

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ  
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ  
ΚΑΙ ΔΙΟΙΚΗΣΗΣ

---

**Περιβάλλουσα Ανάλυση Δεδομένων  
για την Ανάπτυξη Υποδειγμάτων Ταξινόμησης:  
Μια Συγκριτική Ανάλυση**

---



**Σαλάππα Αθηνά**

Μια εργασία που παρουσιάστηκε στο Πολυτεχνείο  
Κρήτης σε εκπλήρωση των απαιτήσεων για την  
απόκτηση διπλώματος στο τμήμα Μηχανικών  
Παραγωγής και Διοίκησης

**Επιβλέπων Καθηγητής:**

**Δούμπος Μιχάλης**

Ιούλιος 2003

## ΠΕΡΙΕΧΟΜΕΝΑ

<b>ΠΕΡΙΕΧΟΜΕΝΑ.....</b>	<b>i</b>
<b>ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ.....</b>	<b>iii</b>
<b>ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ.....</b>	<b>iv</b>
<b>ΕΥΧΑΡΙΣΤΙΕΣ.....</b>	<b>v</b>
<b>ΚΕΦΑΛΑΙΟ 1<sup>ο</sup>.....</b>	<b>1</b>
ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΣΥΓΚΡΙΤΙΚΗΣ ΑΝΑΛΥΣΗΣ.....	1
1.1 ΕΙΣΑΓΩΓΗ.....	1
1.2 ΑΝΑΠΤΥΞΗ ΥΠΟΔΕΙΓΜΑΤΩΝ ΤΑΞΙΝΟΜΗΣΗΣ.....	2
1.2.1 Το πρόβλημα της ταξινόμησης.....	4
1.2.2 Σπουδαιότητα της διαδικασίας ταξινόμησης.....	5
1.3 ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΚΑΙ ΣΤΟΧΟΙ ΑΥΤΗΣ.....	6
1.4 ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ.....	7
<b>ΚΕΦΑΛΑΙΟ 2<sup>ο</sup>.....</b>	<b>8</b>
ΠΕΡΙΒΑΛΛΟΥΣΑ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ.....	8
2.1 ΕΙΣΑΓΩΓΗ.....	8
2.2 ΜΕΘΟΔΟΣ DEA ΚΑΙ ΜΑΘΗΜΑΤΙΚΕΣ ΔΙΑΤΥΠΩΣΕΙΣ.....	10
2.2.1 Ιστορικό Υπόβαθρο και Φιλοσοφία της DEA.....	10
2.2.2 Βασικές μαθηματικές διατυπώσεις της DEA.....	11
2.2.2.1 Το μοντέλο CCR των Charnes, Cooper, Rhodes.....	11
2.2.2.2 Το μοντέλο BCC των Banker, Charnes and Cooper.....	15
2.2.3 Σύγκριση της DEA με Στατιστικές Μεθόδους Γραμμικής Παλινδρόμησης.....	16
2.2.4 Επεκτάσεις της μεθόδου DEA.....	19
2.2.4.1 Το προσθετικό μοντέλο.....	19
2.2.4.2 Το πολλαπλασιαστικό μοντέλο.....	20
2.2.4.3 Το μοντέλο MED.....	20
2.2.4.4 Το μοντέλο ασφαλούς πεδίου.....	20
2.2.4.5 Μοντέλα εξωγενών παραγόντων.....	21
2.2.4.6 Το κατηγορικό μοντέλο.....	21
2.2.5 Κριτικές, περιορισμοί και πλεονεκτήματα της DEA.....	22
2.3 ΠΕΡΙΛΗΨΗ.....	24
<b>ΚΕΦΑΛΑΙΟ 3<sup>ο</sup>.....</b>	<b>26</b>
ΠΡΟΣΑΡΜΟΓΗ ΤΗΣ DEA ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ.....	26
3.1 ΕΙΣΑΓΩΓΗ.....	26
3.2 ΤΕΧΝΙΚΕΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΣΤΟΧΩΝ ΓΙΑ ΤΗΝ ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΩΝ ΤΑΞΙΝΟΜΗΣΗΣ.....	26
3.3 ΜΟΝΤΕΛΑ ΤΗΣ DEA ΣΤΗ ΔΙΑΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ.....	30
3.3.1 Γενική θεώρηση του προβλήματος.....	31
3.3.2 Ομοιότητες και διαφορές μεταξύ των μοντέλων DEA και DA.....	35

---

3.3.3 Μαθηματική διατύπωση του μοντέλου DEA-DA .....	38
3.4 ΠΕΡΙΛΗΨΗ .....	43
<b>ΚΕΦΑΛΑΙΟ 4<sup>ο</sup></b> .....	<b>44</b>
ΠΕΙΡΑΜΑΤΙΚΗ ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ .....	44
4.1 ΣΚΟΠΟΣ ΤΗΣ ΕΡΕΥΝΑΣ .....	44
4.2 ΕΞΕΤΑΖΟΜΕΝΕΣ ΜΕΘΟΔΟΙ .....	44
4.2.1 Γραμμική Διακριτική Ανάλυση .....	45
4.2.2 Γραμμικός Προγραμματισμός .....	47
4.2.3 Ο Αλγόριθμος Πλησιέστερου Γείτονα .....	49
4.2.4 Δέντρα Τάξινόμησης και Παλινδρόμησης .....	50
4.2.5 Μηχανές Διανύσματος Υποστήριξης .....	53
4.3 ΣΧΕΔΙΑΣΜΟΣ ΤΟΥ ΠΕΙΡΑΜΑΤΟΣ .....	57
4.3.1 Εξεταζόμενοι παράγοντες .....	57
4.3.2 Διαδικασία παραγωγής των δεδομένων και υλοποίηση της πειραματικής ανάλυσης .....	59
4.4 ΑΝΑΛΥΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ .....	61
4.5 ΒΑΣΙΚΕΣ ΕΠΙΣΗΜΑΝΣΕΙΣ .....	70
4.6 ΠΕΡΙΛΗΨΗ .....	72
<b>ΚΕΦΑΛΑΙΟ 5<sup>ο</sup></b> .....	<b>74</b>
ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ .....	74
<b>ΒΙΒΛΙΟΓΡΑΦΙΑ</b> .....	<b>77</b>

---

## ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΣΧΗΜΑ 1.1: ΓΕΝΙΚΟ ΠΕΡΙΓΡΑΦΜΑ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ ΑΝΑΠΤΥΞΗΣ ΥΠΟΔΕΙΓΜΑΤΩΝ ΤΑΞΙΝΟΜΗΣΗΣ .....	3
ΣΧΗΜΑ 3.1: ΜΙΑ ΟΠΤΙΚΗ ΔΟΜΗ ΤΗΣ ΔΙΑΚΡΙΤΙΚΗΣ ΑΝΑΛΥΣΗΣ .....	32
ΣΧΗΜΑ 3.2: ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΗΝ ΠΕΡΙΠΤΩΣΗ ΕΠΙΚΑΛΥΨΗΣ ΤΩΝ ΚΑΤΗΓΟΡΙΩΝ .....	33
ΣΧΗΜΑ 3.3: Η ΔΙΑΔΙΚΑΣΙΑ ΤΑΞΙΝΟΜΗΣΗΣ ΤΗΣ ΔΙΑΚΡΙΤΙΚΗΣ ΑΝΑΛΥΣΗΣ .....	34
ΣΧΗΜΑ 3.4: ΔΙΑΔΙΚΑΣΙΑ ΤΑΞΙΝΟΜΗΣΗΣ ΕΝΟΣ ΝΕΟΥ ΔΕΙΓΜΑΤΟΣ ΔΕΔΟΜΕΝΩΝ .....	34
ΣΧΗΜΑ 3.5: ΠΕΡΙΒΑΛΛΟΥΣΑ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (DEA) .....	36
ΣΧΗΜΑ 3.6: ΔΙΑΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ (DA) .....	37
ΣΧΗΜΑ 4.1: ΣΧΗΜΑΤΙΚΗ ΑΠΕΙΚΟΝΙΣΗ ΤΟΥ ΚΑΝΟΝΑ ΤΑΞΙΝΟΜΗΣΗΣ ΤΗΣ ΓΡΑΜΜΙΚΗΣ ΔΙΑΚΡΙΤΙΚΗΣ ΑΝΑΛΥΣΗΣ .....	46
ΣΧΗΜΑ 4.2: ΓΡΑΦΙΚΗ ΑΠΕΙΚΟΝΗΣΗ ΤΩΝ SVM .....	53
ΣΧΗΜΑ 4.3: ΜΕΣΟ ΠΟΣΟΣΤΟ ΕΣΦΑΛΜΕΝΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ ΣΥΝΑΡΤΗΣΕΙ ΤΗΣ ΟΜΟΙΟΓΕΝΕΙΑΣ ΣΤΙΣ ΔΙΑΚΥΜΑΝΣΕΙΣ ΤΩΝ ΚΑΤΗΓΟΡΙΩΝ .....	62
ΣΧΗΜΑ 4.4: ΜΕΣΟ ΠΟΣΟΣΤΟ ΕΣΦΑΛΜΕΝΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ ΣΥΝΑΡΤΗΣΕΙ ΤΟΥ ΣΥΝΤΕΛΕΣΤΗ ΣΥΣΧΕΤΙΣΗΣ .....	63
ΣΧΗΜΑ 4.5: ΜΕΣΟ ΠΟΣΟΣΤΟ ΕΣΦΑΛΜΕΝΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ ΤΩΝ ΕΞΕΤΑΖΟΜΕΝΩΝ ΤΑΧΝΙΚΩΝ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΟ ΔΕΙΓΜΑ ΕΛΕΓΧΟΥ .....	64

---

## ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΙΝΑΚΑΣ 4.1: ΕΞΕΤΑΖΟΜΕΝΟΙ ΠΑΡΑΓΟΝΤΕΣ ΣΤΟΝ ΠΕΙΡΑΜΑΤΙΚΟ ΣΧΕΔΙΑΣΜΟ ΓΙΑ .....	57
ΠΙΝΑΚΑΣ 4.2: ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΤΟΥ ΠΕΙΡΑΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ ΤΟΥ ΔΕΙΓΜΑΤΟΣ ΕΛΕΓΧΟΥ .....	61
ΠΙΝΑΚΑΣ 4.3: ΜΕΣΟ ΠΟΣΟΣΤΟ ΕΣΦΑΛΜΕΝΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ ΤΩΝ ΕΞΕΤΑΖΟΜΕΝΩΝ ΠΡΟΣΕΓΓΙΣΕΩΝ ΤΑΞΙΝΟΜΗΣΗΣ ΣΥΝΑΡΤΗΣΕΙ ΤΗΣ ΟΜΟΙΟΓΕΝΕΙΑΣ ΤΩΝ ΔΙΑΚΥΜΑΝΣΕΩΝ .....	65
ΠΙΝΑΚΑΣ 4.4: ΜΕΣΟ ΠΟΣΟΣΤΟ ΕΣΦΑΛΜΕΝΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ ΤΩΝ ΕΞΕΤΑΖΟΜΕΝΩΝ ΠΡΟΣΕΓΓΙΣΕΩΝ ΤΑΞΙΝΟΜΗΣΗΣ ΣΥΝΑΡΤΗΣΕΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΤΑΝΟΜΗΣ .....	66
ΠΙΝΑΚΑΣ 4.5: ΜΕΣΟ ΠΟΣΟΣΤΟ ΕΣΦΑΛΜΕΝΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ ΣΥΝΑΡΤΗΣΕΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΤΑΝΟΜΗΣ ΚΑΙ ΤΗΣ ΟΜΟΙΟΓΕΝΕΙΑΣ ΤΩΝ ΔΙΑΚΥΜΑΝΣΕΩΝ .....	67
ΠΙΝΑΚΑΣ 4.6: ΜΕΣΟ ΠΟΣΟΣΤΟ ΕΣΦΑΛΜΕΝΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ ΤΩΝ ΕΞΕΤΑΖΟΜΕΝΩΝ ΠΡΟΣΕΓΓΙΣΕΩΝ ΤΑΞΙΝΟΜΗΣΗΣ ΣΥΝΑΡΤΗΣΕΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΤΑΝΟΜΗΣ ΚΑΙ ΤΗΣ ΟΜΟΙΟΓΕΝΕΙΑΣ ΤΩΝ ΔΙΑΚΥΜΑΝΣΕΩΝ .....	68
ΠΙΝΑΚΑΣ 4.7: ΔΙΜΕΡΗΣ ΣΥΓΚΡΙΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΤΩΝ ΠΡΟΣΕΓΓΙΣΕΩΝ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΟ ΔΕΙΓΜΑ ΕΛΕΓΧΟΥ .....	71
ΠΙΝΑΚΑΣ 4.8: ΣΥΝΟΛΙΚΑ ΠΟΣΟΣΤΑ ΤΗΣ ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΤΗΤΑΣ ΤΗΣ ΚΑΘΕ ΠΡΟΣΕΓΓΙΣΗΣ ΤΑΞΙΝΟΜΗΣΗΣ .....	72

---

## ΕΥΧΑΡΙΣΤΙΕΣ

Με το πέρας της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω τον καθηγητή Δούμπο Μιχάλη τόσο για την επίβλεψη του στη διεκπεραίωση της εργασίας, όσο και για τις γνώσεις που μου προσέφερε κατά τη διάρκεια της φοιτητικής μου πορείας.

Επίσης, ευχαριστώ την οικογένειά μου για την υλική και κυρίως ηθική υποστήριξή της όλο αυτά τα χρόνια που λείπω από κοντά της και στην οποία αφιερώνεται η παρούσα διπλωματική εργασία.

Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου, τωρινούς και παλαιούς, οι οποίοι δημιούργησαν ένα ευχάριστο κλίμα κατά τη διάρκεια των φοιτητικών μου χρόνων.

Σαλάππα Αθηνά  
Χανιά  
Ιούλιος 2003

## ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΣΥΓΚΡΙΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

### 1.1 Εισαγωγή

Η ταξινόμηση ενός συνόλου εναλλακτικών παρατηρήσεων ή αντικειμένων σε προκαθορισμένες ομοιογενείς κατηγορίες είναι ένα πρόβλημα ιδιαίτερου πρακτικού και ερευνητικού ενδιαφέροντος. Αυτού του είδους τα προβλήματα αναφέρονται ως προβλήματα απλής ταξινόμησης (classification) ή διατεταγμένης ταξινόμησης (sorting), ανάλογα με το εάν οι κατηγορίες ταξινόμησης ορίζονται ονομαστικά (nominal) ή είναι διατεταγμένες (ordinal). Η προβληματική της ταξινόμησης παρέχει το πλαίσιο μιας εναλλακτικής θεώρησης των προβλημάτων σε σύγκριση με άλλες προβληματικές, όπως η ομαδοποίηση (clustering), η επιλογή της καλύτερης εναλλακτικής (choice), η κατάταξη των εξεταζόμενων εναλλακτικών από τις καλύτερες στις χειρότερες βάσει των χαρακτηριστικών τους (ranking) και η περιγραφή των εναλλακτικών δραστηριοτήτων για τον εντοπισμό των βασικών τους ιδιοτήτων (description). Για τη μελέτη των προβλημάτων ταξινόμησης έχουν αναπτυχθεί κατά καιρούς διάφορες μεθοδολογίες από τους χώρους της στατιστικής και της οικονομετρίας, της τεχνικής νοημοσύνης και της επιχειρησιακής έρευνας. Η ανάπτυξη και χρήση ποσοτικών τεχνικών ταξινόμησης κρίνεται απαραίτητη τόσο για την καλύτερη αντιμετώπιση του εξεταζόμενου προβλήματος, όσο και για την σημαντική μείωση του χρόνου και του κόστους που απαιτούνται για την αντιμετώπισή του. Γενικά, όπως αναφέρεται και παρακάτω, η αντιμετώπιση του προβλήματος της ταξινόμησης βάσει των διαθέσιμων μεθοδολογικών προσεγγίσεων συνίσταται στην ανάπτυξη ποσοτικών υποδειγμάτων, τα οποία

υποστηρίζουν την διαδικασία επίλυσης προβλημάτων στη βάση της προβληματικής της ταξινόμησης.

## 1.2 Ανάπτυξη υποδειγμάτων ταξινόμησης

Οι μεθοδολογικές προσεγγίσεις για την ανάπτυξη υποδειγμάτων ταξινόμησης ακολουθούν την γενική φιλοσοφία της παλινδρόμησης, προσπαθώντας να αξιοποιήσουν τη διαθέσιμη γνώση και πληροφορία που απορρέει από το γεγονός ότι οι κατηγορίες είναι προκαθορισμένες.

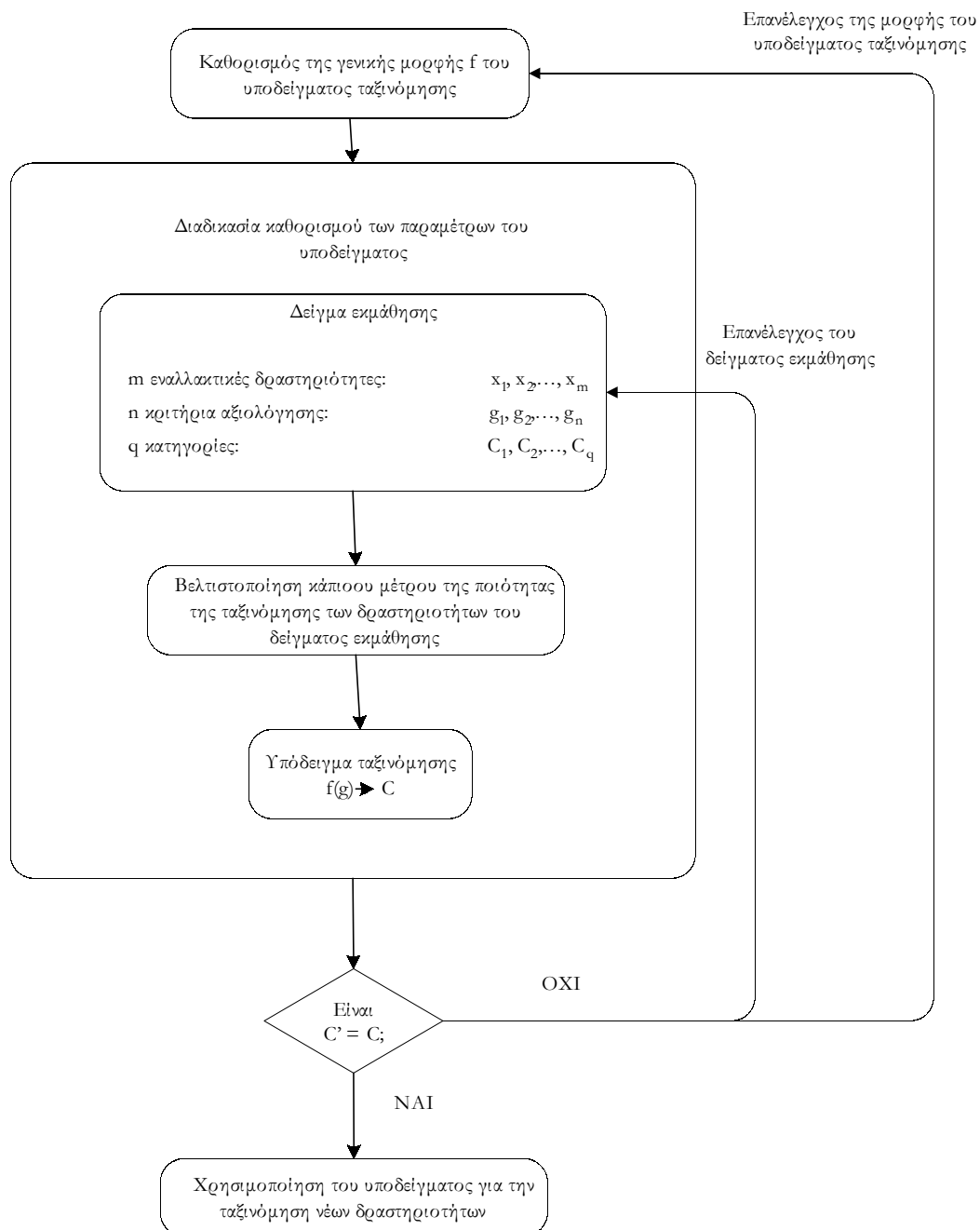
Στην ανάλυση παλινδρόμησης στόχος είναι ο εντοπισμός της συναρτησιακής σχέσης που συνδέει μια εξαρτημένη μεταβλητή  $Y$  με ένα διάνυσμα μεταβλητών  $X$  βάσει της ανάλυσης ενός συνόλου δεδομένων παρατηρήσεων  $(X, Y)$ . Κατά ανάλογο τρόπο αντιμετωπίζεται και το πρόβλημα της ταξινόμησης με την μόνη διαφορά ότι η εξαρτημένη μεταβλητή δεν είναι συνεχής, αλλά αφορά ένα περιορισμένο σύνολο διακριτών επιπέδων καθένα από τα οποία αντιστοιχεί σε μια κατηγορία. Το δείγμα των παρατηρήσεων που χρησιμοποιείται για την ανάπτυξη των υποδειγμάτων ταξινόμησης ονομάζεται δείγμα εκμάθησης και περιλαμβάνει ζεύγη της μορφής  $(X, C)$ , όπου ως  $C$  συμβολίζεται η εξαρτημένη μεταβλητή που υποδηλώνει την ταξινόμηση των εναλλακτικών δραστηριοτήτων σε ένα σύνολο κατηγοριών  $q$ . Η επίλυση του προβλήματος της ταξινόμησης συνίσταται στην ανάπτυξη ενός υποδείγματος της μορφής  $f(X) \rightarrow C$  το οποίο ελαχιστοποιεί ένα μέτρο των διαφορών που εντοπίζονται μεταξύ της εκτιμώμενης ταξινόμησης  $\hat{C}$  και της δεδομένης ταξινόμησης  $C$ . Εφόσον, ολοκληρωθεί η ανάπτυξη του υποδείγματος ταξινόμησης μπορεί πλέον αυτό να χρησιμοποιηθεί για την ταξινόμηση οποιονδήποτε άλλων εναλλακτικών, οι οποίες δεν συμπεριλαμβάνονται στο δείγμα εκμάθησης. Η χρησιμότητα της παραπάνω διαδικασίας βασίζεται στην εκμετάλλευση της υπάρχουσας γνώσης από το δείγμα εκμάθησης, με σκοπό την μοντελοποίηση και αναπαράστασή της σε ένα υπόδειγμα ταξινόμησης, το οποίο θα διαθέτει την απαραίτητη ικανότητα γενίκευσης.

Γενικά, τα αναπτυσσόμενα υποδείγματα ταξινόμησης είναι μια συνάρτηση, η οποία συνδυάζει όλα τα επιμέρους χαρακτηριστικά των εναλλακτικών δραστηριοτήτων σε έναν ολικό ποσοτικό δείκτη βάσει του οποίου λαμβάνονται οι αποφάσεις για την ταξινόμηση των εναλλακτικών δραστηριοτήτων. Ο δείκτης αυτός μπορεί να αναπαριστά την πιθανότητα να



ανήκει μια εναλλακτική σε μια κατηγορία ή κάποια αυθαίρετη «βαθμολογία» η οποία σε συνδυασμό με κατάλληλους κανόνες οδηγεί στην ταξινόμηση των εναλλακτικών.

Ένα γενικό περίγραμμα της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης φαίνεται παρακάτω:



Σχήμα 1.1: Γενικό περίγραμμα της διαδικασίας ανάπτυξης υποδειγμάτων ταξινόμησης

### 1.2.1 Το πρόβλημα της ταξινόμησης

Το πρόβλημα της ταξινόμησης βασίζεται στην πραγματοποίηση απόλυτων συγκρίσεων (absolute comparisons). Κάθε εναλλακτική δραστηριότητα εντάσσεται σε μια εκ των προκαθορισμένων κατηγοριών βάσει ενός συγκεκριμένου κανόνα, ο οποίος συνήθως αναφέρεται στη σύγκριση με συγκεκριμένα πρότυπα τα οποία διαχωρίζουν τις κατηγορίες. Η εφαρμογή του κανόνα ταξινόμησης, καθώς επίσης και το αποτέλεσμα της αξιολόγησής της, δεν επηρεάζεται από το σύνολο των εξεταζόμενων δραστηριοτήτων. Για παράδειγμα, μια εκτίμηση της μορφής «το προϊόν X είναι ακατάλληλο» είναι απόλυτη, καθώς δεν επηρεάζεται με κανέναν τρόπο από το σύνολο των ομοειδών προϊόντων του X. Βέβαια μερικές φορές, τέτοιου είδους εκτιμήσεις δεν είναι πάντα απόλυτες, καθώς καθορίζονται μέσα στο γενικό πλαίσιο που χαρακτηρίζει το περιβάλλον του προβλήματος ή της απόφασης. Για παράδειγμα, μια επιχείρηση μπορεί να πληροί τις προϋποθέσεις χρηματοδότησης από ένα χρηματοπιστωτικό ίδρυμα υπό συγκεκριμένες συνθήκες του γενικότερου οικονομικού και επιχειρηματικού περιβάλλοντος. Καθώς όμως οι συνθήκες αυτές εξελίσσονται δυναμικά, μεταβάλλονται ανάλογα και οι προϋποθέσεις χρηματοδότησης, με αποτέλεσμα η ίδια επιχείρηση να κριθεί ακατάλληλη προς χρηματοδότηση. Γενικά, παρά τις όποιες μεταβολές που είναι δυνατόν να πραγματοποιηθούν στο χρησιμοποιούμενο κανόνα ταξινόμησης, ο καθορισμός του πραγματοποιείται πάντα ανεξάρτητα από το σύνολο των εξεταζόμενων εναλλακτικών δραστηριοτήτων. Αυτή είναι και η βασικότερη διαφορά του προβλήματος της ταξινόμησης από άλλες προβληματικές (επιλογή, κατάταξη).

Ένας γενικός ορισμός της ταξινόμησης όπως προτάθηκε από τον Mirkin (1996) είναι ο εξής:

*«Ταξινόμηση είναι η ρεαλιστική ή ιδεατή τοποθέτηση μαζί παρόμοιων αντικειμένων και ο διαχωρισμός των αντικειμένων τα οποία διαφέρουν, με απώτερο σκοπό: (α) τη διαμόρφωση, οργάνωση και διατήρηση της γνώσης, (β) την ανάλυση της δομής του φαινομένου που εξετάζεται, (γ) τη συσχέτιση των διαφόρων πλευρών του υπό εξέταση φαινομένου.»*

Οι συνηθέστεροι όροι που χρησιμοποιούνται για την αναφορά στο πρόβλημα της ταξινόμησης είναι οι εξής:

- Ταξινόμηση (Classification)
- Διατεταγμένη ταξινόμηση (Sorting)

Ο πρώτος όρος αναφέρεται σε προβλήματα όπου οι κατηγορίες στις οποίες θα ενταχθούν οι εξεταζόμενες εναλλακτικές δραστηριότητες, ορίζονται κατά ονομαστικό τρόπο. Χαρακτηριστικά παραδείγματα αυτής της κατηγορίας είναι τα προβλήματα αναγνώρισης προτύπων (pattern recognition) τα οποία παρουσιάζουν σημαντικές εφαρμογές σε τεχνολογίες αναγνώρισης φωνής, γραφής, ακόμα και στην ιατρική. Αντίθετα, ο δεύτερος όρος αναφέρεται σε προβλήματα όπου οι οριζόμενες κατηγορίες, στις οποίες θα πραγματοποιηθεί η ταξινόμηση των εναλλακτικών δραστηριοτήτων, είναι διατεταγμένες από τις καλύτερες στις χειρότερες. Τέτοια προβλήματα απαντώνται συχνά στην πράξη σε χώρους, όπως στη χρηματοοικονομική διοίκηση, το μάρκετινγκ, τη διοίκηση ανθρωπίνου δυναμικού, στη περιβαλλοντολογική διαχείριση, κά.

### 1.2.2 Σπουδαιότητα της διαδικασίας ταξινόμησης

Η σημασία του προβλήματος ταξινόμησης δεν περιορίζεται μόνο στην πολυπλοκότητα που παρουσιάζει ως ένα επιστημονικό πεδίο έρευνας, αλλά επεκτείνεται και σε πρακτικό επίπεδο. Χαρακτηριστικές είναι οι παρακάτω πρακτικές εφαρμογές:

- *Ιατρική*: Πραγματοποίηση ιατρικών διαγνώσεων ταξινομώντας τους ασθενείς σε κατηγορίες με βάση τα συμπτώματα που παρουσιάζουν (Tsumoto (1998), Belacel(2000)).
- *Αναγνώριση προτύπων*: Διερεύνηση των χαρακτηριστικών φυσικών προσώπων ή αντικειμένων και ταξινόμησή τους σε ανάλογες κατηγορίες. Χαρακτηριστικά παραδείγματα της αναγνώρισης βασικών ανθρώπινων χαρακτηριστικών είναι η αναγνώριση φωνής, δακτυλικών αποτυπωμάτων και οι εφαρμογές τους στην ασφάλεια καίριων συστημάτων (Ripley (1996), Young και Fu (1997), Nieddu και Patrizi (2000)).
- *Διαχείριση ανθρωπίνου δυναμικού*: Αξιολόγηση του ανθρώπινου δυναμικού βάσει των προσόντων του, με απώτερο σκοπό τον προσδιορισμό της κατάλληλης θέσης (Rulon et al. (1967) και Gochet et al. (1997)).
- *Διαχείριση Τεχνικών Συστημάτων και Τεχνική Διάγνωση*: Παρακολούθηση της λειτουργίας πολύπλοκων συστημάτων παραγωγής για την έγκαιρη διάγνωση πιθανών βλαβών (Catelani και Ford (2000), Shen et al. (2000)).

- *Μάρκετινγκ*: Μέτρηση της ικανοποίησης πελατών, μελέτη των επιμέρους χαρακτηριστικών διαφορετικών κατηγοριών καταναλωτών, ανάπτυξη κατάλληλων πολιτικών για την διείσδυση προϊόντων στην αγορά, κ.ά. (Dutka (1995), Siskos et al. (1998)).
- *Περιβαλλοντική και ενεργειακή διαχείριση, οικολογία*: Ανάλυση και έγκαιρη διάγνωση των περιβαλλοντικών επιπτώσεων διαφόρων ενεργειακών πολιτικών, διερεύνηση της αποτελεσματικότητας ενεργειακών πολιτικών σε κρατικό επίπεδο (Diakoulaki et al. (1999)).
- *Χρηματοοικονομική Διοίκηση και Οικονομική Θεωρία*: Πρόβλεψη της πτώχευσης επιχειρήσεων, εκτίμηση του πιστωτικού κινδύνου επιχειρήσεων και καταναλωτών, επιλογή και διαχείριση χαρτοφυλακίων επενδύσεων, αξιολόγηση δανειοληπτικής ικανότητας χωρών (Zorounidis (1998), Doumpos και Zorounidis (1998)).

