# 10 Εναλλακτικές Μέθοδοι και ειδικά θέματα Κατηγοριοποίησης

## Σύνοψη

*Στο δέκατο Κεφάλαιο ολοκληρώνεται η κάλυψη της θεματικής ενότητας της Κατηγοριοποίησης. Παρουσιάζονται τρεις πρόσθετες μέθοδοι κατηγοριοποίησης, οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ), οι k-Πλησιέστεροι Γείτονες και η Λογιστική Παλινδρόμηση. Βασική ιδέα των ΜΔΥ είναι η κατασκευή ενός υπερεπιπέδου, το οποίο διαχωρίζει τις κλάσεις. Οι δύο κλάσεις θεωρούνται γραμμικά διαχωρίσιμες. Για τον καθορισμό του βέλτιστου υπερεπιπέδου εισάγεται η έννοια του περιθωρίου. Το βέλτιστο υπερεπίπεδο διαχωρισμού είναι αυτό που εξασφαλίζει το μέγιστο περιθώριο. Οι πραγματικές περιπτώσεις γραμμικού διαχωρισμού των κλάσεων είναι μάλλον σπάνιες. Για τον λόγο αυτό, τα σημεία προβάλλονται σε έναν χώρο περισσότερων διαστάσεων με τη βοήθεια μιας διανυσματικής συνάρτησης. Στον χώρο αυτόν τα σημεία είναι γραμμικώς διαχωρίσιμα. Η συνάρτηση πυρήνα ορίζει το εσωτερικό γινόμενο της διανυσματικής συνάρτησης, το οποίο απαιτείται για τον υπολογισμό της συνάρτησης απόφασης. Στη μέθοδο των k-Πλησιέστερων Γειτόνων η μάθηση βασίζεται στην αναλογία. Κάθε παρατήρηση θεωρείται ως ένα σημείο μέσα σε έναν πολυδιάστατο χώρο. Για την πρόβλεψη της κλάσης μιας νέας παρατήρησης, εντοπίζονται τα k πλησιέστερα σημεία και η νέα παρατήρηση εκχωρείται στην κλάση που πλειοψηφεί μεταξύ των k γειτονικών σημείων. Η Λογιστική (ή Λογαριθμική) Παλινδρόμηση είναι η κλασική μέθοδος την οποία χρησιμοποιούν οι οικονομολόγοι για την αντιμετώπιση προβλημάτων κατηγοριοποίησης. Σε ένα πρόβλημα δυαδικής κλάσης, ο λογάριθμος του λόγου των πιθανοτήτων να ανήκει η παρατήρηση στις δύο τιμές της κλάσης εκφράζεται ως γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών. Άλλες εκδοχές της Παλινδρόμησης χρησιμοποιούνται για την πρόβλεψη αριθμητικών τιμών. Στο παρόν Κεφάλαιο παρουσιάζονται η Απλή Γραμμική Παλινδρόμηση, η Πολλαπλή Γραμμική Παλινδρόμηση και η Πολυωνυμική Παλινδρόμηση.*
*Τα τελευταία χρόνια ιδιαίτερη άνθηση γνωρίζουν οι λεγόμενοι σύνθετοι κατηγοριοποιητές. Οι σύνθετοι κατηγοριοποιητές μπορούν να χωριστούν σε δύο βασικές κατηγορίες, στους συνδυασμούς κατηγοριοποιητών και στους υβριδικούς κατηγοριοποιητές. Στους συνδυασμούς κατηγοριοποιητών δημιουργείται ένας αριθμός ατομικών κατηγοριοποιητών, οι οποίοι προβλέπουν την κλάση και η τελική απόφαση υπολογίζεται με συνάθροιση των ατομικών αποφάσεων. Στους υβριδικούς κατηγοριοποιητές εφαρμόζονται ετερογενείς τεχνικές, κάθε μια από τις οποίες επιλύει ένα διαφορετικό πρόβλημα. Η τελική απόφαση κατηγοριοποίησης λαμβάνεται από έναν μόνο κατηγοριοποιητή. Η ακρίβεια ενός μοντέλου πρέπει να εκτιμάται έναντι παρατηρήσεων, οι οποίες δεν ανήκουν στο σύνολο εκπαίδευσης. Η διαδικασία αυτή ονομάζεται επικύρωση του μοντέλου και έχουν προταθεί σχετικές τεχνικές. Στο παρόν Κεφάλαιο παρουσιάζονται η μέθοδος holdout, η διασταυρούμενη επικύρωση 10 τμημάτων, η μέθοδος «άφησε ένα έξω» και η μέθοδος bootstrap. Δύο σημαντικά θέματα, τα οποία συναντώνται συχνά σε ρεαλιστικά προβλήματα κατηγοριοποίησης, είναι το πρόβλημα της ανισοκατανομής των κλάσεων και το πρόβλημα του διαφορετικού κόστους σφάλματος. Τα δύο αυτά θέματα αναλύονται με σύντομο, αλλά ουσιαστικό τρόπο. Για την εκτίμηση και παρουσίαση της ικανότητας των μοντέλων να προβλέπουν συγκεκριμένη τιμή κλάσης έχουν προταθεί ειδικές τεχνικές. Στο παρόν κεφάλαιο παρουσιάζονται ο πίνακας σύγχυσης (confusion matrix) και η καμπύλες ROC (Receiver Operating Characteristics). Τέλος, παρατίθεται μια μελέτη περίπτωσης, όπου εφαρμόζονται τεχνικές κατηγοριοποίησης για την πρόβλεψη του τύπου του εξωτερικού ελεγκτή. Εφαρμόζονται Δένδρα Αποφάσεων, Νευρωνικά Δίκτυα και k-Πλησιέστεροι Γείτονες, καθώς και συνδυασμοί κατηγοριοποιητών τύπου bagging, οι οποίοι βελτιώνουν περαιτέρω τις επιδόσεις των ατομικών τεχνικών.*

