμένων, το συνδυασμό του μέτρου ομοιότητας του Gowda και της σύγκριση των διαμέτρων τους $d_1,d_2,d_3,...,d_q$. Για κάθε μία από τις διαμέτρους, συγκρίνουμε το μήκος και την θέση τους.

Δηλαδή μετατρέπουμε το μέτρο ομοιότητας του Gowda, ώστε να είναι δυνατή όχι μόνο η σύγκριση διαστημάτων (μία μεταβλητή) αλλά και η σύγκριση περιοχών που ορίζονται από p μεταβλητές. Έτσι συγκρίνουμε τις τρεις συνιστώσες θέση, εύρος, περιεχόμενο, στις p διαστάσεις.

Για δύο διαμέτρους $d_a=(a_l,a_u)$, $d_b=(b_l,b_u)$, εξαρτώμενες από p μεταβλητές, ο τύπος θα είναι:

$$S_{p\text{-}g.d}(d_a , d_b)=S_p(d_a , d_b) +S_s(d_a , d_b)+S_c(d_a , d_b)$$

όπου Θέση : $S_p(d_a , d_b) =1-\sqrt{\sum_i \frac{(a_{li}-b_{li})^2}{\max(v_{da},v_{db})}}$

Εύρος: $S_s(d_a , d_b)=|v_{da} +v_{db}|/2v_s$ Περιεχόμενο: $S_c(d_a , d_b)=v_{inters}/v_s$

v_{da}= υπερ-όγκος της $d_a=\Pi(a_{ui}-a_{li})$ v_{db}= υπερ-όγκος της $d_b=\Pi(b_{ui}-b_{li})$

v_s= εύρος υπερ-όγκου των d_a και $d_b=|\Pi(\max(a_{ui}-b_{ui})-\min(a_{li}-b_{li}))|$

v_{inters}= υπερ-όγκος της τομής των d_a και d_b

Το νέο μέτρο ομοιότητας για τα δύο συμβολικά αντικείμενα A και B (που περιγράφονται με τον τύπο 3) θα είναι:

$$S_{new}(A , B)=\Sigma S_{g.d}(y_{Ai} , y_{Bi}) +\Sigma S_{p\text{-}g.d}(d_{Ai} , d_{Bi}) \quad (4)$$

7.2 Εφαρμογή του νέου μέτρου ομοιότητας για pixels

Για τα δεδομένα των φωτογραφιών η περιγραφή μιας κλάσης pixels με ένα συμβολικό αντικείμενο, είναι:

$S=[\text{κόκκινο}=[r_l,r_u]]\wedge[\text{πράσινο}=[g_l,g_u]]\wedge[\text{μπλε}=[b_l,b_u]]\wedge$

$[d_1=[d_{1l},d_{1u}]]\wedge[d_2=[d_{2l},d_{2u}]]\wedge[d_3=[d_{3l},d_{3u}]]\wedge[d_4=[d_{4l},d_{4u}]]\wedge[x=[x_l,x_u]]\wedge[y=[y_l,y_u]]$ (5)

όπου

r_l, r_u, g_l, g_u ,b_l, b_u είναι οι αντίστοιχες μικρότερες και μεγαλύτερες τιμές για τα χρώματα κόκκινο, πράσινο και μπλε στην κλάση S.

d_{1l}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (0,0,0).

d_{1u}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (255,255,255).

d_{2l}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (255,0,0).
d_{2u}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (0,255,255).
d_{3l}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (0,255,0).
d_{3u}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (255,0,255).
d_{4l}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (0,0,255).
d_{4u}=η τριάδα χρωμάτων (κόκκινο, πράσινο, μπλε) που είναι κοντύτερα στην (255,255,0).
x_l, x_u, y_l, y_u είναι οι αντίστοιχες μικρότερες και μεγαλύτερες συντεταγμένες για τον x και y άξονα.

Όταν αναζητούνται οι ομοιογενείς περιοχές στην εικόνα, χρειάζεται να βρεθούν τα τμήματα εκείνα, όπου τα τρία χρώματα παρουσιάζονται ταυτόχρονα με παρόμοιες τιμές, λαμβάνοντας υπόψιν και την θέση τους στην φωτογραφία.

Επομένως η ομοιότητα μεταξύ δύο κλάσεων A και B, υπολογίζεται σε δύο βήματα:

1. Συγκρίνοντας τις τιμές για τα χρώματα κόκκινο, πράσινο, μπλε και επίσης για τις 4 διαμέτρους, με τον τύπο 4.

και

2. Λαμβάνοντας υπόψιν την φυσική γειτνίαση ή όχι, υπολογίζοντας την ευκλείδεια απόσταση μεταξύ των σημείων (x_A,y_A) και (x_B,y_B).

8. Αποτελέσματα

Εφαρμόσαμε τον παραπάνω αλγόριθμο σε μια έγχρωμη φωτογραφία όπως φαίνεται στο σχήμα 4. Πρώτα κάναμε ένα διαμελισμό όλων των pixels σε 100 κλάσεις. Κάθε κλάση παριστάνεται από ένα συμβολικό αντικείμενο του τύπου (1). Κατόπιν χρησιμοποιήσαμε τον αλγόριθμο ιεραρχικής ταξινόμησης και το νέο μέτρο ομοιότητας, δημιουργώντας το δενδρόγραμμα ιεράρχησης των κλάσεων. Χρησιμοποιώντας το κριτήριο επιλογής του βέλτιστου διαμελισμού, επιλέξαμε τον διαμελισμό εκείνο, με τον καλύτερο αριθμό κλάσεων. Προβάλλουμε τις κλάσεις αυτού του διαμελισμού σε μια νέα εικόνα (σχήμα 4).

σχήμα 4. Πραγματική φωτογραφία (αριστερά) και ξανακατασκευασμένη (δεξιά)

Στο σχήμα 4 παρουσιάζουμε την πραγματική φωτογραφία και την φωτογραφία που φτιάξαμε, χρησιμοποιώντας μόνο 8 ομογενείς κλάσεις.
Μπορούμε να διατηρήσουμε την πληροφορία μιας φωτογραφίας, έχοντας μόνο μικρό αριθμό (8-12) περιοχών.

9. Συμπεράσματα

Περιγράψαμε την ιεραρχική Συμβολική Ανάλυση Δεδομένων για την αντιμετώπιση σύνθετων δεδομένων, που δεν μπορούν να επεξεργαστούν αποτελεσματικά με τις κλασικές μεθόδους της Ανάλυσης Δεδομένων. Το μεγαλύτερο πλεονέκτημα της ανάλυσης αυτής, είναι η δυνατότητα επεξεργασίας σύνθετων δεδομένων και σχέσεων μεταξύ τους.

Το κύριο βάρος της επεξεργασίας τέτοιων παρατηρήσεων, βρίσκεται στον τρόπο επιλογής της μετρικής απόστασης ή ομοιότητας μεταξύ δύο παρατηρήσεων. Για παρατηρήσεις που χαρακτηρίζονται από p αριθμητικές μεταβλητές, ορίσαμε ένα νέο μέτρο ομοιότητας με την εισαγωγή 2^{p-1} διαμέτρων, που βασίζεται στο μέτρο ομοιότητας των Gowda-Diday.

Το νέο αυτό μέτρο ομοιότητας υπολογίζεται εύκολα και είναι αποτελεσματικότερο γιατί χρησιμοποιούνται όχι μόνο οι τιμές κάθε μεταβλητής ξεχωριστά αλλά και οι ταυτόχρονες τιμές πολλών μεταβλητών.

Σαν εφαρμογή, χρησιμοποιήσαμε το νέο αυτό μέτρο στην επεξεργασία εγχρώμων φωτογραφιών μέσω υπολογιστή, λαμβάνοντας ικανοποιητικά αποτελέσματα στην εύρεση των ομοιογενών περιοχών μιας φωτογραφίας.

ΚΕΦΑΛΑΙΟ 8
ΑΝΤΙΣΤΟΙΧΙΣΗ ΑΚΕΡΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ ΣΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΙΕΡΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ

1. Τοποθέτηση του προβλήματος

Στο κεφάλαιο αυτό αντιμετωπίζουμε το πρόβλημα, της επεξεργασίας και ανάλυσης όχι μόνο αντικειμένων που χαρακτηρίζονται από μεταβλητές αλλά κλάσεων αντικειμένων. Δηλαδή γνωρίζουμε ότι τα αντικείμενα είναι οργανωμένα σε κλάσεις και θέλουμε να προχωρήσουμε σε μια νέα ανάλυση, χρησιμοποιώντας αυτό το γεγονός. Έχουμε δει ότι με την ιεραρχική ταξινόμηση, καταλήγουμε σε κλάσεις ομογενών αντικειμένων.

Τί γίνεται όμως όταν θέλουμε να χρησιμοποιήσουμε τις κλάσεις αυτές για μια περαιτέρω ανάλυση; Πρέπει να βρούμε ένα αριθμό για κάθε κλάση που να την αντιπροσωπεύει όσο το δυνατό καλύτερα, και να τη διακρίνει από τις υπόλοιπες. Πρέπει δηλαδή να ορίσουμε μια νέα μεταβλητή, (συνήθως με ακέραιες τιμές), για κάθε αντικείμενο των κλάσεων. Στα αντικείμενα της ίδιας κλάσης δίνουμε την ίδια τιμή, ακριβώς για να διατηρήσουμε την ιδιότητα ότι τα αντικείμενα αυτά μοιάζουν και ανήκουν στην ίδια κλάση. Στα αντικείμενα δύο διαφορετικών κλάσεων, δίνουμε τέτοιες τιμές, που να αντιστοιχούν στο γεγονός ότι τα αντικείμενα διακρίνονται μεταξύ τους και η διαφορά των τιμών αυτών, πρέπει να προσεγγίζει τη διαφορά ομοιότητας μεταξύ των κλάσεων.

Για παράδειγμα, έστω ότι θέλουμε να μελετήσουμε και να αναλύσουμε την απόδοση διαφόρων ποικιλιών σταριού, σε σχέση με το έδαφος, τον τρόπο εκμετάλλευσης και τον χρόνο σποράς. Έχουμε καταλήξει στο συμπέρασμα (μετά από μια ταξινόμηση), ότι υπάρχουν 3 κλάσεις-ομάδες ποικιλιών. Αυτές που αποδίδουν πολύ καλά, καλά ή καθόλου. Αν στη συνέχεια επιθυμούμε να μελετήσουμε την επίδραση λιπασμάτων, του είδος άδρευσης κλπ, στις ομάδες αυτές, εφαρμόζοντας μια άλλη ανάλυση, χρειαζόμαστε την αντιστοίχιση μιας αριθμητικής τιμής στις ποικιλίες κάθε ομάδας, που να εκφράζει την ομάδα στην οποία ανήκουν. Συνήθως δίνουμε τιμές 1, 2 και 3. Όμως είναι σωστό να θεωρήσουμε ίδια την διαφορά μεταξύ πολύ καλής και καλής απόδοσης, με την διαφορά μεταξύ καλής και καθόλου απόδοσης; Στο ερώτημα αυτό, προσπαθούμε να απαντήσουμε παρακάτω.

2. Εισαγωγή

Έστω ότι έχουμε βρει ένα διαμελισμό των αντικειμένων, σε κάποιο αριθμό κλάσεων και επιθυμούμε να αντιστοιχήσουμε μια νέα ακέραια μεταβλητή σε κάθε αντικείμενο, σύμφωνα με την κλάση στην οποία ανήκει. Το πρόβλημα έγκυται στον προσδιορισμό των τιμών της νέας αυτής μεταβλητής σε κάθε κλάση, καθώς και στην διαφορά των τιμών αυτών από κλάση σε κλάση. Οι τιμές αυτές της μεταβλητής, είναι χρήσιμες στην περαιτέρω ανάλυση των αρχικών δεδομένων, όπως για παράδειγμα στην Συνεπαγωγική ταξινόμηση μεταξύ κλάσεων.