Τα παραπάνω προβλήματα πιστοποιούν την σπουδαιότητα του προβλήματος ταξινόμησης και της ανάπτυξης των αντίστοιχων αποτελεσματικών υποδειγμάτων.

### 1.3 Προτεινόμενη μεθοδολογική προσέγγιση και στόχοι αυτής

Τα τελευταία χρόνια η περιβάλλουσα ανάλυση δεδομένων (data envelopment analysis-DEA) αποτελεί μια σημαντική μεθοδολογική προσέγγιση της επιχειρησιακής έρευνας για την αξιολόγηση της αποτελεσματικότητας επιχειρήσεων και οργανισμών χρησιμοποιώντας τεχνικές γραμμικού προγραμματισμού. Πρόσφατα, το μεθοδολογικό πλαίσιο της DEA έχει χρησιμοποιηθεί και για την ανάπτυξη υποδειγμάτων ταξινόμησης (μοντέλων για την τοποθέτηση ενός συνόλου αντικειμένων σε προκαθορισμένες κατηγορίες). Σκοπός της εργασίας αυτής είναι να διερευνήσει την αποτελεσματικότητα της DEA στην αντιμετώπιση τέτοιων προβλημάτων, σε σύγκριση με άλλες διαδεδομένες τεχνικές από τον χώρο της στατιστικής και της επιχειρησιακής έρευνας, όπως τη γραμμική διακριτική ανάλυση, το γραμμικό προγραμματισμό, τον αλγόριθμο του πλησιέστερου γείτονα, τα δέντρα ταξινόμησης, και τις μηχανές διανύσματος υποστήριξης.. Η αξιολόγηση της αποτελεσματικότητας πραγματοποιείται με βάση το ποσοστό εσφαλμένων ταξινομήσεων κατά τη διαδικασία της ταξινόμησης, καθώς επίσης και του τρόπου με τον οποίο επιδρά κάθε παράγοντας στην αποτελεσματικότητα της κάθε προσέγγισης. Για τον σκοπό αυτό απαιτείται η χρήση πειραματικών δεδομένων.

## 1.4 Δομή της εργασίας

Η υπόλοιπη ανάλυση που πραγματοποιείται στην παρούσα εργασία οργανώνεται σε πέντε κεφάλαια ως εξής:

Στο δεύτερο κεφάλαιο που ακολουθεί περιγράφεται αναλυτικά η Περιβάλλουσα Ανάλυση Δεδομένων (DEA) και το μαθηματικό υπόβαθρο αυτής, όπως επίσης και οι σημαντικότερες παραλλαγές της που σημειώθηκαν κατά καιρούς. Επίσης, πραγματοποιείται μία εκτενή αναφορά στα προβλήματα που εντοπίζονται κατά την υλοποίηση της μεθόδου, καθώς επίσης και στους περιορισμούς που θα πρέπει ο κάθε αναλυτής να έχει υπόψη κατά την εφαρμογή της DEA. Παράλληλα, αναφέρονται όλα εκείνα τα θετικά στοιχεία που την καθιστούν χρήσιμη στον χώρο της διοίκησης.

Στο τρίτο κεφάλαιο παρουσιάζεται η προτεινόμενη μεθοδολογική προσέγγιση ταξινόμησης “Extended DEA-DA” και δίνεται η μαθηματική της διατύπωση. Επίσης, πραγματοποιείται μια γενική σύγκριση της εξεταζόμενης μεθοδολογίας και παρεμφερών τεχνικών προγραμματισμού στόχων, τονίζοντας τις ομοιότητες και διαφορές που εντοπίζονται. Επίσης, δίνεται μια οπτική δομή της διακριτικής ανάλυσης, της διαδικασίας ταξινόμησης που ακολουθεί και του τρόπου αντιμετώπισης της επικάλυψης των κατηγοριών κατά τη ταξινόμηση.

Στο τέταρτο κεφάλαιο πραγματοποιείται ένας εκτενής πειραματικός σχεδιασμός με σκοπό τη διερεύνηση της αποτελεσματικότητας της DEA έναντι εναλλακτικών μεθοδολογικών προσεγγίσεων ταξινόμησης (γραμμική διακριτική ανάλυση, γραμμικός προγραμματισμός, αλγόριθμος πλησιέστερου γείτονα, δέντρα ταξινόμησης και μηχανές διανύσματος υποστήριξης). Επίσης, πραγματοποιείται ανάλυση των αποτελεσμάτων και τονίζονται τα βασικά συμπεράσματα του πειραματικού σχεδιασμού που πραγματοποιήθηκε.

Τέλος, στο πέμπτο κεφάλαιο παρουσιάζονται τα βασικά αποτελέσματα που επιτεύχθηκαν από την έρευνα που πραγματοποιήθηκε και προτείνονται μελλοντικές ερευνητικές κατευθύνσεις, οι οποίες θα συμβάλλουν στην καλύτερη αντιμετώπιση του προβλήματος της ταξινόμησης, αλλά και στην περαιτέρω διερεύνηση των ιδιοτήτων, ομοιοτήτων και διαφορών που χαρακτηρίζουν τις εξεταζόμενες μεθόδους ταξινόμησης.

## ΠΕΡΙΒΑΛΛΟΥΣΑ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

### 2.1 Εισαγωγή

Η Περιβάλλουσα Ανάλυση Δεδομένων (Data Envelopment Analysis, Cooper et al.(2000)) είναι μία μεθοδολογία η οποία τις τελευταίες δύο δεκαετίες έχει γνωρίσει σημαντική διάδοση για την εκτίμηση της αποτελεσματικότητας (efficiency) διαφόρων λειτουργικών μονάδων (Decision Making Units, DMUs). Κάθε μονάδα θεωρείται ότι χρησιμοποιεί κάποιες εισόδους για την επίτευξη ορισμένων αποτελεσμάτων (έξοδοι). Η έννοια της μονάδας είναι γενική. Για παράδειγμα ως μονάδα μπορεί να θεωρηθεί μια επιχείρηση, ένας οργανισμός, ή κάποιο τμήμα αυτών. Τα τελευταία χρόνια έχει παρουσιαστεί ένας σημαντικός αριθμός εφαρμογών της DEA στην εκτίμηση της αποτελεσματικότητας διαφόρων μονάδων. Χαρακτηριστικά αναφέρονται εφαρμογές της μεθόδου DEA στη μέτρηση της αποτελεσματικότητας ασφαλιστικών και χρηματοπιστωτικών οργανισμών, μονάδων τοπικής αυτοδιοίκησης όπως, σχολεία, νοσοκομεία, εμπορικά καταστήματα, κ.ά (Arnold et al. (1996), Banker et al. (1986), Chaffai (1997)).

Ερευνητές σε μία πλειάδα επιστημονικών πεδίων αναγνώρισαν στην DEA μία εξαιρετική και συνάμα εύκολη στην χρήση μέθοδο για την μοντελοποίηση επιχειρησιακών διαδικασιών σχετικών με την εκτίμηση της αποτελεσματικότητας. Βασικό πλεονέκτημα της DEA και κύριο στοιχείο για την ευρεία διάδοσή της αποτελεί η ευελιξία που παρέχει και η απουσία οποιοδήποτε υποθέσεων σχετικών με την ανάλυση σε αντίθεση με άλλες μεθοδολογίες, όπως η στατιστική παλινδρόμηση..

Στην αρχική τους μελέτη, οι Charnes et al. (1978) περιέγραψαν την DEA ως ένα μαθηματικό μοντέλο γραμμικού προγραμματισμού το οποίο εφαρμόζόμενο σε εμπειρικά δεδομένα παρέχει ένα νέο τρόπο για την αξιολόγηση των σχέσεων της παραγωγικής λειτουργίας και/ή δυνατότητας αποτελεσματικής παραγωγής ως θεμελιώδη λίθο της μοντέρνας οικονομικής θεωρίας.

Γενικότερα, η DEA είναι μία μεθοδολογία που τείνει περισσότερο προς την ερμηνεία των αποτελεσματικών ορίων και όχι προς τις κεντρικές κλίσεις. Δηλαδή δεν επιχειρείται η προσαρμογή ενός μοντέλου παλινδρόμησης στα διαθέσιμα δεδομένα, αλλά αντίθετα αναπτύσσεται μια κατά τμήματα γραμμική επιφάνεια στην κορυφή όλων των παρατηρήσεων. Από αυτή την άποψη η DEA παρέχει τη δυνατότητα εντοπισμού σχέσεων στα δεδομένα οι οποίες δεν είναι δυνατόν να βρεθούν μέσω άλλων μεθοδολογιών. Για την επεξήγηση και κατανόηση της παραπάνω θέσης αρκεί να σημειωθεί ότι σε κάποιες εφαρμογές επιθυμείται η εύρεση του αποτελεσματικού ορίου, ενώ σε κάποιες άλλες η εκτίμηση του πιο αποτελεσματικού DMU. Αυτό επιτυγχάνεται με απλό και άμεσο τρόπο μέσω της DEA χωρίς να απαιτούνται προβλέψεις και αποκλίσεις με διαφόρου τύπου μοντέλα, όπως γραμμικής και μη γραμμικής παλινδρόμησης.

Βασικό στοιχείο στο μεθοδολογικό πλαίσιο της DEA είναι η εισαγωγή της έννοιας της σχετικής αποτελεσματικότητας (relative efficiency). Ένας πρώτος ορισμός που μπορεί να χρησιμοποιηθεί για την περιγραφή της σχετικής αποτελεσματικότητας είναι ο ακόλουθος:

*1<sup>ος</sup> ορισμός: (Efficiency – Extended Pareto Koopmans Definition)*

Ένα DMU θεωρείται αποτελεσματικό εάν και μόνο εάν μία τυχόν βελτίωση των εισόδων ή εξόδων του δεν χειροτερεύει κάποιες από τις υπόλοιπες εισόδους ή εξόδους του.

Στις περισσότερες εφαρμογές της διοικητικής επιστήμης τα θεωρητικά πιθανά επίπεδα αποτελεσματικότητας δεν είναι γνωστά. Επομένως, ο προηγούμενος ορισμός μπορεί να αντικατασταθεί, δίνοντας έμφαση στα εμπειρικά δεδομένα τα οποία είναι διαθέσιμα στην κάθε περίπτωση.

*2<sup>ος</sup> ορισμός: (Relative Efficiency)*

Ένα DMU θεωρείται αποτελεσματικό εάν και μόνο εάν οι αποτελεσματικότητες των υπολοίπων DMU δεν δείχνουν ότι μία τυχόν

βελτίωση των εισόδων ή των εξόδων του θα χειροτερέψει κάποιες από τις υπόλοιπες εισόδους ή εξόδους.

Ας σημειωθεί ότι αυτός ο ορισμός δεν απαιτεί την ύπαρξη πληροφοριών σχετικών με τιμές/κόστη και επιπλέον δεν βασίζεται στον καθορισμό της σημαντικότητας των διαφόρων εισόδων και εξόδων. Επίσης, αποφεύγει τον σαφή προσδιορισμό των τυπικών σχέσεων μεταξύ εισόδων και εξόδων. Αυτό το είδος της αποτελεσματικότητας αναφέρεται στην οικονομική θεωρία ως τεχνική αποτελεσματικότητα (technical efficiency) και μπορεί να επεκταθεί και σε άλλα είδη αποτελεσματικότητας όταν πρόκειται για δεδομένα όπως τιμές, μοναδιαία κόστος κ.α., τα οποία είναι διαθέσιμα προς χρήση στην DEA.

## 2.2 Μέθοδος DEA και μαθηματικές διατυπώσεις

### 2.2.1 Ιστορικό Υπόβαθρο και Φιλοσοφία της DEA

Η ανάλυση αποτελεσματικότητας (efficiency analysis) έχει απασχολήσει αρκετούς ερευνητές εξαιτίας της σχετικής δυσκολίας που συναντάται στην αξιολόγηση της αποτελεσματικότητας μιας επιχείρησης ή ενός οργανισμού. Ο Farrell (1957) προσπάθησε να μετρήσει την αποτελεσματικότητα μιας μονάδας παραγωγής στην περίπτωση που υπάρχει μία μόνο είσοδος και έξοδος. Η μελέτη του Farrell περιλαμβάνει την μέτρηση και εκτίμηση αποτελεσματικών τεχνικών και την δημιουργία μιας αποτελεσματικής παραγωγικής συνάρτησης.

Ο Farrell εφάρμοσε το μοντέλο του στην εκτίμηση της αποτελεσματικότητας του αγροτικού τομέα στην ΗΠΑ σε σύγκριση με άλλες χώρες. Εντούτοις, ο Farrell δεν κατάφερε να ορίσει μία μέθοδο που θα συνδυάζει πολλαπλές εισόδους και εξόδους.

Οι Charnes et al. (1978) διεύρυναν την ιδέα του Farrell και πρότειναν ένα μοντέλο το οποίο γενικεύει την εκτίμηση της αποτελεσματικότητας στην περίπτωση πολλαπλών εισόδων και εξόδων. Ειδικότερα, οι Charnes et al. καθόρισαν την τεχνική αποτελεσματικότητα ως τον λόγο των εξόδων που παράγονται προς τις εισόδους που χρησιμοποιούνται:

$$\text{Τεχνική αποτελεσματικότητα} = \frac{\text{σταθμισμένο άθροισμα των δεδομένων εξόδου}}{\text{σταθμισμένο άθροισμα των δεδομένων εισόδου}}$$



Η DEA επιτρέπει σε κάθε DMU να επιλέγει το σύνολο των βαρών  $u_o$  (έξοδοι) και  $v_o$  (είσοδοι). Το αποτελεσματικό σκορ (efficiency score) που σημειώνεται προκύπτει από την ανάλυση ενός δείγματος από διάφορα DMU δοθέντος ότι το σύνολο των βαρών πρέπει να είναι προσιτό από άλλες μονάδες και καμία από αυτές τις μονάδες δεν θα πρέπει να έχει αποτελεσματικό σκορ μεγαλύτερο της μονάδας.

Σε σύγκριση με την ανάλυση παλινδρόμησης (regression analysis), η οποία παρέχει μία μέση κατανομή των DMU, η DEA οδηγεί στην ανάπτυξη μιας κατά τμήματα γραμμικής εμπειρικής επιφάνειας παραγωγής η οποία στην οικονομική θεωρία αναπαριστά το αποτελεσματικότερο όριο παραγωγής. Προβάλλοντας κάθε παραγωγική μονάδα πάνω στο αποτελεσματικό όριο δίνεται η δυνατότητα να εκτιμηθεί το επίπεδο της μη αποτελεσματικότητας σε σύγκριση με ένα σύνολο αναφοράς ή ένα κυρτό συνδυασμό από διάφορα σύνολα αναφοράς. Η προβολή αυτή αναφέρεται σε ένα DMU το οποίο είναι συνδυασμός ενός ή περισσότερων DMU.

### **2.2.2 Βασικές μαθηματικές διατυπώσεις της DEA**

Παρακάτω παρατίθενται μερικές μαθηματικές διατυπώσεις της μεθόδου DEA, όπως αναπτύχθηκαν κατά καιρούς από διάφορους ερευνητές.

#### **2.2.2.1 Το μοντέλο CCR των Charnes, Cooper, Rhodes**

Το αρχικό μοντέλο της DEA όπως διατυπώθηκε από τους Charnes, Cooper και Rhodes (1978, 1979) παρουσιάζεται παρακάτω:

$$\left. \begin{aligned}
 & \text{Max} \left\{ h_o = \frac{\sum_{r=1}^s u_r y_{ro}}{\sum_{i=1}^m v_i x_{io}} \right\} \\
 & \text{υπό:} \\
 & \frac{\sum_{r=1}^s u_r y_{rj}}{\sum_{i=1}^m v_i x_{ij}} \leq 1 \quad \text{για } j = 1, 2, \dots, n \\
 & \frac{u_r}{\sum_{i=1}^m v_i x_{io}} > \varepsilon \quad \text{για } r = 1, 2, \dots, s \\
 & \frac{v_i}{\sum_{i=1}^m v_i x_{io}} > \varepsilon \quad \text{για } i = 1, 2, \dots, m \\
 & \varepsilon > 0
 \end{aligned} \right\} \quad (2.1)$$

Η ανάπτυξη του παραπάνω μοντέλου θεωρεί ότι υπάρχουν διαθέσιμα στοιχεία για  $n$  DMUs καθένα από τα οποία χρησιμοποιεί  $m$  εισόδους και παράγει  $s$  εξόδους. Ως  $y_{rj}$  και  $x_{ij}$  συμβολίζονται αντίστοιχα η έξοδος  $r$  και η είσοδος  $i$  του DMU  $j$ . Δεδομένων αυτών των πληροφοριών η επίλυση του παραπάνω προβλήματος για ένα  $DMU_o$  οδηγεί στον προσδιορισμό της μέγιστης αποτελεσματικότητάς του  $h_o^*$  σε μια κλίμακα από το 0 έως το 1. Η περίπτωση  $h_o^* = 1$  υποδεικνύει ότι το συγκεκριμένο DMU είναι πλήρως αποτελεσματικό, ενώ αντίθετα η περίπτωση  $h_o^* < 1$  υποδεικνύει ότι το DMU δεν αξιοποιεί πλήρως τις διαθέσιμες εισόδους για την παραγωγή μέγιστων αποτελεσμάτων (εξόδων) και συνεπώς είναι αναποτελεσματικό.

Η αντικειμενική συνάρτηση του μοντέλου (2.1), έχει στον αριθμητή το σύνολο των επιθυμητών εξόδων, ενώ ο παρονομαστής αναπαριστά το σύνολο των εισόδων που χρησιμοποιούνται για την παραγωγή αυτών των εξόδων. Ως  $u_r$  και  $v_i$  συμβολίζονται αντίστοιχα τα βάρη της εξόδου  $r$  και της εισόδου  $i$ . Οι βέλτιστες τιμές των βαρών αυτών  $u_r^*$  και  $v_i^*$  προσδιορίζονται από την επίλυση του μοντέλου. Βάσει των βαρών αυτών ορίζεται μία εικονική έξοδος  $Y_o = \sum u_r^* y_{ro}$  για  $r = 1, 2, \dots, s$  και μία εικονική είσοδος

$X_o = \sum v_i^* x_{io}$  για  $i=1,2,\dots,m$ , οι οποίες επιτρέπουν την μέτρηση της αποτελεσματικότητας βάσει του λόγου  $h_o = Y_o / X_o$ . Τα προσδιοριζόμενα βέλτιστα βάρη είναι εκείνα που αποδίδουν τη μέγιστη αποτελεσματικότητα  $h_o^*$  για ένα DMU<sub>o</sub>.

Ο πρώτος περιορισμός του προβλήματος προσδιορίζει το άνω όριο για το βαθμό αποτελεσματικότητας ενός DMU, ενώ οι υπόλοιποι δύο περιορισμοί χρησιμοποιούνται ώστε να αποφευχθεί ο προσδιορισμός μηδενικού βάρους για κάποια είσοδο ή έξοδο (ως  $\varepsilon$  συμβολίζεται μια μικρή θετική σταθερά).

Εκείνο που πρέπει να σημειωθεί, είναι ότι η DEA είναι μία μέθοδος σχετικής αποδοτικότητας, η οποία και προσδιορίζεται με την εφαρμογή της παραπάνω βελτιστοποίησης. Επομένως, για κάθε εκτιμώμενο DMU, η μέγιστη αποτελεσματικότητά του προσδιορίζεται σε σχέση με ένα υποσύνολο  $K$  αποτελούμενο από αποτελεσματικά DMU για τα οποία ισχύει:

$$\left. \begin{aligned} \frac{\sum_{r=1}^s u_r^* y_{rk}}{\sum_{i=1}^m v_i^* x_{ik}} &= 1, k \in K \end{aligned} \right\} \quad (2.2)$$

Το κύριο πρόβλημα στην εφαρμογή του μοντέλου αφορά τη μη γραμμική μορφή της αντικειμενικής συνάρτησης. Για την αντιμετώπιση του θέματος αυτού, οι Charnes et al. (1978) πρότειναν την μετατροπή του μοντέλου (2.1) σε ένα πρόβλημα γραμμικού προγραμματισμού η μαθηματική διατύπωση του οποίου είναι η ακόλουθη:

$$\left. \begin{aligned} &Max \left\{ \sum_{r=1}^s u_r y_{ro} \right\} \\ &υπό: \\ &\sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} \leq 0 \\ &\sum_{i=1}^m v_i x_{io} = 1 \\ &-u_r \leq -\varepsilon \\ &-v_i \leq -\varepsilon \end{aligned} \right\} \quad (2.3)$$

Η δυϊκή μορφή του παραπάνω γραμμικού προβλήματος δίνεται παρακάτω:

$$\left. \begin{array}{l} \text{Min} \left\{ \theta - \varepsilon \left[ \sum_{i=1}^m s_i^- + \sum_{r=1}^s s_r^+ \right] \right\} \\ \text{υπό:} \\ \theta x_{io} - \sum_{j=1}^n x_{ij} \lambda_j - s_i^- = 0 \\ y_{ro} = \sum_{j=1}^n y_{rj} \lambda_j - s_r^+ \\ \lambda_j \geq 0 \quad \text{για} \quad j=1, \dots, n \\ s_i^- \geq 0 \quad \text{για} \quad i=1, \dots, m \\ s_r^+ \geq 0 \quad \text{για} \quad r=1, \dots, s \\ \theta \text{ ελεύθερο προσήμου} \end{array} \right\} \quad (2.4)$$

Κάθε αποδεκτή επιλογή των  $\lambda_j$  προσδιορίζει ένα άνω όριο για τις εξόδους και ένα κάτω όριο για τις εισόδους ενός DMU<sub>o</sub> και με βάση αυτά τα όρια και τις βέλτιστες τιμές  $\lambda_j^*, s_i^-, s_r^+ \geq 0$ , το  $\theta$  ελαχιστοποιείται. Η επιλογή αυτών των λύσεων οδηγεί στον προσδιορισμό ενός άνω ορίου, το οποίο περιλαμβάνει όλες τις παρατηρήσεις.

Μία βέλτιστη λύση επιτυγχάνεται στην περίπτωση που η τιμή της μεταβλητής  $\theta$  κυμαίνεται μεταξύ του μηδενός και της μονάδας ( $0 \leq \theta^* \leq 1$ ) και πάντα θα υπάρχει μία λύση για την οποία θα ισχύουν τα εξής:  $\theta = 1$ ,  $\lambda_o = 1$  και  $\lambda_j, s_i^-, s_r^+ = 0$ . Επειδή το γραμμικό πρόβλημα (2.4) έχει μία πεπερασμένη βέλτιστη λύση, από τη θεωρία δυϊκότητας προκύπτει ότι:

$$h_o^* = \theta^* - \varepsilon \left( \sum_{i=1}^m s_i^{-*} + \sum_{r=1}^s s_r^{+*} \right) = \sum_{r=1}^s u_r^* y_{ro} \quad \left. \right\} \quad (2.5)$$

Στην περίπτωση όπου η τιμή της μεταβλητής  $\theta$  είναι ίση με την μονάδα ( $\theta = 1$ ) αυτό δεν σημαίνει ότι το αποτελεσματικό σκορ  $h_o^*$  θα πάρει την τιμή 1. Για να συμβεί αυτό θα πρέπει οι τιμές των μεταβλητών  $s_r^{+*}$  και  $s_i^{-*}$  να είναι ίσες με το μηδέν για όλες τις τιμές των  $r, i$ . Αντίστοιχα, εάν  $s_r^{+*}, s_i^{-*} = 0$  για όλες τις τιμές των  $r, i$ , αυτό δεν

συνεπάγεται ότι  $h_o^* = 1$  παρὰ μόνο εάν  $\theta = 1$ . Συνεπώς, ένα  $DMU_o$  θεωρείται ως 100% αποτελεσματικό εάν και μόνο εάν όλες οι τιμές των μεταβλητών  $s_r^{+*}$  και  $s_i^{-*}$  είναι μηδενικές και  $\theta = 1$ .

### 2.2.2.2 Το μοντέλο BCC των Banker, Charnes and Cooper

Το BCC είναι μία άλλη διατύπωση της DEA που διατυπώθηκε από τους Banker, Charnes και Cooper (1984). Η βασική διαφορά αυτού του μοντέλου και του CCR είναι η χρήση της κλίμακας αποδόσεων. Το μοντέλο CCR στηρίζει την εκτίμηση της αποτελεσματικότητας στην υπόθεση ότι οποιαδήποτε αύξηση των εισόδων οδηγεί σε μια ανάλογη αύξηση των εξόδων (constant returns to scale) παρέχοντας μια συνολική αξιολόγηση της αποτελεσματικότητας. Το μοντέλο BCC βασίζεται στην υπόθεση ότι μια αύξηση των εισόδων δεν οδηγεί απαραίτητα σε μια ανάλογη αύξηση των εξόδων (variable returns to scale) και παρέχει μια εκτίμηση της τεχνικής αποτελεσματικότητας μιας μονάδας βάσει της λειτουργίας της (pure technical efficiency).

Παρακάτω διατυπώνεται το μοντέλο BCC, το οποίο είναι ισοδύναμο της διατύπωσης (2.4) του μοντέλου CCR:

$$\left. \begin{array}{l} \text{Min} \left\{ \theta - \varepsilon \left[ \sum_{i=1}^m s_i^- + \sum_{r=1}^s s_r^+ \right] \right\} \\ \text{υπό:} \\ \theta x_{io} - \sum_{j=1}^n x_{ij} \lambda_j - s_i^- = 0 \\ y_{ro} = \sum_{j=1}^n y_{rj} \lambda_j - s_r^+ \\ \sum_{j=1}^n \lambda_j = 1 \\ \lambda_j \geq 0 \quad \text{για } j=1, \dots, n \\ s_i^- \geq 0 \quad \text{για } i=1, \dots, m \\ s_r^+ \geq 0 \quad \text{για } r=1, \dots, s \end{array} \right\} \quad (2.6)$$

Η διαφορά μεταξύ του μοντέλου CCR (διατύπωση (2.4)) και του μοντέλου BCC (διατύπωση (2.6)) βρίσκεται στις μεταβλητές  $\lambda_j$  το άθροισμα των οποίων θα πρέπει να είναι ίσο με τη μονάδα.

Ερμηνεύοντας την προσέγγιση BCC σε σχέση με τη θεώρηση του μοντέλου CCR, θα πρέπει να σημειωθεί ότι στην περίπτωση του μοντέλου CCR η αποτελεσματικότητα ενός DMU εξαρτάται τόσο από την λειτουργία του όσο και από τις συνθήκες στις οποίες λειτουργεί. Αντίθετα, το μοντέλο BCC εξετάζει τη λειτουργία του ίδιου του DMU ανεξάρτητα από τις συνθήκες στις οποίες λειτουργεί παρέχοντας έτσι μια εκτίμηση της καθαρά τεχνικής αποτελεσματικότητάς του.

Η δυϊκή μορφή του μοντέλου BCC δίνεται παρακάτω:

$$\left. \begin{array}{l} \text{Max} \left\{ \sum_{r=1}^s u_r y_{ro} - u_o \right\} \\ \text{υπό:} \\ \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} - u_o \leq 0 \\ \sum_{i=1}^m v_i x_{io} = 1 \\ -u_r \leq -\varepsilon \\ -v_i \leq -\varepsilon \end{array} \right\} \quad (2.7)$$

Σε αυτό το μοντέλο, η μεταβλητή  $u_o^*$  εκφράζει τον τρόπο με τον οποίο μεταβάλλονται οι έξοδοι του DMU<sub>o</sub> όταν μεταβάλλονται οι εισοδοί του. Ειδικότερα, η περίπτωση  $u_o^* = 0$  υποδεικνύει ότι υπάρχει μια σταθερή σχέση μεταξύ εισόδων και εξόδων (η αύξηση των εισόδων οδηγεί σε ανάλογη αύξηση των εξόδων). Η περίπτωση  $u_o^* < 0$  υποδεικνύει ότι αύξηση των εξόδων είναι μια αύξουσα συνάρτηση της αύξησης στις εισόδους (για παράδειγμα οικονομίες κλίμακας), ενώ τέλος η περίπτωση  $u_o^* > 0$  υποδεικνύει ότι η αύξηση των εξόδων είναι φθίνουσα συνάρτηση της αύξησης των εισόδων.

### 2.2.3 Σύγκριση της DEA με Στατιστικές Μεθόδους Γραμμικής Παλινδρόμησης

Στο παρελθόν, οι στατιστικές προσεγγίσεις παλινδρόμησης χρησιμοποιούνταν για τη μέτρηση της αποτελεσματικότητας. Συνεπώς είναι χρήσιμο να εξεταστούν οι διαφορές της DEA σε σχέση με τις στατιστικές προσεγγίσεις ώστε να τονιστούν τα ιδιαίτερα χαρακτηριστικά της DEA.

Ένα πρώτο βασικό χαρακτηριστικό της DEA είναι ότι σε αντίθεση με τις παραδοσιακές στατιστικές προσεγγίσεις παλινδρόμησης, δεν απαιτεί σαφή προσδιορισμό των συναρτησιακών εξισώσεων που συνδέουν τις εισόδους με τις εξόδους. Τα αποτελέσματα της DEA μπορούν να ερμηνευτούν ορίζοντας μια διαδικασία προσέγγισης σε οποιαδήποτε συνάρτηση είναι εφαρμόσιμη σε γειτονικά σημεία από τις αρχικές εκτιμώμενες συντεταγμένες ενός DMU. Συνεπώς, η μέθοδος DEA είναι περισσότερο ευέλικτη στο να αναγνωρίζει τις διαφορές των παραγωγικών λειτουργιών ανάμεσα στα DMUs.

Δεύτερον, η DEA προσανατολίζεται προς εκείνα τα DMUs τα οποία θεωρούνται υπεύθυνα για τη χρησιμοποίηση των εισόδων προς την παραγωγή των κατάλληλων εξόδων. Συνεπώς, χρησιμοποιεί  $n$  βελτιστοποιήσεις, μία για κάθε DMU, αντί της απλής βελτιστοποίησης η οποία συνήθως συνδέεται με τα στατιστικά μοντέλα παλινδρόμησης. Συνεπώς στη DEA προσδιορίζεται μια λύση για κάθε εκτιμώμενο DMU.

Τρίτον, το μειονέκτημα όλων των άλλων στατιστικών προσεγγίσεων παλινδρόμησης σχετίζεται άμεσα με την αδυναμία τους να προσδιορίσουν το επίπεδο της αναποτελεσματικότητας και τις πηγές της με αποτέλεσμα να μην υπάρχουν ενδείξεις όσον αφορά τα διορθωτικά μέτρα που πρέπει να ληφθούν. Αντίθετα, η DEA παρέχει πληροφορίες τόσο για τα αίτια των αναποτελεσματικοτήτων όσο και για το επίπεδό τους, στοιχεία σημαντικά για την αντιμετώπιση των προβλημάτων που προσδιορίζονται (Chaffai, 1997).