## Προηγούμενη γνώση

*Στο παρόν Κεφάλαιο παρουσιάζονται μέθοδοι και αναπτύσσονται ειδικά θέματα κατηγοριοποίησης. Για την κατανόηση αυτών των θεμάτων απαιτείται η προηγούμενη ανάγνωση του ένατου κεφαλαίου και ειδικά των υποκεφαλαίων από 9.1 έως και 9.6, τα οποία εισάγουν τον αναγνώστη σε βασικές έννοιες κατηγοριοποίησης. Επίσης, χρήσιμη είναι η προηγούμενη ανάγνωση του Κεφαλαίου 6, το οποίο αποτελεί εισαγωγή στην Εξόρυξη Δεδομένων και του Κεφαλαίου 7, το οποίο αναφέρεται στην προεπεξεργασία των δεδομένων. Πρόσθετη πληροφόρηση για τα παραπάνω θέματα μπορεί να αναζητήσει ο ενδιαφερόμενος αναγνώστης σε ένα από τα πολλά συγγράμματα Εξόρυξης Δεδομένων. Ενδεικτικά αναφέρουμε τα βιβλία των Han, Kamber and Pei (2011) και των Maimon and Rokach (2010). Για μεθόδους που χρησιμοποιούν συναρτήσεις πυρήνα και ειδικότερα και τις Μηχανές Διανυσμάτων Υποστήριξης, ενδιαφέρον παρουσιάζει η ιστοθέση της kernel-machines.org.*

## 10.1 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) (Support Vector Machines (SVM)) προτάθηκαν από τον Vapnik (1995) και γρήγορα γνώρισαν μεγάλη διάδοση λόγω της στιβαρής θεωρητικής θεμελίωσης τους και των υψηλών επιδόσεων τους. Οι ΜΔΥ αποτέλεσαν αντικείμενο ενδιαφέροντος πολλών ερευνητών και εφαρμόστηκαν για την ανάπτυξη μοντέλων σε πλήθος προβλημάτων κατηγοριοποίησης.