Η απλούστερη διαδικασία είναι να θέσουμε τιμές της μεταβλητής 1 έως r στην περίπτωση που τα δεδομένα οργανώνονται σε r κλάσεις [Καραπιστόλης,1996]. Για παράδειγμα, έστω ότι η ιεράρχηση κάποιων αντικειμένων ως προς ένα σύνολο μεταβλητών, μας οδήγησε σε τρεις ομάδες: {άριστα αντικείμενα} {μέτρια αντικείμενα} {ακατάλληλα αντικείμενα}. Αν ακολουθήσουμε την λογική της αντιστοίχησης 1 έως r, στα αντικείμενα της πρώτης ομάδας θα αντιστοιχίσουμε την τιμή 3, της δεύτερης 2 και της τρίτης 1. Η μέθοδος αυτή όμως δεν είναι πάντα η καλύτερη. Προϋποθέτει ότι η απόσταση μεταξύ των ιδιοτήτων που χαρακτηρίζουν δύο διαδοχικές κλάσεις, είναι πάντα 1. Δηλαδή παραδεχόμαστε ότι δύο οποιεσδήποτε διαδοχικές κλάσεις απέχουν εξίσου μεταξύ τους. Στην πραγματικότητα δύο διαδοχικές κλάσεις είναι δυνατό να απέχουν περισσότερο ή λιγότερο από ότι δύο άλλες. Η έννοια της απόστασης, έχει σχέση με την ομοιότητα ή διαφοροποίηση μεταξύ κλάσεων.

Πρέπει λοιπόν η νέα μεταβλητή, που θα ορίσουμε να έχει τέτοιες τιμές για τις διάφορες κλάσεις, ώστε να εκφράζουν την αναλογία των αποστάσεων μεταξύ διαφορετικών κλάσεων, που να αντιστοιχεί στην ουσιαστική διαφορά μεταξύ των ιδιοτήτων που τις χαρακτηρίζουν.

Στην αρχή παρουσιάζουμε την μαθηματική ανάλυση για την εύρεση μιας τέτοιας μεταβλητής, που βασίζεται σε ένα κριτήριο διατήρησης της διασποράς. Κατόπιν, προτείνουμε ένα εύκολο, γρήγορο, όσο και αποτελεσματικό τρόπο εύρεσης μιας τέτοιας μεταβλητής, χρησιμοποιώντας τα αποτελέσματα του δενδρογράμματος μιας ταξινόμησης.

3. Κριτήριο βασισμένο στη διατήρηση της αδράνειας (διασποράς)

Θεωρούμε ότι έχουμε n αντικείμενα τα οποία χαρακτηρίζονται από J μεταβλητές. Με x_{ij} παριστάνεται η τιμή της j μεταβλητής για το i αντικείμενο. Η αδράνεια (διασπορά) του N νέφους αυτού των αντικειμένων, θα δίνεται από τον τύπο:

$$I(N)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}(x_{ij}-\bar{x}_i)^2$$

όπου με $\bar{x}$ παριστάνεται το κέντρο βάρους των δεδομένων.

Αν τώρα τα αντικείμενα έχουν ομαδοποιηθεί σε ένα διαμελισμό με r κλάσεις, δημιουργούμε μια νέα μεταβλητή m με r τιμές $m_1,m_2,..m_r$. Στο i αντικείμενο, αντιστοιχεί η m_k τιμή αν το αντικείμενο ανήκει στην k κλάση.

Η αδράνεια διασπορά του νέφους των αντικειμένων, μόνο ως προς την μεταβλητή m, θα δίνεται από τον τύπο:

$$I_m(N)=\frac{1}{n}\sum_{i=1}^{r}p_i(m_i-m)^2$$

όπου με m παριστάνεται το κέντρο βάρους των δεδομένων και με p_i το βάρος κάθε κλάσης, δηλαδή το πλήθος των αντικειμένων της i κλάσης.

Σκοπός μας είναι ο προσδιορισμός των τιμών των $m_1,m_2,..m_r$, έτσι ώστε η αδράνεια του νέφους ως προς την m μεταβλητή να πλησιάζει όσο το δυνατό περισσότερο την αρχική ολική αδράνεια, ως προς το σύνολο των αρχικών μεταβλητών. Τότε λέμε ότι η νέα μεταβλητή είναι αντιπροσωπευτική των κλάσεων.

3.1 Υπόθεση αριθμητικής προόδου

Υποθέτουμε ότι οι τιμές $m_1,m_2,..m_r$, είναι όροι μιας αριθμητικής προόδου με λόγο ω και αρχικό όρο α. Τότε μπορούμε να γράψουμε

$m_1=a+k_1\omega$ $m_2=a+k_2\omega$ ... $m_r=a+k_r\omega$ οπότε η μέση τιμή m θα δίνεται από τον τύπο:

$$m=\frac{na+(p_1k_1+p_2k_2+..+p_rk_r)\omega}{n}=a+\frac{p_1k_1+p_2k_2+..+p_rk_r}{n}\omega$$

Η αδράνεια του νέφους των n σημείων ως προς την μεταβλητή m μπορεί να εκφραστεί με τον παρακάτω τύπο:

$$I_m(N)=\frac{1}{n}\sum_{i=1}^{r}p_i\left(k_i\omega-\frac{\sum_{j=1}^{r}p_jk_j}{n}\omega\right)^2=\frac{1}{n}\sum_{i=1}^{r}\omega^2p_i\left(k_i-\frac{\sum_{j=1}^{r}p_jk_j}{n}\right)^2=\frac{1}{n}\sum_{i=1}^{r}\frac{\omega^2}{n^2}p_i\left(nk_i-\sum_{j=1}^{r}p_jk_j\right)^2$$

Στον τελευταίο τύπο, άγνωστα είναι τα ω, $k_1,k_2,..k_r$ για το οποία ξέρουμε ότι είναι θετικοί ακέραιοι αριθμοί, διατεταγμένοι σε αύξουσα σειρά. Χωρίς βλάβη της

γενικότητας και εφόσον μας ενδιαφέρει η αναλογία μεταξύ των $k_1,k_2,..k_r$, μπορούμε να υποθέσουμε ότι ω=1.

Δεν υπάρχει αναλυτική μέθοδος για τον προσδιορισμό των $k_1,k_2,..k_r$ αφού ισχύει μόνο μια εξίσωση. Ισχύουν όμως και οι ανισότητες $0<k_1<k_2<..<k_r$. Μπορούμε λοιπόν να αντιμετωπίσουμε το πρόβλημα σαν πρόβλημα βελτιστοποίησης με περιορισμούς. Η ποσότητα που θέλουμε να μεγιστοποιήσουμε είναι ο λόγος των δύο αδρανειών, η αλλιώς το ποσοστό ερμηνευόμενης αδράνειας.

$$\max \frac{I_m(N)}{I(N)}$$

με περιορισμούς $0<k_1<k_2<..<k_r$

Πρόκειται για ένα μη γραμμικό πρόβλημα και για να βρούμε μια λύση του, είναι απαραίτητη η θεώρηση αρχικής τιμής για τους αγνώστους. Θέτουμε λοιπόν σαν αρχική τιμή για τα $k_1,k_2,..k_r$ τις τιμές 1,2,.. r, και εφαρμόζουμε ένα αλγόριθμο μη γραμμικής βελτιστοποίησης.

4. Αδράνεια (διασπορά) ενός διαμελισμού δεδομένων

Η ολική αδράνεια των δεδομένων ως προς το κέντρο βάρους τους g, συμβολίζεται με I(g) και υπολογίζεται από τον τύπο $I(g) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}(x_{ij} - \bar{x}_i)^2$.

Σύμφωνα με το θεώρημα του Huygens [Roux,1985], η ολική αδράνεια μπορεί να αναλυθεί σε δύο μέρη. Σε αδράνεια μέσα στις κλάσεις και σε αδράνεια μεταξύ των κλάσεων. Έτσι αν τα δεδομένα μας ομαδοποιούνται σε Q κλάσεις με κέντρο βάρους g_q και βάρος p_q για την κάθε μία, τότε η ολική αδράνεια γράφεται:

$$I(g) = \sum_{q=1}^{Q} I(g_q) + \sum_{q=1}^{Q} p_q d^2(g_q,g) = \sum_{q=1}^{Q}\sum_{i=1}^{p_q}(x_q - g_q)^2 + \sum_{q=1}^{Q} p_q d^2(g_q,g)$$

Όσο πιο ομογενή είναι τα στοιχεία μιας κλάσης τόσο μικρότερη είναι η αδράνεια μέσα στις κλάσεις και τόσο μεγαλύτερη η αδράνεια μεταξύ των κλάσεων.

Αν τώρα θέλουμε να αντιστοιχήσουμε ένα αριθμό σε κάθε κλάση που να αντιπροσωπεύει όσο γίνεται καλύτερα τον διαμελισμό, με την έννοια της ελαχιστοποίησης της εσωτερικής αδράνειας $\left(\sum_{q=1}^{Q}\sum_{i=1}^{p_q}(x_q - g_q)^2\right)$ και άρα

μεγιστοποίησης της αδράνειας μεταξύ κλάσεων $\left(\sum_{q=1}^{Q} p_q d^2(g_q, g)\right)$, πρέπει ο αριθμός κάθε κλάσης να είναι το κέντρο βάρος της.

Για το πρόβλημά μας λοιπόν, δηλαδή την αναζήτηση ακεραίων αριθμών που να αντιπροσωπεύουν όσο γίνεται καλύτερα την διαφορά μεταξύ δύο κλάσεων, και έτσι ώστε να προσεγγίζουν την ολική αδράνεια των δεδομένων όσο το δυνατόν περισσότερο, προτείνουμε να αντιστοιχήσουμε σε κάθε κλάση, τον ακέραιο αριθμό που είναι κοντύτερα στο κέντρο βάρους της.

Το κέντρο βάρους της κλάσης όμως, συνήθως είναι ένα διάνυσμα και όχι απλός αριθμός. Έτσι γίνεται δύσκολη η εύρεση ενός μόνου αριθμού που να το προσεγγίζει.

Για το πρόβλημα όμως που αντιμετωπίζουμε, όπως αναφέραμε, μας ενδιαφέρουν κυρίως οι διαφορές μεταξύ των κλάσεων. Φροντίζουμε λοιπόν οι διαφορές αυτές να είναι ίσες, ή ανάλογες με τις διαφορές μεταξύ των κέντρων βάρους τους. Με τον τρόπο αυτό ικανοποιούνται οι προϋποθέσεις που θέσαμε προηγουμένως, δηλαδή αντιπροσώπευση κάθε κλάσης από ένα αριθμό ώστε να έχουμε το μέγιστο ποσοστό διατήρησης αδράνειας και παράλληλα την καλύτερη διακριτοποίηση μεταξύ των κλάσεων.

Η διαφορά μεταξύ των κέντρων βάρους κάθε κλάσης είναι εύκολο να υπολογιστεί όταν γνωρίζουμε τα στοιχεία κάθε κλάσης. Προτείνουμε να υπολογίζεται σαν αριθμός για την διαφορά δύο κλάσεων, η απόσταση μεταξύ των διανυσμάτων των κέντρων βάρους τους. Ξεκινώντας λοιπόν από την πρώτη κλάση, αντιστοιχούμε σ'αυτή τον αριθμό 1 και στη συνέχεια προσθέτουμε την απόσταση πρώτης-δεύτερης κλάσης (με ακέραια προσέγγιση) και αντιστοιχούμε τον νέο ακέραιο στη δεύτερη κλάση και συνεχίζουμε έτσι για τις υπόλοιπες κλάσεις.

Το πλεονέκτημα της μεθόδου αυτής, είναι ο υπολογισμός ακεραίων θετικών (αφού αντιστοιχούν σε αποστάσεις) αριθμών και οι οποίοι μάλιστα υπολογίζονται ανεξάρτητα από το πλήθος μεταβλητών που χαρακτηρίζουν τα δεδομένα μας. Επιπλέον, αν έχει προηγηθεί μια ταξινόμηση, έχει γίνει ήδη ο χωρισμός των δεδομένων σε κλάσεις και ο υπολογισμός των αποστάσεων των κέντρων βάρους, και έτσι δεν απαιτούνται νέοι υπολογισμοί.

Οι αποστάσεις μεταξύ των κέντρων βάρους των κλάσεων, προκύπτουν εύκολα και από το δενδρόγραμμα της ταξινόμησης των δεδομένων και αντιστοιχούν στο ύψος μεταξύ δύο κόμβων του δενδρογράμματος, όπως περιγράφουμε στην επόμενη παράγραφο.