Τέταρτον, οι συντελεστές ή τα βάρη που προσδιορίζονται μέσω της DEA είναι μοναδικά για κάθε DMU που εξετάζεται, σε αντίθεση με τους συντελεστές ενός μοντέλου παλινδρόμησης οι οποίοι αφορούν ένα σύνολο DMUs. Συνεπώς, η DEA, σε αντίθεση με τα στατιστικά μοντέλα παλινδρόμησης, δεν οδηγεί στην ανάπτυξη ενός μοντέλου πρόβλεψης της αποτελεσματικότητας. Συνεπώς, η ανάλυση παλινδρόμησης είναι πιο κατάλληλη εάν ένας αναλυτής επιθυμεί να χρησιμοποιήσει τα αποτελέσματα για την εκτίμηση ή την πρόβλεψη μελλοντικών επιδόσεων μιας ομάδας παραγωγικών μονάδων ή εάν τα αποτελέσματα της ανάλυσης πρόκειται να χρησιμοποιηθούν για την πρόβλεψη των επιδόσεων νέων παραγωγικών μονάδων. Επίσης, οι μεταβλητές της γραμμικής παλινδρόμησης μπορούν να προσδιορίσουν και να αποδώσουν στατιστική ακρίβεια κάτι το οποίο δεν υφίσταται με τα βάρη της DEA. Επίσης, στην DEA δεν λαμβάνεται υπόψη το τυχαίο σφάλμα, όπως στα στατιστικά μοντέλα. Για το λόγο αυτό, υπάρχει πιθανότητα η DEA να συγχέει την τυχαιότητα με τα μη αποτελεσματικά σύνολα.

Παρά τις παραπάνω διαφοροποιήσεις είναι σημαντικό να τονιστεί ότι πρόσφατα έχουν προταθεί διάφοροι τρόποι για το συνδυασμό των δύο προσεγγίσεων. Σχετικές έρευνες στο θέμα αυτό έχουν παρουσιαστεί από τους Charnes et al. (1986), Arnold et al. (1996), Bardhan et al. (1997).

Τέλος χρήσιμο είναι να αναφερθούν τα αποτελέσματα ορισμένων χαρακτηριστικών οι οποίες εξέτασαν την αποτελεσματικότητα της DEA σε σχέση με την αποτελεσματικότητα στατιστικών προσεγγίσεων παλινδρόμησης.

Οι Bowlin et al. (1985) σε μια πρώτη σχετική έρευνα στο θέμα αυτό, ανέπτυξαν ένα σύνολο υποθετικών δεδομένων για νοσοκομειακές μονάδες έχοντας προκαθορίσει ποια από τα δεδομένα είναι αποτελεσματικά ή μη. Αυτό το σύνολο δεδομένων χρησιμοποιήθηκε για να εξεταστεί η DEA σε σύγκριση με την αναλογική ανάλυση (ratio analysis) και την ανάλυση παλινδρόμησης (regression analysis). Από αυτή την έρευνα διαπιστώθηκε ότι η DEA υπερτερεί έναντι των άλλων στατιστικών προσεγγίσεων καθώς μπορεί να προσδιορίσει τις μη αποτελεσματικές μονάδες. Επιπλέον, διαπιστώθηκε ότι τα αποτελέσματα των προσεγγίσεων παλινδρόμησης χειροτερεύουν όταν χρησιμοποιούνται διαστήματα εμπιστοσύνης για τη διάκριση μεταξύ των αποτελεσματικών και των μη αποτελεσματικών μονάδων.

Ο Banker et al. (1988) ολοκλήρωσε μία μελέτη προσομοίωσης συγκρίνοντας τους υπολογισμούς της DEA για ένα γνωστό όριο παραγωγικότητας με τους υπολογισμούς που προέκυψαν από μία μέθοδο υπερλογαριθμικής παλινδρόμησης. Με βάση τα αποτελέσματα από μία λογαριθμική – γραμμική μέθοδο για την εκτίμηση τυχαίων γενικευμένων παρατηρήσεων, διαπιστώθηκε ότι η DEA υπερτερεί έναντι κάθε υπερλογαριθμικής μεθόδου στην εκτίμηση και τον προσδιορισμό των μη αποτελεσματικών μονάδων.

Οι Banker, Conrad και Strauss (1986) χρησιμοποίησαν πρακτικά δεδομένα από ένα σύνολο πληροφοριών που παρείχαν νοσοκομεία της Βόρειας Καρολίνας με απώτερο σκοπό να συγκρίνουν τα αποτελέσματα της DEA σε σχέση με οικονομετρικά μοντέλα. Βάσει των αποτελεσμάτων της ανάλυσης διαπίστωσαν ότι η DEA είναι σε θέση να εντοπίσει αναποτελεσματικότητες οι οποίες δεν ήταν δυνατόν να προσδιοριστούν βάσει των οικονομετρικών προσεγγίσεων.



## 2.2.4 Επεκτάσεις της μεθόδου DEA

Μετά την αρχική διατύπωση της μεθόδου της DEA από τους Charnes et al. (1978), υπήρξαν διάφορες παραλλαγές του μοντέλου οι οποίες δημιουργήθηκαν για να καλύψουν καινούργιες και ποικίλες ανάγκες. Παρακάτω περιγράφονται κάποια από αυτά τα μοντέλα, καθώς επίσης και ο λόγος για τον οποίο σχεδιάστηκαν. Πρόσθετες πληροφορίες για αυτά τα μοντέλα δίνονται από τους Ahn et al. (1988), οι οποίοι αναλύουν τα αποτελέσματα των διαφόρων μοντέλων της DEA βάσει μαθηματικών διατυπώσεων.

### 2.2.4.1 Το προσθετικό μοντέλο

Η ιδέα για την ανάπτυξη του προσθετικού μοντέλου παρουσιάστηκε αρχικά από τους Charnes et al. (1985) και τεκμηριώθηκε αργότερα από τους Banker et al. (1989). Το προσθετικό μοντέλο διατυπώνεται ως εξής:

$$\left. \begin{array}{l}
 \text{Max} \left\{ \sum_{i=1}^m s_i^- + \sum_{r=1}^s s_r^+ \right\} \\
 \text{υπό:} \\
 x_{io} - \sum_{j=1}^n x_{ij} \lambda_j - s_i^- = 0 \\
 y_{ro} = \sum_{j=1}^n y_{rj} \lambda_j - s_r^+ \\
 \sum_{j=1}^n \lambda_j = 1 \\
 \lambda_j \geq 0 \quad \text{for } j=1, \dots, n \\
 s_i^- \geq 0 \quad \text{for } i=1, \dots, m \\
 s_r^+ \geq 0 \quad \text{for } r=1, \dots, s
 \end{array} \right\} \quad (2.8)$$

Αυτή η διατύπωση αντιστοιχεί στις παραδοχές του μοντέλου BCC. Παρόμοια, εάν από το μοντέλο (2.8) αφαιρεθεί ο περιορισμός  $\sum_j \lambda_j = 1$ , τότε το νέο μοντέλο που διαμορφώνεται αντιστοιχεί στις παραδοχές του μοντέλου CCR. Η διαφορά μεταξύ του προσθετικού μοντέλου και του BCC είναι ότι το  $\theta$  παραλείπεται από το μοντέλο και όλες οι μη αποτελεσματικές μονάδες εκφράζονται με την βοήθεια των μεταβλητών  $s_r^+$  και  $s_i^-$ . Επομένως, ο μόνος τρόπος για την εκτίμηση της αποτελεσματικότητας είναι να εξεταστεί εάν αυτές οι μεταβλητές είναι μηδενικές.

Ένα DMU το οποίο είναι αποτελεσματικό για το μοντέλο BCC είναι επίσης αποτελεσματικό και για το προσθετικό μοντέλο και αντιστρόφως (Ahn et al., 1988). Εντούτοις, όταν ένα DMU είναι μη αποτελεσματικό, οι πηγές και τα επίπεδα της αποτελεσματικότητας διαφέρουν για τις διάφορες διατυπώσεις της DEA εξαιτίας των διαφορετικών τρόπων μέτρησης της αποτελεσματικότητας.

#### **2.2.4.2 Το πολλαπλασιαστικό μοντέλο**

Τα μοντέλα που συζητήθηκαν προηγουμένως – CCR, BCC και προσθετικό – ορίζουν ένα προσθετικό συνδυασμό εισόδων και εξόδων. Εναλλακτικά, οι Charnes et al. (1982, 1983) πρότειναν τη χρησιμοποίηση ενός πολλαπλασιαστικού συνδυασμού. Η βασική διαφορά μεταξύ αυτού του μοντέλου και των άλλων μοντέλων της DEA είναι ότι οι έξοδοι και οι εισοδοί προσδιορίζονται πολλαπλασιαστικά και όχι αθροιστικά.. Επίσης, τα διανύσματα  $X$  και  $Y$  των εισόδων και εξόδων είναι λογάριθμοι. Το αποτέλεσμα είναι ότι το αποτελεσματικό όριο είναι λογαριθμικό-γραμμικό και όχι απλά γραμμικό όπως στην περίπτωση των μοντέλων CCR και BCC. Επιπλέον, όπως και στο προσθετικό μοντέλο παραλείπεται το  $\theta$  και όλες οι μη αποτελεσματικές μονάδες εκφράζονται με την βοήθεια των μεταβλητών  $s_r^+$  και  $s_i^-$ .

#### **2.2.4.3 Το μοντέλο MED**

Τα CCR, BCC και άλλα μοντέλα της DEA επιτρέπουν την σύγκριση ενός DMU που δεν ανήκει στο αποτελεσματικό όριο με ένα γραμμικό συνδυασμό εκείνων των DMUs που ανήκουν στο αποτελεσματικό σύνολο. Σε μερικές περιπτώσεις, είναι επιθυμητή η σύγκριση μεταξύ των ίδιων των DMUs και όχι σε σχέση με κάποιο γραμμικό συνδυασμό τους. Το μοντέλο MED (Measures of Efficiency Dominance) το οποίο προτάθηκε από τους Bardhan et al. (1996) επιτρέπει την ανάλυση του θέματος αυτού.

#### **2.2.4.4 Το μοντέλο ασφαλούς πεδίου**

Οι Charnes et al. (1990) επισήμαναν ότι η δομή μιας εμπειρικής παραγωγικής συνάρτησης συχνά είναι αναξιόπιστη, καθώς είναι πιθανό να μην είναι σε θέση να περιγράψει τις πραγματικές δυνατότητες ενός DMU. Ως αποτέλεσμα αυτού του προβλήματος, είναι πιθανόν μονάδες που εντοπίζονται μέσω της DEA ως αποτελεσματικές στη ουσία να μην είναι.

Για την αντιμετώπιση του θέματος αυτού οι Thompson et al. (1986, 1990) πρότειναν την προσέγγιση του ασφαλούς πεδίου (assurance region approach) αναπτύσσοντας και το αντίστοιχο μοντέλο. Η φιλοσοφία της συγκεκριμένης μεθοδολογίας βασίζεται στον προσδιορισμό του πεδίου τιμών των βαρών που μπορούν να αποδοθούν στις εισόδους και στις εξόδους, περιορίζοντας έτσι τα αποδεκτά επίπεδα των αποτελεσματικών εισόδων και εξόδων. Ειδικότερα, στην προσέγγιση αυτή ορίζονται άνω και κάτω όρια στις επιτρεπόμενες τιμές των μεταβλητών. Τα όρια αυτά έχουν την παρακάτω μορφή:

$$\left. \begin{aligned} \alpha_r &\leq \frac{v_r}{v_{ro}} \leq \beta_r, & r=1, \dots, s \\ \delta_i &\leq \frac{u_i}{u_{io}} \leq \gamma_i, & i=1, \dots, m \end{aligned} \right\} \quad (2.9)$$

Τα  $v_{ro}$  και  $u_{io}$  είναι οι δυϊκές μεταβλητές οι οποίες χρησιμοποιούνται για τον προσδιορισμό των άνω ( $\alpha_r$  και  $\beta_r$ ) και κάτω ( $\delta_i$  και  $\gamma_i$ ) ορίων για τις δυϊκές μεταβλητές που συνδέονται με κάθε είσοδο και έξοδο και  $\alpha_{ro}, \beta_{ro}, \delta_{io}, \gamma_{io} = 1$ .

#### 2.2.4.5 Μοντέλα εξωγενών παραγόντων

Το μοντέλο της DEA παρέχει πληροφορίες σχετικά με τον βαθμό, στον οποίο ένα DMU που ανήκει στο μη αποτελεσματικό σύνολο, μπορεί να μειώσει μία είσοδο ή να αυξήσει μία έξοδο έτσι ώστε να γίνει αποτελεσματικό. Εντούτοις, μερικές φορές μεταβλητές της ανάλυσης (είσοδοι ή έξοδοι) προσδιορίζονται βάσει εξωγενών παραγόντων οι οποίοι δεν μπορούν να ελέγχουν από τους υπεύθυνους για την λειτουργία του DMU. Συνεπώς, τέτοιες μεταβλητές θα Για παράδειγμα, κατά την λειτουργία των πτήσεων μιας αεροπορικής εταιρείας θα πρέπει να λαμβάνονται υπόψη οι καιρικές συνθήκες (εξωγενής μεταβλητή), καθώς αναλόγως με τις καιρικές συνθήκες μπορεί να υπάρχουν θετικές ή αρνητικές επιπτώσεις για την υπάρχουσα πτήση. Οι Charnes et al. (1985) καθώς και οι Banker και Morey (1986b) παρουσίασαν κατάλληλες επεκτάσεις της DEA στις οποίες έμφαση δίνεται στις μεταβλητές οι οποίες μπορούν να ελεγχθούν ενώ αντίθετα οι εξωγενείς μεταβλητές θεωρούνται σταθερές.

#### 2.2.4.6 Το κατηγορικό μοντέλο

Οι Banker και Morey (1986a) πρότειναν μία προσαρμογή στο μοντέλο της DEA η οποία επιτρέπει των συνυπολογισμό εκείνων των μεταβλητών ποιοτικής (κατηγορικής) μορφής.

Αυτή η προέκταση της DEA εξαλείφει την ανάγκη για την εκτίμηση των μεταβλητών σε μία συνεχή κλίμακα και επιτρέπει την ενσωμάτωση των κατηγορικών/ποιοτικών μεταβλητών στην ανάλυση. Για παράδειγμα σε μία μελέτη που πραγματοποιήθηκε από τους Banker και Morey (1986a ) για τα εστιατόρια των Ηνωμένων Πολιτειών, χρησιμοποιήθηκε ως μεταβλητή στην ανάλυση το σύστημα λειτουργίας των εστιατορίων.

### **2.2.5 Κριτικές, περιορισμοί και πλεονεκτήματα της DEA**

Όπως είναι κατανοητό υψηλή αποτελεσματικότητα μπορεί να επιτευχθεί με πολλούς τρόπους. Στην περίπτωση της μεθόδου DEA αυτό αποτυπώνεται στις διαφοροποιήσεις που πιθανόν να υπάρχουν στα βάρη που αποδίδονται στις διάφορες εισόδους και εξόδους. Είναι γεγονός ότι με την κατάλληλη επιλογή βαρών ένα μεγάλο ποσοστό από το σύνολο των υπό μελέτη μονάδων μπορεί να αποδειχθούν αποτελεσματικές και έτσι η DEA θα έχει μικρή διακριτική ικανότητα. Ένα σημαντικό σημείο το οποίο θα πρέπει να τονιστεί είναι ότι μια μονάδα η οποία έχει τον υψηλότερο δείκτη αποδοτικότητας προϊόν/πόρος σε ένα ζεύγος από τις μεταβλητές εισόδου/εξόδου θα αποδειχθεί αποτελεσματική ή θα έχει αποτελεσματικότητα που θα προσεγγίζει τη μονάδα λόγω του γεγονότος ότι με βάσει την προσέγγιση της DEA τοποθετείται το μέγιστο βάρος σε αυτό το δείκτη και το ελάχιστο στις υπόλοιπες μεταβλητές εισόδου/εξόδου.

Το παραπάνω γεγονός αποτελεί ένα σημείο το οποίο απαιτεί προσοχή κατά την εφαρμογή και χρήση της DEA. Δηλαδή ότι μια μονάδα μπορεί να αποδειχθεί αποτελεσματική, όχι διότι είναι πραγματικά ο τρόπος που λειτουργεί αποδοτικός, αλλά διότι αυτό προέκυψε από μια ευνοϊκή επιλογή βαρών κατά τη διαδικασία επίλυσης με βάση την προσέγγιση DEA. Ένας τρόπος να αντιμετωπιστεί το παραπάνω πρόβλημα είναι να τεθούν κατάλληλοι περιορισμοί στα προσδιοριζόμενα βάρη. Αυτό γίνεται με τον προσδιορισμό ενός ελάχιστου βάρους για κάθε μεταβλητή του μοντέλου έτσι, ώστε να διασφαλιστεί το γεγονός ότι κάθε μεταβλητή εισόδου ή εξόδου θα διαδραματίζει κάποιο ρόλο στην αξιολόγηση της αποτελεσματικότητας των μονάδων. Κατά ανάλογο τρόπο θα μπορούσε να τοποθετηθεί και ένα μέγιστο όριο στα βάρη έτσι ώστε να αποκλειστεί η περίπτωση αποδόσης ιδιαίτερα υψηλής σημασίας σε συγκεκριμένες μεταβλητές εισόδου ή εξόδου. Βέβαια αυτοί οι περιορισμοί δε θα πρέπει να είναι ιδιαίτερα αυστηροί καθότι τότε θα παραβιάζονταν η βασική προϋπόθεση της μεθόδου που θέλει τις μονάδες να αξιολογούνται με βάση ένα κοινό σύνολο βαρών. Δηλαδή το ζητούμενο είναι να γίνει ένας συμβιβασμός μεταξύ της ευελιξίας

στην επιλογή βαρών και της εξασφάλισης μιας κοινής βάσης αξιολόγησης για όλες τις μονάδες.

Ερευνώντας την επίδραση των βαρών στην εφαρμογή της DEA, οι Charnes et al. (1994) αναφέρουν δύο καταστάσεις στις οποίες απαιτείται συμπληρωματικός έλεγχος όσον αφορά τα προσδιοριζόμενα βάρη:

- Όταν στην ανάλυση δεν λαμβάνονται υπόψη συμπληρωματικές πληροφορίες οι οποίες δεν μπορούν να ενσωματωθούν απευθείας στο μοντέλο.
- Όταν ο αριθμός των παραγόντων είναι υπερβολικά μεγάλος συγκρινόμενος με τον αριθμό των DMUs.

Γενικότερα, η DEA μπορεί να λειτουργήσει ως ένα πολύ ισχυρό εργαλείο όταν βέβαια χρησιμοποιείται με το σωστό τρόπο. Μερικοί από τους λόγους για τους οποίους η μέθοδος αυτή είναι τόσο χρήσιμη στο χώρο της διοίκησης είναι οι ακόλουθοι :

- Η DEA μπορεί να συμπεριλάβει πολλαπλά δεδομένα εισόδου και εξόδου κατά την αξιολόγηση της αποτελεσματικότητας μιας μονάδας.
- Δεν απαιτείται ο προσδιορισμός της μορφής συσχέτισης μεταξύ των δεδομένων εισόδου/εξόδου.
- Οι μονάδες συγκρίνονται απευθείας με ένα σύνολο ανταγωνιστικών ομοειδών μονάδων.
- Τα δεδομένα εισόδου/εξόδου μπορεί να μετρώνται σε διαφορετικές μονάδες. Για παράδειγμα μια είσοδος μπορεί να αφορά την ποσότητα των διαθέσιμων υλικών πόρων, ενώ μια άλλη μπορεί να αναφέρεται σε διαθέσιμα κεφάλαια.

Τα ίδια, όμως, χαρακτηριστικά τα οποία καθιστούν την DEA ισχυρό εργαλείο, ταυτόχρονα μπορούν να δημιουργήσουν και προβλήματα κατά την εφαρμογή της μεθόδου. Ένας λοιπόν αναλυτής θα πρέπει να έχει υπόψη τους παρακάτω περιορισμούς κατά τη χρήση της DEA.

- Η DEA απαιτεί θετικά ορισμένες μεταβλητές εισόδου και εξόδου. Εάν μία μεταβλητή δεν είναι θετική, υπάρχουν δύο τρόποι για την αντιμετώπιση αυτής της κατάστασης. Μια πρώτη προσέγγιση είναι να προστεθεί ένας θετικός αριθμός στην

αρνητική τιμή έτσι, ώστε η τιμή της συγκεκριμένης μεταβλητής εισόδου ή εξόδου να γίνει θετική. Μια δεύτερη προσέγγιση είναι η αντικατάσταση της αρνητικής τιμής με μια πολύ μικρή θετική τιμή.

- Η αύξηση της τιμής της μεταβλητής μιας οποιαδήποτε εισόδου θα μπορούσε να συμβάλλει στην αύξηση των τιμών μερικών εξόδων, αλλά όχι στην μείωση μιας οποιαδήποτε εξόδου.
- Η DEA απαιτεί σχετική ομοιογένεια στο σύνολο των παραγωγικών μονάδων. Αυτό σημαίνει όλες οι μονάδες θα πρέπει έχουν τον ίδιο αριθμό εισόδων και εξόδων.
- Κατά ένα γενικό κανόνα, για κάθε έξοδο και είσοδο του μοντέλου θα πρέπει να χρησιμοποιούνται τρεις λειτουργικές μονάδες έτσι ώστε να διασφαλιστούν ικανοποιητικοί βαθμοί ελευθερίας. Εάν σε ένα σύνολο δεδομένων συμπεριλαμβάνονται λιγότερα DMU για κάθε μεταβλητή εισόδου και εξόδου, υπάρχει κίνδυνος ένας αρκετά DMU να θεωρηθούν αποτελεσματικά εξαιτίας του ότι υπάρχει ελλιπής αριθμός βαθμών ελευθερίας.

Γενικά, θεωρώντας τη DEA ως μια μέθοδο ακραίου σημείου, είναι πιθανό τα σφάλματα μέτρησης να προκαλέσουν σημαντικά προβλήματα. Είναι καλή μέθοδος για την εκτίμηση της σχετικής αποτελεσματικότητας μιας μονάδας αλλά συγκλίνει δύσκολα σε μια ακριβή εκτίμηση της απόλυτης αποτελεσματικότητας. Με άλλα λόγια, μπορεί να δώσει αξιόπιστη περιγραφή για το πόσο καλά τα καταφέρνει μια μονάδα σε σχέση με τους ανταγωνιστές της, αλλά δεν μπορεί να συγκρίνει την αποτελεσματικότητας της μονάδας αυτής όσον αφορά μια θεωρητικά μέγιστη απόδοση. Επειδή η DEA είναι μια μη παραμετρική τεχνική, η πραγματοποίηση εφαρμογών στατιστικών υποθέσεων είναι δύσκολη και σε αυτό τον τομέα έχει στραφεί το ερευνητικό ενδιαφέρον.

## 2.3 Περίληψη

Σε αυτό το κεφάλαιο έγινε μια γενική περιγραφή της μεθόδου DEA, γνωστής ως Περιβάλλουσα Ανάλυση Δεδομένων. Παρουσιάστηκαν διάφορες μαθηματικές διατυπώσεις της DEA, και πραγματοποιήθηκε μία σύγκριση της προσέγγισης αυτή με την μέθοδο της στατιστικής παλινδρόμησης. Επίσης, αναφέρθηκαν κάποια πλεονεκτήματα της μεθόδου και

παρουσιάστηκαν κάποιοι γενικότεροι περιορισμοί οι οποίοι απαιτούν προσοχή κατά τη χρήση της DEA. Στο επόμενο κεφάλαιο, γίνεται μια αναφορά στο μοντέλο DEA-DA, το οποίο χρησιμοποιείται αργότερα στην πειραματική ανάλυση, για την ανάπτυξη υποδειγμάτων ταξινόμησης.

## ΠΡΟΣΑΡΜΟΓΗ ΤΗΣ DEA ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ

### 3.1 Εισαγωγή

Στη βάση της προσέγγισης της DEA, η οποία αναλύθηκε στο προηγούμενο κεφάλαιο, πρόσφατα ξεκίνησαν προσπάθειες για την ανάπτυξη νέων μοντέλων της DEA με σκοπό την αντιμετώπιση του προβλήματος της ταξινόμησης, όπως αυτό αναλύθηκε στο πρώτο κεφάλαιο της παρούσας εργασίας. Ένα τέτοιο μοντέλο της DEA είναι και αυτό που προτάθηκε πρόσφατα από τον Sueyoshi (1999). Το συγκεκριμένο μοντέλο αναφέρεται ως Extended DEA-Discriminant Analysis (Extended DEA-DA) και σχεδιάστηκε έτσι, ώστε να προσδιορίζει την ύπαρξη ή μη επικάλυψης (overlap) μεταξύ δύο κατηγοριών και να ταξινομεί ένα καινούργιο δείγμα παρατηρήσεων σε δύο προκαθορισμένες κατηγορίες. Παράλληλα γίνεται μια λεπτομερή ανάλυση αυτής της προσέγγισης και της προσαρμογής της στην ταξινόμηση.

### 3.2 Τεχνικές προγραμματισμού στόχων για την ανάπτυξη μοντέλων ταξινόμησης

Όπως αναφέρθηκε στο εισαγωγικό κεφάλαιο, η ταξινόμηση αναφέρεται στην τοποθέτηση ενός συνόλου μονάδων ή παρατηρήσεων σε προκαθορισμένες κατηγορίες. Οι ίδιοι σχετικοί παράγοντες εκτιμώνται για όλες τις μονάδες ή τις παρατηρήσεις και βάσει ενός ξεχωριστού αποτελέσματος αποφασίζεται σε ποια κατηγορία ανήκει η κάθε μονάδα.



Τις τελευταίες δύο δεκαετίες ένα σημαντικό τμήμα της επιστημονικής έρευνας έχει επικεντρωθεί στην ανάπτυξη μοντέλων ταξινόμησης χρησιμοποιώντας τεχνικές προγραμματισμού στόχων (goal programming, GP). Τα μοντέλα ταξινόμησης που μπορούν να αναπτυχθούν δύνανται να έχουν διάφορες μορφές. Στην απλούστερη περίπτωση χρησιμοποιούνται γραμμικά μοντέλα στα οποία οι παράγοντες που περιγράφουν τις παρατηρήσεις συνδυάζονται γραμμικά επιλέγοντας κατάλληλους αριθμητικούς συντελεστές. Στόχος στην περίπτωση αυτή είναι η επιλογή των κατάλληλων αριθμητικών συντελεστών ώστε να επιτευχθεί η ταξινόμηση των μονάδων/παρατηρήσεων με τον πλέον ακριβή τρόπο. Στη συνέχεια το μοντέλο που αναπτύσσεται μπορεί να χρησιμοποιηθεί για να προβλέψει σε ποια κατηγορία ανήκει ένα νέο σύνολο από μονάδες.

Γενικότερα, η ταξινόμηση μπορεί να αφορά δύο ή περισσότερες κατηγορίες, με τους ερευνητές να επικεντρώνουν το ενδιαφέρον τους στην περίπτωση των δύο κατηγοριών. Στην συγκεκριμένη περίπτωση, ανάλυση επικεντρώνεται αποκλειστικά στην χρήση δύο κατηγοριών, καθώς αυτό το σενάριο είναι σχετικό με την DEA. Θεωρείται ότι υπάρχουν  $n$  μονάδες προς ταξινόμηση ( $j=1, \dots, n$ ) και  $k$  ανεξάρτητοι παράγοντες ή μεταβλητές ( $i=1, \dots, k$ ) οι οποίοι έχουν καταγραφεί για κάθε μονάδα και συμβολίζονται με  $Z_{ij}$ . Είναι δεδομένο ότι  $n_1$  μονάδες ανήκουν στην κατηγορία 1,  $n_2$  μονάδες ανήκουν στην κατηγορία 2 και  $n_1 + n_2 = n$ . Στόχος της ανάλυσης είναι να καθοριστεί ένα σύνολο από βάρη  $\{W_i\}$  και μια οριακή τιμή  $\{T\}$ , βάσει των οποίων θα καθοριστεί ένα υπερεπίπεδο το οποίο θα διαχωρίζει τις δύο κατηγορίες ως ακολούθως:

$$\left. \begin{array}{l} \sum_{i=1}^k W_i Z_{ij} \leq T, \quad j \in G_1 \\ \sum_{i=1}^k W_i Z_{ij} > T, \quad j \in G_2 \\ W_i \quad \text{για} \quad i=1, \dots, k \\ T \end{array} \right\} \text{ελεύθερα προσήμου} \quad (3.1)$$

Για τον προσδιορισμό των βέλτιστων παραμέτρων του μοντέλου απαιτείται ο καθορισμός του κατάλληλου κριτηρίου το οποίο θα αναπαριστά την ποιότητα του μοντέλου. Για το σκοπό αυτό έχουν προταθεί διάφορα κριτήρια όπως:

- (α) Ελαχιστοποίηση της μέγιστης εξωτερικής απόκλισης, όπου ως εξωτερική απόκλιση θεωρείται η παραβίαση των κανόνων (3.1) που οδηγεί στην εσφαλμένη ταξινόμηση μιας μονάδας.
- (β) Μεγιστοποίηση της ελάχιστης εσωτερικής απόκλισης, όπου ως εσωτερική απόκλιση εννοείται ο βαθμός στον οποίο ικανοποιούνται οι κανόνες ταξινόμησης (3.1) από κάποια μονάδα.
- (γ) Ελαχιστοποίηση των συνολικών εξωτερικών αποκλίσεων.
- (δ) Μεγιστοποίηση των συνολικών εσωτερικών αποκλίσεων.

Μία εξωτερική απόκλιση ορίζεται με την μεταβλητή  $d_i$  και είναι η απόσταση μιας εσφαλμένα ταξινομημένης μονάδας που τοποθετείται έξω από την κατηγορία στην οποία ανήκει. Μία εσωτερική απόκλιση ορίζεται με την μεταβλητή  $B_j$  και είναι η απόσταση μιας σωστά ταξινομημένης μονάδας που τοποθετείται στην κατηγορία στην οποία ανήκει. Στην ουσία, για κάθε μονάδα υπολογίζεται ένα σκορ το οποίο είναι μεγαλύτερο, μικρότερο ή ίσο της βέλτιστης οριακής τιμής  $T^*$  και το οποίο καθορίζει την κατηγορία στην οποία εντάσσεται η συγκεκριμένη μονάδα. Ένα νέο σύνολο μονάδων μπορεί να ταξινομηθεί με βάση τις προσδιοριζόμενες βέλτιστες τιμές  $(W^*, T^*)$ .