Βασική ιδέα των ΜΔΥ είναι η κατασκευή ενός **υπερεπιπέδου** (hyperplane), το οποίο διαχωρίζει τις κλάσεις και λειτουργεί ως συνάρτηση απόφασης. Οι νέες παρατηρήσεις κατηγοριοποιούνται ανάλογα με την πλευρά του υπερεπιπέδου στην οποία βρίσκονται. Ας θεωρήσουμε μια απλή περίπτωση όπου η κλάση είναι δυαδική και οι παρατηρήσεις είναι γραμμικά διαχωρίσιμες. Το **κυρτό περίβλημα** (convex hull) ενός συνόλου σημείων είναι το μικρότερο κυρτό πολύγωνο, το οποίο περικλείει όλα τα σημεία του συνόλου. Οι δύο κλάσεις είναι **γραμμικά διαχωρίσιμες**, όταν τα κυρτά περιβλήματα τους δεν επικαλύπτονται. Παράδειγμα παρατηρήσεων δυαδικής κλάσης, οι οποίες είναι γραμμικά διαχωρίσιμες, απεικονίζεται στο Σχήμα 10.1.Α. Οι παρατηρήσεις συμβολίζονται ως μικροί κύκλοι, ενώ το διαφορετικό χρώμα συμβολίζει τις διαφορετικές κλάσεις. Η μια τιμή κλάσης μπορεί να οριστεί ως θετική και να συμβολιστεί με την τιμή +1, ενώ η άλλη τιμή να οριστεί ως αρνητική και να συμβολιστεί με την τιμή -1.

Το γενικό υπερεπίπεδο διαχωρισμού ορίζεται από την Εξίσωση 10.1

$$w^T x + b = 0 \tag{10.1}$$

όπου *w* είναι ένα διάνυσμα βαρών, το οποίο είναι κάθετο στο επίπεδο και ορίζει τον προσανατολισμό του και *b* είναι το κατώφλι. Η μεταβολή της τιμής του *b* έχει σαν αποτέλεσμα την παράλληλη μετατόπιση του επιπέδου. Για μια παρατήρηση $x_1$ θετικής κλάσης ισχύει ότι

$$w^T x_1 + b > 0 \tag{10.2}$$

ενώ για μια παρατήρηση $x_2$ αρνητικής κλάσης ισχύει ότι

$$w^T x_2 + b < 0 \tag{10.3}$$

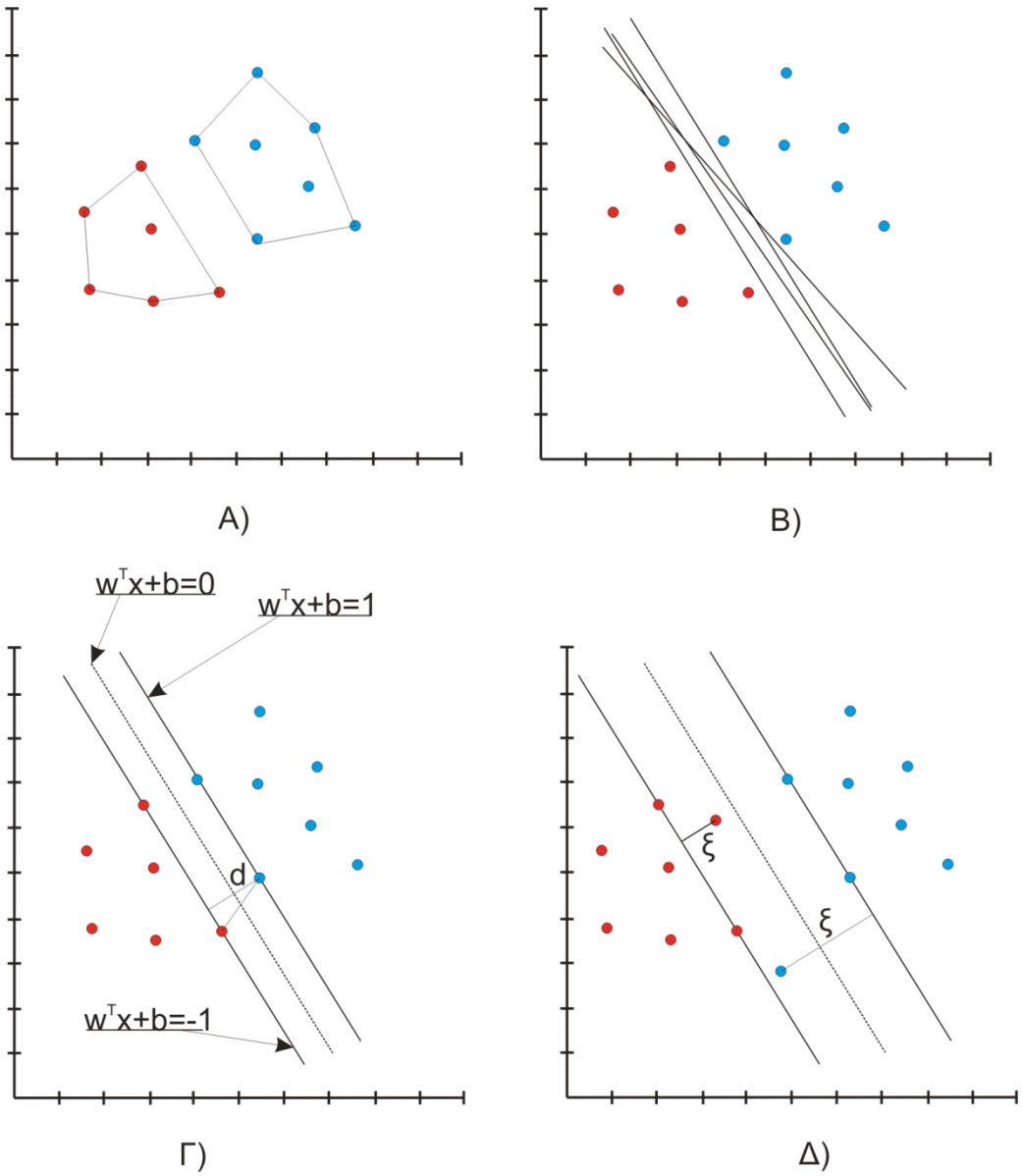
Πλέον το πρόβλημα της κατηγοριοποίησης ανάγεται σε πρόβλημα καθορισμού του υπερεπιπέδου διαχωρισμού. Όπως φαίνεται στο σχήμα 10.1.Β υπάρχουν πολλά υπερεπίπεδα, τα οποία θα μπορούσαν να χρησιμοποιηθούν, και το ερώτημα είναι ποιο από αυτά είναι το καλύτερο. Για τον υπολογισμό του βέλτιστου επιπέδου εισάγεται η έννοια του περιθωρίου (margin). Ως **περιθώριο** ορίζεται η μικρότερη απόσταση ενός σημείου από το υπερεπίπεδο διαχωρισμού. Η κλίμακα του περιθωρίου επηρεάζεται από το διάνυσμα βαρών *w*. Θεωρούμε τα σημεία $x_i$, τα οποία είναι πλησιέστερα στο υπερεπίπεδο. Μπορούμε να ρυθμίσουμε τις τιμές των *w* και *b* έτσι ώστε η απόσταση των σημείων αυτών από το υπερεπίπεδο να είναι ίση με 1 (Εξίσωση 10.4)