5. Κριτήριο βασισμένο στο δενδρόγραμμα της ταξινόμησης

Έστω ότι ξεκινώντας από τα αρχικά δεδομένα, έχουμε καταλήξει με τις μεθόδους των προηγουμένων κεφαλαίων, στη δημιουργία των σημαντικότερων ομάδων των δεδομένων, οι οποίες παριστάνονται στο δενδρόγραμμα ιεράρχησης. Κάθε ομάδα, αντιστοιχεί σε ένα κόμβο του δενδρογράμματος αυτού, και τα ύψη των γραμμών μεταξύ των κόμβων, αντιστοιχούν ανάλογα, στις αποστάσεις μεταξύ των ομάδων.

Χρησιμοποιώντας την ιδιότητα αυτή, μπορούμε να ορίσουμε μια ακέραια μεταβλητή, η οποία θα έχει τιμές τέτοιες για δύο ομάδες, που θα αναλογούν στη απόσταση των κόμβων των ομάδων.

Για να ελέγξουμε την ικανότητα διάκρισης της παραπάνω οριζόμενης μεταβλητής, συγκρίνουμε την αδράνεια των κλάσεων ως προς την νέα αυτή μεταβλητή, με την αρχική αδράνεια των δεδομένων. Την αδράνεια δηλαδή που προκύπτει θεωρώντας τα αρχικά δεδομένα με όλες τις μεταβλητές τους.

Όσο κοντινότερα βρίσκεται η αδράνεια της νέας μεταβλητής, στην αρχική αδράνεια, τόσο καλύτερα προσαρμοζόμενες στα δεδομένα είναι οι τιμές της.

5.1 Παράδειγμα

Έχουμε τα δεδομένα που παριστάνει ο πίνακας 1.

Πίνακας 1

TYP	0	4	0	0	2	0
SAL	0	2	0	0	1	0
ARO	8	0	5	3	0	3
PEN	7	0	5	2	0	3
MEN	2	2	0	0	0	1
ORI	4	0	2	0	0	6
STA	3	0	1	3	0	0

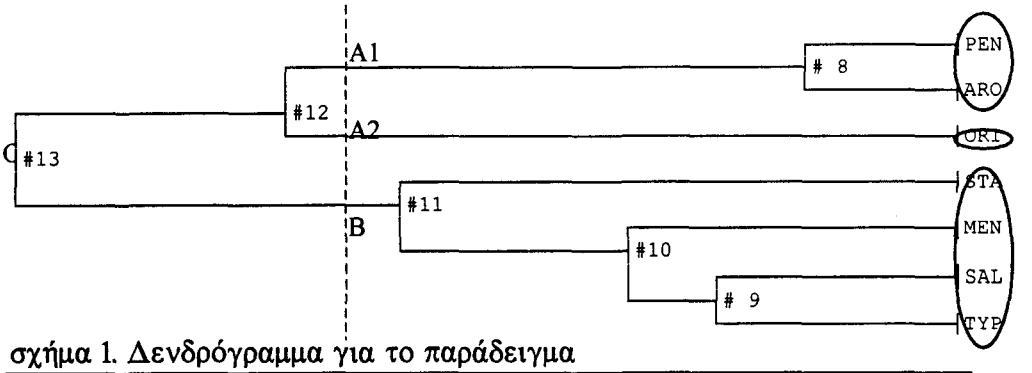
Εφαρμόσαμε ιεραρχική ταξινόμηση στα δεδομένα αυτά και πήραμε το δενδρόγραμμα που φαίνεται στο σχήμα 1. Ο καλύτερος διαμέλισμός των δεδομένων με το κριτήριο επιλογής βέλτιστου διαμελισμού, είναι αυτός που περιλαμβάνει τις τρεις κλάσεις A1 με στοιχεία {PEN, ARO}, B με στοιχεία {STA,MEN,SAL,TYP} και A2 που περιέχει το {ORI}.

πίνακας 2
Αποστάσεις κόμβων

κόμβος				βάρος	απόσταση
8	A1	PEN	ARO	2	1.414214
9		SAL	TYP	2	2.236068
10		MEN	9	3	3.027521
11	B	STA	10	4	5.047992
12	A2	8	ORI	3	6.062601
13	O	12	11	7	8.459352

πίνακας 3
Αδράνεια διαμελισμών

κλάσεις	αδράνεια	ποσοστό
2	14.5391	0.6658
3	18.0153	0.8250
4	20.5510	0.9411
5	21.3367	0.9771
6	21.6939	0.9935
7	21.8367	1.0000



σχήμα 1. Δενδρόγραμμα για το παράδειγμα

Αν θελήσουμε να αντιστοιχήσουμε μια μεταβλητή στις κλάσεις αυτές, ώστε να προχωρήσουμε στην συνέχεια σε κάποια άλλη ανάλυση μεταξύ των κλάσεων, συνήθως δίνουμε για τιμές στην μεταβλητή αυτή, τις τιμές 1,2,3 αντίστοιχα για κάθε κλάση. Με τον τρόπο αυτό, η αδράνεια των κλάσεων ως προς την μεταβλητή αυτή είναι 0.4081.

Η αρχική ολική αδράνεια των δεδομένων, όπως υπολογίζεται από τον τύπο:

$$I(N) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}(x_{ij} - \bar{x}_i)^2$$

είναι 21.8367. Αντιστοιχώντας λοιπόν τους αριθμούς 1, 2 και 3 στις 3 κλάσεις, δεν προσεγγίζουμε την αρχική ολική αδράνεια.

Αν τώρα χρησιμοποιήσουμε τους κόμβους του δενδρογράμματος, παρατηρούμε (πίνακας 2) ότι η απόσταση μεταξύ των κλάσεων A1 ({PEN,ARO}) και B ({STA,MEN,SAL,TYP}) είναι 3.6 ενώ μεταξύ κλάσεων B και A2 ({ORI}) είναι 1.02. Έτσι τώρα προτείνουμε για τιμές της νέας μεταβλητής, τις ακέραιες τιμές 1, 5, 6 για τις κλάσεις A1, B και A2 αντίστοιχα. Η αδράνεια του διαμελισμού όπως αυτός περιγράφεται από τη νέα μεταβλητή, είναι 3.714 και παρατηρούμε ότι πλησιάζει περισσότερο στην αρχική ολική αδράνεια, απ' ότι αν αντιστοιχίσουμε τις τιμές 1, 2, 3 στις παραπάνω κλάσεις.

Φαίνεται λοιπόν και από το παράδειγμα πως από τα αποτελέσματα μιας ιεραρχικής ταξινόμησης, και με πολύ εύκολο και γρήγορο τρόπο, μπορούμε να αντιστοιχήσουμε μια νέα ακέραια μεταβλητή στις κλάσεις ενός οποιουδήποτε

διαμελισμού των δεδομένων. Η νέα αυτή μεταβλητή, περιγράφει καλύτερα τη διακριτοποίηση μεταξύ των κλάσεων.

6. Συμπεράσματα

Παρουσιάσαμε στο κεφάλαιο αυτό, την πρότασή μας που αφορά στην αντιστοίχηση τιμών μιας νέας ακέραιας μεταβλητής στα αντικείμενα που είναι ομαδοποιημένα σε κλάσεις ενός διαμελισμού. Σκοπός είναι η εύρεση τέτοιων τιμών που να διακρίνουν μεταξύ τους τις κλάσεις σύμφωνα με τις ιδιότητές τους, και παράλληλα να διατηρούν όσο γίνεται καλύτερα την αρχική ολική αδράνεια των δεδομένων.

Αναλύσαμε την μαθηματική προσέγγιση με την μέθοδο της βελτιστοιποίησης του γραμμικού προγραμματισμού, και επίσης την στατιστική προσέγγιση χρησιμοποιώντας το θεώρημα του Huygens.

Τέλος προτείναμε ένα εύκολο και αποτελεσματικό τρόπο για την λύση του παραπάνω προβλήματος, χρησιμοποιώντας το δενδρόγραμμα των αποτελεσμάτων μιας ιεραρχικής ταξινόμησης. Με τον τρόπο αυτό, αντιστοιχούμε ακέραιους θετικούς αριθμούς στα αντικείμενα των κλάσεων, οι οποίοι υπολογίζονται από τις αποστάσεις μεταξύ των κόμβων του δενδρογράμματος και διατηρούν όσο γίνεται περισσότερο την αρχική αδράνεια των δεδομένων.

Θα θέλαμε να παρατηρήσουμε ακόμη, ότι είναι δυνατό να μην αντιστοιχήσουμε μόνο ακέραιες τιμές στα αντικείμενα των κλάσεων αλλά οποιαδήποτε θετική τιμή. Η θεωρητική διαδικασία είναι ακριβώς όμοια. Προτείνουμε την πλησιέστερη ακέραια τιμή, για να αυξήσουμε την διακριτότητα των κλάσεων με τη νέα μεταβλητή.

ΕΠΙΛΟΓΟΣ

Με την παρούσα διατριβή, επιθυμία μας ήταν να παρουσιάσουμε τόσο τη θεωρητική μελέτη όσο και την πρακτική διαδικασία που πρέπει να ακολουθήσει κάποιος, προκειμένου να προβεί σε μια ανάλυση με τις μεθόδους της Ανάλυσης Δεδομένων και ειδικότερα με την Ιεραρχική Ταξινόμηση.

Καταρχήν όσον αφορά στην παρουσίαση των πινάκων δεδομένων μπορούμε να έχουμε μια αμεσότερη και χρησιμότερη εικόνα της σχέσης δύο ή περισσοτέρων μεταβλητών, χρησιμοποιώντας τους πίνακες των υπό συνθήκη σχετικών συχνοτήτων, τον απλό πίνακα Burt ή τον γενικευμένο πίνακα Burt για ποσοτικές μεταβλητές. Αν τα δεδομένα που επιθυμούμε να επεξεργαστούμε, περιγράφονται από τριδιάστατους πίνακες, πρέπει να ακολουθήσουμε τη διαδικασία που προτείνουμε, για την επεξεργασία αυτών των πινάκων.

Λόγω του ότι η επιλογή της μετρικής απόστασης μεταξύ αντικειμένων ή μεταβλητών παίζει σπουδαίο λόγω στην Ιεραρχική Ταξινόμηση, παρουσιάσαμε τα προταθέντα μέτρα απόστασης ή ομοιότητας και τον τρόπο επιλογής του καταλληλότερου κάθε φορά. Αυτό εξαρτάται τόσο από τη φύση των δεδομένων όσο και από τις ανάγκες της έρευνας.

Εκτός από την επιλογή μετρικής απόστασης, για την πορεία της Ταξινόμησης χρειάζεται και η επιλογή της καλύτερης μεθόδου ομαδοποίησης, ώστε οι ομάδες δεδομένων που προκύπτουν, να προσεγγίζουν όσο γίνεται καλύτερα, την πραγματικότητα. Προτείναμε ένα κριτήριο για την επιλογή της καλύτερης μεθόδου, αλλά και για την επιλογή του βέλτιστου διαμελισμού των δεδομένων σε μη καθορισμένο εκ των προτέρων πλήθος κλάσεων. Επίσης το κριτήριο μπορεί να εφαρμοσθεί για την εύρεση του καλύτερου διαμελισμού σε ένα εκ των προτέρων καθορισμένο πλήθος κλάσεων.

Ένα άλλο πολύ χρήσιμο πεδίο έρευνας είναι ο τρόπος παρουσίασης και ερμηνείας των αποτελεσμάτων μιας Ιεραρχικής Ταξινόμησης. Δηλαδή η εύρεση των μεταβλητών που είναι υπεύθυνες για την δημιουργία ή για τον χραρακτηρισμό των δημιουργούμενων ομάδων. Θεωρούμε το στάδιο αυτό απαραίτητο για την εξαγωγή των σωστών συμπερασμάτων. Ακόμη μπορούμε να εντοπίσουμε τις σχέσεις συνεπαγωγής μεταξύ των μεταβλητών αυτών, χρησιμοποιώντας τη μέθοδο της Συνεπαγωγικής Ταξινόμησης μεταξύ μεταβλητών.

Στην περίπτωση που θέλουμε να προχωρήσουμε σε μια περαιτέρω ανάλυση των αποτελεσμάτων με μια άλλη μέθοδο Ανάλυσης Δεδομένων, χρειάζεται η

εύρεση μιας ακέραιας μεταβλητής για το χαρακτηρισμό των υπάρχουσων ομάδων. Προτείνουμε ένα τρόπο προσδιορισμού μιας τέτοιας μεταβλητής, ώστε αυτή να χαρακτηρίζει από την μια τις ιδιότητες της κάθε ομάδας, και από την άλλη τις διαφορές μεταξύ των ομάδων.