Βάσει των κανόνων ταξινόμησης (3.1) ο προσδιορισμός των βέλτιστων παραμέτρων  $(W^*, T^*)$  του μοντέλου μπορεί να πραγματοποιηθεί διαμορφώνοντας ένας πρόβλημα γραμμικού προγραμματισμού. Η προσέγγιση αυτή έχει εξεταστεί από διάφορους ερευνητές (Freed και Glover (1981), Freed και Glover (1986), Glover et al. (1988), Koehler (1989), Glover (1990), κ.α.). Μια γενική διατύπωση του προβλήματος προτάθηκε από τον Glover (1990) είναι η ακόλουθη:

$$\left. \begin{aligned}
& \text{Min} \left( h_o d_o + \sum_{j=1}^n h_j d_j - k_o B_o - \sum_{j=1}^n k_j B_j \right) \\
& \text{υπό:} \\
& \sum_{i=1}^k W_i Z_{ij} - d_o - d_j + B_o + B_j = T, j \in G_1 \\
& \sum_{i=1}^k W_i Z_{ij} + d_o + d_j - B_o - B_j = T, j \in G_2 \\
& 2n_1 n_2 (B_o - d_o) + n_2 \sum_{j \in G_1} (B_j - d_j) + n_1 \sum_{j \in G_2} (B_j - d_j) = c \\
& d_o, B_o, d_j, B_i \geq 0, j = 1, \dots, n \\
& \left. \begin{array}{l} W_i, i = 1, \dots, k \\ T \end{array} \right\} \text{ελεύθερα προσήμου}
\end{aligned} \right\} \quad (3.2)$$

Οι μεταβλητές  $d_o$  και  $B_o$  αναπαριστούν αντίστοιχα τη μέγιστη εξωτερική και εσωτερική απόκλιση. Στην αντικειμενική συνάρτηση, οι σταθερές  $h_o, h_j, k_o, k_j$  επιλέγονται έτσι, ώστε να μεγιστοποιούνται οι εσωτερικές αποκλίσεις και να ελαχιστοποιούνται οι εξωτερικές αποκλίσεις. Ένα πλεονέκτημα αυτής της μεθόδου είναι η ικανότητα να προσδιορίζει τις στατιστικές αποκλίσεις των διαφόρων μονάδων με διαφορετικό τρόπο, όπως επίσης, και να επιτρέπει τον προσδιορισμό των εξωτερικών και εσωτερικών αποκλίσεων εάν αυτό είναι επιθυμητό. Ένα άλλο πλεονέκτημα αυτής της μεθόδου είναι να συνδυάζει και τα τέσσερα προαναφερθέντα κριτήρια για την αξιολόγηση της ποιότητας του μοντέλου που αναπτύσσεται. Ακριβώς όμως λόγω της χρησιμοποίησης και των τεσσάρων αυτών κριτηρίων, το πρόβλημα αποκτά αυξημένη πολυπλοκότητα, ενώ επιπλέον παρουσιάζεται δυσκολία στον προσδιορισμό των σχετικών συντελεστών βαρύτητας των κριτηρίων (σταθερές  $h_o, h_j, k_o, k_j$ ).

Μία απλούστερη προσέγγιση, η οποία και προτάθηκε από τους Ragsdale και Stam (1991), Stam και Ragsdale (1992) και Glover et al. (1988), περιλαμβάνει την χρήση ενός μικρού αριθμού  $e$  για να την εκμετάλλευση του διαχωριστικού περιθωρίου μεταξύ των κατηγοριών. Θεωρώντας την ελαχιστοποίηση των συνολικών εξωτερικών αποκλίσεων το πρόβλημα που διαμορφώνεται έχει την ακόλουθη μορφή:

$$\left. \begin{array}{l}
 \text{Min} \left( \sum_{j=1}^n d_j \right) \\
 \text{υπό:} \\
 \sum_{i=1}^k W_i Z_{ij} - d_j \leq T - e, j \in G_1 \\
 \sum_{i=1}^k W_i Z_{ij} + d_j \geq T, j \in G_2 \\
 d_j \geq 0, j = 1, \dots, n \\
 \left. \begin{array}{l} W_i, i = 1, \dots, k \\ T \end{array} \right\} \text{ελεύθερα προσήμου}
 \end{array} \right\} (3.3)$$

Γενικά, τις τελευταίες δύο δεκαετίες έχουν παρουσιαστεί διάφορες τεχνικές προγραμματισμού στόχων για την ανάπτυξη μοντέλων ταξινόμησης, οι οποίες παρουσιάζουν τα δικά τους πλεονεκτήματα ή μειονεκτήματα. Αναλυτική καταγραφή των μέχρι σήμερα ερευνών στο χώρο αυτό δίνεται από τον Stam (1997). [Stam, A. (1997), “Nontraditional approaches to statistical classification: Some perspectives on  $L_p$ -norm methods”, *Annals of Operations Research*, 74, 1–36.]

### 3.3 Μοντέλα της DEA στη διακριτική ανάλυση

Ένα βασικό χαρακτηριστικό των μέχρι σήμερα ερευνών σχετικά με τη χρησιμοποίηση τεχνικών προγραμματισμού στόχων για την ανάπτυξη μοντέλων ταξινόμησης είναι η χρησιμοποίηση γραμμικών διακριτικών συναρτήσεων, κυρίως λόγω του περιορισμένου υπολογιστικού φόρτου που απαιτεί η προσέγγιση αυτή. Σε πολλές περιπτώσεις όμως η χρήση μη γραμμικών μοντέλων παρέχει πιο ακριβή αποτελέσματα ταξινόμησης και υψηλότερη δυνατότητα γενίκευσης. Εντούτοις, η ανάπτυξη μη γραμμικών μοντέλων συχνά απαιτεί σημαντικό υπολογιστικό φόρτο στοιχείο που περιορίζει την εφαρμογή αυτής της εναλλακτικής προσέγγισης.

Για να αντιμετωπιστεί η παραπάνω δυσκολία προτάθηκε από τους Sueyoshi και Kirihaara (1998) και Sueyoshi (1999) μια νέα μοντελοποίηση του προβλήματος κατά την οποία προσδιορίζεται ένα σύνολο βαρών από κατά τμήματα γραμμικές διακριτικές συναρτήσεις οι οποίες αποδίδουν ένα εκτιμώμενο σκορ για τον προσδιορισμό των κατηγοριών ταξινόμησης. Αυτή η προσέγγιση βασίζεται στην προσαρμογή της DEA στο πρόβλημα της ταξινόμησης και αναφέρεται ως “Περιβάλλουσα Ανάλυση Δεδομένων-Διακριτική Ανάλυση” (DEA-DA),

καθώς πιστοποιούνται οι διακριτικές της ικανότητες, ενσωματώνοντας τη μη παραμετρική μορφή της DEA στα ήδη υπάρχοντα μοντέλα προγραμματισμού στόχων. Γενικά, διαπιστώθηκε ότι η κατά τμήματα διακριτική συνάρτηση είναι πιο ευέλικτη από ότι οι συμβατικές γραμμικές διακριτικές συναρτήσεις όσον αφορά της διακριτικές τους δυνατότητες.

Δυστυχώς, η προσέγγιση DEA-DA παρουσιάζει δύο βασικές αδυναμίες όσον αφορά την εφαρμογή της:

- (α) Δεν μπορεί να χρησιμοποιηθεί για ένα αρνητικό σύνολο δεδομένων.
- (β) Παράγει δύο διαχωριστικές διακριτικές συναρτήσεις, μία για κάθε κατηγορία.

Το πρώτο μειονέκτημα συντελεί αρνητικά στην εφαρμογή της μεθόδου σε διάφορα πρακτικά πεδία όπου οι εξεταζόμενοι παράγοντες μπορούν να έχουν και αρνητικές τιμές (για παράδειγμα, χρηματοοικονομικά στοιχεία). Το δεύτερο μειονέκτημα παρουσιάζει δυσκολία στην εφαρμογή των δύο προσδιοριζόμενων διακριτικών συναρτήσεων. Προκειμένου βέβαια να αντιμετωπιστούν όλες αυτές οι δυσκολίες προτάθηκε από τους Sueyoshi και Kirihiara (1998) και Sueyoshi (1999) μια εναλλακτική μορφή μη παραμετρικής προσέγγισης, αναφερόμενη ως “Extended DEA-DA”. Η νέα αυτή τεχνική σχεδιάστηκε για την κατά τμήματα γραμμική ταξινόμηση των μονάδων χρησιμοποιώντας μια κατάλληλη διακριτική συνάρτηση. Η νέα αυτή προσέγγιση μπορεί να χρησιμοποιηθεί και για αρνητικές τιμές δεδομένων.

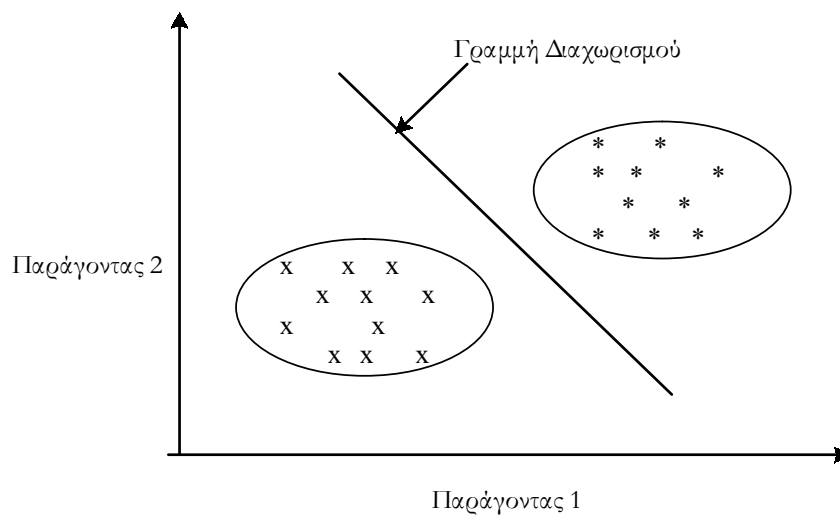
### 3.3.1 Γενική θεώρηση του προβλήματος

Για την κατανόηση της αναλυτικής δομής του εκτεταμένου μοντέλου της DEA-DA είναι απαραίτητη η σύγκριση αυτού με άλλες μεθόδους της διακριτικής ανάλυσης. Ο καλύτερος τρόπος για την πραγματοποίηση αυτής της σύγκρισης είναι η κατασκευή ενός εικονικού και γενικού πλαισίου εργασίας, το οποίο και θα περιγράφει το πως οι διάφορες προσεγγίσεις σχετίζονται μεταξύ τους, χρησιμοποιώντας την γενική δομή της διακριτικής ανάλυσης (DA).

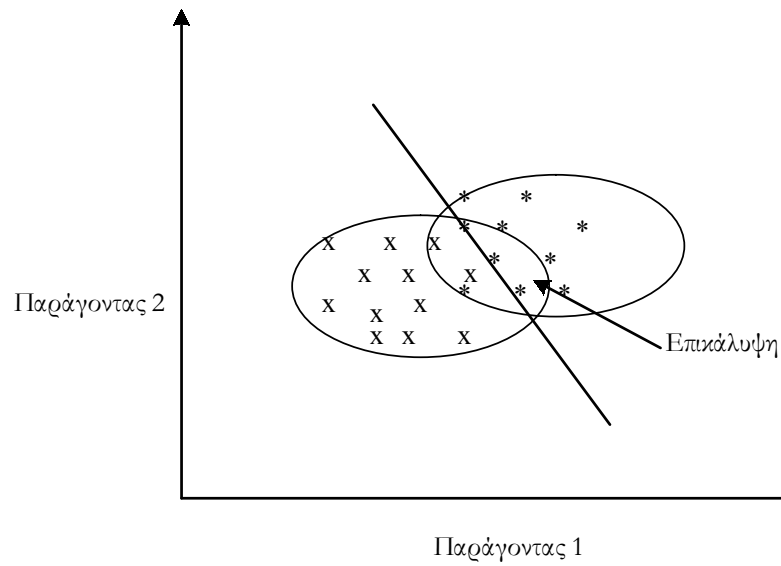
Στο πρόβλημα της ταξινόμησης, θεωρείται ότι υπάρχουν διαθέσιμα στοιχεία για  $n$  παρατηρήσεις κάθε μία από τις οποίες περιγράφεται από  $k$  παράγοντες ( $i = 1, \dots, k$ ). Η περιγραφή της παρατήρησης  $j$  βάσει του παράγοντα  $i$  συμβολίζεται ως  $Z_{ij}$ . Στην πιο διαδεδομένη μορφή του προβλήματος της ταξινόμησης, θεωρείται ότι κάθε μία από τις

παρατηρήσεις ταξινομείται σε μία εκ των δύο κατηγοριών που εξετάζονται στην ανάλυση ( $G_1$  ή  $G_2$ ). Οι προκαθορισμένες δύο κατηγορίες περιέχουν  $n_1$  και  $n_2$  παρατηρήσεις έτσι, ώστε  $n = n_1 + n_2$  και  $G = G_1 \cup G_2$ .

Με βάση αυτές τις υποθέσεις, το σχήμα 3.1 αναπαριστά το πώς ένα υπερεπίπεδο διαχωρίζει τις δύο κατηγορίες ( $G_1$  και  $G_2$ ) στην απλούστερη περίπτωση των δύο παραγόντων. Όπως φαίνεται στο σχήμα μία γραμμή διαχωρισμού τοποθετείται μεταξύ των δύο κατηγοριών έτσι, ώστε να διακρίνονται καθαρά τα δύο σύνολα ταξινόμησης, στην περίπτωση όπου δεν υπάρχει επικάλυψη (overlap) των δύο κατηγοριών. Το σχήμα 3.2 αναπαριστά τη γενικότερη περίπτωση κατά την οποία παρατηρείται επικάλυψη των δύο κατηγοριών με αποτέλεσμα να μην είναι δυνατός ο απόλυτος διαχωρισμός των παρατηρήσεων στις προκαθορισμένες κατηγορίες.



Σχήμα 3.1: Μία οπτική δομή της διακριτικής ανάλυσης

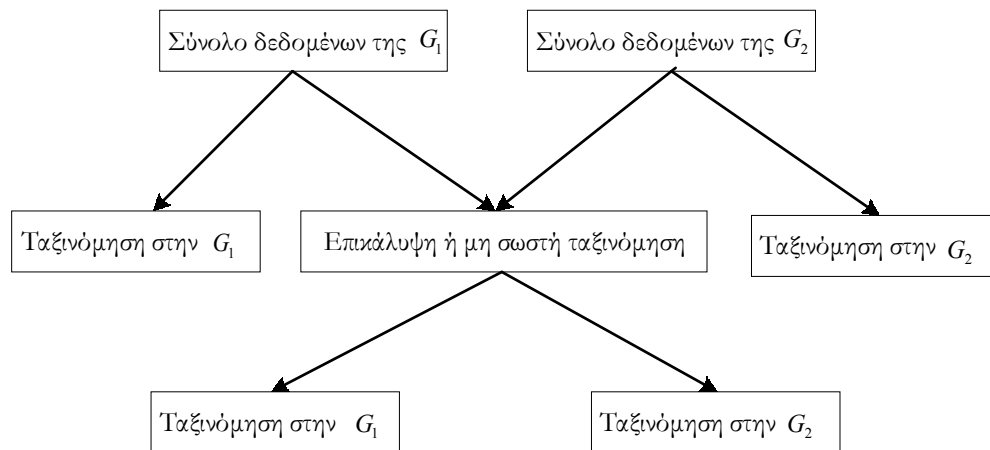


Σχήμα 3.2: Το πρόβλημα της ταξινόμησης στην περίπτωση επικάλυψης των κατηγοριών

Μια προσέγγιση για την αντιμετώπιση φαινομένων επικάλυψης των κατηγοριών η οποία προτάθηκε από τον Sueyoshi (2001) είναι η πραγματοποίηση της ανάλυσης σε δύο στάδια ως εξής:

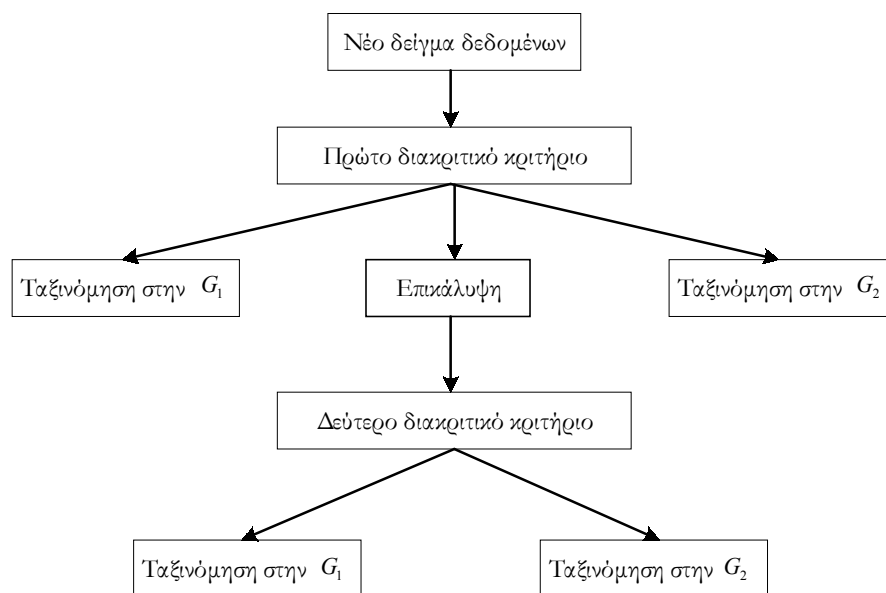
- (α) Στάδιο 1<sup>ο</sup>: Η ταξινόμηση του συνολικού συνόλου των παρατηρήσεων στις κατηγορίες  $G_1$ ,  $G_2$  ή  $(G_1 \cap G_2 = \text{επικάλυψη})$ .
- (β) Στάδιο 2<sup>ο</sup>: Την επαναταξινόμηση των μη σωστά ταξινομημένων παρατηρήσεων  $(G_1 \cap G_2)$  σε μία από τις κατηγορίες  $G_1, G_2$ .

Το σχήμα 3.3 αναπαριστά ένα διάγραμμα ροής της παραπάνω διαδικασίας ταξινόμησης, όπου κάθε στάδιο παράγει τα δικά του κριτήρια ταξινόμησης. Το σχήμα 3.4 περιγράφει τη διαδικασία που χρησιμοποιείται για την ταξινόμηση νέων παρατηρήσεων στις δύο προκαθορισμένες κατηγορίες.



---

Σχήμα 3.3: Η διαδικασία ταξινόμησης της διακριτικής ανάλυσης



---

Σχήμα 3.4: Διαδικασία ταξινόμησης ενός νέου δείγματος δεδομένων



Εν συνεχεία, θα εξεταστούν οι προαναφερθείσες διαδικασίες ταξινόμησης σε σχέση με τις δυνατότητες της διακριτικής ανάλυσης:

- (α) *Μη γραμμική διακριτική συνάρτηση*: Όπως προαναφέρθηκε, τα σχήματα 3.1 και 3.2 αναπαριστούν μία διαχωριστική γραμμή, η οποία εκφράζεται από μία γραμμική διακριτική συνάρτηση. Γενικότερα, μία μη γραμμική συνάρτηση θεωρείται πιο ευέλικτη, όσον αφορά την ταξινόμηση, σε σχέση με μία γραμμική συνάρτηση. Επομένως, η διακριτική ικανότητα μιας μη γραμμικής συνάρτησης υπερτερεί έναντι μιας γραμμικής. Εντούτοις, η χρήση μιας μη γραμμικής διακριτικής συνάρτησης απαιτεί λεπτομερή προσδιορισμό των όρων της, γεγονός που την καθιστά δύσκολη στην χρήση. Για την αντιμετώπιση της παραπάνω δυσκολίας είναι απαραίτητη η ανάπτυξη μιας αναλυτικής προσέγγισης.
- (β) *Διαδικασία προσέγγισης πολλαπλών σταδίων*: Η διακριτική ικανότητα εξαρτάται από τον τρόπο με τον οποίο θα αντιμετωπιστεί η επικάλυψη των κατηγοριών. Στα σχήματα 3.3 και 3.4, χρησιμοποιούνται πολλαπλά στάδια για την αντιμετώπιση αυτού του προβλήματος. Βέβαια, στην περίπτωση όπου οι κατηγορίες δεν επικαλύπτονται, τότε το πρώτο στάδιο της ανάλυσης είναι αρκετό για την ταξινόμηση των παρατηρήσεων (Σχήμα 3.1).

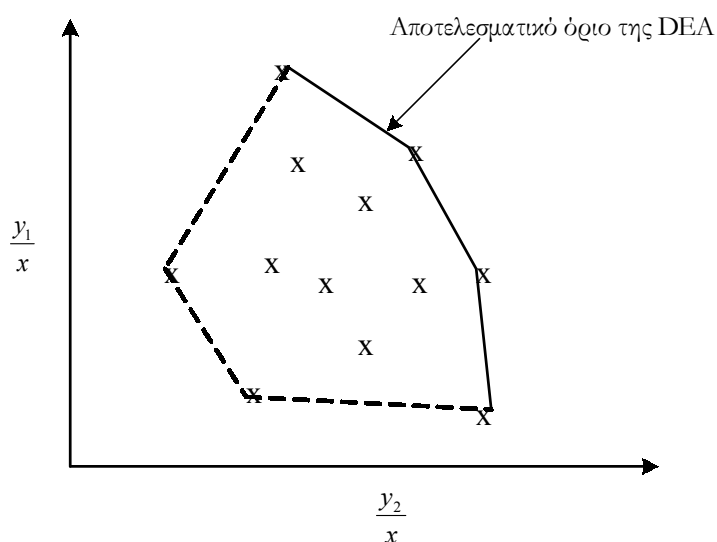
### 3.3.2 Ομοιότητες και διαφορές μεταξύ των μοντέλων DEA και DA

Προτού γίνει η μαθηματική διατύπωση του μοντέλου DEA-DA για την ανάπτυξη υποδειγμάτων ταξινόμησης, είναι απαραίτητο να παρουσιαστούν κάποιες ομοιότητες και διαφορές που εντοπίζονται ανάμεσα στο μοντέλο της DEA και στα μοντέλα προγραμματισμού στόχων για την αντιμετώπιση προβλημάτων ταξινόμησης (DA). Στα παρακάτω σχήματα 3.1 και 3.2 αναπαριστάται η δομή των μοντέλων DEA και DA, αντίστοιχα.

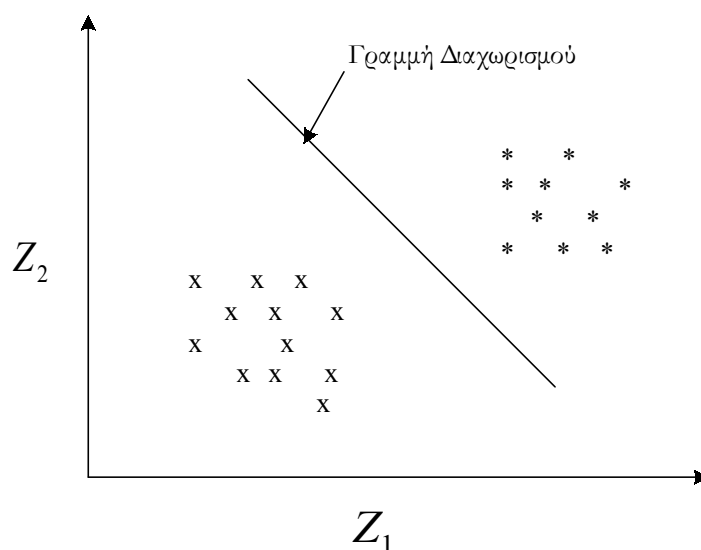
Στο σχήμα 3.1 απεικονίζεται η περίπτωση της DEA θεωρώντας δύο εξόδους ( $y_1, y_2$ ) και μία είσοδο ( $x$ ), όπου ο οριζόντιος άξονας συμβολίζει τον λόγο  $y_1/x$  και ο κάθετος άξονας συμβολίζει τον λόγο  $y_2/x$ . Στο επίπεδο των δύο αυτών αξόνων αναπαριστώνται οι εκτιμώμενες μονάδες. Ο αλγόριθμος του μοντέλου της DEA προσδιορίζει ένα αποτελεσματικό όριο και εκτιμά την αποτελεσματικότητα της κάθε μονάδας (DMU) συγκρίνοντάς την με το αποτελεσματικό όριο. Το αποτελεσματικό όριο της DEA είναι ένα

κατά τμήματα γραμμικό περίγραμμα, που καλύπτει τα DMU. Γενικά, η DEA είναι εύκολη στην χρήση της, καθώς δεν απαιτεί τον προσδιορισμό των συναρτήσεων μεταξύ των μεταβλητών εισόδου και εξόδου.

Το σχήμα 3.2 δίνει μια οπτική απεικόνιση της σχέσης που υπάρχει μεταξύ μιας διακριτικής συνάρτησης και των παρατηρήσεων (μονάδων). Η διακριτική συνάρτηση προσδιορίζεται ελαχιστοποιώντας το άθροισμα των αποκλίσεων των DMU. Η διαχωριστική γραμμή τοποθετείται μεταξύ των δύο κατηγοριών των παρατηρήσεων και είναι γραμμικής μορφής. Στην περίπτωση που υπήρχαν περισσότεροι από δύο παράγοντες τότε η διαχωριστική γραμμή θα ήταν της μορφής γραμμικού υπερεπιπέδου.



Σχήμα 3.5: Περιβάλλουσα Ανάλυση Δεδομένων (DEA)



Σχήμα 3.6: Διακριτική Ανάλυση (DA)

Και οι δύο προσεγγίσεις αποδίδουν ένα σκορ σε κάθε εξεταζόμενη μονάδα, το οποίο προσδιορίζεται βάσει του σταθμισμένου αθροίσματος των παραγόντων και σχετίζεται με κάποια οριακή τιμή. Το κάθε σκορ προσδιορίζει ένα υπερεπίπεδο προκειμένου να καθορίσει τις κατηγορίες των αποτελεσματικών και μη πληροφοριών. Στην μέθοδο DEA κάθε μονάδα επιλέγει και το δικό της υπερεπίπεδο. Εντούτοις, και στις δυο μεθόδους το εκτιμώμενο σταθμισμένο σύνολο των παραγόντων σχετίζεται αποκλειστικά με τις συγκεκριμένες μονάδες του συνόλου. Στην προσέγγιση της DA, κάθε νέα μονάδα εκτιμάται βάσει του συνόλου των βαρών που προσδιορίστηκαν από τις προηγούμενες μονάδες. Μία άλλη ομοιότητα των δύο προσεγγίσεων είναι ότι και οι δύο μπορεί να παρουσιάσουν εκφυλισμένες λύσεις λόγω του μεγάλου αριθμού των περιορισμών (ένα για κάθε μονάδα) όσον αφορά τον αριθμό των DMU (ένα για κάθε παράγοντα). Τέλος, όπως προαναφέρθηκε, και οι δύο προσεγγίσεις, DEA και DA, βασίζονται στη χρήση τεχνικών προγραμματισμού στόχου.

Η DA χρησιμοποιεί δύο εκ των προτέρων κατηγορίες ταξινόμησης, σε αντίθεση με την DEA στην οποία οι δύο κατηγορίες προκύπτουν ως αποτέλεσμα της ανάλυσης. Επιπλέον, εντός της κάθε κατηγορίας η DA αποδίδει διαφορετική βαθμολογία σε κάθε μονάδα, ενώ αντίθετα στη DEA όλες οι αποτελεσματικές μονάδες έχουν την ίδια βαθμολογία η οποία είναι ίση με τη μονάδα.

Ενώ και οι δύο μέθοδοι ταξινομούν τις παρατηρήσεις σε δύο κατηγορίες, η κύρια διαφορά τους είναι εάν οι κατηγορίες είναι γνωστές εκ των προτέρων. Στην DA η ταξινόμηση των παραγωγικών μονάδων βασίζεται σε ένα ευκρινές αποτέλεσμα, το οποίο είναι αποκλειστικά υπεύθυνο για την ταξινόμηση των μονάδων σε μια από τις δύο κατηγορίες. Στην DEA δεν είναι γνωστό σε ποια κατηγορία ανήκει η κάθε μονάδα. Βασικός στόχος αυτής της μεθόδου είναι η ταξινόμηση των μονάδων διαμέσου σχετικών συγκρίσεων με άλλες μονάδες.

### 3.3.3 Μαθηματική διατύπωση του μοντέλου DEA-DA

Βάσει του παραπάνω μεθοδολογικού πλαισίου ο Sueyoshi (2001) πρότεινε ένα μοντέλο γραμμικού προγραμματισμού για την ανάπτυξη υποδειγμάτων ταξινόμησης μέσω της DEA (μοντέλο DEA-DA). Αναλυτικότερα, η ανάλυση των δύο σταδίων που παρουσιάστηκε στην προηγούμενη ενότητα υλοποιείται στο μοντέλο DEA-DA ως εξής:

*1<sup>ο</sup> Στάδιο: Ταξινόμηση και Προσδιορισμός Εσφαλμένων Ταξινόμήσεων*

$$\left. \begin{array}{l}
 \text{Min} \quad \sum_{j \in G_1} S_{1j}^+ + \sum_{j \in G_2} S_{2j}^- \\
 \text{υπό:} \\
 \sum_{i=1}^k \lambda_i Z_{ij} + S_{1j}^+ - S_{1j}^- = d + 1, \quad j \in G_1 \\
 \sum_{i=1}^k \lambda_i Z_{ij} + S_{2j}^+ - S_{2j}^- = d, \quad j \in G_2 \\
 \sum_{i=1}^k |\lambda_i| = 1, \\
 S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \geq 0, \quad \forall j \\
 \lambda_i, d \text{ ελεύθερα προσήμου}
 \end{array} \right\} \quad (3.4)$$

Στόχος του προβλήματος αυτού είναι ανάπτυξη μιας κατάλληλης διακριτικής συνάρτησης  $f = \sum_j \lambda_i Z_{ij}$  για το διαχωρισμό των δύο κατηγοριών  $G_1$  και  $G_2$ . Οι μεταβλητές  $S_{1j}^+$  και  $S_{1j}^-$  είναι θετικές και αρνητικές αποκλίσεις των σκορ διάκρισης των παρατηρήσεων της πρώτης κατηγορίας  $G_1$  από το σημείο διαχωρισμού  $d+1$ . Αντίστοιχα, οι μεταβλητές  $S_{2j}^+$  και  $S_{2j}^-$  είναι θετικές και αρνητικές αποκλίσεις των σκορ διάκρισης των παρατηρήσεων της δεύτερης κατηγορίας  $G_2$  από το σημείο διαχωρισμού  $d$ . Οι

μεταβλητές  $S_{1j}^+ (j \in G_1)$  και  $S_{2j}^- (j \in G_2)$  δηλώνουν την ύπαρξη εσφαλμένων ταξινομήσεων, ενώ οι εναπομένουσες μεταβλητές αναφέρονται στη σωστή ταξινόμηση. Όλοι οι παράγοντες  $Z_{ij}$  συνδυάζονται στην αναπτυσσόμενη διακριτική συνάρτηση με κάποιους αριθμητικούς συντελεστές  $\lambda_i$ , οι οποίοι υποδηλώνουν τη σημαντικότητα του κάθε παράγοντα στην ταξινόμηση των παρατηρήσεων. Το άθροισμα των απόλυτων τιμών των  $\lambda_i, i=1, \dots, k$  είναι ίσο με την μονάδα.