$$|(w^T x_i) + b| = 1 \tag{10.4}$$

Θεωρούμε δύο σημεία $x_1$ και $x_2$ τα οποία είναι πλησιέστερα στο υπερεπίπεδο, δηλαδή η απόσταση τους από αυτό είναι ίση με 1 και τα οποία βρίσκονται εκατέρωθεν του υπερεπιπέδου, δηλαδή η τιμή κλάσης του ενός είναι +1 και του άλλου -1. Από τα σημεία αυτά μπορούμε να ορίσουμε το περιθώριο ως την απόσταση τους *d*, μετρημένη κάθετα στο υπερεπίπεδο, όπως φαίνεται και στο σχήμα 10.1.Γ. Το περιθώριο υπολογίζεται σύμφωνα με τη Σχέση 10.5

$$\left(\frac{w}{\|w\|}(x_1 - x_2)\right) = \frac{2}{\|w\|}$$

(10.5)




**Σχήμα 10.1** *Μηχανές Διανυσμάτων Υποστήριξης*

Το βέλτιστο υπερεπίπεδο διαχωρισμού των κλάσεων είναι αυτό που εξασφαλίζει το **μέγιστο περιθώριο**. Τα σημεία, τα οποία βρίσκονται στο όριο του περιθωρίου, ονομάζονται **διανύσματα υποστήριξης**. Προφανώς η κάθετη απόσταση από το υπερεπίπεδο των σημείων $x_1$ και $x_2$ είναι ίση με το μισό του περιθωρίου, δηλαδή $1/\|w\|$. Το πρόβλημα μετατρέπεται σε ένα πρόβλημα βελτιστοποίησης. Η ποσότητα $1/\|w\|$ πρέπει να μεγιστοποιηθεί για κάθε σημείο, με τον περιορισμό ότι η απόσταση του πλησιέστερου σημείου θα είναι ίση με 1. Για $n$ σημεία $x_i$ το παραπάνω πρόβλημα διατυπώνεται ως εξής:

$$Maximize\ \frac{1}{\|w\|}$$

(10.6)