Σήμερα όλο και περισσότερο, υπάρχουν δεδομένα τα οποία αφορούν αντικείμενα που περιγράφονται από λογικές σχέσεις μεταβλητών. Αυτά ονομάζονται "συμβολικά" αντικείμενα. Παρουσιάσαμε μια νέα μέθοδο Ταξινόμησης, τη Συμβολική Ταξινόμηση, η οποία επεξεργάζεται τα αντικείμενα αυτά. Προτείναμε μια νέα μετρική απόστασης μεταξύ τέτοιων αντικειμένων, για τη μέτρηση της μεταξύ τους προσέγγισης.

Τέλος, στην περίπτωση που έχουμε πάρα πολλά δεδομένα ή όταν δεν μπορούμε να υπολογίσουμε αποστάσεις μεταξύ αντικειμένων, προτάθηκαν πρόσφατα οι λεγόμενοι "Γενετικοί αλγόριθμοι", όπου με τυχαίο και επεναλαμβανόμενο τρόπο δημιουργούμε νέους διαμελισμούς από τους ήδη υπάρχοντες. [Kettaf,1996] Στην παρούσα διατριβή δεν ασχοληθήκαμε με τους αλγορίθμους αυτούς, αλλά ελπίζουμε να δώσαμε το έναυσμα για κάποια άλλη μελλοντική έρευνα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Δριτσάκης Ν. Παπαδημητρίου Γ., 1993, "Ταξινόμηση και διαχρονική εξέλιξη των 12 χωρών της Ε.Ο.Κ. ως προς τα μη νομισματικά μεγέθη του τομέα της υγείας για την περίοδο 1960-1987", Πρακτικά 6ου Πανελληνίου Συνεδρίου Στατιστικής, Θεσσαλονίκη, σελ 71-83.

Ζαχαροπούλου Χ., 1989, "Στατιστική μέθοδοι-εφαρμογές", Θεσσαλονίκη.

Καρακώστας Κ., "Επαγωγική Στατιστική", Πανεπιστήμιο Ιωαννίνων.

Καραπιστόλης Δ., 1996, "Δημιουργία λογισμικού για την κατάρτιση Φερέγγυου Χαρτοφυλακίου με μεθόδους της Ανάλυσης Δεδομένων", Διδακτορική Διατριβή, Πανεπιστήμιο Μακεδονίας.

Κουνιά Σ. Κολυβά-Μαχαίρα Φ. Μπαγιάτη Κ. Μπόρα-Σέντα Ε., 1985, "Εισαγωγή στη Στατιστική", Θεσσαλονίκη.

Παπαδημητρίου Γ. Φλώρου Γ., 1992, "Στατιστική επεξεργασία του τρισδιάστατου πίνακα δεδομένων των πρωτοετών φοιτητών του Πανεπιστημίου Μακεδονίας με τις μεθόδους της Ανάλυσης Δεδομένων. 5ο Πανελλήνιο Συνέδριο Στατιστικής", Βόλος.

Παπαδημητρίου Γ. Φλώρου Γ., 1993, "Προσδιορισμός της ιδανικότερης μεθόδου ιεράρχησης μεταξύ των min, max, moyenne με την ευκλείδεια μετρική", 6ο Πανελλήνιο Συνέδριο Στατιστικής, σελ. 286-294, ΕΣΙ.

Παπαδημητρίου Γ. Φλώρου Γ., 1994, "Μελέτη τριών διαφορετικών αποστάσεων για την ταξινόμηση σε ποιοτικές μεταβλητές (Ταξινόμηση σε λογικά δεδομένα)", 7ο Πανελλήνιο Συνέδριο Στατιστικής, ΕΣΙ.

Παπαδημητρίου Γ. Φλώρου Γ., 1995, "Ανασκόπηση των αποστάσεων για την ταξινόμηση", 8ο Πανελλήνιο Συνέδριο Στατιστικής, ΕΣΙ.

Παπαδημητρίου Γ. Φλώρου Γ., 1996, "Συμβολή της ευκλείδειας και Χ2 μετρικής στον προσδιορισμό της ιδανικότερης ταξινόμησης κατά αύξουσα ιεραρχία", Επιστημονική Επετηρίδα Πανεπιστημίου Μακεδονίας,Τιμητικός τόμος Ιωάννου Λιάκη, Θεσσαλονίκη, σελ. 545-581.

Παπαδημητρίου Γ., 1989, "Στατιστική: Τεύχος ΙΙ Επαγωγική Στατιστική", Εκδ. Παρατηρητής, Θεσσαλονίκη.

Παπαδημητρίου Γ., 1990, "Στατιστική: Τεύχος Ι Περιγραφική Στατιστική", Εκδ. Παρατηρητής, Θεσσαλονίκη.

Παπαδημητρίου Γ., 1990, "Συστήματα εισαγωγικών εξετάσεων και επίδρασή τους στη δομή του φοιτητικού πληθυσμού των οικονομικών σχολών", Παρατηρητής, Θεσσαλονίκη.

Παπαδημητρίου Γ., 1991, "Στατιστική επεξεργασία τρισδιάστατων πινάκων με την Παραγοντική Ανάλυση των Αντιστοιχιών. 4ο Πανελλήνιο Συνέδριο Στατιστικής", Πάτρα.

Παπαδημητρίου Γ., 1994, "Μέθοδοι ανάλυσης δεδομένων (Πανεπιστημιακές Παραδόσεις)", Εκδ. Πανεπιστημίου Μακεδονίας, Θεσσαλονίκη.

Παπαϊωάννου Τ. Φερεντίνος Κ., 1994, "Ιατρικά Μαθηματικά και Στοιχεία Βιομαθηματικών", Ιωάννινα.

Ξένη

Agresti A., 1990, "Categorical Data Analysis", New York: Wiley.

Anderberg M., 1973, "Cluster analysis for applications", Academic Press London.

Anderson T., 1958, "Introduction to multivariate statistical analysis", Wiley.

Andrews D.F. Herzberg A.M., 1985, "Data A collection of Problems from Many Fields for the Student and Research Worker", Springer-Verlag, New-York.

Ball G.H. Hall D.J., 1965, "Isodata a novel method of data analysis and pattern classification", Techn. report SRI Project 5533, Stanford Research Institute, California.

Ball G.H., 1971, "Classification analysis", SRI Project 5533, Stanford Research Institute, California.

Batagelg V., 1990, "Comparing Resemblance Measures", Journal of Classification, pp 73-90.

Benzecri F, 1980, "Introduction a la classification automatique d'apres un exemple des données medicales", Les cahiers de l'analyse des données, no 3, pp 311-340, Dunod, Paris.

Benzecri J.P, 1982, "Construction d'une classification ascendante hierarchique par la recherche en chaine de voisins reciproques", Les cahiers de l'analyse des données, no 7, Dunod, Paris.

Benzecri J.P. Benzecri F. Cheung Y.L. Maiza S., 1985, "Aides a l'interpretation et etiquetage des arbres en classification ascendante hierarchique: listage Facor Vacor et Insup", Les cahiers de l'analyse des données, no 3, Dunod, Paris.

Benzecri J.P. et collaborateurs, 1973, "L'analyse des données: L'analyse des correspondance", Dynod, Paris.

Benzecri J.P. et collaborateurs, 1973, "L'analyse des données: La taxinomie", Dunod, Paris.

Benzecri J.P. Lebreaux M.O. Jambu M., 1980, "Aide a l'interpretation en classification automatique", Les cahiers de l'analyse des données, nol, pp.101-123, Dunod, Paris.

Benzecri J.P.et collaborateurs, 1980, "Pratique de l' Analyse des données", Vol 1, Dunod, Paris.

Bertier P. Bouroche J.-M., 1975, "Analyse des données multidimensionnelles", Presses Universitaires de France, Paris.

Bochenski I.M., 1970, "A history of formal logic", I.Thomas, trans., New York:Chelsea Publishing Co.

Bock H., 1994, "Classification and Clustering: Problems for the Future", in New approaches in Classification and Data Analysis,Springer-Verlag, pp. 3-24.

Bouroche J.-M. Saporta G., 1980, "L'analyse des données", Collection Que je--sais?, no 1854, Presses Universitaires de France, Paris.

Cazes P., 1986, "Correspondance entre deux ensembles et partition de ces deux ensembles", Les cahiers de l'analyse des données, no 11, pp. 335-340, Dunod, Paris.

Celeux G. Diday E. Govaert G. Ralambondrainy H., 1989, "Classification automatique des données", Dunod, Paris.

Chambers J.M. Cleveland W.S. Kleiner B. Tukey P.A., 1983, "Graphical Methods for Data Analysis", Pacific Grove, Wadsworth.

Cibois P., 1983, "L'analyse factorielle", Presses Universitaires de France, Paris.

Cohen D., 1993, "Les genes de l'espoir", Robert Laffont, Paris.

Daude F, 1992, "Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hierarchique par A.V.L.", These, Universite de Rennes 1.

De Leeuw J. Heiser W. Meulman J. Critchley F., 1989, "Multidimensional Data Analysis", Leiden, DSWO Press.

Denimal J, 1994, "Analyse des interactions entre k partitions prises 2 a 2 Theorie et Application en Biologie", Revue statistique appliquee, XLII(1), pp.19-40.

Der Megreditchian G., 1988, "Le traitement statistique des données multidimensionnelles", Tome 1, Ecole Nationale de la Meteorologie, Paris.

Diday E. Bochi S. Brossier G. Celeux G., 1979, "Optimisation en classification automatique", INRIA.

Diday E. Lemaire J. Pouget J. Testu F., 1982, "Eléments d' analyse des données", Dunod, Paris.

Diday E., 1971, "La méthode des nuées dynamiques", Rev. Stat. App. Vol 19, pp.19-34

Diday E., 1993, "An introduction to symbolic data analysis", INRIA RR1936

Diday E., 1994, "Science of objects attempting a historical review", Journees symbolique-numerique, Orsay.

Droesbeke J. fichet B. Tassi P., 1992, "Modeles pour l'analyse des données Multidimensionnelles", Econimica, Paris.

Escofier B. Pages J., 1990, "Analyses factorielles simples et multiples", Dunod.

Escofier B., 1981, "Quelque indices pour comparer des tableaux de contingence. Statistique et analyse des données", no 1.

Florou G., 1996, "Colour image processing using symbolic data analysis", Applications of Digital Image Processing XIX, SPIE Proceedings Vol. 2847, Editor(s): Andrew G. Tescher, Lockheed Martin Palo Alto Research Lab., Saratoga, CA, USA, pp 252-260.

Foucart T., 1984, " Analyse factorielle de tableaux multiples", Masson, Paris.

Fowlkes E.B. Gnanadesikan R. Kettenring J.R., 1988, "Variable Selection in Clustering", Journal of classification, pp205-228.

Gifi A., 1990, "Nonlinear Multivariate Analysis", Chichester, Wiley.

Gnanadesikan R., Kettenring J.R. Landwehr J.M., 1977, "Interpreting and assessing the results of cluster analyses", Bulletin of the International Statistical Institute, no 47, pp. 451-463.

Good I.J., 1965, "Categorization of classification", Mathematics and computer science in Biology and Medicine,

Goodman L.A. Kruskal W.H., 1954, "Measures of association for cross-classification", Journal of American Statistical Association, no 49, pp. 723-764.

Goodman L.A., 1991, "Measures, models and graphical displays in the analysis of cross-classified data (with discussion)", Journal of American Statistical Association, no 86, pp. 1085-1137.

Govaert G., 1984, "Classification simultanee de tableaux binaires", in E. Diday et col. (eds), Data Analysis and Informatics, III, pp. 223-236, Amsterdam, North-Holland.

Gowda K..C. Diday E., 1992, "Symbolic clustering using a new similarity measure", IEEE Transsystem, man, cybernetics, vol 22, no 2, pp. 368-378

Gras R. Briand H. Peter P., 1995, "Hierarchie d'implications. Conference internationale sur L'analyse des données ordinales et symboliques", Paris.

Gras R. Larher A., 1992, "L'implication statistique, une nouvelle methode d'analyse des données", Mathematique, Informatique et Sciences Humaines, no 120.

Gras R., 1993, "Une methode de classification non symetrique: implication statistique", bulletin no 1 de la Societe Francophone de Classification.