Εφόσον, ο γραμμικός προγραμματισμός δεν μπορεί απευθείας να επεξεργαστεί μια απόλυτη τιμή, θα πρέπει να ανασχηματιστεί το παραπάνω γραμμικό πρόβλημα εισάγοντας τις παρακάτω μεταβλητές:

$$\lambda_i^+ = \frac{(|\lambda_i| + \lambda_i)}{2} \quad \text{και} \quad \lambda_i^- = \frac{(|\lambda_i| - \lambda_i)}{2} \quad \text{για} \quad i=1, \dots, k \quad (3.5)$$

Δεδομένου ότι:  $|\lambda_i| = \lambda_i^+ + \lambda_i^-$  και  $\lambda_i = \lambda_i^+ - \lambda_i^-$ .

Ο παραπάνω ορισμός για τις μεταβλητές  $\lambda_i^+$  και  $\lambda_i^-$  δηλώνει ότι η μη γραμμική συνθήκη  $(\lambda_i^+ \lambda_i^- = 0)$  πάντα ικανοποιείται από τις εξισώσεις (3.5), καθώς

$$\lambda_i^+ \lambda_i^- = \frac{(|\lambda_i|^2 - \lambda_i^2)}{4} = 0.$$

Εισάγοντας τις μεταβλητές  $\lambda_i^+$  και  $\lambda_i^-$  στο γραμμικό πρόβλημα (3.4) προκύπτει το παρακάτω:

$$\left. \begin{aligned}
 & \text{Min} \quad \sum_{j \in G_1} S_{1j}^+ + \sum_{j \in G_2} S_{2j}^- \\
 & \text{υπό:} \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} + S_{1j}^+ - S_{1j}^- = d + 1, \quad j \in G_1 \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} + S_{2j}^+ - S_{2j}^- = d, \quad j \in G_2 \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) = 1, \\
 & \quad S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \geq 0, \quad \forall j \\
 & \quad \lambda_i^+ \geq 0 \quad \& \quad \lambda_i^- \geq 0, \quad \forall i \\
 & \quad d \text{ ελεύθερο προσήμου}
 \end{aligned} \right\} \quad (3.6)$$

Από την επίλυση του παραπάνω γραμμικού προγράμματος προσδιορίζονται οι βέλτιστες τιμές των συντελεστών  $\lambda_i^* (= \lambda_i^{+*} - \lambda_i^{-*})$ , καθώς και η βέλτιστη τιμή του σημείου διαχωρισμού  $d^*$ . Βάσει αυτής της λύσης οι παρατηρήσεις μπορούν να ταξινομηθούν ως εξής::

- (α) Εάν ισχύει:  $\sum_{i=1}^k \lambda_i^* Z_{im} \geq d^* + 1$ , τότε η παρατήρηση  $m$  ανήκει στην ομάδα  $G_1$ .
- (β) Εάν ισχύει:  $d^* < \sum_{i=1}^k \lambda_i^* Z_{ij} < d^* + 1$ , τότε δεν μπορεί να προσδιοριστεί σε ποια ομάδα ανήκει η παρατήρηση  $m$ . Θεωρείται ότι ανήκει και στο κοινό υποσύνολο των δύο ομάδων  $(G_1 \cap G_2)$ .
- (γ) Εάν ισχύει:  $\sum_{i=1}^k \lambda_i^* Z_{im} \leq d^*$ , τότε η παρατήρηση  $m$  ανήκει στην ομάδα  $G_2$ .

Με τη χρήση των παραπάνω κανόνων οι παρατηρήσεις ταξινομούνται στα ακόλουθα υποσύνολα:

$$1. \quad R_1 = \left\{ j \in G \mid \sum_{i=1}^k \lambda_i^* Z_{ij} \geq d^* + 1 \right\},$$

$$2. \quad R_0 = \left\{ j \in G \mid d^* < \sum_{i=1}^k \lambda_i^* Z_{ij} < d^* + 1 \right\},$$

$$3. \quad R_2 = \left\{ j \in G \mid \sum_{i=1}^k \lambda_i^* Z_{ij} \leq d^* \right\},$$

$$4. \quad C_1 = \{ j \in R_1 \mid j \in G_1 \},$$

$$5. \quad C_2 = \{ j \in R_2 \mid j \in G_2 \},$$

Το σύνολο  $G$  χωρίζεται σε δύο υποσύνολα  $G_1$  &  $G_2$ , όπου  $G = G_1 \cup G_2 = R_1 \cup R_0 \cup R_2$ . Το υποσύνολο  $C_1$  αναφέρεται σε ένα σύνολο από παρατηρήσεις που σωστά ταξινομήθηκαν στην ομάδα  $G_1$ . Ομοίως, το υποσύνολο  $C_2$  αναφέρεται σε εκείνες τις παρατηρήσεις που σωστά ταξινομήθηκαν στην ομάδα  $G_2$ . Επιπλέον, η ύπαρξη των εσφαλμένων ταξινόμήσεων προσδιορίζεται με τα υποσύνολα  $G_1'$  &  $G_2'$ , όπου το πρώτο υποσύνολο ( $G_1' = G_1 \cap (R_0 \cup R_2)$ ) αναφέρεται σε εκείνο το σύνολο των παρατηρήσεων που ανήκουν στην ομάδα  $G_1$ , αλλά ταξινομήθηκαν εσφαλμένα στην ομάδα  $G_2$  και αντιστρόφως, το δεύτερο υποσύνολο ( $G_2' = G_2 \cap (R_0 \cup R_1)$ ) αναφέρεται σε εκείνο το σύνολο των παρατηρήσεων που ανήκουν στην ομάδα  $G_2$ , αλλά ταξινομήθηκαν εσφαλμένα στην ομάδα  $G_1$ .

Εφόσον παρατηρήθηκε η ύπαρξη λανθασμένων ταξινόμήσεων, ενσωματώνεται ένα νέο κατώφλι  $\epsilon$  στο μοντέλο (3.6), το οποίο χρησιμοποιείται για την εκ νέου ταξινόμηση των δύο υποομάδων ( $G_1' \cup G_2'$ ). Η μαθηματική διατύπωση του δεύτερου σταδίου της διαδικασίας ταξινόμησης διαμορφώνεται ως εξής:

2<sup>ο</sup> Στάδιο: Διαδικασία Επαναπροσδιορισμού των Εσραλμένων Ταξινομήσεων

$$\begin{aligned}
 & \text{Min} \quad \sum_{j \in G_1} S_{1j}^+ + \sum_{j \in G_2} S_{2j}^- \\
 & \text{υπό:} \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} \geq d + 1, \quad j \in C_1 \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} + S_{1j}^+ - S_{1j}^- = c, \quad j \in G_1' = G_1 \cap (R_0 \cup R_2) \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} + S_{2j}^+ - S_{2j}^- = c, \quad j \in G_2' = G_2 \cap (R_0 \cup R_1) \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} \leq d, \quad j \in C_2 \\
 & \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) = 1, \\
 & \quad d \leq c \leq d + 1, \\
 & \quad S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \geq 0, \quad \forall j \\
 & \quad \lambda_i^+ \geq 0 \quad \& \quad \lambda_i^- \geq 0, \quad \forall i \\
 & \quad c, d \text{ ελεύθερα προσήμους}
 \end{aligned} \tag{3.7}$$

Τα δύο σημαντικότερα χαρακτηριστικά του γραμμικού μοντέλου (3.7) τεκμηριώνονται παρακάτω:

- (α) Η ταξινόμηση των παρατηρήσεων που ταξινομήθηκαν σωστά στο πρώτο στάδιο δεν μεταβάλλεται καθώς οι παρατηρήσεις αυτές υπόκεινται στους ακόλουθους περιορισμούς:

$$\sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} \geq d + 1, \quad j \in C_1 \quad \text{και} \quad \sum_{i=1}^k (\lambda_i^+ - \lambda_i^-) Z_{ij} \leq d, \quad j \in C_2$$

- (β) Θέτοντας ως άνω κατώφλι την τιμή  $d + 1$  για το υποσύνολο  $C_1$  και ως κάτω κατώφλι την τιμή  $d$  για το υποσύνολο  $C_2$ , η νέα μεταβλητή  $c$  συμβάλλει στην



ελαχιστοποίηση του συνολικού αθροιστικού σφάλματος  $\sum_{j \in G_1} S_{1j}^+ + \sum_{j \in G_2} S_{2j}^-$  των

δύο υποσυνόλων  $G_1 \cap (R_0 \cup R_2)$  και  $G_2 \cap (R_0 \cup R_1)$ .

Οι βέλτιστες λύσεις που προκύπτουν από την επίλυση του (3.7) γραμμικού μοντέλου είναι:  $\lambda_i^* (= \lambda_i^{+*} - \lambda_i^{-*})$  και  $c^*$ . Βάσει αυτής της λύσης η ταξινόμηση των παρατηρήσεων του δείγματος υλοποιείται με τους παρακάτω κανόνες:

(α) Εάν ισχύει:  $\sum_{i=1}^k \lambda_i^* Z_{im} \geq c^*$ , τότε η παρατήρηση  $m$  ανήκει στην ομάδα  $G_1$ .

(β) Εάν ισχύει:  $\sum_{i=1}^k \lambda_i^* Z_{im} < c^*$ , τότε η παρατήρηση  $m$  ανήκει στην ομάδα  $G_2$ .

Στην περίπτωση που  $\sum_{i=1}^k \lambda_i^* Z_{im} = c^*$ , τότε η παραπάνω μέθοδος δεν θεωρείται ικανή για την ταξινόμηση των παρατηρήσεων. Σε αυτό το ενδεχόμενο, θα ήταν καλύτερο να χρησιμοποιηθεί μια νέα τεχνική διακριτικής ανάλυσης.

### 3.4 Περίληψη

Όπως παρουσιάστηκε παραπάνω, η μεθοδολογική προσέγγιση Extended DEA-DA είναι μία νέα μορφή μη παραμετρικού προγραμματισμού στόχων για την αντιμετώπιση προβλημάτων ταξινόμησης και έχει ως σκοπό τον προσδιορισμό της επικάλυψης η οποία μπορεί να υφίσταται μεταξύ δύο κατηγοριών από παρατηρήσεις και την επαναταξινόμηση ενός νέου δείγματος παρατηρήσεων. Το βασικό χαρακτηριστικό αυτής της προσέγγισης είναι ότι μπορεί να αποφύγει τον προσδιορισμό μιας διαχωριστικής συνάρτησης, δημιουργώντας δύο γραμμικές κατά τμήματα διακριτικές συναρτήσεις, μια για τον προσδιορισμό της επικάλυψης και μία για την ταξινόμηση. Έχοντας ολοκληρώσει την ανάλυση της μεθοδολογικής προσέγγισης Extended DEA-DA, για την ανάπτυξη υποδειγμάτων ταξινόμησης, θα πραγματοποιηθεί μια συγκριτική πειραματική ανάλυση της αποτελεσματικότητάς της σε σύγκριση με άλλες μεθόδους. Παρακάτω παρουσιάζονται αναλυτικά τα αποτελέσματα του πειράματος, αφού γίνεται αρχικά κάποια αναφορά στις εξεταζόμενες μεθόδους και στο σχεδιασμό του πειράματος.

## ΠΕΙΡΑΜΑΤΙΚΗ ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ

### 4.1 Σκοπός της έρευνας

Βασικός σκοπός της έρευνας που παρουσιάζεται στο κεφάλαιο αυτό είναι η εξέταση της αποτελεσματικότητας της μεθόδου DEA στην αντιμετώπιση προβλημάτων ταξινόμησης σε σύγκριση με άλλες εναλλακτικές διαδεδομένες τεχνικές. Για κάθε προσέγγιση υπολογίζεται το ποσοστό εσφαλμένων ταξινομήσεων συναρτήσει της επίδρασης κάποιων προκαθορισμένων παραγόντων ή των αλληλεπιδράσεων αυτών. Με βάση το ποσοστό εσφαλμένων ταξινομήσεων αξιολογείται η αποτελεσματικότητα της μεθόδου κατά τη διαδικασία της ταξινόμησης και εξετάζεται ο τρόπος με τον οποίο επιδρά ο κάθε παράγοντας στην αποτελεσματικότητα της κάθε προσέγγισης. Τα αποτελέσματα της ανάλυσης δείχνουν ότι οι κύριοι προσδιοριστικοί παράγοντες είναι εκείνοι που σχετίζονται με τις στατιστικές ιδιότητες των εξεταζόμενων δεδομένων. Παρακάτω αναλύονται τόσο οι μέθοδοι που χρησιμοποιήθηκαν στην παρούσα έρευνα, καθώς και η πειραματική ανάλυση και τα αποτελέσματά της.

### 4.2 Εξεταζόμενες μέθοδοι

Σε κάθε συγκριτική ανάλυση μεταξύ εναλλακτικών μεθοδολογικών προσεγγίσεων ενός προβλήματος, οι εξεταζόμενες προσεγγίσεις θα πρέπει να επιλέγονται έτσι ώστε να είναι αντιπροσωπευτικές όλων των διαθέσιμων επιλογών και ευρείας αποδοχής στον χώρο του προβλήματος. Στην παρούσα συγκριτική ανάλυση η διερεύνηση της αποτελεσματικότητας της μεθόδου DEA, πραγματοποιείται μέσω της σύγκρισής της με άλλες τέσσερις

διαδεδομένες τεχνικές: (1) τη γραμμική διακριτική ανάλυση, (2) μια διατύπωση γραμμικού προγραμματισμού, (3) τον αλγόριθμο του πλησιέστερου γείτονα, (4) τα δέντρα ταξινόμησης, και (5) τις μηχανές διανύσματος υποστήριξης. Στις ενότητες που ακολουθούν παρουσιάζονται τα βασικά χαρακτηριστικά των τεχνικών αυτών.

#### 4.2.1 Γραμμική Διακριτική Ανάλυση

Η γραμμική διακριτική ανάλυση (LDA) αποτέλεσε την πρώτη πολυδιάστατη μέθοδο ταξινόμησης και αναπτύχθηκε αρχικά από τον Fisher (1936). Σκοπός της μεθόδου είναι η ανάπτυξη μιας σειράς διακριτικών συναρτήσεων οι οποίες μεγιστοποιούν τη διακύμανση μεταξύ των κατηγοριών σε σχέση με την διακύμανση εντός των κατηγοριών, χρησιμοποιώντας ως δείγμα εκμάθησης ένα σύνολο εναλλακτικών δραστηριοτήτων η ταξινόμηση των οποίων είναι γνωστή. Στην περίπτωση των δύο κατηγοριών ( $C_1$  και  $C_2$ ), η οποία εξετάζεται στην παρούσα ανάλυση, η LDA οδηγεί στην ανάπτυξη μιας διακριτικής συνάρτησης της μορφής::

$$F(\mathbf{z}) = a + b_1 Z_1 + b_2 Z_2 + \dots + b_k Z_k$$

όπου  $\mathbf{z} = (Z_1, Z_2, \dots, Z_k)$  είναι το διάνυσμα των χαρακτηριστικών που περιγράφουν τις εναλλακτικές δραστηριότητες,  $a$  είναι μια σταθερά και  $b_1, b_2, \dots, b_k$  είναι οι συντελεστές των χαρακτηριστικών στη συνάρτηση.

Ο υπολογισμός του σταθερού όρου  $a$  και του διανύσματος  $\mathbf{b} = (b_1, b_2, \dots, b_k)^T$  βασίζεται στην υπόθεση ότι οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών είναι ίσοι και ότι οι επιδόσεις των εναλλακτικών δραστηριοτήτων στα εξεταζόμενα χαρακτηριστικά ακολουθούν την πολυμεταβλητή κανονική κατανομή. Βάσει των υποθέσεων αυτών οι υπολογισμοί των  $a$  και  $\mathbf{b}$  πραγματοποιούνται ως εξής:

$$\mathbf{b} = \Sigma^{-1} \cdot [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2] \quad \text{και} \quad a = -[\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2]' \cdot \mathbf{b} / 2$$

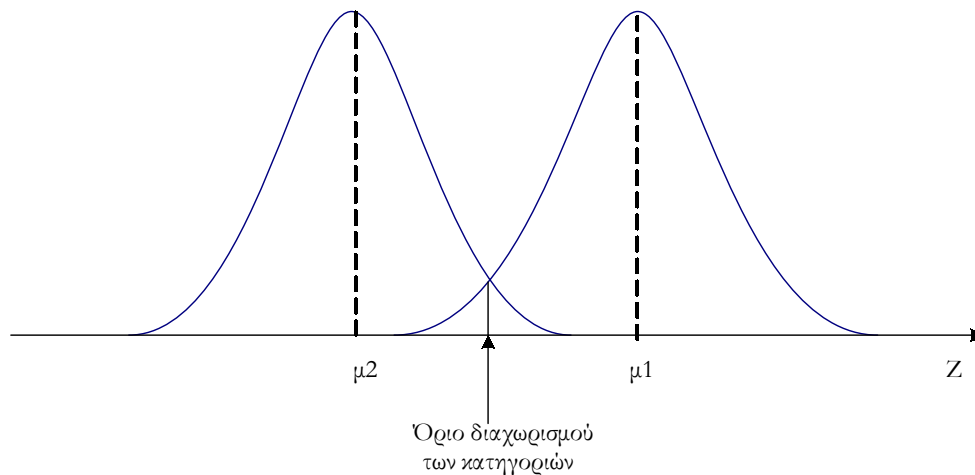
όπου  $\boldsymbol{\mu}_1$  και  $\boldsymbol{\mu}_2$  είναι τα διανύσματα των μέσων τιμών των χαρακτηριστικών για τις εναλλακτικές δραστηριότητες των κατηγοριών  $C_1$  και  $C_2$ , αντίστοιχα, και  $\Sigma$  είναι ο πίνακας διακύμανσης-συνδιακύμανσης μεταξύ των κατηγοριών.

Η ταξινόμηση κάθε αντικειμένου (μονάδας)  $j$  σε μια εκ των προκαθορισμένων κατηγοριών πραγματοποιείται βάσει του σκορ διάκρισης  $F(\mathbf{z}_j)$  του αντικειμένου όπως αυτό υπολογίζεται από τη διακριτική συνάρτηση. Συγκεκριμένα, ο κανόνας ταξινόμησης έχει την ακόλουθη μορφή:

$$F(z_j) \geq \ln \frac{K(1/2)\pi_1}{K(2/1)\pi_2} \Rightarrow j \in C_1$$

$$F(z_j) < \ln \frac{K(1/2)\pi_1}{K(2/1)\pi_2} \Rightarrow j \in C_2$$

Ως  $K(1/2)$  συμβολίζεται το κόστος της εσφαλμένης ταξινόμησης ενός αντικειμένου, της κατηγορίας  $C_1$  στην κατηγορία  $C_2$ , ενώ ως  $\pi_1$  συμβολίζεται η εκ των προτέρων πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα στην κατηγορία  $C_1$ . Θεωρώντας ίσα τα κόστη εσφαλμένων ταξινομήσεων και τις εκ των προτέρων πιθανότητες, ο γραμμικός κανόνας ταξινόμησης για τη διάκριση μεταξύ δύο κατηγοριών μπορεί να αποδοθεί γραφικά μέσω του παρακάτω σχήματος.



Σχήμα 4.1: Σχηματική απεικόνιση του κανόνα ταξινόμησης της γραμμικής διακριτικής ανάλυσης

Δεδομένου ότι ο καθορισμός των εκ των προτέρων πιθανοτήτων και του κόστους των εσφαλμένων ταξινομήσεων είναι συχνά δύσκολος, το όριο που διαχωρίζει τις κατηγορίες καθορίζεται συνήθως μέσω διαδικασιών δοκιμής και λάθους, ώστε να ελαχιστοποιηθεί ο

συνολικός αριθμός των εσφαλμένων ταξινόμησεων και παράλληλα να υπάρχει μια ισορροπία στον αριθμό των εσφαλμένων ταξινόμησεων ανά κατηγορία.

#### 4.2.2 Γραμμικός Προγραμματισμός

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο τεχνικές γραμμικού προγραμματισμού (linear programming, LP) έχουν ήδη χρησιμοποιηθεί για την ανάπτυξη μοντέλων ταξινόμησης. Η διατύπωση που χρησιμοποιείται στην παρούσα έρευνα προτάθηκε από τους Gochet et al. (1997).

Στην προσέγγιση αυτή θεωρείται ένα πεπερασμένο σύνολο  $S = \{1, \dots, s\}$  από κατηγορίες αντικειμένων, όπου κάθε αντικείμενο ανήκει σε μια και μόνο μια κατηγορία. Τα δείγματα των αντικειμένων που ανήκουν σε κάθε κατηγορία είναι μεγέθους  $n_j, j \in S$  και η κατηγορία του κάθε αντικείμενου στο δείγμα εκμάθησης είναι γνωστή. Το πλήθος των αντικειμένων του δείγματος εκμάθησης είναι  $n$  και η αναλογία των αντικειμένων κάθε κατηγορίας  $j$  στο δείγμα εκμάθησης συμβολίζεται με  $P_j = \{1, \dots, n_j\}$ . Κάθε αντικείμενο  $i$ , για το οποίο δεν έχει προσδιοριστεί η κατηγορία στην οποία ανήκει, χαρακτηρίζεται από ένα σύνολο από  $k$  χαρακτηριστικά που περιλαμβάνονται σε ένα διάνυσμα στήλης  $\mathbf{z}_i = (Z_{i0}, Z_{i1}, \dots, Z_{ik})'$ , όπου  $Z_{i0} = 1$ .

Σύμφωνα με την μέθοδο του γραμμικού προγραμματισμού υπολογίζονται  $s$  διανύσματα γραμμής  $\mathbf{a}^j = (a_{j0}, a_{j1}, \dots, a_{jk})$  και τα γραμμικά σκορ ταξινόμησης  $\mathbf{a}^j \mathbf{z}_i (j = 1, \dots, s)$  για κάθε αντικείμενο  $i$  σε σχέση με την κατηγορία  $j, j \in S$  που ανήκει. Η διακριτική συνάρτηση ταξινόμησης  $\mathbf{a}^m \mathbf{z}_i = \max_{j \in S} \{\mathbf{a}^j \mathbf{z}_i\}$  ταξινομεί ένα αντικείμενο σε μια κατηγορία

για την οποία αυτό επιτυγχάνει το υψηλότερο σκορ ταξινόμησης. Τα διανύσματα  $\mathbf{a}_j, j \in S$  προσδιορίζονται βελτιστοποιώντας την ποιότητα του μοντέλου και τα σφάλματα που πραγματοποιεί στην ταξινόμηση των αντικειμένων του δείγματος εκμάθησης. Η ποιότητα προσαρμογής για ένα αντικείμενο  $i \in P_r, r \in S$  μπορεί να μετρηθεί από την συνάρτηση

$G_{ij}^i(\mathbf{a}^r, \mathbf{a}^j) = (\mathbf{a}^r \mathbf{z}_i - \mathbf{a}^j \mathbf{z}_i)^+, i \in P_r, j \in S_{-r}, r \in S$ , στην οποία το σκορ ταξινόμησης  $\mathbf{a}^r \mathbf{z}_i$  του αντικείμενου  $i$  σε σχέση με την κατηγορία στην οποία ανήκει συγκρίνεται με τα σκορ ταξινόμησης  $\mathbf{a}^j \mathbf{z}_i$  του αντικείμενου  $i$  σε σχέση με τις υπόλοιπες κατηγορίες  $j \in S_{-r}$ . Για

την ποιότητα προσαρμογής προτιμώνται μεγάλες και θετικές τιμές. Αντιθέτως, για το σφάλμα προσαρμογής για ένα αντικείμενο  $i \in P$  σε σχέση με την κατηγορία  $j$  προτιμώνται μικρές τιμές. Το σφάλμα αυτό ορίζεται από την σχέση  $B_{rj}^i(a^r, a^j) = (a^r z_i - a^j z_i)^-, i \in P_r, j \in S_{-r}, r \in S$ . Το άθροισμα της ποιότητας και του σφάλματος ενός αντικειμένου  $i \in P_r$  δίνεται από τις παρακάτω σχέσεις:

$$\left. \begin{aligned} G_r^i(a) &= G_r^i(a^1, \dots, a^s) = \sum_{j \in S_{-r}} G_{rj}^i(a^r, a^j), i \in P_r, r \in S \\ B_r^i(a) &= B_r^i(a^1, \dots, a^s) = \sum_{j \in S_{-r}} B_{rj}^i(a^r, a^j), i \in P_r, r \in S \end{aligned} \right\} \quad (4.1)$$

Η ποιότητα και το σφάλμα συνδυάζοντας όλα τα αντικείμενα  $i$  στην κατηγορία  $r$  δίνονται από τις παρακάτω σχέσεις:

$$\left. \begin{aligned} G_r(a) &= G_r(a^1, \dots, a^s) = \sum_{i \in P_r} G_r^i(a^1, \dots, a^s) = \sum_{i \in P_r} \sum_{j \in S_{-r}} G_{rj}^i(a^r, a^j), r \in S \\ B_r(a) &= B_r(a^1, \dots, a^s) = \sum_{i \in P_r} B_r^i(a^1, \dots, a^s) = \sum_{i \in P_r} \sum_{j \in S_{-r}} B_{rj}^i(a^r, a^j), r \in S \end{aligned} \right\} \quad (4.2)$$

Οι εκτιμήσεις της συνολικής ποιότητας και σφάλματος για όλες τις κατηγορίες  $r \in S$  δίνονται από τις παρακάτω σχέσεις:

$$\left. \begin{aligned} G(a) &= G(a^1, \dots, a^s) = \sum_{r \in S} G_r(a) = \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} G_{rj}^i(a^r, a^j) \\ B(a) &= B(a^1, \dots, a^s) = \sum_{r \in S} B_r(a) = \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} B_{rj}^i(a^r, a^j) \end{aligned} \right\} \quad (4.3)$$

Οι εκτιμήσεις της συνολικής ποιότητας και σφάλματος προσαρμογής (4.3) είναι αντίστοιχες των εσωτερικών και εξωτερικών αποκλίσεων και μη αρνητικές για κάθε  $\alpha$ . Μια μη μηδενική λύση  $a^r = a^*$ , για όλα τα  $r \in S$ , μηδενίζει την ποιότητα και το σφάλμα προσαρμογής, αλλά δεν παρέχει καμία χρήσιμη πληροφορία όσον αφορά την ταξινόμηση, καθώς κάθε αντικείμενο μπορεί να ταξινομηθεί αυθαίρετα σε μια από τις κατηγορίες. Μια λύση  $\alpha$  για την οποία το συνολικό σφάλμα ξεπερνά την συνολική ποιότητα προσαρμογής ( $G(a) - B(a) < 0$ ), δεν είναι ικανοποιητική. Μπορεί εύκολα να επαληθευτεί ότι για κάθε  $\alpha$  ισχύει  $G(a) = B(-a)$ , επομένως για λύση  $\alpha$  με  $G(a) - B(a) = -q < 0$ , ισχύει επίσης

$G(-a) - B(-a) = q > 0$ . Συνεπώς, μη επιθυμητές λύσεις με συνολικό σφάλμα μεγαλύτερο της συνολικής ποιότητας προσαρμογής μπορούν να απορριφθούν βάσει της εξής κανονικοποίησης:  $G(a) - B(a) = \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} (a^r - a^j) \mathbf{z}_i = q$ , όπου  $q$  είναι ένας θετικός αριθμός.

Βάσει αυτής της μοντελοποίησης το βέλτιστο μοντέλο ταξινόμησης (προσδιορισμός των διανυσμάτων  $\alpha$ ) μπορεί να αναπτυχθεί μέσω της επίλυσης ενός γραμμικού προγράμματος της ακόλουθης γενικής μορφής (γραμμικό πρόγραμμα  $LP^q$ ).

$$\left. \begin{array}{l} \text{Min} B(a) \\ \text{υπό:} \\ G(a) - B(a) = q \\ a \text{ ελεύθερο προσήμου} \end{array} \right\} \quad (4.4)$$

Το παραπάνω γραμμικό πρόβλημα μπορεί να αναδιατυπωθεί εισάγοντας τις μεταβλητές  $\beta_{rj}^i$  και  $\gamma_{rj}^i$  που αναπαριστούν το σφάλμα ταξινόμησης  $B_{rj}^i(a^r, a^j)$  και την ποιότητα  $G_{rj}^i(a^r, a^j)$  ταξινόμησης ενός αντικειμένου  $i \in P_r$  σε σχέση με την κατηγορία  $j \in S_{-r}$ .

$$\left. \begin{array}{l} \text{Min} \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} \beta_{rj}^i \\ \text{υπό:} \\ \beta_{rj}^i + (a^r - a^j) \mathbf{z}_i - \gamma_{rj}^i = 0 \quad \text{για} \quad i \in P_r, j \in S_{-r}, r \in S, \\ \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} (\gamma_{rj}^i - \beta_{rj}^i) = q, \\ \beta_{rj}^i, \gamma_{rj}^i \geq 0, \quad \text{για} \quad i \in P_r, j \in S_{-r}, r \in S. \end{array} \right\} \quad (4.5)$$

Αυτό το γραμμικό πρόβλημα είναι παρόμοιο με το υβριδικό μοντέλο (αναλύεται στο 2<sup>ο</sup> κεφάλαιο) χωρίς τις minimax αποκλίσεις.

#### 4.2.3 Ο Αλγόριθμος Πλησιέστερου Γείτονα

Αντικειμενικός σκοπός των αλγορίθμων των πλησιέστερων γειτόνων (nearest neighbors, NN) είναι η εκτίμηση της υπό συνθήκη πιθανότητας ένα αντικείμενο  $\mathbf{z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{ik})$  να ανήκει στην κατηγορία  $j$ . Ο υπολογισμός αυτής της πιθανότητας πραγματοποιείται βάσει του πλήθους των μονάδων του δείγματος εκμάθησης οι οποίες ανήκουν στην κατηγορία  $j$  και

βρίσκονται στον γειτονικό χώρο του  $\mathbf{z}_i$ . Ο προσδιορισμός των γειτονικών μονάδων του  $\mathbf{z}_i$  μπορεί εύκολα πραγματοποιηθεί προσδιορίζοντας την απόσταση κάθε μονάδας του δείγματος εκμάθησης από το  $\mathbf{z}_i$ , χρησιμοποιώντας για παράδειγμα την Ευκλείδεια απόσταση, και θεωρώντας ότι όλες οι μεταβλητές (χαρακτηριστικά) έχουν την ίδια σπουδαιότητα. Ταυτόχρονα, θα πρέπει να καθοριστεί και το εύρος του γειτονικού χώρου του  $\mathbf{z}_i$  με τον καθορισμό μιας παραμέτρου  $K$  η οποία προσδιορίζει το πλήθος των γειτονικών μονάδων που θα εξεταστούν.

Δεδομένων των  $K$  γειτονικών μονάδων του  $\mathbf{z}_i$ , η ταξινόμησή του μπορεί εύκολα να πραγματοποιηθεί εξετάζοντας την κατηγορία στην οποία ανήκουν τα γειτονικά του στοιχεία. Ειδικότερα, μέσω του απλού κανόνα της πλειοψηφίας, αποφασίζεται η ταξινόμηση του  $\mathbf{z}_i$  στην κατηγορία στην οποία ανήκει η πλειοψηφία των  $K$  πλησιέστερων γειτόνων του.