με τον περιορισμό ότι

$$\min_{i=1,2,..,n} |w^T x_i + b| = 1$$

(10.7)

Με δεδομένο ότι η κλάση $y_i$ μιας παρατήρησης $x_i$ μπορεί να πάρει τιμές +1 ή -1, καθώς και ότι το $w^T x_i + b$ θα έχει τιμή >=1 για παρατηρήσεις θετικής κλάσης και <=-1 για παρατηρήσεις αρνητική κλάσης, προκύπτει ότι το γινόμενο του $(w^T x_i + b)$ με την τιμή της κλάσης θα δίνει αποτέλεσμα μεγαλύτερο ή ίσο του 1

$$y_i * (w^T x_i + b) \geq 1$$

(10.8)

Το πρόβλημα επαναδιατυπώνεται ως:

$$Minimize \frac{1}{2} \|w^2\|$$

(10.9)

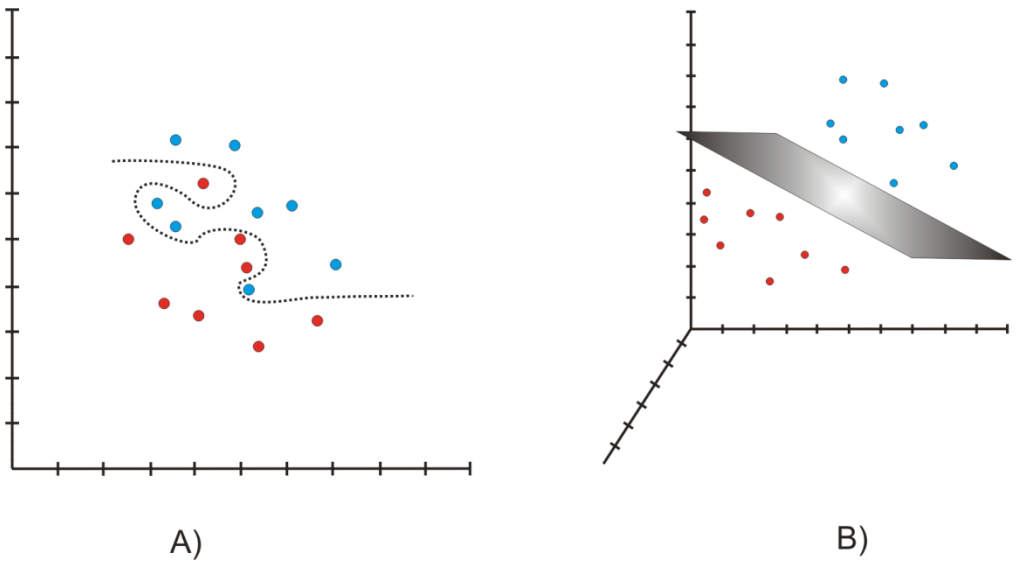
με τον περιορισμό της Σχέσης 10.8.

Το πρόβλημα μπορεί να λυθεί με τη χρήση του τετραγωνικού προγραμματισμού. Διαθέσιμα λογισμικά επιλύουν προβλήματα τετραγωνικού προγραμματισμού. Σε προβλήματα του πραγματικού κόσμου μπορεί να μην είναι όλες οι παρατηρήσεις γραμμικά διαχωρίσιμες. Για να ξεπεράσει το πρόβλημα του απόλυτου γραμμικού διαχωρισμού, ο Vapnik εισήγαγε τις μεταβλητές χαλαρότητας $\xi_i$. Με τη συμμετοχή των μεταβλητών χαλαρότητας, η Σχέση 10.8 τροποποιείται ως ακολούθως:

$$y_i * (w^T x_i + b) \geq 1 - \xi_i$$

(10.10)

*όπου* $\xi_i$ >=0.