Greenacre M., 1988, "Clustering the rows and columns of a contingency table", Journal of Classification, no 5, pp.39-51.

Greenacre M., 1993, "Correspondence Analysis in Practice", Academic Press, London.

Guigou J-L, 1977, "Méthodes multidimensionnelles", Dunod, Paris.

Guttman L., 1941, "The Quantification of a Class of Attributes: a Theory and Method of Scale Construction", in The Prediction of Personal Adjustment, Social Science Research Council, New York, pp. 319-348.

Hirschfeld H.O., 1935, "A Connection between correlation and Contingency", Cambridge Philosophical Soc. Proc., 31, pp. 520-524.

Hotteling H., 1933, "Analysis of a complex of statistical variables into principal components", Journal of Educational Psychology, vol 24, pp.417-441, 498-520.

Jambu M. Benzecri J.P. Lebeaux M.O., 1980, "Aide a l' interpretation en classification automatique", Les Cahiers de l' Analyse des données, no 1.

Jambu M., 1978, "Classification Automatique pour l' Analyse des données I Methodes et algorithmes", Dunod, Paris.

Kaufman L. Rousseeuw P., 1990, "Finding Groups in Data: an introduction to Cluster Analysis", John Wiley & Sons, Inc.

Kettaf F.Z. Asselin de Beauville J-P., 1996, "Sur les algorithmes genetiques en classification supervisee par partition", Societe Francophone de Classification, bulletin no7.

KIT-Etudiant SPAD.N, 1993, CISIA France.

Larher A., 1991, "Implication statistique et applications a l'analyse de demarches de preuve mathematique", These de l'Universite de Rennes I.

Lavit Ch., 1988, " Analyse conjointe de tableaux quantitatifs", Masson, Paris.

Lebart L. Morineau A. Fenelon J.P., 1982, "Traitement des données statistiques", Dunod, Paris.

Lebart L. Morineau A. Tabard N., 1977, "Techniques de la description statistique", Dunod, Paris.

Lerman I., 1970, "Les bases de la classification automatique", Gauthier-Villars, Paris.

Lerman I., 1981, "Classification et Analyse ordinale des données", Dunod, Paris.

Louis P., 1956, "Aristote, les parties des animaux", reprinted by Les belles lettres.

Marchetti F., 1989, "Contribution a la Classification des données Binaires et Qualitatives", These, Universite de Metz.

Moussaoui A., 1987, "Qualite de la representation des classes d'une CAH sur un tableau construit en cumulant des blocs de variables", Les cahiers de l'analyse des données, vol XII no2, pp.237-242, Dunod, Paris.

Ngouenet R., 1992, "Une nouvelle famille d'indices de dissimilarite pour la MDS", Publication Interne no 766, Irisa, Rennes, France.

Nishisato S., 1980, "Analysis of Categorical Data: Dual Scaling and its Applications", University of Toronto Press.

Ouadrani A., 1994, "Generalisation du tableau de Burt", Revue de statistiques appliquees vol XIX no 2, pp.229-246.

Pages J-P. Caillez F., 1976, "Introduction a l'analyse des données", SMASH, Paris.

Pearson K., 1901, "On lines and planes of closest fit to system of points in space", Phil. Mag., vol2 no 11, pp. 559-572.

Pottier P., 1994, "Mesures de la liaison entre deux variables qualitatives: Relation entre un coefficient de correlation generalise et le x2", Revue statistique appliquee, XLII(1),41-61.

Roux M., 1985, "Algorithmes de classification", Masson, Paris.

Sneath Socal, 1972, "Numerical Taxonomy", Freeman, San Francisco, London.

Socal R. Michener C., 1958, "A statistical method for evaluating systematic relationships", Univ. Kansas Sci. Bull., 38, pp.1409-1438.

Spearman C., 1904, "General intelligence objectively determined and mesured", American Journal of Psychology, vol 15, pp. 201-292.

Tallur B., 1983, "Methode d'interpretation d'une classification hierarchique d'attributs-modalites pour l'explication d'une variable", Revue de statistiques appliquees vol XXXI no 1, pp. 25-43.

Tallur B., 1986, "Un nouvel indice d' agregation en classification ascendante hierarchique", Revue de statistiques appliquees vol XXXIV no 2, pp. 53-62.

Tallur B., 1988, "Contribution a l'analyse exploratoire de tableaux de contingence par la classification", Thèse, Universitè de Rennes 1.

Tenenbaus, 1994, "Methodes statistiques en gestion", Dunod, Paris.

The panelist, 1996, Newsletter published by the Office for Official Publications of European Communitites.

Torrens-Ibern J., 1972, "Modeles et Methodes de L'analyse Factorielle", Dunod, Paris.

Van Cutsem, 1994, "Classification and Dissimilarity Analysis", Lecture Notes in Statistics, Springer-Verlag, New-York.

Volle M., 1981, "Analyse des données", Economica, Paris.

ΠΑΡΑΡΤΗΜΑ
ΠΑΡΟΥΣΙΑΣΗ ΤΟΥ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΤΗΝ ΤΑΞΙΝΟΜΗΣΗ

Γενικά

Η δημιουργία του λογισμικού που υλοποιεί τις διάφορες μεθόδους Ταξινόμησης και τη Συνεπαγωγική Ταξινόμηση, καθώς και τις διαδικασίες για τη δημιουργία και επεξεργασία πινάκων δεδομένων, συνοδεύει τη διδακτορική διατριβή της Φλώρου Γιαννούλας που υποβλήθηκε στο Πανεπιστήμιο Μακεδονίας.

Το πρόγραμμα έχει γραφεί με τη γλώσσα προγραμματισμού Turbo Pascal και μπορεί να χρησιμοποιηθεί από οποιοδήποτε υπολογιστή PC που διαθέτει τουλάχιστον μνήμη 512 KB, οθόνη 14΄΄ με κάρτα VGA.

Έγινε προσπάθεια το πρόγραμμα να είναι φιλικό προς τον χρήστη και να αντιμετωπίζει τυχόν λάθη που αυτός κάνει, εκτυπώνοντας το κατάλληλο μήνυμα. Οι επιλογές των διαδικασιών γίνονται μέσω διαφόρων menu επιλογών. Ο χρήστης μπορεί να εγκαταλείψει οποιαδήποτε διαδικασία, πατώντας το πλήκτρο ESC.

Εγκατάσταση του προγράμματος στον υπολογιστή

Το πρόγραμμα μπορεί να δουλέψει κατευθείαν από την δισκέτα και δεν χρειάζεται ιδιαίτερη εγκατάσταση. Στην περίπτωση που κάποιος θέλει να το εγκαταστήσει στο σκληρό δίσκο, αρκεί να δημιουργήσει ένα κατάλογο με το όνομα ierarxia και μέσα σ' αυτό να αντιγράψει όλα τα αρχεία της δισκέτας.

Οι εργασίες αυτές από το λειτουργικό MS-DOS, αν υποθέσουμε ότι η δισκέτα βρίσκεται στο A: drive, μπορούν να γίνουν με τις εντολές:

```
C:\>MD IERARXIA
C:\>CD IERARXIA
C:\IERARXIA> COPY A:\*.*
```

Αν χρησιμοποιούμε WINDOWS, δεν έχουμε παρά να αντιγράψουμε τα αρχεία της δισκέτας στον κατάλογο ierarxia, χρησιμοποιώντας το αντίστοιχο πρόγραμμα της διαχείρισης αρχείων.

Διαδικασίες του προγράμματος

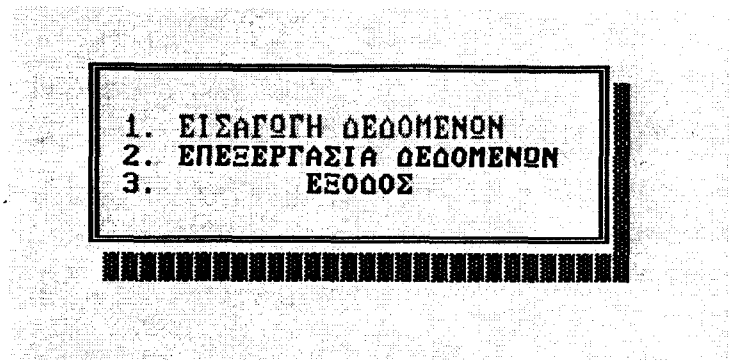
Οι διαδικασίες που εκτελεί το πρόγραμμα είναι

* *Εισαγωγή δεδομένων*
  - είτε απευθείας από το πληκτρολόγιο
  - είτε από ένα αρχείο ASC
* *Επεξεργασία ενός τρισδιάστατου πίνακα δεδομένων*
  - Δηλαδή μετατροπή του σε κατάλληλους διδιάστατους πίνακες (όπως περιγράφεται στο κεφάλαιο 2 της διατριβής).]
  - Δημιουργούνται πίνακες φέτες, αθροίσματα από πίνακες φέτες, αθροίσματα πινάκων και πίνακες διατεταγμένοι οριζόντια ή κάθετα.
* *Ιεραρχική ταξινόμηση των δεδομένων η οποία περιλαμβάνει*
  - Επιλογή μετρικής απόστασης μεταξύ αντικειμένων
  - Επιλογή μεθόδου ιεράρχησης
  - Έξοδο του πίνακα των κόμβων της ιεράρχησης
  - Έξοδο του δενδρογράμματος της ιεράρχησης
  - Έξοδο του ιστογράμματος με το ποσοστό ερμηνευόμενης αδράνειας όλων των διαμελισμών
  - Ερμηνεία των ανωτέρων κόμβων με τη μέθοδο vacor
* *Συνεπαγωγική ταξινόμηση μεταξύ δυαδικών μεταβλητών* κατά την οποία εμφανίζονται τα αποτελέσματα για
  - Πίνακα εντάσεων συνεπαγωγής
  - Πίνακα κόμβων της ιεράρχησης
  - Δενδρόγραμμα της ιεράρχησης με τις προσανατολισμένες σχέσεις συνεπαγωγής μεταξύ των μεταβλητών

Εκτέλεση ("τρέξιμο") του προγράμματος

Για να ξεκινήσουμε το πρόγραμμα, αρκεί να πληκτρολογήσουμε το όνομά του, στο περιβάλλον MS-DOS, ή χρησιμοποιώντας την εντολή RUN του περιβάλλοντος WINDOWS.

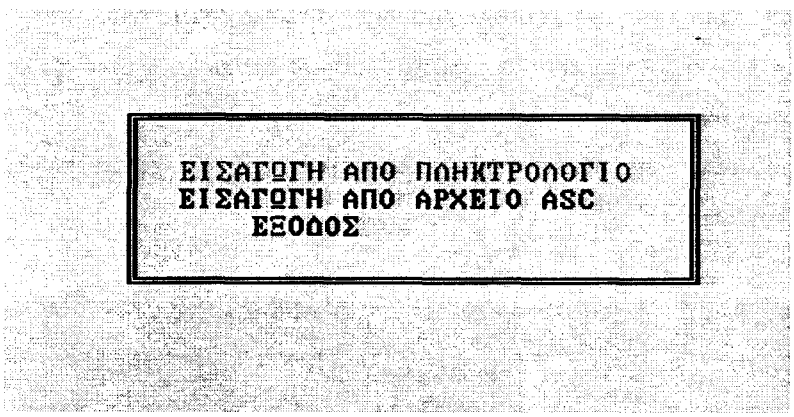
Στη συνέχεια τα μηνύματα και οι ενέργειες για την πορεία του προγράμματος είναι όμοια, είτε η εκκίνησή του έγινε από περιβάλλον MS-DOS, είτε από περιβάλλον WINDOWS. Η πρώτη οθόνη που εμφανίζεται είναι η παρακάτω οθόνη 1.