Στην παρούσα έρευνα χρησιμοποιείται ο αλγόριθμος του πλησιέστερου γείτονα, θεωρώντας  $K=1$ . Περαιτέρω ανάλυση των ιδιοτήτων, των χαρακτηριστικών και των υπολογιστικών μεθόδων των αλγορίθμων των πλησιέστερων γειτόνων δίδονται στο βιβλίο του Hand (1997).

#### **4.2.4 Δέντρα Ταξινόμησης και Παλινδρόμησης**

Η μέθοδος CART (Classification and Regression Trees, CART, Yohannes Y. Και Webb P. (1999)) είναι μια μη παραμετρική προσέγγιση που αναπτύχθηκε για την ανάλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Σε κάθε περίπτωση το μοντέλο ταξινόμησης ή παλινδρόμησης που αναπτύσσεται μέσω της μεθόδου CART αναπαριστά τη μορφή ενός δέντρου αποφάσεων. Στην περίπτωση της ταξινόμησης κύριος σκοπός της μεθόδου CART είναι να παράγει ένα ακριβές σύνολο από κανόνες ταξινόμησης βάσει των οποίων θα προβλέπεται σε ποια κατηγορία θα ανήκει κάθε μελλοντική παρατήρηση, σύμφωνα με τα αντίστοιχα χαρακτηριστικά της. Η δομή ενός κανόνα ταξινόμησης της μεθόδου CART επικεντρώνεται στους ορισμούς τριών κύριων παραγόντων: (α) του κανόνα διαχωρισμού του δείγματος των παρατηρήσεων, (β) των κριτηρίων αξιολόγησης της ποιότητας του διαχωρισμού, (γ) των κριτηρίων για την επιλογή του βέλτιστου δέντρου για ανάλυση. Τα βασικά βήματα για την δημιουργία ενός δέντρου ταξινόμησης είναι: (α) δημιουργία ενός δέντρου με μεγάλο αριθμό από κόμβους, (β) ένωση μερικών διακλαδώσεων για την παραγωγή μιας σειράς από μικρότερα δέντρα διαφορετικού μεγέθους, (γ) επιλογή ενός βέλτιστου δέντρου μέσω της μέτρησης της ακρίβειας του δέντρου.



Για την ανάπτυξη ενός δέντρου ταξινόμησης, η μέθοδος CART χρησιμοποιεί μια πιθανοθεωρητική προσέγγιση η οποία μπορεί να υλοποιηθεί με τρεις τρόπους: (α) προσδιορισμός των a priori πιθανοτήτων των κατηγοριών από τα δεδομένα:  $\pi_i = n_i/n$ , όπου  $\pi_i$  η a priori της κατηγορίας  $C_i$ ,  $n$  ο αριθμός των αντικειμένων στο δείγμα, και  $n_i$  ο αριθμός των αντικειμένων της κατηγορίας  $C_i$ , (β) θεώρηση των a priori πιθανοτήτων των κατηγοριών ως ίσων, και (γ) προσδιορισμός των a priori πιθανοτήτων των κατηγοριών μέσω μιας υβριδικής προσέγγισης θεωρώντας τον μέσο όρο των δύο εκτιμήσεων που υπολογίζονται από τις προηγούμενες δύο προσεγγίσεις.

Η ανάπτυξη ενός δέντρου ταξινόμησης απαιτεί τον καθορισμό τριών στοιχείων: (α) ενός συνόλου ερωτήσεων η απάντηση των οποίων οδηγεί στην ταξινόμηση των αντικειμένων, (β) των κανόνων αξιολόγησης της ποιότητας των ερωτήσεων που αναπτύσσονται, και (γ) των κανόνων για το προσδιορισμό της κατηγορίας σε κάθε τερματικό κόμβο του δέντρου.

Αρχικά, όλες οι παρατηρήσεις τοποθετούνται σε έναν αρχικό κόμβο, ο οποίος είναι ανομοιογενής καθώς περιέχει παρατηρήσεις από διάφορες κατηγορίες. Ο κύριος στόχος είναι η εύρεση εκείνων των κανόνων που θα διαχωρίσουν τις παρατηρήσεις δημιουργώντας νέους κόμβους σε κατώτερα επίπεδα του δέντρου, οι οποίοι θα είναι περισσότερο ομοιογενείς σε σχέση με τους προηγούμενους κόμβους.

Σε κάθε κόμβο  $t$  του δέντρου οι παρατηρήσεις (αντικείμενα) του δείγματος διαχωρίζονται σε δύο επιμέρους κόμβους  $t_L$  και  $t_R$  στο αμέσως κατώτερο επίπεδο του δέντρου, ανάλογα με τον εάν ικανοποιούν ή όχι έναν κανόνα (ερώτηση) της μορφής  $Z_{ij} \leq d_j$ , όπου  $Z_{ij}$  είναι ένα χαρακτηριστικό και  $d_j$  είναι ένα όριο διαχωρισμού. Ειδικότερα, μια παρατήρηση  $i$  τοποθετείται στον κόμβο  $t_L$  εάν  $Z_{ij} \leq d_j$ , διαφορετικά τοποθετείται στον κόμβο  $t_R$ . Ο βέλτιστος κανόνας διαχωρισμού καθορίζεται μεγιστοποιώντας τη μείωση της ανομοιογένειας (impurity) που αποφέρει ο διαχωρισμός. Ένας διαχωρισμός θεωρείται ομοιογενής εάν οι δύο κόμβοι που δημιουργούνται από αυτόν περιλαμβάνουν (ο καθένας) παρατηρήσεις από διαφορετικές κατηγορίες. Εάν κάποιος κόμβος περιλαμβάνει παρατηρήσεις από διαφορετικές κατηγορίες, τότε ο διαχωρισμός θεωρείται ως ανομοιογενής. Βάσει αυτής της θεώρησης ως κριτήριο επιλογής του κατάλληλου διαχωρισμού (επιλογή χαρακτηριστικού  $Z_j$  και τιμής ορίου  $d_j$ ) θεωρείται η μεγιστοποίηση της ακόλουθης συνάρτησης:

$$\Delta_i(s, t) = i(t) - p_L[i(t_L)] - p_R[i(t_R)],$$

όπου:  $s$  ο διαχωρισμός των παρατηρήσεων από τον κανόνα που αναπτύσσεται,  $p_L$  η αναλογία των περιπτώσεων του κόμβου  $t$  που καταλήγουν στον αριστερό κόμβο  $t_L$ ,  $p_R$  η αναλογία των περιπτώσεων του κόμβου  $t$  που καταλήγουν στο δεξί κόμβο  $t_R$ ,  $i(t_L)$  η ομοιογένεια του αριστερού κόμβου και  $i(t_R)$  η ομοιογένεια του δεξιού κόμβου.

Αυτή η διαδικασία διαχωρισμού ξεκινά από τον αρχικό κόμβο του δέντρου στον οποίο εντάσσονται όλες οι παρατηρήσεις και συνεχίζεται επαναληπτικά για κάθε νέο κόμβο που κατασκευάζεται. Εάν η διαδικασία εφαρμοστεί χωρίς κάποιο κριτήριο τερματισμού, τότε θα ολοκληρωθεί με την ανάπτυξη ενός μεγάλου και περίπλοκου δέντρου στο οποίο κάθε τελικός κόμβος θα περιέχει μόνο μια παρατήρηση του δείγματος εκμάθησης. Για να αποφευχθεί αυτό το φαινόμενο συνήθως χρησιμοποιούνται τεχνικές μείωσης των διαστάσεων του δέντρου οι οποίες υλοποιούνται είτε με την εισαγωγή κριτηρίων έγκαιρου τερματισμού της διαδικασίας ανάπτυξης του δέντρου, είτε με την «περικοπή» (pruning) του δέντρου μετά την πλήρη ανάπτυξή του.

Οι Breiman et al. (1984) και Steinberg και Colla (1995) τονίζουν ως βασικά πλεονεκτήματα της μεθόδου CART τα ακόλουθα σημεία:

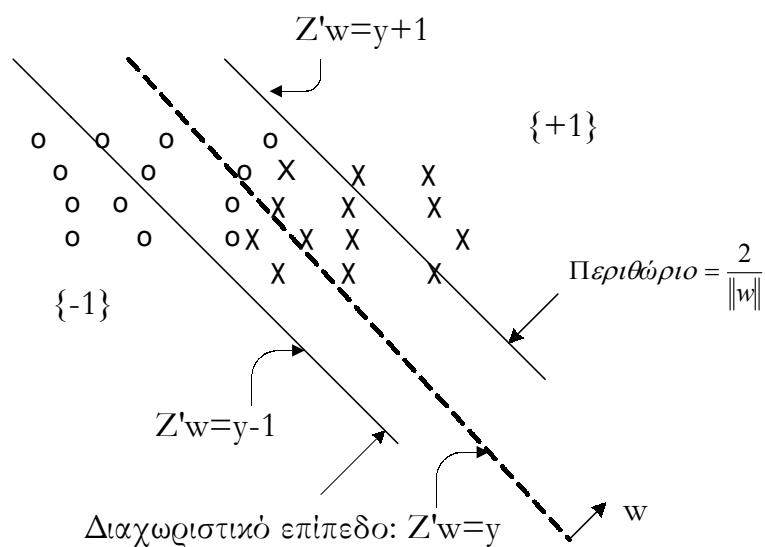
1. Δεν πραγματοποιείται καμία στατιστική υπόθεση όσον αφορά τα χαρακτηριστικά.
2. Είναι δυνατή η χρησιμοποίηση τόσο ποιοτικών όσο και ποσοτικών χαρακτηριστικών.
3. Είναι δυνατή η ανάπτυξη δέντρων ακόμα και από δεδομένα που δεν είναι πλήρη.
4. Τα αποτελέσματα της μεθόδου CART δεν επηρεάζονται από την ύπαρξη μη φυσιολογικών δεδομένων (outliers), από φαινόμενα πολυσυγγραμμικότητας (multicollinearity) ή άλλα στατιστικά προβλήματα.
5. Η CART έχει τη δυνατότητα να αναζητά και να αποκαλύπτει τις αλληλεπιδράσεις των μεταβλητών μέσα στο σύνολο των δεδομένων.
6. Τα αποτελέσματα της μεθόδου παραμένουν αμετάβλητα ανεξάρτητα από πιθανούς μονότονους μετασχηματισμούς των δεδομένων.

7. Μπορεί να παράγει χρήσιμα αποτελέσματα από ένα μεγάλο αριθμό μεταβλητών που παρέχονται προς ανάλυση, χρησιμοποιώντας μόνο ελάχιστες σημαντικές μεταβλητές.
8. Η κατανόηση των δέντρων ταξινόμησης της μεθόδου CART είναι ιδιαίτερα εύκολη.

#### 4.2.5 Μηχανές Διανύσματος Υποστήριξης

Οι μηχανές διανύσματος υποστήριξης (support vector machines, SVM, Vapnik (2000) και Burges (1998)) έχουν αναπτυχθεί τα τελευταία χρόνια ως μια από τις σημαντικότερες μεθόδους για την ανάπτυξη μοντέλων ταξινόμησης. Κύριο χαρακτηριστικό τους αποτελεί το σημαντικό θεωρητικό υπόβαθρο πάνω στο οποίο βασίζονται οι SVM, καθώς και η πληθώρα επιτυχημένων πρακτικών εφαρμογών.

Η λογική των SVM παρουσιάζεται συνοπτικά στο σχήμα 4.2 όπου απεικονίζεται ένα πρόβλημα ταξινόμησης  $n$  μονάδων οι οποίες περιγράφονται βάσει  $k$  χαρακτηριστικών, σε δύο κατηγορίες οι οποίες συμβολίζονται ως  $+1$  και  $-1$ .



Σχήμα 4 2: Γραφική απεικόνιση των SVM

Στόχος των SVM, στην απλή γραμμική περίπτωση, είναι η ανάπτυξη του βέλτιστου υπερεπιπέδου της μορφής  $Zw - y$  για την ταξινόμηση των παρατηρήσεων, όπου ως  $Z$  συμβολίζεται ένας πίνακας διαστάσεων  $n \times k$  με τα στοιχεία των παρατηρήσεων του δείγματος εκμάθησης. Συμβολίζοντας ως  $D$  ένα διαγώνιο πίνακα διαστάσεων  $n \times n$  με την κύρια

διαγώνιο να έχει τιμές +1 ή -1 ανάλογα με την ταξινόμηση των παρατηρήσεων του δείγματος εκμάθησης, και ως  $e$  το μοναδιαίο διάνυσμα διαστάσεων  $n \times 1$ , ο εντοπισμός του βέλτιστου υπερεπιπέδου επιτυγχάνεται με την επίλυση του ακόλουθου τετραγωνικού προγράμματος (ως  $v$  συμβολίζεται μια αυστηρά θετική σταθερά):

$$\left. \begin{array}{l} \text{Min}_{w,y,d} \quad ve'd + \frac{1}{2} w'w \\ \text{υπό:} \\ D(Zw - ey) + d \geq e \\ d \geq 0 \end{array} \right\} \quad (4.6)$$

Ο τετραγωνικός όρος  $w'w$  στην αντικειμενική συνάρτηση του προβλήματος (4.6) μεγιστοποιεί το περιθώριο μεταξύ των δυο υπερεπιπέδων  $Zw - y = +1$  και  $Zw - y = -1$ , το οποίο ισούται με  $2/\|w\|$ . Εκτός της μεγιστοποίησης του περιθωρίου των κατηγοριών, το πρόβλημα (4.6) λαμβάνει υπόψη και το σφάλμα ταξινόμησης με τις μεταβλητές  $d$  (η σταθερά  $v > 0$  αναπαριστά τη σχετική βαρύτητα που αποδίδεται στην ελαχιστοποίηση των σφαλμάτων). Όταν όλες οι μεταβλητές  $d$  είναι ίσες με το μηδέν, τότε οι δύο κατηγορίες είναι αυστηρά γραμμικά διαχωρισμένες και το επίπεδο  $Z'w = y + 1$  περιλαμβάνει όλες τις μονάδες της κατηγορίας +1, ενώ το επίπεδο  $Z'w = y - 1$  περιλαμβάνει όλες τις μονάδες της κατηγορίας -1. Εάν οι κατηγορίες είναι γραμμικά διαχωρίσιμες (σχήμα 4.2) τα δύο επίπεδα καθορίζουν τα όρια των δύο κατηγοριών με ένα μη αρνητικό σφάλμα της μεταβλητής  $d$ .

$$\left. \begin{array}{l} Z'_i w + d_i \geq y + 1, \quad \text{για} \quad D_{ii} = +1 \\ Z'_i w - d_i \leq y - 1, \quad \text{για} \quad D_{ii} = -1 \end{array} \right\} \quad (4.7)$$

Με την επίλυση του προβλήματος (4.6) και τον προσδιορισμό των  $w$  και  $y$  που καθορίζουν το βέλτιστο υπερεπίπεδο, η ταξινόμηση κάθε αντικειμένου μπορεί εύκολα να πραγματοποιηθεί ως εξής:

$$\text{Εάν} \quad Z'_i w - y \begin{cases} > 0, & \text{τότε} \quad Z_i \in \{+1\}, \\ < 0, & \text{τότε} \quad Z_i \in \{-1\}, \\ = 0, & \text{τότε} \quad Z_i \in \{+1\} \quad \text{ή} \quad x \in \{-1\} \end{cases} \quad (4.8)$$

Το κύριο μειονέκτημα του προβλήματος βελτιστοποίησης (4.6) για τον προσδιορισμό του βέλτιστου μοντέλου ταξινόμησης αφορά τον αυξημένο υπολογιστικό φόρτο που απαιτεί η επίλυσή του καθώς πρόκειται για ένα πρόβλημα τετραγωνικού προγραμματισμού. Για την αντιμετώπιση του προβλήματος αυτού οι Fung και Mangasarian (2001) πρότειναν μια εναλλακτική διατύπωση του προβλήματος ως εξής:

$$\left. \begin{array}{l} \text{Min}_{w,y,d} v \frac{1}{2} \|d\|^2 + \frac{1}{2} (w'w + y^2) \\ \text{υπό:} \\ D(Zw - ey) + d \geq e \end{array} \right\} \quad (4.9)$$

Η αντικειμενική συνάρτηση του νέου προβλήματος αφορά την ελαχιστοποίηση της νόρμας δεύτερης τάξης του διανύσματος  $d$  (σε αντίθεση με την νόρμα πρώτης τάξεως που χρησιμοποιείται στο πρόβλημα (4.6)). Επιπλέον, η μεγιστοποίηση του περιθωρίου πραγματοποιείται τόσο σε σχέση με τη διεύθυνση  $w$  του διαχωριστικού υπερεπιπέδου, όσο και σε σχέση με τη σχετική του θέση  $y$  ως προς την αρχή των αξόνων. Στην εναλλακτική αυτή διατύπωση δεν απαιτείται ο περιορισμός μη αρνητικότητας του  $d$ , καθώς εάν υπάρχει κάποιο  $d$  αρνητικό τότε η αντικειμενική συνάρτηση μπορεί να μειωθεί θέτοντας  $d=0$ , κάτι που δεν παραβιάζει τον περιορισμό ανισότητας.

Το πρόβλημα (4.9) απλοποιείται ακόμα περισσότερο εάν ο περιορισμός ανισότητας μετατραπεί σε περιορισμό ισότητας ως εξής::

$$\left. \begin{array}{l} \text{Min}_{w,y,d} v \frac{1}{2} \|d\|^2 + \frac{1}{2} (w'w + y^2) \\ \text{υπό:} \\ D(Zw - ey) + d = e \end{array} \right\} \quad (4.10)$$

Το πρόβλημα αυτό μπορεί πλέον να λυθεί εύκολα με τη μέθοδο των πολλαπλασιαστών Lagrange, διαμορφώνοντας την ακόλουθη συνάρτηση:

$$L(w, y, d, u) = \frac{v}{2} \|d\|^2 + \frac{1}{2} \left\| \begin{bmatrix} w \\ y \end{bmatrix} \right\|^2 - u' (D(Zw - ey) + d - e) \quad (4.11)$$

όπου  $u$  είναι το διάνυσμα με τους πολλαπλασιαστές Langrage που αντιστοιχούν στους περιορισμούς ισότητας του προβλήματος (4.10). Θέτοντας τις μερικές παραγώγους της συνάρτησης  $L$  ίσες με το μηδέν προκύπτουν οι παρακάτω συνθήκες:

$$\left. \begin{aligned} w - Z' Du &= 0 \\ y + e' Du &= 0 \\ vd - u &= 0 \\ D(Zw - ey) + d - e &= 0 \end{aligned} \right\} \quad (4.12)$$

Εκφράζοντας τις μεταβλητές  $(w, y, d)$  συναρτήσει των  $u$  προκύπτουν:

$$w = Z' Du, y = -e' Du, d = \frac{u}{v} \quad (4.13)$$

και αντικαθιστώντας αυτές στην τελευταία ισότητα της (4.12) υπολογίζεται τα  $u$  συναρτήσει των  $Z$  και  $D$ :

$$u = \left( \frac{I}{v} + D(ZZ' + ee')D' \right)^{-1} e = \left( \frac{I}{v} + HH' \right)^{-1} e \quad (4.14),$$

όπου ο πίνακας  $H$  ορίζεται ως:  $H = D[Z - e]$  (4.15).

Αντικαθιστώντας την (4.14) στις (4.13) υπολογίζεται η βέλτιστη λύση  $(w, y, d)$  του προβλήματος (4.10).

Γενικά, ο αλγόριθμος της γραμμικής μεθόδου SVM σύμφωνα με την παραπάνω μοντελοποίηση για την ταξινόμηση  $n$  μονάδων ορίζεται ως εξής:

- (α) Ορίζεται το  $H$  από την σχέση (4.15), όπου  $e$  είναι ένα διάνυσμα από μονάδες διάστασης  $n \times 1$  και υπολογίζονται οι πολλαπλασιαστές Langrage  $u$  από την σχέση (4.14) για θετικές τιμές της μεταβλητής  $v$ .
- (β) Προσδιορίζονται τα  $(w, y)$  από τη σχέση (4.13).
- (γ) Ταξινομείται το νέο δείγμα χρησιμοποιώντας τη σχέση (4.8).

### 4.3 Σχεδιασμός του πειράματος

#### 4.3.1 Εξεταζόμενοι παράγοντες

Η σύγκριση όλων των προαναφερθέντων μεθόδων βασίζεται στην πραγματοποίηση μιας εκτεταμένης προσομοίωσης Monte Carlo. Η προσομοίωση αυτή επιτρέπει την πραγματοποίηση της σύγκρισης σε δεδομένα τα οποία διαθέτουν συγκεκριμένες στατιστικές ιδιότητες, συμβάλλοντας με τον τρόπο αυτό στην εξαγωγή των αντίστοιχων συμπερασμάτων όσον αφορά την αποτελεσματικότητα των εξεταζόμενων μεθοδολογιών στην αντιμετώπιση του προβλήματος της ταξινόμησης.

Κατά την παρούσα συγκριτική ανάλυση, η αποτελεσματικότητα των εξεταζόμενων μεθόδων ελέγχεται βάσει των ακόλουθων παραγόντων:

1. Στατιστική κατανομή των επιδόσεων των αντικειμένων (μονάδων) στα κριτήρια αξιολόγησης.
2. Μέγεθος του δείγματος εκμάθησης.
3. Ομοιογένεια των διακυμάνσεων κάθε κατηγορίας.
4. Βαθμός συσχέτισης των χαρακτηριστικών (κριτηρίων αξιολόγησης).

Ο Πίνακας 4.1 παρουσιάζει τον τρόπο με τον οποίο χρησιμοποιούνται οι παραπάνω παράγοντες στον πειραματικό σχεδιασμό που πραγματοποιείται με σκοπό την αξιολόγηση της αποτελεσματικότητας των εξεταζόμενων μεθόδων ταξινόμησης.

Πίνακας 4.1: Εξεταζόμενοι παράγοντες στον πειραματικό σχεδιασμό για την σύγκριση των μεθόδων ταξινόμησης

Παράγοντες		Επίπεδα
$\Pi_1$	Στατιστική κατανομή	1. Κανονική κατανομή (κύρτωση=0) 2. Ομοιόμορφη κατανομή (κύρτωση=-1) 3. Λογιστική κατανομή (κύρτωση=1) 4. Κατανομή Laplace (κύρτωση=3)
$\Pi_2$	Μέγεθος δείγματος εκμάθησης	1. 200 αντικείμενα, 5 κριτήρια 2. 400 αντικείμενα, 5 κριτήρια

		3. 600 αντικείμενα, 5 κριτήρια
$\Pi_3$	Ομοιογένεια διακυμάνσεων $\sigma_1^2, \sigma_2^2$	1. $\sigma_1^2=\sigma_2^2=1$ 2. $\sigma_1^2=1, \sigma_2^2=2$ 3. $\sigma_1^2=1, \sigma_2^2=4$
$\Pi_4$	Βαθμός συσχέτισης κριτηρίων	1. $\rho_{ij} \in [0, 0.3]$ 2. $\rho_{ij} \in [0.4, 0.8]$
$\Pi_5$	Τεχνικές ταξινόμησης	1. Περιβάλλουσα ανάλυση δεδομένων (DEA) 2. Γραμμική διακριτική ανάλυση (LDA) 3. Γραμμικός προγραμματισμός (LP) 4. Αλγόριθμος Πλησιέστερου γείτονα (NN) 5. Δέντρα ταξινόμησης και παλινδρόμησης (CART) 6. Μηχανές διανύσματος υποστήριξης (SVM)

Ο παράγοντας  $\Pi_1$  είναι από τους πρώτους παράγοντες που καθορίζουν την μορφή των εξεταζόμενων δεδομένων. Ο παράγοντας αυτός αναφέρεται στην στατιστική κατανομή των δεδομένων που εξετάζονται. Ο προτεινόμενος πειραματικός σχεδιασμός βασίζεται στην παραγωγή δεδομένων που ακολουθούν τέσσερις πολυμεταβλητές κατανομές οι οποίες καθορίζονται βάσει της κύρτωσής τους. Οι επιλεγμένες τιμές κύρτωσης είναι -1, 0, 1 και 3. Όταν η κύρτωση είναι ίση με το -1, τα δεδομένα κατανέμονται σχεδόν ομοιόμορφα. Όταν η κύρτωση είναι ίση με 0, τα δεδομένα ακολουθούν την κανονική κατανομή. Όταν η κύρτωση είναι ίση με 1, τα δεδομένα ακολουθούν (σχεδόν) την λογιστική κατανομή. Τέλος, όταν η κύρτωση είναι ίση με 3, τα δεδομένα ακολουθούν μια κατανομή, η οποία μοιάζει με την κατανομή Laplace (διπλή εκθετική κατανομή). Και στις τέσσερις περιπτώσεις οι εξεταζόμενες κατανομές είναι συμμετρικές.

Ο επόμενος παράγοντας  $\Pi_2$  καθορίζει το πλήθος των εναλλακτικών δραστηριοτήτων του δείγματος εκμάθησης. Στον παρόντα πειραματικό σχεδιασμό εξετάζονται τρία επίπεδα για τον παράγοντα αυτό, σύμφωνα με τα οποία το δείγμα εκμάθησης περιλαμβάνει από 200 έως 600 αντικείμενα, τα οποία περιγράφονται από ένα σύνολο πέντε χαρακτηριστικών (κριτηρίων αξιολόγησης). Καθώς το δείγμα εκμάθησης αυξάνεται προστίθεται και νέα πληροφορία, αλλά παράλληλα αυξάνεται και η πολυπλοκότητα του προβλήματος ταξινόμησης. Συνεπώς, η εξέταση του παράγοντα αυτού περιλαμβάνει την αξιολόγηση της αποτελεσματικότητας των παραπάνω προσεγγίσεων σε περιπτώσεις που η πληροφορία που παρέχεται είναι



περιορισμένη και άρα και η πολυπλοκότητα, αλλά και σε περιπτώσεις που υπάρχει αυξημένη πληροφορία και άρα και αυξημένη πολυπλοκότητα.

Ο επόμενος παράγοντας  $\Pi_3$  αναφέρεται στην ομοιογένεια των διακυμάνσεων  $\sigma_1^2$  και  $\sigma_2^2$  των χαρακτηριστικών για την κάθε κατηγορία. Στον πειραματικό αυτό σχεδιασμό εξετάζονται τρεις περιπτώσεις: (α) η περίπτωση που οι διακυμάνσεις των δύο κατηγοριών είναι ίσες (ομοιογενείς) και (β) δύο περιπτώσεις όπου οι διακυμάνσεις των δύο κατηγοριών διαφέρουν (είναι ανομοιογενείς). Συμβολίζοντας ως  $\sigma_{ki}^2$  την διακύμανση του χαρακτηριστικού  $Z_i$  για την κατηγορία  $C_k$  με  $k=1$  ή  $2$ , εξετάζονται οι παρακάτω τρεις περιπτώσεις:

$$(\alpha) \sigma_{1i}^2 = \sigma_{2i}^2 = 1, \forall i=1, 2, \dots, 5$$

$$(\beta) \sigma_{1i}^2 = 1, \sigma_{2i}^2 = 2, \forall i=1, 2, \dots, 5$$

$$(\gamma) \sigma_{1i}^2 = 1, \sigma_{2i}^2 = 4, \forall i=1, 2, \dots, 5$$

Ο τελευταίος παράγοντας  $\Pi_4$  σχετίζεται με τον βαθμό συσχέτισης των χαρακτηριστικών. Για τον παράγοντα αυτό εξετάζονται δύο περιπτώσεις. Στην πρώτη περίπτωση, θεωρείται ότι ο συντελεστής συσχέτισης είναι μικρός και κυμαίνεται μεταξύ 0 και 0.3. Αντίθετα, στη δεύτερη περίπτωση θεωρείται ένας μεγαλύτερος βαθμός συσχέτισης που κυμαίνεται μεταξύ των τιμών 0.4 και 0.8. Και στις δύο περιπτώσεις, κάθε συντελεστής συσχέτισης  $r_{ij}$  μεταξύ των χαρακτηριστικών  $Z_i$  και  $Z_j$  καθορίζεται ως μια ομοιόμορφη κατανομημένη τυχαία μεταβλητή στο αντίστοιχο διάστημα, ανάλογα με το επιθυμητό επίπεδο συσχέτισης.

#### 4.3.2 Διαδικασία παραγωγής των δεδομένων και υλοποίηση της πειραματικής ανάλυσης

Βασικό σημείο αυτού του πειραματικού σχεδιασμού είναι η παραγωγή των δεδομένων των εναλλακτικών δραστηριοτήτων έτσι ώστε αυτές να ακολουθούν όλες τις παραπάνω στατιστικές ιδιότητες, όπως την ομοιογένεια των διακυμάνσεων των κατηγοριών και τις τέσσερις προκαθορισμένες στατιστικές κατανομές. Στην περίπτωση της παραγωγής των δεδομένων που ακολουθούν κανονική κατανομή δεν υφίσταται καμία δυσκολία καθώς υπάρχουν κατάλληλες γνωστές διαδικασίες. Το πρόβλημα παρουσιάζεται στην περίπτωση της παραγωγής των δεδομένων που ακολουθούν μη κανονικές κατανομές. Για αυτή την περίπτωση χρησιμοποιείται η διαδικασία που προτάθηκε από τους Vale και Maurelli (1983). Απώτερος στόχος της υλοποίησης αυτής της διαδικασίας είναι η παραγωγή ενός διάνυσματος αποτελούμενου από  $n$  τυχαίες μεταβλητές, οι οποίες διαθέτουν τις παραπάνω στατιστικές ιδιότητες. Στο συγκεκριμένο πειραματικό σχεδιασμό στο διάνυσμα αυτό

αντιστοιχεί το σύνολο των πέντε κριτηρίων αξιολόγησης, καθένα από τα οποία θεωρείται ως μια τυχαία μεταβλητή.

Για κάθε συνδυασμό των κατηγοριών  $\Pi_1$  έως και  $\Pi_4$ , η παραπάνω διαδικασία χρησιμοποιείται για την παραγωγή δύο συνόλων δεδομένων. Το πρώτο χρησιμοποιείται ως δείγμα εκμάθησης και το δεύτερο ως δείγμα ελέγχου. Το μέγεθος του δείγματος εκμάθησης καθορίζεται από τον παράγοντα  $\Pi_2$  και περιλαμβάνει 200, 400 ή 600 εναλλακτικές, ενώ το δείγμα ελέγχου αποτελείται σε κάθε περίπτωση από 200 εναλλακτικές δραστηριότητες. Επίσης, τόσο στο δείγμα ελέγχου όσο και στο δείγμα εκμάθησης, ο αριθμός των εναλλακτικών δραστηριοτήτων των δύο κατηγοριών είναι ίδιος.

Ο παραπάνω πειραματικός έλεγχος επαναλαμβάνεται 20 φορές για κάθε συνδυασμό των παραγόντων  $\Pi_1$  έως και  $\Pi_4$  (72 δυνατοί συνδυασμοί). Συνολικά ελέγχονται 1440 διαφορετικά δείγματα εκμάθησης, στα οποία και αντιστοιχούν ισάριθμα δείγματα ελέγχου. Κάθε δείγμα εκμάθησης εφαρμόζεται στο γραμμικό πρόβλημα (3.6) με σκοπό την σωστή ταξινόμηση των παρατηρήσεων. Στην περίπτωση όπου παρατηρηθούν εσφαλμένες ταξινομήσεις τότε χρησιμοποιείται η διαδικασία επαναπροσδιορισμού των εσφαλμένων ταξινομήσεων, σύμφωνα με το γραμμικό πρόβλημα (3.7). Αυτή η διαδικασία των δύο σταδίων οδηγεί στην ανάπτυξη του βέλτιστου μοντέλου ταξινόμησης μέσω της προσέγγισης της DEA. Στη συνέχεια, το μοντέλο αυτό χρησιμοποιείται για την ταξινόμηση των παρατηρήσεων του δείγματος ελέγχου στις δύο κατηγορίες. Με βάση το δείγμα ελέγχου εξετάστηκε η αποτελεσματικότητα της μεθόδου DEA στην ταξινόμηση του δείγματος ελέγχου. Με τον ίδιο τρόπο εφαρμόζονται αντίστοιχα και όλες οι υπόλοιπες τεχνικές ταξινόμησης που αναφέρονται στον παράγοντα  $\Pi_5$  (πίνακας 4.1).

Η υλοποίηση αυτού του πειραματικού σχεδιασμού πραγματοποιήθηκε στο περιβάλλον του Microsoft Visual C++ 6.0, όπου κατασκευάστηκε ο αλγόριθμος της μεθόδου DEA με την βοήθεια της γλώσσας προγραμματισμού C και του Matlab 5.3, για την εφαρμογή των υπολοίπων μεθόδων. Το περιβάλλον του Matlab χρησιμοποιήθηκε επίσης και για την παραγωγή των δεδομένων. Η στατιστική ανάλυση και επεξεργασία των αποτελεσμάτων πραγματοποιήθηκε στο στατιστικό πακέτο SPSS 10.

#### 4.4 Ανάλυση των αποτελεσμάτων

Η ανάλυση των αποτελεσμάτων που επιτυγχάνεται για κάθε μέθοδο ταξινόμησης αφορά μόνο το ποσοστό των εσφαλμένων ταξινομήσεων του δείγματος ελέγχου, καθώς αυτά εξηγούν πληρέστερα την ικανότητα των αντίστοιχων μεθόδων στην αντιμετώπιση προβλημάτων ταξινόμησης. Η επεξεργασία των αποτελεσμάτων βασίζεται στην ανάλυση διασποράς (ANOVA) σε συνδυασμό με τον στατιστικό έλεγχο του Tukey ο οποίος επιτρέπει τη διαμόρφωση ομοιογενών ομάδων όπου κάθε μια περιλαμβάνει τα επίπεδα ενός παράγοντα για τα οποία δεν παρουσιάζουν στατιστικά σημαντικές διαφορές μεταξύ τους, ως προς το ποσοστό των εσφαλμένων ταξινομήσεων. Συνεπώς, η αξιολόγηση των αποτελεσμάτων του συγκεκριμένου πειραματικού σχεδιασμού επιτρέπει την εξαγωγή αντικειμενικών εκτιμήσεων όσον αφορά την αναμενόμενη αποτελεσματικότητα των εξεταζόμενων μεθόδων στην πρακτική αντιμετώπιση των προβλημάτων ταξινόμησης.

Ο πίνακας 4.2 παρουσιάζει τα αποτελέσματα της στατιστικής ανάλυσης διασποράς των πέντε εξεταζόμενων παραγόντων, στα αποτελέσματα ταξινόμησης του δείγματος ελέγχου.

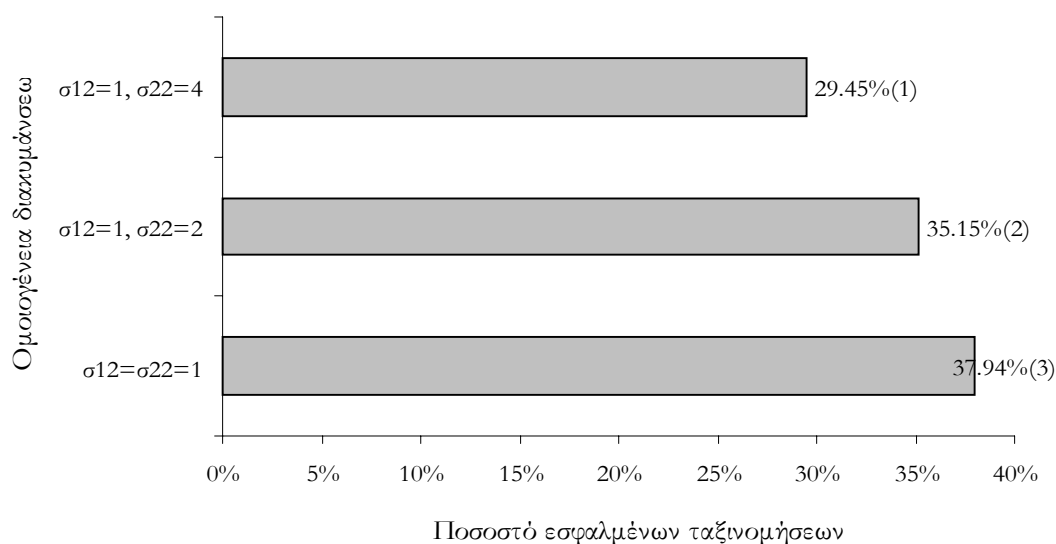
Πίνακας 4.2: Αποτελέσματα στατιστικής ανάλυσης του πειραματικού σχεδιασμού του δείγματος ελέγχου

1-Κατανομή, 2-Μέγεθος, 3-Διασπορές, 4-Συσχετίσεις, 5-Μέθοδοι				
Κύριες επιδράσεις	Βαθμοί ελευθερίας	Μέσο άθροισμα Τετραγώνων	F	$\omega^2$
3	2	4.877	3966.596	22.77%
35	10	0.870	707.720	20.29%
4	1	4.957	4030.864	11.57%
15	15	0.265	215.316	9.23%
5	5	0.478	388.976	5.57%
135	30	0.026	21.385	1.76%
2	2	0.181	146.957	0.84%
13	6	0.058	47.046	0.79%
1	3	0.086	70.114	0.60%

Ο πίνακας 4.2 επικεντρώνεται στην παρουσίαση των κύριων επιδράσεων καθώς και των αλληλεπιδράσεων των εξεταζόμενων παραγόντων, οι οποίες είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 1% και εξηγούν τουλάχιστον το 0.5% της συνολικής διακύμανσης των αποτελεσμάτων (βάσει του στατιστικού  $\omega^2$ ). Οι υπόλοιπες αλληλεπιδράσεις δεν εξηγούν

περισσότερο από το 0.5% της συνολικής διακύμανσης των αποτελεσμάτων. Συνεπώς, η πειραματική ανάλυση που θα πραγματοποιηθεί βασίζεται μόνο στις πλέον σημαντικές επιδράσεις και αλληλεπιδράσεις του παραπάνω πίνακα, έτσι ώστε να διευκολυνθεί η διαδικασία ανάλυσης των αποτελεσμάτων του πειραματικού σχεδιασμού. Καθένας από τους παραπάνω παράγοντες επιδρά σημαντικά στην αποτελεσματικότητα με την οποία μπορεί να αντιμετωπιστεί το πρόβλημα ταξινόμησης όσον αφορά τη δυνατότητα γενίκευσης των αναπτυσσόμενων υποδειγμάτων στο δείγμα ελέγχου.

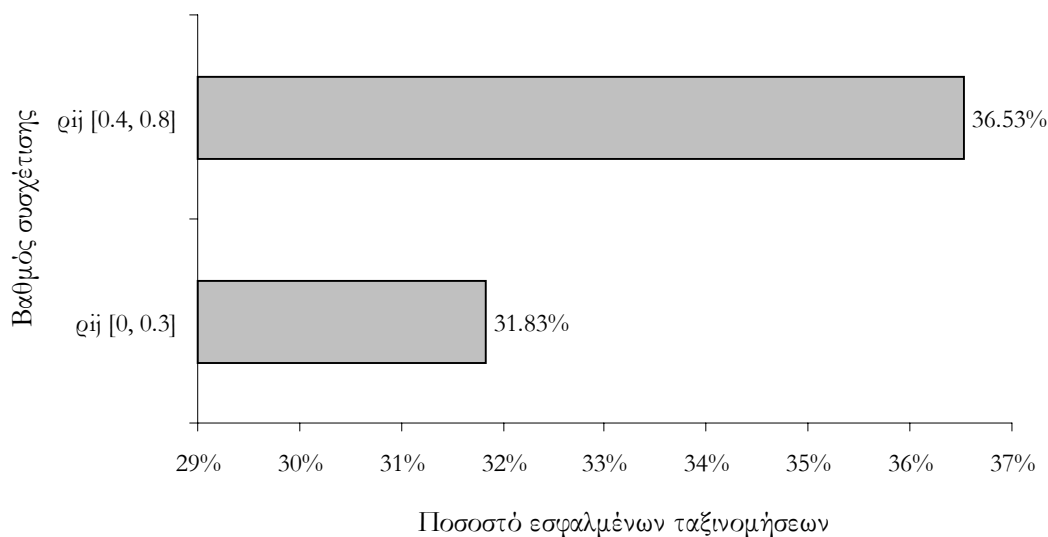
Βάσει των αποτελεσμάτων του Πίνακα 4.2, διαπιστώνεται ότι ο παράγοντας που αφορά την ομοιογένεια των διακυμάνσεων των κατηγοριών είναι ο πλέον σημαντικός στην ανάλυση των αποτελεσμάτων, καθώς εξηγεί περισσότερο από το 22% της συνολικής διακύμανσης των αποτελεσμάτων του πειραματικού σχεδιασμού. Στο σχήμα 4.3 απεικονίζονται τα αντίστοιχα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό εσφαλμένων ταξινομήσεων στο δείγμα ελέγχου (όλων των μεθόδων) για τα επιμέρους επίπεδα του παράγοντα αυτού. Στις αντίστοιχες παρενθέσεις παρουσιάζεται η ομαδοποίηση των διαφόρων μορφών διακύμανσης σύμφωνα με τον στατιστικό έλεγχο Tukey, σε επίπεδο σημαντικότητα 5%.



Σχήμα 4 3: Μέσο ποσοστό εσφαλμένων ταξινομήσεων συναρτήσει της ομοιογένειας στις διακυμάνσεις των κατηγοριών

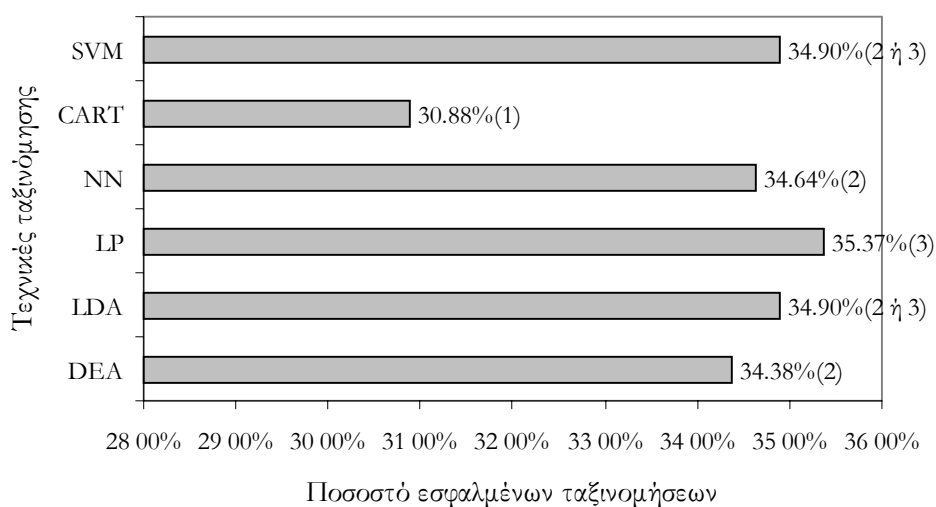
Από το σχήμα 4.3 φαίνεται ότι το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων παρουσιάζεται στην περίπτωση όπου οι διακυμάνσεις των δύο κατηγοριών διαφέρουν σημαντικά, ενώ μεγαλύτερο σφάλμα εντοπίζεται στην περίπτωση όπου οι διακυμάνσεις των δύο κατηγοριών είναι ίσες. Συνεπώς, όσο αυξάνεται η διαφορά των διακυμάνσεων μεταξύ των κατηγοριών ταξινόμησης, τόσο αυξάνεται και η αποτελεσματικότητα των εξεταζόμενων μεθόδων. Αυτό βέβαια ισχύει στην περίπτωση που η διακύμανση εξετάζεται ανεξάρτητα από τους υπόλοιπους παράγοντες. Περαιτέρω ανάλυση του θέματος αυτού θα πραγματοποιηθεί παρακάτω θεωρώντας τη διακύμανση σε συνδυασμό με άλλους παράγοντες που εξετάζονται στον πειραματικό σχεδιασμό.

Μία δεύτερη κύρια επίδραση που αφορά το βαθμό συσχέτισης των επιδόσεων των εναλλακτικών δραστηριοτήτων στα επιμέρους κριτήρια αξιολόγησης εξηγεί περισσότερο από το 11% της συνολικής διακύμανσης των αποτελεσμάτων του πειραματικού σχεδιασμού. Στο σχήμα 4.4 απεικονίζονται τα αντίστοιχα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό εσφαλμένων ταξινομήσεων των διαφόρων βαθμών συσχέτισης στο δείγμα ελέγχου. Από το σχήμα φαίνεται ότι το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων εντοπίζεται στην περίπτωση που έχουμε μικρό βαθμό συσχέτισης.



Σχήμα 4 4: Μέσο ποσοστό εσφαλμένων ταξινομήσεων συναρτήσει του συντελεστή συσχέτισης

Μία τρίτη κύρια επίδραση, με αυξημένο ενδιαφέρον, αφορά τη σύγκριση της αποτελεσματικότητας των εξεταζόμενων τεχνικών ταξινόμησης και εξηγεί περισσότερο από το 6% της συνολικής διακύμανσης των αποτελεσμάτων του πειραματικού σχεδιασμού. Στο σχήμα 4.6 συνοψίζονται τα αποτελέσματα του πειραματικού σχεδιασμού σχετικά με το μέσο ποσοστό εσφαλμένων ταξινομήσεων όλων των εξεταζόμενων τεχνικών στο δείγμα ελέγχου. Τα αποτελέσματα δείχνουν την υψηλή αποτελεσματικότητα που παρουσιάζει η προσέγγιση CART έναντι των υπολοίπων μεθοδολογικών προσεγγίσεων στο σωστό εντοπισμό της ταξινόμησης των εναλλακτικών δραστηριοτήτων σε σύγκριση με την προκαθορισμένη ταξινόμησή τους. Με σημαντικά μεγαλύτερο σφάλμα ταξινόμησης από την CART ακολουθούν οι προσεγγίσεις DEA και NN και στη συνέχεια LDA και SVM που αποδίδουν παρόμοια αποτελέσματα. Την χαμηλότερη αποτελεσματικότητα παρουσιάζει η μέθοδος LP. Επομένως, βάσει αυτών των συμπερασμάτων η DEA ταξινομείται στη δεύτερη θέση με το δεύτερο μικρότερο σφάλμα ταξινόμησης σε σχέση με τις υπόλοιπες μεθοδολογικές προσεγγίσεις ταξινόμησης.



Σχήμα 4 5: Μέσο ποσοστό εσφαλμένων ταξινομήσεων των εξεταζόμενων τεχνικών ταξινόμησης στο δείγμα ελέγχου

Τα αποτελέσματα που προκύπτουν από τις επιδράσεις των εξεταζόμενων παραγόντων στο σύνολο του πειραματικού σχεδιασμού αποδίδουν την γενική εικόνα της αποτελεσματικότητας του κάθε παράγοντα χωριστά, αλλά δεν συμβάλλουν στον τρόπο με τον οποίο αλληλεπιδρούν οι εξεταζόμενοι παράγοντες. Για αυτό κατά την διάρκεια της

πειραματικής ανάλυσης θα συνδυαστούν κάποιοι παράγοντες, η αλληλεπίδραση των οποίων είναι σημαντική στην αποτελεσματικότητα του προβλήματος ταξινόμησης.

Η πιο σημαντική από τις αλληλεπιδράσεις αυτές είναι η αλληλεπίδραση της ομοιογένειας των διακυμάνσεων  $\sigma_1^2$  και  $\sigma_2^2$  των επιδόσεων των εναλλακτικών δραστηριοτήτων κάθε κατηγορίας στα κριτήρια αξιολόγησης με τις εξεταζόμενες προσεγγίσεις. Η αλληλεπίδραση αυτή εξηγεί το 20.29% της συνολικής διακύμανσης των αποτελεσμάτων του πειραματικού σχεδιασμού. Τα σχετικά αποτελέσματα παρουσιάζονται στον πίνακα 4.3. Από τα αποτελέσματα του πίνακα 4.3 φαίνεται ότι η DEA βρίσκεται στην κατηγορία με το χαμηλότερο ποσοστό εσφαλμένων ταξινομήσεων μόνο στην περίπτωση που υπάρχει ίση ομοιογένεια στις διακυμάνσεις  $\sigma_1^2$  και  $\sigma_2^2$  κάθε κατηγορίας. Στην περίπτωση που οι διακυμάνσεις διαφέρουν, τότε η αποτελεσματικότητα της DEA σε σχέση με τις μεθόδους CART και NN μειώνεται, ενώ αντίθετα συγκριτικά με τις άλλες μεθόδους η αποτελεσματικότητα της DEA βελτιώνεται. Αξιοσημείωτη είναι η σημαντική βελτίωση της αποτελεσματικότητας των μεθόδων CART και NN με την αύξηση της ομοιογένειας των διακυμάνσεων, γεγονός που δικαιολογείται από τον έντονα μη γραμμικό χαρακτήρα των μεθόδων. Θα πρέπει πάντως να σημειωθεί ότι οι αυτές οι δύο τεχνικές (CART και NN) αποδίδουν συγκριτικά χειρότερα αποτελέσματα όταν υπάρχει ομοιογένεια στις διακυμάνσεις των κατηγοριών.

Πίνακας 4.3: Μέσο ποσοστό εσφαλμένων ταξινομήσεων των εξεταζόμενων προσεγγίσεων ταξινόμησης συναρτήσει της ομοιογένειας των διακυμάνσεων

<i>Εξεταζόμενες προσεγγίσεις</i>	<i>Ομοιογένεια διακυμάνσεων</i>		
	$\sigma_1^2=\sigma_2^2=1$	$\sigma_1^2=1, \sigma_2^2=2$	$\sigma_1^2=1, \sigma_2^2=4$
DEA	36.36% (1)	35.16% (2)	31.62% (3)
LDA	36.22% (1)	35.56% (2)	32.90% (4)
LP	36.27% (1)	35.81% (2 ή 3)	34.05% (5)
NN	42.99% (3)	36.71% (3)	24.21% (2)
CART	39.58% (2)	32.07% (1)	20.79% (1)
SVM	36.21% (1)	35.57% (2)	32.91% (4)

Δεύτερη ιδιαίτερα σημαντική αλληλεπίδραση είναι αυτή μεταξύ των εξεταζόμενων προσεγγίσεων και της στατιστικής κατανομής των επιδόσεων των εναλλακτικών δραστηριοτήτων του δείγματος ελέγχου στα κριτήρια αξιολόγησης. Η αλληλεπίδραση αυτή

εξηγεί περισσότερο από το 9% της συνολικής διακύμανσης των αποτελεσμάτων του πειραματικού σχεδιασμού. Τα σχετικά αποτελέσματα παρουσιάζονται στον πίνακα 4.4. Τα αποτελέσματα αυτά δείχνουν ότι η μέθοδος DEA παρουσιάζει το μικρότερο σφάλμα ταξινόμησης στις περιπτώσεις της λογιστικής κατανομή και της κατανομής Laplace η οποία προσομοιώνει την ύπαρξη μη φυσιολογικών δεδομένων στο δείγμα (outliers). Στην κανονική ή την ομοιόμορφη κατανομή το ποσοστό σφάλματος της DEA αυξάνεται κατά 1% ή 2% με αποτέλεσμα να μειώνεται η αποτελεσματικότητα ταξινόμησης της μεθόδου. Στην κανονική κατανομή, η CART είναι η προσέγγιση με το μικρότερο σφάλμα, ενώ οι υπόλοιπες προσεγγίσεις ταξινομούνται στην δεύτερη θέση χωρίς ιδιαίτερες διαφοροποιήσεις μεταξύ τους. Παρόμοια αποτελέσματα παρατηρούνται και στην περίπτωση της ομοιόμορφης κατανομής. Στην περίπτωση όμως αυτή οι διαφορές μεταξύ της μεθόδου CART και των υπολοίπων τεχνικών αυξάνονται. Αξιοσημείωτο είναι επίσης το γεγονός, ότι η αύξηση της κύρωσης των δεδομένων οδηγεί σε βελτίωση της αποτελεσματικότητας των DEA, LP, LDA και SVM, ενώ αντίθετα τα ποσοστά σφάλματος για τις μεθόδους CART και NN παρουσιάζουν αύξηση.

Πίνακας 4.4: Μέσο ποσοστό εσφαλμένων ταξινόμησης των εξεταζόμενων προσεγγίσεων ταξινόμησης συναρτήσει της στατιστικής κατανομής

Εξεταζόμενες προσεγγίσεις	Στατιστική κατανομή			
	Κανονική κατανομή (κύρωση=0)	Ομοιόμορφη κατανομή (κύρωση=-1)	Λογιστική κατανομή (κύρωση=1)	Κατανομή Laplace (κύρωση=3)
DEA	34.66% (2)	35.97% (3)	33.84% (1 ή 2)	33.05% (1)
LDA	35.30% (2)	36.58% (3)	34.37% (1 ή 2)	33.34% (1)
LP	35.63% (2)	37.04% (3)	34.97% (2 ή 3)	33.86% (1)
NN	35.13% (2)	31.67% (2)	36.02% (3)	35.73% (2)
CART	32.99% (1)	23.30% (1)	33.52% (1)	33.52% (1)
SVM	35.29% (2)	36.58% (3)	34.37% (1 ή 2)	33.36% (1)

Μια τελευταία σημαντική αλληλεπίδραση δευτέρου βαθμού είναι αυτή μεταξύ της στατιστικής κατανομής των δεδομένων και της ομοιογένειας στις διακυμάνσεις των κατηγοριών. Τα σχετικά αποτελέσματα παρουσιάζονται στον πίνακα 4.5. Όπως παρατηρείται, βάσει των αποτελεσμάτων, μικρότερο ποσοστό ταξινόμησης επιτυγχάνεται στην περίπτωση που συνδυάζεται οποιαδήποτε μορφή στατιστικής κατανομής με άνισες διακυμάνσεις μεταξύ των κατηγοριών ταξινόμησης. Μάλιστα, όσο μεγαλύτερη είναι η



διαφορά μεταξύ των διακυμάνσεων τόσο μικρότερο σφάλμα ταξινόμησης επιτυγχάνεται. Επίσης, όπως παρατηρήθηκε από την ανάλυση της κύριας επίδρασης που αφορά την ομοιογένεια των διακυμάνσεων (σχήμα 4.3) το μικρότερο ποσοστό εσφαλμένων ταξινόμησης παρουσιάζεται στην περίπτωση όπου οι διακυμάνσεις των δύο κατηγοριών διαφέρουν σημαντικά. Μεγαλύτερες βελτιώσεις στα σφάλματα ταξινόμησης παρατηρούνται στις περιπτώσεις της κανονικής, ομοιόμορφης και λογιστικής κατανομής, όπου η διαφορά μεταξύ των περιπτώσεων  $\sigma_1^2=\sigma_2^2=1$  και  $\sigma_1^2=1, \sigma_2^2=4$  είναι περίπου της τάξης του 10%. Αντίθετα στην περίπτωση της κατανομής Laplace η επίδραση της ανομοιογένειας των διακυμάνσεων εμφανίζεται περιορισμένη συγκριτικά με τις υπόλοιπες κατανομές.

Πίνακας 4.5: Μέσο ποσοστό εσφαλμένων ταξινόμησης συναρτήσει της στατιστικής κατανομής και της ομοιογένειας των διακυμάνσεων

	Στατιστική κατανομή			
	Κανονική κατανομή (κύρτωση=0)	Ομοιόμορφη κατανομή (κύρτωση=-1)	Λογιστική κατανομή (κύρτωση=1)	Κατανομή Laplace (κύρτωση=3)
$\sigma_1^2=\sigma_2^2=1$	38.90% (3)	38.32% (3)	38.17% (3)	36.37% (3)
$\sigma_1^2=1, \sigma_2^2=2$	35.51% (2)	34.47% (2)	35.40% (2)	35.21% (2)
$\sigma_1^2=1, \sigma_2^2=4$	30.09% (1)	27.84% (1)	29.98% (1)	29.85% (1)

Εκτός των παραπάνω αλληλεπιδράσεων δευτέρου βαθμού, η ανάλυση διασποράς υποδεικνύει ως σημαντική και μία αλληλεπίδραση τρίτου βαθμού για την εξήγηση των αποτελεσμάτων του παρόντος πειραματικού σχεδιασμού. Η αλληλεπίδραση αυτή αφορά τους παράγοντες που σχετίζονται με τις εξεταζόμενες προσεγγίσεις ταξινόμησης, την στατιστική κατανομή και την ομοιογένεια στις διακυμάνσεις των κατηγοριών. Τα αντίστοιχα αποτελέσματα συνοψίζονται στον πίνακα 4.6 και επιτρέπουν την πληρέστερη ανάλυση των συμπερασμάτων που παρουσιάστηκαν προηγουμένως όσον αφορά την επίδραση της στατιστικής κατανομής των δεδομένων και της ομοιογένειας των κατηγοριών ταξινόμησης στην αποτελεσματικότητα των εξεταζόμενων προσεγγίσεων ταξινόμησης.

Πίνακας 4.6: Μέσο ποσοστό εσφαλμένων ταξινόμησεων των εξεταζόμενων προσεγγίσεων ταξινόμησης συναρτήσει της στατιστικής κατανομής και της ομοιογένειας των διακυμάνσεων

Κανονική κατανομή (κύρτωση=0)			
	Ομοιογένεια διακυμάνσεων		
Εξεταζόμενες προσεγγίσεις	$\sigma_1^2=\sigma_2^2=1$	$\sigma_1^2=1, \sigma_2^2=2$	$\sigma_1^2=1, \sigma_2^2=4$
DEA	36.72% (1)	35.18% (1)	32.09% (3)
LDA	36.79% (1)	35.54% (1 ή 2)	33.56% (3 ή 4)
LP	36.75% (1)	35.60% (1 ή 2)	34.56% (4)
NN	44.02% (3)	37.15% (2)	24.21% (2)
CART	42.37% (2)	34.04% (1)	22.55% (1)
SVM	36.77% (1)	35.54% (1 ή 2)	33.55% (3 ή 4)
Ομοιόμορφη κατανομή (κύρτωση=-1)			
	Ομοιογένεια διακυμάνσεων		
Εξεταζόμενες προσεγγίσεις	$\sigma_1^2=\sigma_2^2=1$	$\sigma_1^2=1, \sigma_2^2=2$	$\sigma_1^2=1, \sigma_2^2=4$
DEA	38.11% (2)	36.67% (3)	33.13% (3)
LDA	37.68% (1 ή 2)	37.31% (3)	34.74% (4)
LP	37.80% (2)	37.47% (3)	35.84% (4)
NN	42.57% (3)	34.13% (2)	18.32% (2)
CART	36.08% (1)	23.93% (1)	8.677% (1)
SVM	37.67% (1 ή 2)	37.32% (3)	34.75% (4)
Λογιστική κατανομή (κύρτωση=1)			
	Ομοιογένεια διακυμάνσεων		
Εξεταζόμενες προσεγγίσεις	$\sigma_1^2=\sigma_2^2=1$	$\sigma_1^2=1, \sigma_2^2=2$	$\sigma_1^2=1, \sigma_2^2=4$
DEA	36.01% (1)	34.51% (1)	31% (3)
LDA	36.08% (1)	34.92% (1)	32.10% (3 ή 4)
LP	36.05% (1)	35.55% (1)	33.31% (4)
NN	43.76% (3)	37.57% (2)	26.74% (2)
CART	41.03% (2)	34.92% (1)	24.60% (1)
SVM	36.06% (1)	34.92% (1)	32.12% (3 ή 4)
Κατανομή Laplace (κύρτωση=3)			
	Ομοιογένεια διακυμάνσεων		

Εξεταζόμενες προσεγγίσεις	$\sigma_1^2=\sigma_2^2=1$	$\sigma_1^2=1, \sigma_2^2=2$	$\sigma_1^2=1, \sigma_2^2=4$
DEA	34.61% (1)	34.27% (1)	30.28% (2)
LDA	34.34% (1)	34.48% (1)	31.20% (2 ή 3)
LP	34.48% (1)	34.61% (1)	32.48% (3)
NN	41.62% (2)	37.99% (2)	27.59% (1)
CART	38.83% (2)	35.40% (1)	26.34% (1)
SVM	34.33% (1)	34.51% (1)	31.22% (2 ή 3)

Η συνδυασμένη εξέταση των παραγόντων που αφορούν την στατιστική κατανομή και την ομοιογένεια των διακυμάνσεων, επιβεβαιώνει τη βελτίωση της αποτελεσματικότητας της μεθόδου DEA στις περιπτώσεις της λογιστικής κατανομής και της κατανομής Laplace, κάτι το οποίο παρατηρήθηκε και κατά την ανάλυση της των εξεταζόμενων προσεγγίσεων συναρτήσει μόνο της στατιστικής κατανομής (πίνακας 4.4). Συνεπώς, πιστοποιείται για ακόμη μια φορά ότι όσο αυξάνεται η τιμή της κύρτωσης, τόσο αυξάνεται η αποτελεσματικότητα της DEA παρουσιάζοντας όλο και μικρό ποσοστό εσφαλμένων ταξινόμησεων. Γενικότερα, μεγαλύτερη τιμή κύρτωσης δεν συνεπάγεται μόνο μικρό σφάλμα της DEA, αλλά και όλων των άλλων μεθόδων. Επίσης, παρατηρείται ότι η ομοιογένεια των δύο κατηγοριών ή η ανομοιογένειά τους με μικρή βέβαια διαφορά διακυμάνσεων μεταξύ τους σε συνδυασμό με τιμή κύρτωσης μεγαλύτερη ή ίση του μηδενός συνεπάγεται υψηλή αποτελεσματικότητα της DEA σχέση με όλες τις άλλες προσεγγίσεις. Αυτό είναι αναμενόμενο, καθώς η DEA βάσει των αποτελεσμάτων που προέκυψαν από την ανάλυση των εξεταζόμενων προσεγγίσεων συναρτήσει της ομοιογένειας των διακυμάνσεων (πίνακας 4.4) παρουσίασε μικρό σφάλμα ταξινόμησης, μόνο στην περίπτωση των ίσων διακυμάνσεων. Συνεπώς, ο συνδυασμός μεγάλης τιμής κύρτωσης και ίσων διακυμάνσεων μεταξύ των κατηγοριών ή ακόμα και άνισων διακυμάνσεων μικρής διαφοράς, δηλαδή με τιμές  $\sigma_1^2=1$  και  $\sigma_2^2=2$ , βελτιώνει ακόμα περισσότερο την αποτελεσματικότητα ταξινόμησης της μεθόδου και συμβάλλει στην καλύτερη εφαρμογή και ταξινόμηση του δείγματος ελέγχου. Για την περίπτωση των ύπαρξης μεγάλων διακυμάνσεων μεταξύ των κατηγοριών κατά την ταξινόμηση του δείγματος ελέγχου, η μέθοδος CART και ο αλγόριθμος NN θεωρούνται ως οι καλύτερες προσεγγίσεις με το μικρότερο σφάλμα ταξινόμησης, οι οποίες όμως παρουσιάζουν σημαντική «ευαισθησία» στην ομοιογένεια των διακυμάνσεων σε όλες τις περιπτώσεις στατιστικών κατανομών.