οθόνη 1. Αρχικό menu του προγράμματος

Διαδικασία εισαγωγής δεδομένων

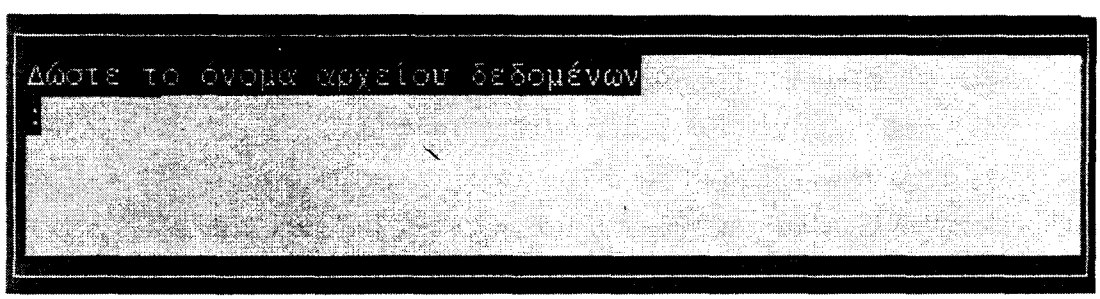
Από το αρχικό menu, όπως εμφανίζεται στην οθόνη 1, επιλέγουμε ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ και στη συνέχεια, στην οθόνη 2, επιλέγουμε ΕΙΣΑΓΩΓΗ ΑΠΟ ΠΛΗΚΤΡΟΛΟΓΙΟ ή ΕΙΣΑΓΩΓΗ ΑΠΟ ΑΡΧΕΙΟ ASC ανάλογα με το πως επιθυμούμε να εισάγουμε τα δεδομένα.



οθόνη 2. Εισαγωγή δεδομένων

ΕΙΣΑΓΩΓΗ ΑΠΟ ΠΛΗΚΤΡΟΛΟΓΙΟ

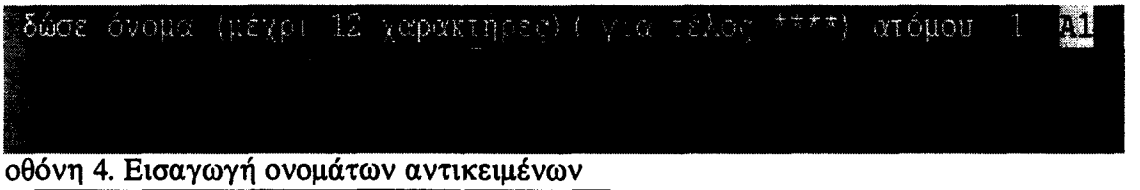
Εμφανίζεται η παρακάτω οθόνη 3, στην οποία πρέπει να πληκτρολογήσουμε το όνομα του αρχείου δεδομένων



οθόνη 3. Εισαγωγή ονόματος αρχείου

Πληκτρολογούμε το όνομα του αρχείου στο οποίο πρόκειται να εισάγουμε τα δεδομένα. Μπορούμε να πληκτρολογήσουμε το πλήρες όνομα, αν το αρχείο βρίσκεται σε άλλο κατάλογο ή σε δισκέτα. Στην περίπτωση που πληκτρολογήσουμε όνομα αρχείου που ήδη υπάρχει, εμφανίζεται προειδοποιητικό μήνυμα για το αν θέλουμε να αλλάξουμε όνομα ή να γράψουμε πάνω στο υπάρχον αρχείο.

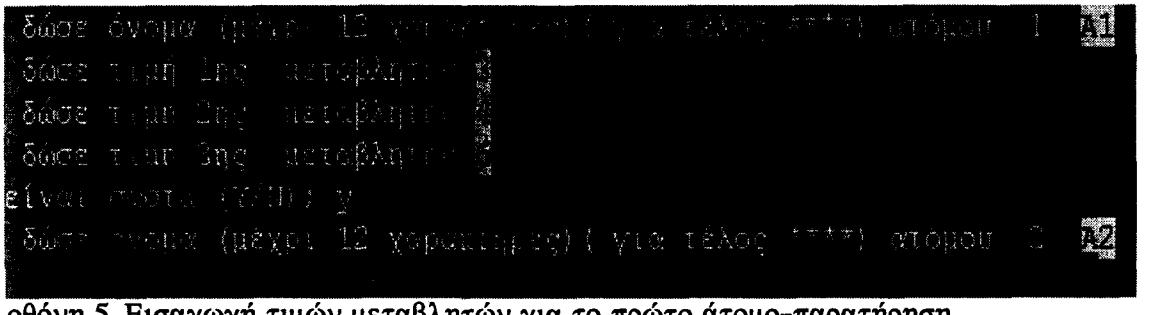
Κατόπιν πρέπει να δώσουμε τον αριθμό των μεταβλητών, που περιγράφουν καθεμία από τις παρατηρήσεις μας και στη συνέχεια εμφανίζεται η οθόνη 4,



οθόνη 4. Εισαγωγή ονομάτων αντικειμένων

στην οποία μπορούμε να πληκτρολογήσουμε ένα όνομα για κάθε άτομο (παρατήρηση). Αν δεν δώσουμε όνομα και πατήσουμε απλά το Enter το άτομο θα έχει εξ ορισμού, σαν όνομα A1.

Κατόπιν δίνουμε τις τιμές κάθε μεταβλητής για το άτομο αυτό, όπως παρουσιάζεται στην οθόνη 5.



οθόνη 5. Εισαγωγή τιμών μεταβλητών για το πρώτο άτομο-παρατήρηση

Στην ερώτηση είναι σωστά (Y/N); απαντάμε Y αν πληκτρολογήσαμε σωστά τα δεδομένα ή N αν υπαρχει κάποιο λάθος που πρέπει να διορθώσουμε.

Η διαδικασία αυτή επαναλαμβάνεται μέχρι να πληκτρολογήσουμε τις τιμές όλων των μεταβλητών για όλα τα άτομα (παρατηρήσεις).

Όταν τελειώσουμε, αντί για όνομα ατόμου πληκτρολογούμε ****, κλείνει το αρχείο δεδομένων και επιστρέφουμε στο menu που φαίνεται στην οθόνη 2.

ΕΙΣΑΓΩΓΗ ΑΠΟ ΑΡΧΕΙΟ ASC

Εμφανίζεται η παρακάτω οθόνη 6, στην οποία πρέπει να πληκτρολογήσουμε το όνομα του αρχείου δεδομένων.

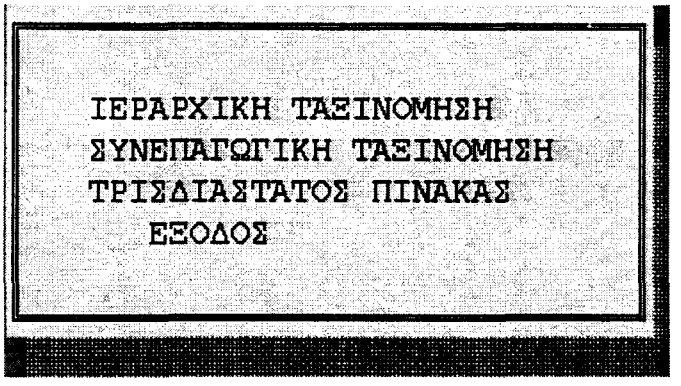
οθόνη 6. Εισαγωγή ονόματος αρχείου ASC

Γράφουμε το όνομα του ASC αρχείου, το οποίο μέσω του προγράμματος μετατρέπεται σε αρχείο binary με το ίδιο όνομα αλλά χωρίς προέκταση, και εμφανίζεται το μήνυμα που βλέπουμε στην οθόνη 7.

οθόνη 7. Επιτυχής μετατροπή ASC αρχείου σε δυαδικό

Διαδικασία επεξεργασίας δεδομένων

Από το αρχικό menu (οθόνη 1) επιλέγουμε ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ και εμφανίζεται η παρακάτω οθόνη 8.



οθόνη 8. Το menu με τις επιλογές της επεξεργασίας δεδομένων

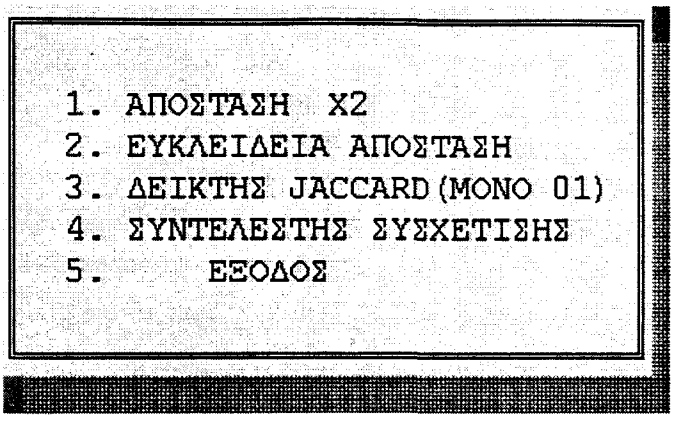
Στο σημείο αυτό μπορούμε να επιλέξουμε μια από τις παρακάτω διαδικασίες:

- ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ
- ΣΥΝΕΠΑΓΩΓΙΚΗ ΤΑΞΙΝΟΜΗΣΗ
- ΕΠΕΞΕΡΓΑΣΙΑ ΤΡΙΣΔΙΑΣΤΑΤΟΥ ΠΙΝΑΚΑ
- ΕΠΙΣΤΡΟΦΗ-ΕΞΟΔΟΣ

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Κατά τη διαδικασία αυτή, ξεκινώντας από ένα πίνακα δεδομένων *αντικείμενα x μεταβλητές* δημιουργείται ιεράρχηση την αντικειμένων και παριστάνεται αυτή, με ένα δενδρόγραμμα. Κατόπιν τέμνοντας το δενδρόγραμμα είναι δυνατό να επιλεγεί κάποιος διαμελισμό των δεδομένων σε κατάλληλο αριθμό κλάσεων. Η όλη διαδικασία περιγράφεται αναλυτικά στο κεφάλαιο 4 της διατριβής.

Μετά την επιλογή της διαδικασίας αυτής από την οθόνη 8, εμφανίζεται η οθόνη 3 για την ανάγνωση του ονόματος αρχείου δεδομένων, και μετά απ' αυτό, η παρακάτω οθόνη 9, για την επιλογή της κατάλληλης μετρικής απόστασης.



οθόνη 9. Επιλογή μετρικής υπολογισμού αποστάσεων

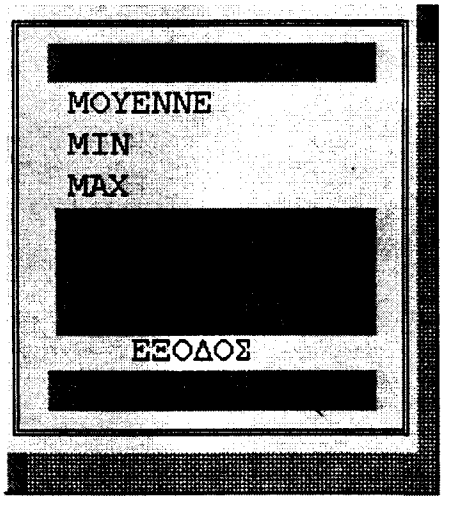
Η μετρική αυτή εξαρτάται από το είδος των δεδομένων μας, ή από την έμφαση που θέλουμε να δώσουμε σ' αυτά. (βλέπε κεφάλαιο 3 της διατριβής).

Η μετρική του X^2 χρησιμοποιείται κυρίως όταν τα δεδομένα μας αποτελούν πίνακα συχνοτήτων, ενώ η ευκλείδεια μετρική κυρίως όταν έχουμε δεδομένα ποσοτικών μεταβλητών, αλλά και σε κάθε άλλη περίπτωση.

Η μετρική του δείκτη Jaccard πρέπει να χρησιμοποιείται μόνο όταν οι τιμές των μεταβλητών μας είναι 0 ή 1. Δηλαδή μόνο όταν οι παρατηρήσεις περιγράφονται από "λογικές" μεταβλητές (παρουσία ή απουσία).

Ο συντελεστής συσχέτισης χρησιμοποιείται στην περίπτωση που έχουμε πίνακα συμπτώσεων μεταβλητών.

Μετά την επιλογή της μετρικής απόστασης, εμφανίζεται η παρακάτω οθόνη 10, για την επιλογή της μεθόδου δημιουργίας της ιεράρχησης.



οθόνη 10. Επιλογή μεθόδου ιεράρχησης

Οι μέθοδοι MOYENNE, MIN, MAX περιγράφονται στο κεφάλαιο 4 της διατριβής, και αφορούν στον τρόπο υπολογισμού αποστάσεων μεταξύ ομάδων αντικειμένων.