## 4.5 Βασικές επισημάνσεις

Ο πειραματικός σχεδιασμός που παρουσιάστηκε παραπάνω συνέβαλε στην αναλυτική διερεύνηση της αποτελεσματικότητας της μεθόδου ταξινόμησης DEA σε σύγκριση με άλλες εναλλακτικές τεχνικές ταξινόμησης. Τα αποτελέσματα του πειραματικού σχεδιασμού βοήθησαν στην εξαγωγή χρήσιμων πληροφοριών σχετικών με την επίδραση διαφόρων παραγόντων στην αποτελεσματικότητα της DEA σε σύγκριση με τις άλλες τεχνικές. Συνολικά, τα βασικά συμπεράσματα και οι κύριες επισημάνσεις του πειραματικού αυτού σχεδιασμού παρουσιάζονται παρακάτω.

Η DEA ως μέθοδος ταξινόμησης μπορεί να θεωρηθεί ως αποτελεσματική, σημειώνοντας την υψηλότερη αποτελεσματικότητα ταξινόμησης με το μικρότερο ποσοστό εσφαλμένων ταξινομήσεων, σε διάφορους συνδυασμούς των εξεταζόμενων παραγόντων. Ειδικότερα, για ίσες διακυμάνσεις των κατηγοριών και για μεγάλες τιμές κύρτωσης, όπως στη λογιστική κατανομή και την κατανομή Laplace, η DEA κατατάσσεται στις καλύτερες μεθόδους ταξινόμησης. Για μεγάλες διακυμάνσεις μεταξύ των κατηγοριών και μικρότερες τιμές κύρτωσης (ομοιόμορφη κατανομή), η μέθοδος CART θεωρείται από τις πιο αποτελεσματικές μεθόδους με τη μέθοδο DEA στη δεύτερη ή τρίτη θέση της κατάταξης των καλύτερων τεχνικών ταξινόμησης. Επίσης, βάσει της παραπάνω πειραματικής ανάλυσης παρατηρείται ότι καμία από τις τεχνικές ταξινόμησης δεν επηρεάζεται από το μέγεθος του δείγματος των παρατηρήσεων. Όσο, λοιπόν και πολύπλοκο γίνεται ένα πρόβλημα ταξινόμησης, τα αποτελέσματα δεν επηρεάζονται σημαντικά από τον αριθμό των παρατηρήσεων που πρέπει να εξεταστούν, ενώ ταυτόχρονα ο υπολογιστικός φόρτος όλων των εξεταζόμενων μεθόδων είναι γενικά περιορισμένος.

Μία συνοπτική ανάλυση όλων των διμερών συγκρίσεων των εξεταζόμενων μεθόδων ταξινόμησης σε όλους τους συνδυασμούς των παραγόντων του πειραματικού σχεδιασμού παρουσιάζεται στον πίνακα 4.7. Για κάθε ζεύγος μεθόδων (A, B) παρουσιάζεται το ποσοστό των επαναλήψεων (1440 επαναλήψεις), στις οποίες η μέθοδος A αποδίδει χαμηλότερο ποσοστό σφάλματος στο δείγμα ελέγχου έναντι της μεθόδου B με την οποία συγκρίνεται. Για παράδειγμα, το στοιχείο (DEA, LDA) υποδεικνύει ότι στο 53.06% του συνόλου των επαναλήψεων (764 επαναλήψεις στις 1440), το ποσοστό των εσφαλμένων ταξινομήσεων της DEA στο δείγμα ελέγχου ήταν μικρότερο από το αντίστοιχο ποσοστό της LDA. Γενικότερα, όπως παρατηρείται βάσει του πίνακα 4.7 η DEA ως προσέγγιση ταξινόμησης

στην πλειοψηφία των περιπτώσεων υπερείχε έναντι των υπολοίπων τεχνικών ταξινόμησης. Η μόνη τεχνική που μπορεί να θεωρείται καλύτερη από την DEA είναι η μέθοδος CART, καθώς στο 57.85% του συνόλου των επαναλήψεων το ποσοστό εσφαλμένων ταξινομήσεων της CART στο δείγμα ελέγχου ήταν μικρότερο από το αντίστοιχο ποσοστό της DEA. Η DEA, λοιπόν, έχει ένα «δυνατό αντίπαλο» όσον αφορά τη αποτελεσματικότητα της ταξινόμησης και αυτός είναι η μέθοδος CART. Σε σύγκριση με τις υπόλοιπες τεχνικές ταξινόμησης η DEA υπερέχει με διαφορά έναντι των υπολοίπων τεχνικών, καθώς υποδεικνύει το μικρότερο ποσοστό ταξινόμησης. Μετά την DEA ως αμέσως αποτελεσματικότερες τεχνικές ταξινόμησης θεωρούνται με την σειρά οι εξής τεχνικές: SVM, LDA, LP και NN.

Πίνακας 4.7: Διμερής σύγκριση των αποτελεσμάτων των προσεγγίσεων ταξινόμησης στο δείγμα ελέγχου

	<i>Εξεταζόμενες προσεγγίσεις</i>					
<i>Εξεταζόμενες προσεγγίσεις</i>	DEA	LDA	LP	NN	CART	SVM
DEA	–	53.06%	59.79%	53.96%	39.38%	53.13%
LDA	33.89%	–	53.13%	52.22%	38.61%	2.92%
LP	30.56%	32.08%	–	51.53%	37.50%	31.94%
NN	43.82%	45.35%	47.43%	–	22.57%	45.35%
CART	57.85%	58.96%	60.21%	73.89%	–	59.10%
SVM	34.17%	2.92%	53.26%	52.01%	38.68%	–

Στον πίνακα 4.8 συνοψίζονται τα ποσοστά της αποτελεσματικότητας της κάθε μεθόδου σύμφωνα με τα συνολικά ποσοστά σφάλματος. Ειδικότερα, σε κάθε επανάληψη του πειράματος (1440 επαναλήψεις) πραγματοποιείται μια κατάταξη των 6 εξεταζόμενων μεθόδων σε μια κλίμακα από 1 έως το 6, όπου το 1 αντιστοιχεί στην μέθοδο με το μικρότερο σφάλμα ταξινόμησης και το 6 στο υψηλότερο σφάλμα ταξινόμησης. Βάσει αυτής της θεώρησης, για την κάθε μέθοδο ο πίνακας 4.8 παρουσιάζει τη συχνότητα (ποσοστό των επαναλήψεων) με την οποία κάθε μέθοδος εμφανίζεται σε κάθε μία από τις 6 θέσης της παραπάνω κλίμακας κατάταξης. Όπως φαίνεται από τα αποτελέσματα, η DEA στο 18.82% του συνόλου επαναλήψεων του πειράματος τοποθετήθηκε στην πρώτη θέση. Η μέθοδος CART σημειώνει την υψηλότερη συχνότητα (49.44%) κατάταξης στην πρώτη θέση μεταξύ

των μεθόδων φέρνοντας της DEA στην δεύτερη θέση ακολουθούμενη με μεγάλη σχετική διαφορά από τις υπόλοιπες προσεγγίσεις. Επιπλέον, πρέπει να σημειωθεί ότι τόσο η DEA όσο και η CART σε πολύ λίγες περιπτώσεις κατατάσσονται τελευταίες, καθώς οι αντίστοιχες συχνότητες είναι περίπου στο 9%. Γενικότερα σε σύγκριση με την CART, η DEA παρουσιάζει πιο σταθερή συμπεριφορά. Για παράδειγμα, η CART με ποσοστό 49.44% βρίσκεται στην πρώτη θέση, με μικρά ποσοστά βρίσκεται στη δεύτερη, τρίτη και τέταρτη θέση και με μεγάλο ποσοστό μπορεί να τοποθετηθεί στην πέμπτη θέση. Αντιθέτως, η DEA παρουσιάζει υψηλότερα συχνότητες εμφάνισης στις τρεις πρώτες θέσεις της κατάταξης. Επίσης, αν αθροιστούν τα ποσοστά για τις τρεις πρώτες κατάταξης των εξεταζόμενων προσεγγίσεων προκύπτουν τα εξής: DEA: 64.23%, LDA: 33.82%, LP: 40.21%, NN: 47.85%, CART: 64.65% και SVM: 33.82%. Επομένως, βάσει των παραπάνω υπολογισμών οι μέθοδοι που έχουν την μεγαλύτερη πιθανότητα να τοποθετηθούν στις τρεις πρώτες θέσεις είναι η CART με πιθανότητα 64.65%, ακολουθεί η DEA με πιθανότητα 64.23%, η NN με 47.85%, η LP με 40.21% και τελευταίες οι LDA και SVM με την χαμηλότερη πιθανότητα 33.82%. Και πάλι παρατηρείται ότι η CART είναι ένας «δυνατός αντίπαλος» της DEA.

Πίνακας 4.8: Συνολικά ποσοστά της αποτελεσματικότητας της κάθε προσέγγισης ταξινόμησης

	<i>Εξεταζόμενες προσεγγίσεις</i>					
<i>Κατάταξη</i>	DEA	LDA	LP	NN	CART	SVM
1	18.82%	14.93%	16.53%	12.64%	49.44%	15.21%
2	15.69%	0.49%	11.60%	30.07%	9.58%	0.90%
3	29.72%	18.40%	12.08%	5.14%	5.63%	17.71%
4	13.82%	22.92%	16.53%	8.26%	9.03%	22.57%
5	13.19%	30.69%	12.85%	16.67%	17.08%	30.69%
6	8.75%	12.57%	30.42%	27.22%	9.24%	12.92%

## 4.6 Περίληψη

Σε αυτό το κεφάλαιο έγινε μια γενική περιγραφή των εξεταζόμενων τεχνικών ταξινόμησης και των παραγόντων που επιδρούν στη διαδικασία της αποτελεσματικότητας της ταξινόμησης. Επίσης, παρουσιάστηκε ο πειραματικός σχεδιασμός των τεχνικών ταξινόμησης και αναλύθηκαν τα αποτελέσματα που προέκυψαν. Στην παραπάνω ανάλυση των

αποτελεσμάτων σκοπός ήταν η σύγκριση της μεθόδου ταξινόμησης DEA με άλλες εναλλακτικές προσεγγίσεις. Τα συμπεράσματα που προέκυψαν επικεντρώθηκαν πάνω στην αποτελεσματικότητα της DEA κατά τη διαδικασία ταξινόμησης και ερευνήθηκε η επίδραση των διαφόρων παραγόντων κατά την εφαρμογή της DEA στην ταξινόμηση. Στο επόμενο κεφάλαιο, θα γίνει μια συνολική αναφορά στα βασικά αποτελέσματα που επιτεύχθηκαν από την έρευνα που πραγματοποιήθηκε και θα προταθούν μελλοντικές ερευνητικές κατευθύνσεις, οι οποίες θα συμβάλλουν στην καλύτερη αντιμετώπιση του προβλήματος της ταξινόμησης.

## ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Το πρόβλημα της ταξινόμησης παρουσίαζε ανέκαθεν αυξημένο ερευνητικό και πρακτικό ενδιαφέρον. Η διαπίστωση αυτή επιβεβαιώνεται από την πληθώρα των πρακτικών εφαρμογών που η προβληματική της ταξινόμησης παρουσιάζει (κεφάλαιο 1<sup>ο</sup>) και οι οποίες μπορούν να αντιμετωπιστούν πραγματοποιώντας απόλυτες συγκρίσεις με προκαθορισμένα δείγματα εκμάθησης και δείγματα ελέγχου. Οι έρευνες για την αντιμετώπιση των προβλημάτων ταξινόμησης επικεντρώνεται στην εφαρμογή κατάλληλων τεχνικών για την ανάπτυξη υποδειγμάτων ταξινόμησης, τα οποία θα συνδυάζουν όλες εκείνες τις επιδράσεις και αλληλεπιδράσεις των παραγόντων του εκάστοτε εξεταζόμενου προβλήματος, με στόχο την αξιολόγηση των εναλλακτικών δραστηριοτήτων και τις διαφοροποιήσεις που παρατηρούνται μεταξύ των κατηγοριών.

Επί δεκαετίες η ανάπτυξη των υποδειγμάτων ταξινόμησης βασίζονταν σε στατιστικές προσεγγίσεις. Οι προσεγγίσεις αυτές συνέβαλαν στην κατανόηση της προβληματικής της ταξινόμησης, των χαρακτηριστικών και των ιδιοτεροτήτων που παρουσιάζει, καθώς επίσης και των κανόνων που ακολουθούν τα αναπτυσσόμενα υποδείγματα ταξινόμησης. Το πρόβλημα, κατά την εφαρμογή των στατιστικών τεχνικών ταξινόμησης, έγκειται στις περιοριστικές υποθέσεις που διέπουν την κάθε μια από αυτές τις τεχνικές, με αποτέλεσμα τη δημιουργία κινήτρου για τη ανάπτυξη εναλλακτικών προσεγγίσεων. Κινούμενη προς αυτή την κατεύθυνση η παρούσα διπλωματική επικεντρώθηκε στη διερεύνηση της αποτελεσματικότητας της μεθόδου DEA (κεφάλαιο 2<sup>ο</sup>), μιας σημαντικής μεθοδολογικής



προσέγγισης της επιχειρησιακής έρευνας για την αξιολόγηση της αποτελεσματικότητας επιχειρήσεων και οργανισμών χρησιμοποιώντας τεχνικές γραμμικού προγραμματισμού, η οποία εφαρμόζεται και στην αντιμετώπιση προβλημάτων ταξινόμησης. Η χρησιμοποίηση τεχνικών γραμμικού προγραμματισμού για την επίτευξη αυτού του στόχου (κεφάλαιο 3<sup>ο</sup>) παρέχει ιδιαίτερη ευελιξία στον τρόπο με τον οποίο μπορεί να αξιολογηθεί η επάρκεια και η αποτελεσματικότητα του αναπτυσσόμενου υποδείγματος ταξινόμησης.

Σε συνδυασμό με την ευελιξία κατά την ανάπτυξη υποδειγμάτων ταξινόμησης, η μέθοδος DEA παρουσιάζει κάποια βασικά πλεονεκτήματα, καθώς οι παράμετροι του υποδείγματος ταξινόμησης είναι εύκολα κατανοητοί και μπορούν να χρησιμοποιηθούν πολλαπλά δεδομένα εισόδου και εξόδου κατά την αξιολόγηση της αποτελεσματικότητας μιας μονάδας, χωρίς να επηρεάζονται τα αποτελέσματα της διαδικασίας ταξινόμησης από την πολυπλοκότητα του προβλήματος. Η πολυπλοκότητα του αναπτυσσόμενου υποδείγματος παραμένει σταθερή ανεξαρτήτως των χαρακτηριστικών του προβλήματος που αντιμετωπίζεται, ενώ επιπλέον δεν πραγματοποιείται καμία υπόθεση σχετικά με τις στατιστικές ιδιότητες του προβλήματος.

Η παρούσα, βέβαια, διπλωματική δεν επικεντρώθηκε μόνο στη μαθηματική διατύπωση του μοντέλου της DEA και των ιδιοτήτων που παρουσιάζει σχετικά με άλλες προσεγγίσεις, αλλά πραγματοποιήθηκε και μια διεξοδική πειραματική ανάλυση (κεφάλαιο 4<sup>ο</sup>) για την αξιολόγηση της αποτελεσματικότητας της μεθόδου αυτής σε σύγκριση με άλλες τεχνικές για τον τρόπο με τον οποίο αντιμετωπίζονται τα διάφορα προβλήματα ταξινόμησης. Η έρευνα αυτή παρέχει τις απαραίτητες βάσεις για την πλήρη κατανόηση των προσεγγίσεων ταξινόμησης, των ιδιοτήτων και χαρακτηριστικών της κάθε προσέγγισης και του τρόπου με τον οποίο μπορούν οι διάφοροι παράγοντες επιδρούν στην αποτελεσματικότητα κάθε μεθόδου. Ο εκτενής αυτός πειραματικός σχεδιασμός επέτρεψε τη σύγκριση της ακρίβειας των εκτιμήσεων των εξεταζόμενων προσεγγίσεων ως προς την ταξινόμηση εναλλακτικών δραστηριοτήτων, οι επιδόσεις των οποίων έχουν προκαθορισμένη και ελεγχόμενη μορφή.

Η έρευνα της αποτελεσματικότητας των τεχνικών ταξινόμησης παρουσιάζει σημαντικές προοπτικές περαιτέρω έρευνας. Οι κύριες από τις προοπτικές αυτές και οι αντίστοιχες ερευνητικές κατευθύνσεις εντοπίζονται στη διερεύνηση της δυνατότητας του συνδυασμού με άλλες εναλλακτικές μεθοδολογίες και των ομοιοτήτων και διαφορών των προσεγγίσεων αυτών. Κύριος στόχος είναι η διερεύνηση εκείνων των συνθηκών υπό τις οποίες οι διάφορες

προσεγγίσεις ταξινόμησης θα παρουσιάζουν παρόμοια αποτελέσματα. Απαραίτητος είναι και ο προσδιορισμός εκείνων των παραγόντων που θα συμβάλλουν στην υψηλότερη αποτελεσματικότητα των τεχνικών ταξινόμησης. Η πειραματική ανάλυση της επίδρασης/αλληλεπίδρασης των παραγόντων στην αξιολόγηση των εναλλακτικών δραστηριοτήτων και στο αποτέλεσμα της ταξινόμησης, συμβάλλουν στην καλύτερη κατανόηση της λειτουργία των εξεταζόμενων τεχνικών ταξινόμησης και στην εξαγωγή χρήσιμων συμπερασμάτων ως προς τις πρακτικές τους εφαρμογές. Επίσης, ο συνδυασμός των τεχνικών ταξινόμησης θα αποτελέσει χρήσιμο εργαλείο για την αντιμετώπιση προβλημάτων ταξινόμησης κατά τα οποία εντοπίζονται προβλήματα τόσο ως προς τον ορισμό των κατηγοριών ταξινόμησης όσο και ως προς τις επιδόσεις των εναλλακτικών δραστηριοτήτων στα κριτήρια αξιολόγησης.

Πολλοί ερευνητές προσπαθούν να βρουν ένα τρόπο να καλύψουν το κενό που υπάρχει μεταξύ της DEA και άλλων τεχνικών. Αυτό έχει παρακινήσει μια εκτενή έρευνα σε διάφορα σημαντικά θέματα που αφορούν πιθανές επεκτάσεις της DEA όπως η στοχαστική μορφή της DEA, και ο συνδυασμός της DEA με μοντέλα παλινδρόμησης, καθώς και σε επιμέρους θέματα που σχετίζονται με την εφαρμογή της μεθόδου, όπως η ανάλυση ευαισθησίας της DEA, και η ευστάθεια των αποτελεσμάτων της DEA σε δυναμικό περιβάλλον (window analysis).

## BIBΛΙΟΓΡΑΦΙΑ

- [1] Ahn, T.; Charnes, A.; and Cooper, W.W. (1988) “Efficiency Characterizations in Different DEA Models,” *Socio-Economic Planning Sciences* 22(6), (pp. 253 – 257).
- [2] Arnold, V.L.; Bardhan, I.R.; Cooper, W.W.; and Kumbhakar, S.C.; (1996) “New Uses of DEA and Statistical Regressions for Efficiency Evaluation and Estimation – with an Application to Public Secondary Schools in Texas,” *Annals of Operations Research* 66.
- [3] Banker, R.D.; Charnes, A.; and Cooper, W.W. (1984) “Models for Estimating Technical and Scale Efficiencies in Data Envelopment Analysis,” *Management Science* 30(9).
- [4] Banker, R.D.; Conrad, R.F.; and Strauss, R.P. (1986) “A Comparative Application of Data Envelopment Analysis and Translog Methods: An Illustrative Study of Hospital Production,” *Management Science* 32(1), (pp. 30 – 44).
- [5] Banker, R.D.; and Morey, R.C. (1986a) “The Use of Categorical Variables in Data Envelopment Analysis,” *Management Science* 32(12), (pp. 1613 – 1627).
- [6] Banker, R.D.; Charnes, A.; Cooper, W.; and Maindiratta, A. (1988) “A Comparison of DEA and Translog Estimates of Production Frontiers,” in A. Dogramaci and R. Fare, eds. *Applications of Modern Production Theory: Efficiency and Productivity* (Norwell, Mass: Kluwer Academic Publishers, Inc.).
- [7] Bardhan, I.; Bowlin, W.F.; Cooper, W.W.; and Sueyoshi, T. (1996) “Models and Measures for Efficiency Dominance in DEA, Part I: Additive Models and MED Measures,” *Journal of the Operations Research Society of Japan* 39(3), (pp. 322 – 332).
- [8] Bardhan, I.; Cooper, W.W.; and Kumbhakar, S.C. (1997) “A Simulation Study of Joint Uses of DEA and Statistical Regressions for Efficiency Evaluation,” *Journal of Productivity Analysis*.
- [9] Belacel, N. (2000) “Multicriteria Assignment Method PROAFTN: Methodology and Medical Applications,” *European Journal of Operational research* 125, (pp. 175-183)

- 
- [10]Bowlin, W.F.; Charnes, A.; Cooper, W.; and Sherman, H.D. (1985) "Data Envelopment Analysis and Regression Approaches to Efficiency Evaluation and Estimation," *Annals of Operations Research* 2, (pp.113 – 138).
- [11]Breiman, L.; Friedman, J.H.; Olshen, R.A.; and Stone, C.J. (1984) "Classification and Regression Trees," Monterey, California, U.S.A, Wadsworth, Inc.
- [12]Burges, C.J.C. (1998) "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery* 2(2), (pp. 121-167).
- [13]Catelani, M.; and Ford, A. (2000) "Fault Diagnosis of Electronic Analog Circuits using a Radial Basis Function Network Classifier," *Measurement* 28 (3), (pp. 147-158).
- [14]Chaffai, M.E. (1997) "Estimating Input Specific Technical Efficiency: The Case of the Tunisian Banking Industry," *European Journal of Operational Research* 98(2).
- [15]Charnes, A.; Cooper, W.W.; and Rhodes, E. (1978) "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research* 2, (pp. 429-444).
- [16]Charnes, A.; Cooper, W.W.; and Rhodes, E. (1979) "Short Communication: Measuring Efficiency of Decision Making Units," *European Journal of operational Research* 4, (pp. 339).
- [17]Charnes, A.; Cooper, W.W.; Seiford, L.; and Stutz, J. (1982) "A Multiplicative Model for Efficiency Analysis," *Socio-Economic Planning Sciences* 16(5), (pp. 223 – 224).
- [18]Charnes, A.; Cooper, W.W.; Golany, B.; Seiford, L.; and Stutz, J. (1983) "Invariant Multiplicative Efficiency and Piecewise Cobb-Douglas Envelopments," *Operations Research Letters* 2, (pp. 101 – 103).
- [19]Charnes, A.; and Cooper, W.W. (1985) "Preface to Topics in Data Envelopment Analysis," *Annals of Operations Research* 2, (pp. 59 – 94).
- [20]Charnes, A.; Cooper, W.W.; Golany, B.; Seiford, L.; and Stutz, J. (1985) "Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions," *Journal of Econometrics* 20, (pp. 91 – 107).
- [21]Charnes, A.; Clark, T.; Cooper, W.W.; and Golany, B. (1985) "A Developmental Study of Data Envelopment Analysis for Measuring the Efficiency of Maintenance Units in the U.S. Air Force," *Annals of Operations Research* 2, (pp. 95 – 112).

- 
- [22]Charnes, A.; Cooper, W.W.; and Sueyoshi, T. (1986) “A Goal Programming/Constrained Regression Review of the Bell System Breakup,” *Management Science*.
- [23]Charnes, A.; Cooper, W.W.; Huang, Z.M.; and Sun, D.B. (1990) “Polyherdal Cone-Ratio DEA Models with an illustrative Application to Large Commercial Banks,” *Journal of econometrics*, 46 (pp. 73 - 91).
- [24]Cooper, W.W.; Seiford, L.M.; and Tone, K. (2000) “Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software,” Kluwer Academic Publishers, Boston.
- [25]Diakoulaki, D.; Zopounidis, C.; Mavrotas, G.; and Doumpos, M. (1999) “The Use of a Preference Disaggregation Method in Energy Analysis and Policy Making,” *Energy* 24(2), (pp. 157-166).
- [26]Donna, L. Retzlaff-Roberts (1996) “Related Discriminant Analysis and Data Envelopment Analysis to One Another,” *Computers Operational Research Vol. 23, No. 4*, (pp. 311-322).
- [27]Doumpos, M.; and Zopounidis, C. (1998) “The Use of the Preference Disaggregation Analysis in the Assessment of Financial Risks,” *Fuzzy Economic Review* 3(1), (pp. 39-57).
- [28]Δούμπος, Μ. και Ζοπουνίδης, Κ. (2001) «Πολυκριτήριες Τεχνικές Ταξινόμησης: Θεωρία και Εφαρμογές».
- [29]Dutka, A. (1995) “AMA Handbook of Customer Satisfaction: A Guide to Research, Planning and Implementation,” *NTC Publishing Group*, Illinois.
- [30]Fisher, R.A. (1936) “The Use of Multiple Measurements in Taxonomic Problems,” *Annals of Eugenics* 7, (pp. 179-188).
- [31]Freed, N.; and Glover, F. (1981) “Simple but powerful Goal Programming Formulations for the Statistical Discriminant Problem,” *European Journal of Operational Research* 7, (pp. 44-60).
- [32]Fung, G.; and Mangasarian, O.L. (2001) “Proximal Support Vector Machine Classifiers,” Data Mining Institute Technical Report 01-02, *Association for Computing Machinery, New York*, (pp.77-86).
- [33]Gochet, W.; Stam, A.; Srinivasan, V.; and Chen, S. (1997) “Multigroup Discriminant Analysis using Linear Programming,” *Operations Research* 45(2), (pp. 213-225).

- 
- [34] Hand, D.J. (1997) "Construction and Assessment of Classification Rules," John Wiley & Sons Ltd., Baffins Lane, Chichester.
- [35] Mirkin, B. (1996) "Mathematical Classification and Clustering," Kluwer Academic Publishers, Dordrecht.
- [36] Nieddu, L.; and Patrizi, G. (2000) "Formal Methods in Pattern Recognition: A review," *European Journal of Operational Research* 120, (pp. 459-495).
- [37] Ripley, B.D. (1996) "Pattern Recognition and Neural Networks," Cambridge University Press, Cambridge.
- [38] Rulon, P.J.; Tiedeman, D.V., Tatsuoaka, M.M., and Langmuir, C.R. (1967) "Multivariate Statistics for Personnel Classification," Wiley, New York.
- [39] Shen, L.; F.E.H.; Qu, L.; and Shen, Y. (2000) "Fault Diagnosis using Rough Sets Theory," *Computer in Industry* 43, (pp. 61-72).
- [40] Siskos, Y.; Grigoroudis, E.; Zopounidis, C.; and Saurais, O. (1998) "Measuring Customer Satisfaction using a Survey Based Preference Disaggregation Model," *Journal of Global Optimization* 12(2), (pp. 175-195).
- [41] Sueyoshi, T.; and Kiriara, Y. (1998) "Efficiency Measurement and Strategic Classification of Japanese Banking Institutions," *International Journal of System Sciences* 29, (pp. 1249-1263).
- [42] Sueyoshi Toshiyuki (1999) "Theory and Methodology: DEA-Discriminant Analysis in the view of the goal programming," *European Journal of Operational Research* 115, (pp. 564-582).
- [43] Sueyoshi Toshiyuki (2001) "Theory and Methodology: Extended DEA-Discriminant Analysis," *European Journal of Operational Research* 131, (pp. 324-351).
- [44] Thompson, R.G.; Singleton, R.D.; Thrall, R.M.; and Smith, B.A. (1986) "Comparative Site Evaluations for Locating a High-Energy Physics Laboratory in Texas," *Interfaces* 16, (pp. 16 – 26).
- [45] Thompson, R.G.; Langemeier, L.N.; Lee, C.T.; and Thrall, R.M. (1990) "The Role of Multiple Bounds in Efficiency of Analysis with Application to Kansas Farming," *Journal of Econometrics* 46 (pp. 93 – 108).
- [46] Tsumoto, S. (1998) "Automated Extraction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory," *Information Sciences* 112, (pp. 67-84).

- 
- [47]Vale, D.C.; and Maurelli, V.A. (1983) "Simulating Multivariate Nonnormal Distributions," *Psychometrika* 48/3, (pp. 465-471).
- [48]Vapnik, V.N. (2000) "The Nature of Statistical Learning Theory," Springer, New York, second edition.
- [49]Yohannes, Y.; and Webb, P (1999) "Classification and dRegression Trees, Cart: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity," Microcomputers in Policy Research 3, International Food Policy Research Institute.
- [50]Young, T.Y.; and Fu, K.S. (1997) "Handbook of Pattern Recognition and Image Processing," *Hanbooks in Science and Technology*, Academic Press, New York.
- [51]Zhu, J. (2002) "Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver," Kluwer Academic Publishers, Boston.
- [52]Zopounidis, C. (1998) "Operational Tools in the Management of Financial Risks," *Kluwer Academic Publishers*, Dordrecht.
- [53]TUTORIAL IN DEA, Moez Hababou Corporation retrieved from the World Wide Web at 17/02/2003,  
<http://members.tripod.com/moezh/DEAtutorial/DEAtutorial.html>
- [54]TUTORIAL IN DEA retrieved from the World Wide Web at 17/02/2003,  
<http://www.warwick.ac.uk/~bsrlu/dea/deat/deat1.htm>
- [55]DEA TUTORIAL retrieved from the World Wide Web at 17/02/2003,  
<http://www.deazone.com/tutorial/issues.htm>