Κατόπιν, ξεκινά η εμφάνιση των αποτελεσμάτων της ιεραρχικής ταξινόμησης. Τα αποτελέσματα εμφανίζονται στην οθόνη, αλλά καταγράφονται επίσης και σε ένα αρχείο με μορφή ASC, ώστε να μπορούμε να τα επεξεργαστούμε ή να τα εκτυπώσουμε σε χαρτί. Το αρχείο αποτελεσμάτων έχει το ίδιο όνομα με το αρχείο δεδομένων αλλά για προέκταση έχει τους χαρακτήρες .EXO.

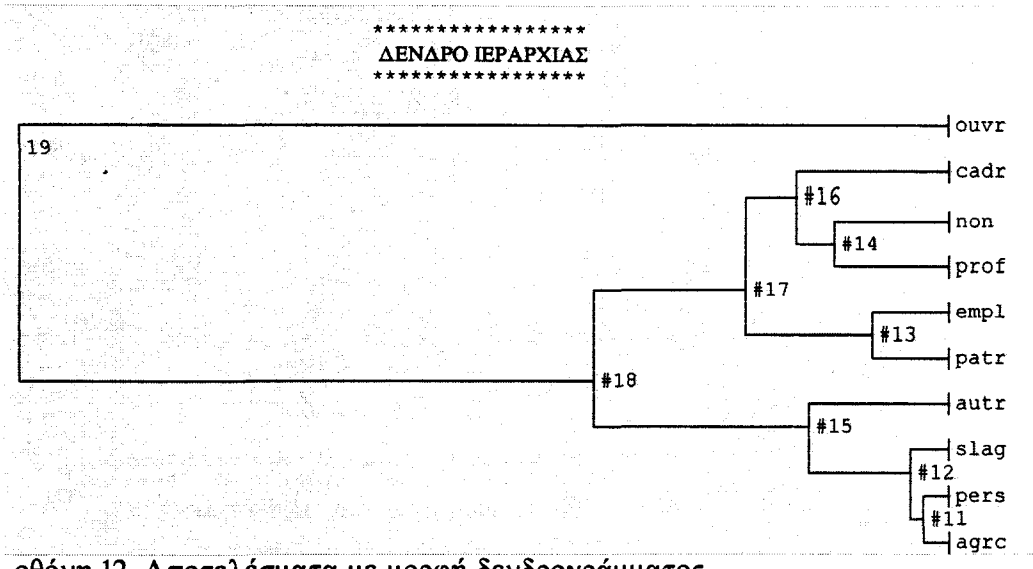
ΚΟΜΒΟΙ ΤΗΣ CLASSIFICATION

κόμβος	αριστερό	δεξιό	βάρος	απόσταση
11	pers	agrc	2	38.288379
12	slagr	11	3	50.795828
13	empl	patr	2	84.829240
14	non	prof	2	119.125984
15	autr	12	4	153.133138
16	cadr	14	3	159.029718
17	16	13	5	207.918071
18	17	15	9	345.243746
19	ouvr	18	10	879.999004

οθόνη 11. Πρώτη οθόνη αποτελεσμάτων

Η πρώτη οθόνη αποτελεσμάτων έχει μορφή παρόμοια με την οθόνη 11, και παρουσιάσει τους διαδοχικούς κόμβους που σχηματίζονται κατά την διαδικασία της ιεράρχησης, μαζί με το "βάρος" τους, δηλαδή πόσα αντικείμενα περιλαμβάνουν. Παρουσιάζεται επίσης η υπολογιζόμενη απόσταση μεταξύ αριστερού και δεξιού υποκόμβου.

Στη συνέχεια εμφανίζεται το δενδρόγραμμα της ιεράρχησης με αριθμημένους τους κόμβους του, παρόμοιο μ' αυτό που εμφανίζεται στην οθόνη 12.

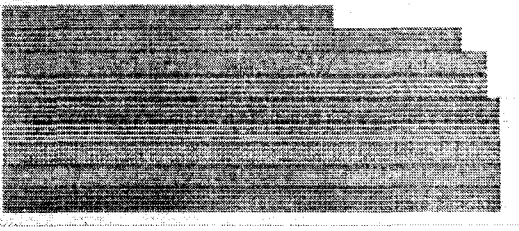


οθόνη 12. Αποτελέσματα με μορφή δενδρογράμματος

Τέλος παρουσιάζονται οι διαμελισμοί για κάθε πλήθος κλάσεων μαζί με την αδράνειά τους καθώς και το ποσοστό της αδράνειας σε σχέση με την αρχική ολική αδράνεια όλων των δεδομένων παρόμοια με την οθόνη 13. Από το ιστόγραμμα που εμφανίζεται μπορούμε να επιλέξουμε τον διαμελισμό με πλήθος κλάσεων καλύτερα προσαρμοζόμενο στα αρχικά δεδομένα όπως περιγράφεται στο κεφάλαιο 4 της διατριβής.

****
ΔΙΑΜΕΛΙΣΜΟΙ ΓΙΑ ΔΙΑΦΟΡΕΤΙΚΟ ΠΛΗΘΟΣ ΚΛΑΣΕΩΝ
****

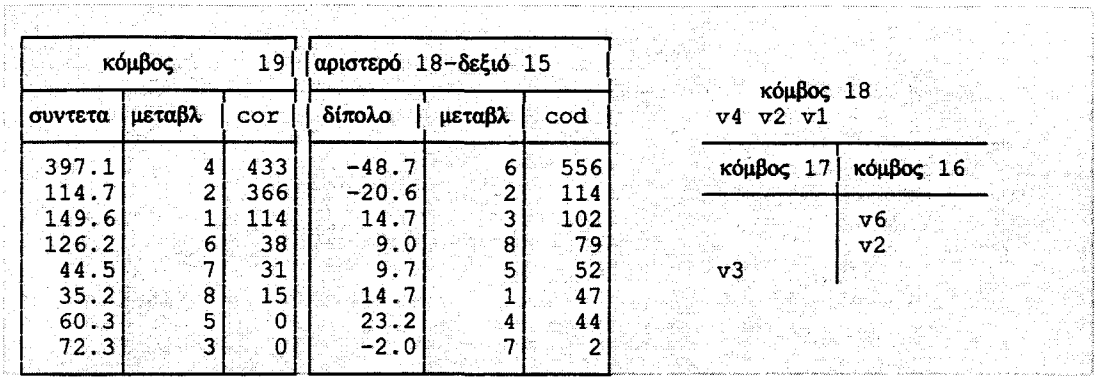
κλάσεις	αδράνεια διαμελισμού	λόγος
2	68868.3378	0.6747
3	93459.5300	0.9157
4	97589.0767	0.9561
5	99044.8600	0.9704
6	100765.3433	0.9873
7	101474.8933	0.9942
8	101834.6933	0.9977
9	101992.6600	0.9993
10	102065.9600	1.0000



οθόνη 13. Αδράνεια όλων των διαμελισμών

Η επόμενη ερώτηση Για πόσους ανώτερους κόμβους επιθυμείτε ερμηνεία; γίνεται για να δοθούν τα αποτελέσματα της μεθόδου vacor για το πλήθος των κόμβων που επιθυμεί ο χρήστης. Αν δεν μας ενδιαφέρουν οι χαρακτηριστικές μεταβλητές κάθε κόμβου, δίνουμε την τιμή 0.

Δίνοντας μια οποιαδήποτε άλλη τιμή, εμφανίζεται μια οθόνη σαν την παρακάτω οθόνη 14.



κόμβος 19			αριστερό 18-δεξιό 15		
συντετα	μεταβλ	cor	δίπολο	μεταβλ	cod
397.1	4	433	-48.7	6	556
114.7	2	366	-20.6	2	114
149.6	1	114	14.7	3	102
126.2	6	38	9.0	8	79
44.5	7	31	9.7	5	52
35.2	8	15	14.7	1	47
60.3	5	0	23.2	4	44
72.3	3	0	-2.0	7	2

οθόνη 14. Αποτελέσματα ερμηνείας κόμβου

Στην οθόνη αυτή αριστερά υπάρχει ένας πίνακας για τον μελετώμενο κάθε φορά κόμβο, όπου παρουσιάζονται οι μεταβλητές σε φθίνουσα σειρά του συνημιτόνου τους. Οι σημαντικότερες μεταβλητές για τον χαρακτηρισμό του κόμβου, προηγούνται των υπολοίπων.

Δίπλα στον πίνακα αυτό υπάρχει ο αντίστοιχος πίνακας για το δίπολο κάθε κόμβου, όπου παρουσιάζονται σε φθίνουσα σειρά οι μεταβλητές που είναι υπεύθυνες για την διάσπαση του κόμβου.

Στα δεξιά της οθόνης παρουσιάζονται με ένα γράφημα σε σχήμα Τ, τόσο οι μεταβλητές που χαρακτηρίζουν τον κόμβο, όσο και οι μεταβλητές που συντελούν στη διάσπασή του και χαρακτηρίζουν τον αριστερό ή δεξιό υποκόμβο.

Αναλυτικότερα η διαδικασία αυτή περιγράφεται στο 5ο κεφάλαιο της διατριβής.

Στο τέλος της διαδικασίας αυτής, επιστρέφουμε στο menu επεξεργασίας δεδομένων.

ΣΥΝΕΠΑΓΩΓΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Κατά τη διαδικασία αυτή, ξεκινώντας από ένα πίνακα δεδομένων (*αντικείμενα* x *μεταβλητές,)* όπου οι τιμές των μεταβλητών είναι 0 ή 1, δημιουργούμε ένα πίνακα (*μεταβλητές x μεταβλητές)* (δηλαδή ένα πίνακα συμπτώσεων Burt). Η διαδικασία και η ερμηνεία του πίνακα αυτού περιγράφεται στο κεφάλαιο 2 της διατριβής.

Κατόπιν με την βοήθεια του προγράμματος δημιουργούμε ιεράρχηση των μεταβλητών, βασιζόμενη στις σχέσεις συνεπαγωγής μεταξύ τους, και παριστάνουμε την ιεράρχηση αυτή με ένα προσανατολισμένο δενδρόγραμμα. Η όλη διαδικασία περιγράφεται αναλυτικά στο κεφάλαιο 6 της διατριβής.

Μετά την επιλογή της διαδικασίας ΣΥΝΕΠΑΓΩΓΙΚΗ ΤΑΞΙΝΟΜΗΣΗ, εμφανίζεται η οθόνη για την ανάγνωση του ονόματος αρχείου δεδομένων, και μετά

απ' αυτό, η παρακάτω οθόνη, με τα αποτελέσματα της Συνεπαγωγικής ταξινόμησης, δηλαδή τον πίνακα σχηματισμού κόμβων και το δενδρόγραμμα συνεπαγωγής μεταβλητών.

****

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΚΑΤΑ ΑΥΞΟΥΣΑ ΙΕΡΑΡΧΙΑ

(έκδοση 1.0)

****

ΕΓΙΝΕ ΕΠΕΞΕΡΓΑΣΙΑ ΤΟΥ ΑΡΧΕΙΟΥ : karb

ΚΑΝΟΝΤΑΣ ΧΡΗΣΗ ΤΗΣ ΣΥΝΕΠΑΓΩΓΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕΤΑΒΛΗΤΩΝ

αριθμός μεταβλητών : 6

ΗΜΕΡΟΜΗΝΙΑ : 13- 2-1997

ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΡΟΓΡΑΜΜΑΤΟΣ ΙΕΡΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ

****

ΚΟΜΒΟΙ ΤΗΣ CLASSIFICATION

κόμβος	αριστερό	δεξιό	βάρος	συνοχή
7	v2	v1	2	0.962484
8	v6	v3	2	0.802866
9	8	v5	3	0.724916
10	v4	7	3	0.562741
11	9	10	6	0.000000

*-

ΠΡΟΣΟΧΗ!!!

ΣΤΟ ΠΑΡΑΚΑΤΩ ΔΕΝΔΡΟΓΡΑΜΜΑ Η ΦΟΡΑ ΣΥΝΕΠΑΓΩΓΗΣ ΕΙΝΑΙ

ΑΠΟ ΚΑΤΩ ΠΡΟΣ ΤΑ ΠΑΝΩ

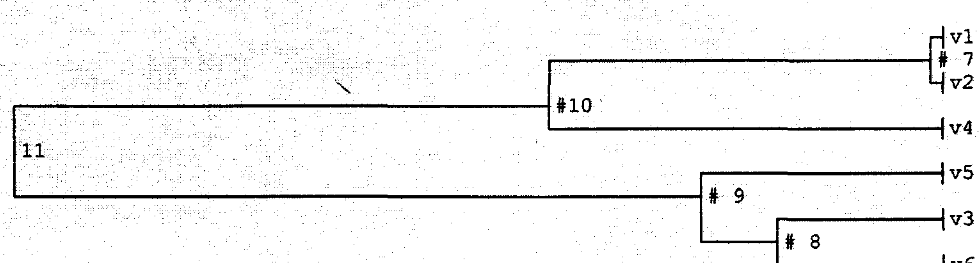
ΑΓΝΟΕΙΣΤΕ ΤΟΥΣ ΚΟΜΒΟΥΣ ΜΕ ΣΥΝΟΧΗ 0

*-

****

ΔΕΝΔΡΟ ΙΕΡΑΡΧΙΑΣ

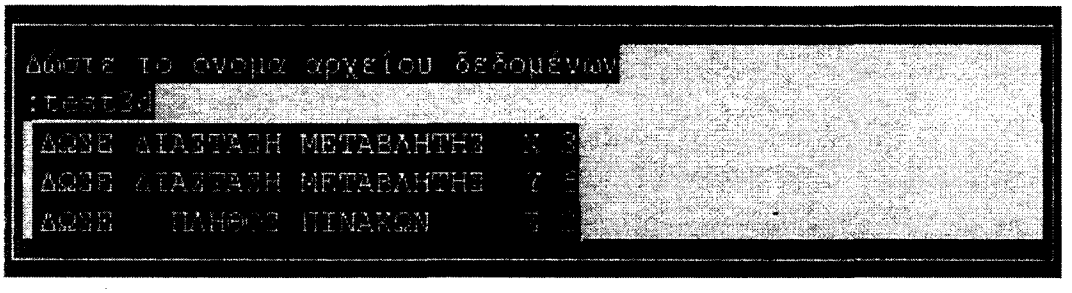
****



Στο τέλος και της διαδικασίας αυτής, επιστρέφουμε στο menu επεξεργασίας δεδομένων.

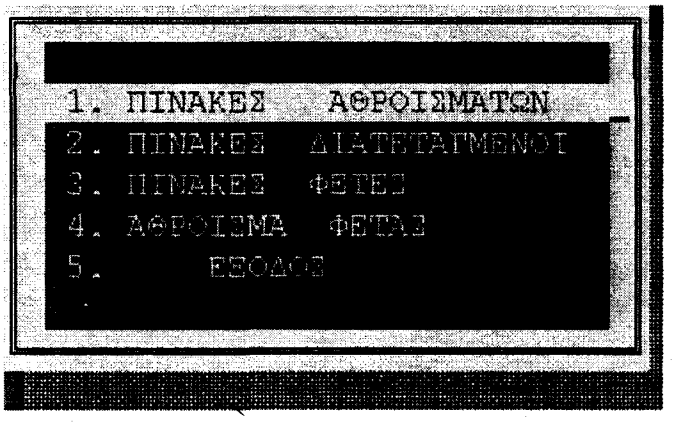
ΕΠΕΞΕΡΓΑΣΙΑ ΤΡΙΣΔΙΑΣΤΑΤΟΥ ΠΙΝΑΚΑ

Στην αρχή επιλέγουμε το όνομα του αρχείου που περιέχει τα δεδομένα, είτε δίνουμε διαδοχικά τα ονόματα αρχείων που περιέχουν τους πίνακες "φέτες", όπως στην περίπτωση που έχουμε ξεχωριστά αρχεία δεδομένων διδιάστατων πινάκων, για κάθε μία από τις τιμές μιας τρίτης διάστασης. Στην οθόνη 15, φαίνεται ο τρόπος που το πρόγραμμα διαβάζει το όνομα αρχείου, καθώς και τις 3 διαστάσεις του τρισδιάστατου πίνακα.



οθόνη 15. Εισαγωγή ονόματος αρχείου και διαστάσεων

Το πρόγραμμα διαβάζει τα δεδομένα και δημιουργεί ένα τρισδιάστατο πίνακα από τον οποίο μπορούμε να επιλέξουμε τους κατάλληλους κάθε φορά διδιάστατων πινάκων, για να προβούμε σε μια άλλη ανάλυση δεδομένων.

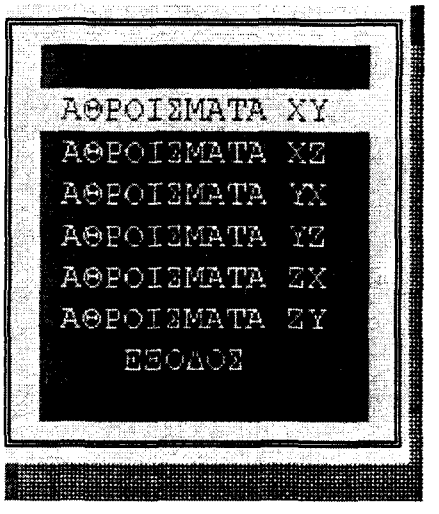


οθόνη 16. Επιλογές επεξεργασίας τρισδιάστατου πίνακα

Έτσι μπορούν να δημιουργηθούν (οθόνη 16) διδιάστατοι πίνακες αθροισμάτων, ως προς μια διάσταση, πίνακες διατεταγμένοι οριζόντια ή κάθετα,

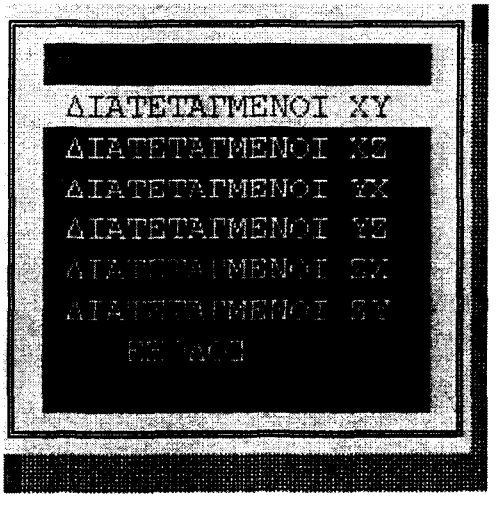
πίνακες "φέτες" ή ακόμη αθροίσματα των πινάκων "φέτες". Η χρήση τέτοιων πινάκων περιγράφηκε αναλυτικά στο 2ο κεφάλαιο της διατριβής.

Μετά την επιλογή ΠΙΝΑΚΕΣ ΑΘΡΟΙΣΜΑΤΩΝ, εμφανίζεται η οθόνη 17, στην οποία μπορούμε να επιλέξουμε τον κατάλληλο πίνακα άθροισμα.



οθόνη 17. Επιλογές για την δημιουργία πίνακα αθροισμάτων

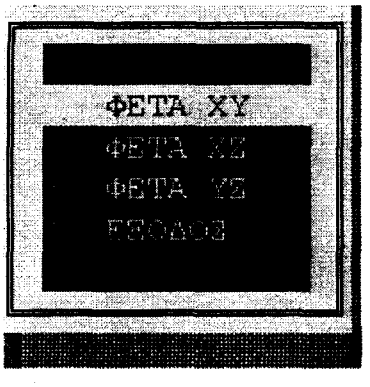
Μετά την επιλογή ΠΙΝΑΚΕΣ ΔΙΑΤΕΤΑΓΜΕΝΟΙ, εμφανίζεται η οθόνη 18 για να διαλέξουμε ως προς ποιες διαστάσεις θέλουμε διάταξη, και κατόπιν η οθόνη 19, για αν αποφασίσουμε αν θέλουμε οριζόντια ή κατακόρυφη διάταξη.



οθόνη 18. Επιλογή διαστάσεων για διατεταγμένο πίνακα

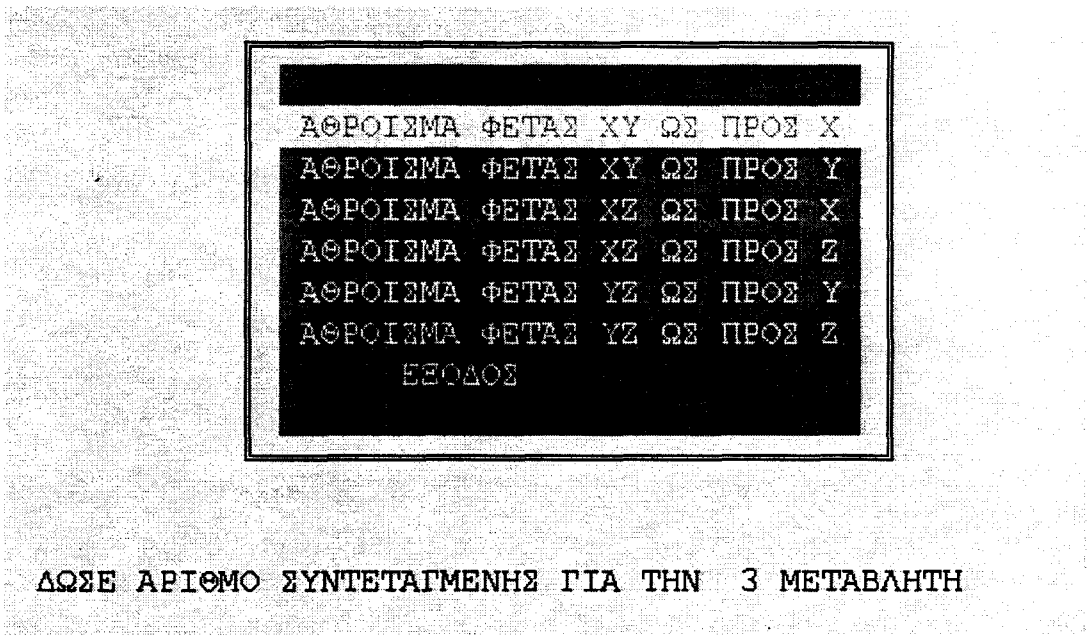
οθόνη 19. Επιλογή οριζόντιας ή κάθετης διάταξης

Μετά την επιλογή ΠΙΝΑΚΕΣ ΦΕΤΕΣ, εμφανίζεται η οθόνη 20 για την επιλογή των διαστάσεων ως προς τις οποίες θα δημιουργηθεί η "φέτα", δηλαδή για μια σταθερή τιμή της τρίτης διάστασης τυπώνονται τα περιεχόμενα κελιών για όλες τις τιμές των δύο άλλων διαστάσεων.



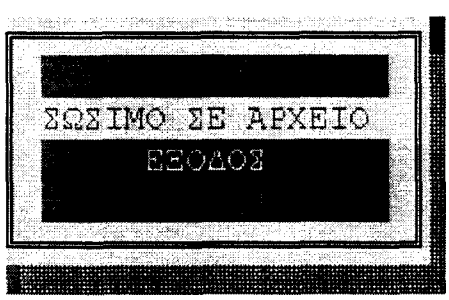
οθόνη 20. Επιλογή διαστάσεων για την "φέτα"

Τέλος, μετά την επιλογή ΑΘΡΟΙΣΜΑ ΦΕΤΑΣ, εμφανίζεται η οθόνη 21 για την επιλογή της φέτας για την οποία θα δημιουργηθεί το άθροισμα καθώς και ως προς ποιας από τις δύο διαστάσεις, θα γίνει η άθροιση.



οθόνη 21. Πίνακας επιλογών για την δημιουργία αθροισμάτων φέτας

Στο τέλος κάθε διαδικασίας της επεξεργασίας του τρισδιάστατου πίνακα εμφανίζεται η οθόνη 22, για να αποθηκεύσουμε τα δεδομένα σε αρχείο ή για να επιστρέψουμε στο προηγούμενο menu.



οθόνη 22. Τελική οθόνη για αποθήκευση ή όχι των δημιουργηθέντων πινάκων

Στο τέλος και της διαδικασίας για τρισδιάστατο πίνακα, επιστρέφουμε στο menu επεξεργασίας δεδομένων.

Επιλέγοντας "ΕΞΟΔΟΣ" ή πατώντας το πλήκτρο ESC, επιστρέφουμε στο αρχικό menu του προγράμματος. Μπορούμε να εγκαταλείψουμε το πρόγραμμα και να επιστρέψουμε στο MS-DOS ή στα Windows, ξαναεπιλέγοντας "ΕΞΟΔΟΣ" ή ξαναπατώντας το πλήκτρο ESC.

ΕΘΝΙΚΟ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