**Σχήμα 10.2** *Γραμμικός διαχωρισμός κλάσεων σε χώρο περισσότερων διαστάσεων*

Αν για ένα σημείο $x_i$ η μεταβλητή $\xi_i$ είναι μεγαλύτερη από 1, τότε το σημείο κατηγοριοποιείται εσφαλμένα, όπως φαίνεται και στο σχήμα 10.1.Δ. Το άθροισμα των $\xi_i$ μπορεί να θεωρηθεί το πλήθος των σφαλμάτων κατηγοριοποίησης. Η 9.10 μπορεί να τροποποιηθεί ώστε να περιλαμβάνει τις μεταβλητές χαλαρότητας.

$$Minimize \frac{1}{2}\|w^2\| + C\sum_{i=1}^{n}\xi_i$$

(10.11)

Η σταθερά *C* είναι μια παράμετρος, που ορίζει το ισοζύγιο μεταξύ πολυπλοκότητας και εμπειρικού σφάλματος. Οι περιπτώσεις δυνατότητας γραμμικού διαχωρισμού των κλάσεων είναι μάλλον σπάνιες σε πραγματικά προβλήματα. Εάν όμως τα σημεία $x_i$ προβληθούν με μία μη γραμμική διανυσματική συνάρτηση $\varphi(x_i)$ σε έναν χώρο περισσότερων διαστάσεων, τότε είναι πιθανό οι απεικονίσεις τους στον νέο χώρο να είναι γραμμικώς διαχωρίσιμες. Στο Σχήμα 10.2.Α απεικονίζονται τα σημεία στον αρχικό δισδιάστατο χώρο. Τα σημεία δεν είναι γραμμικώς διαχωρίσιμα. Στο Σχήμα 10.2.Β τα σημεία προβάλλονται σε έναν τρισδιάστατο χώρο, και εκεί είναι γραμμικώς διαχωρίσιμα. Εφόσον στον χώρο αυτόν ισχύει ο γραμμικός διαχωρισμός, μπορεί να εφαρμοστεί η μέθοδος των διανυσμάτων υποστήριξης που παρουσιάστηκε προηγουμένως. Η συνάρτηση απόφασης επαναδιατυπώνεται ως εξής:

$$f(x) = w^T\varphi(x) + b$$

(10.12)

Ο προσδιορισμός της συνάρτησης φ μπορεί να είναι εξαιρετικά δύσκολος και ο χώρος προβολής μπορεί να έχει πάρα πολλές διαστάσεις. Όμως για τον υπολογισμό της συνάρτησης απόφασης *f*, απαιτείται μόνο ο ορισμός του εσωτερικού γινόμενου $\varphi(x_i)*\varphi(x_j)$. Ορίζουμε μια συνάρτηση $K(x_i, x_j)$, οποία υπολογίζει το εσωτερικό γινόμενο των απεικονίσεων $\varphi(x_i)$ και $\varphi(x_j)$ (Σχέση 10.13). Η συνάρτηση *K* καλείται **συνάρτηση πυρήνα** (kernel function)

$$K(x_i, x_j) = \phi(x_i) * \phi(x_j)$$

(10.13)

Διάφορες συναρτήσεις μπορούν να χρησιμοποιηθούν ως συναρτήσεις πυρήνα. Σε αυτές περιλαμβάνονται η Συνάρτηση Ακτινωτής Βάσης (Radial Base Function – RBF), η Σιγμοειδής, η πολυωνυμική και η αντίστροφη πολυτετραγωνική συνάρτηση. Ο πυρήνας καθορίζει τη μορφή του υπερεπιπέδου διαχωρισμού και συνεπώς επηρεάζει την απόδοση του κατηγοριοποιητή. Η επιλογή της καλύτερης συνάρτησης πυρήνα είναι θέμα το οποίο διερευνάται (Steinwart, 2003).

Έχουν προταθεί κατάλληλες παραλλαγές των Μηχανών Διανυσμάτων Υποστήριξης, που τις καθιστούν ικανές να υπολογίζουν αριθμητικές τιμές και όχι τιμές κλάσης. Η μέθοδος ονομάζεται **Παλινδρόμηση Διανυσμάτων Υποστήριξης** (Support Vector Regression (SVR)). Η κεντρική ιδέα των SVR (Smola & Schoelkopf, 2004) είναι να οριστεί μια συνάρτηση $f(x_i)$, της οποίας το αποτέλεσμα να μην αποκλίνει περισσότερο από μια ποσότητα ε από τις πραγματικές τιμές $y_i$.

Οι ΜΔΥ αρχικά σχεδιάστηκαν για την επίλυση διχότομων προβλημάτων, προβλημάτων δηλαδή με δύο δυνατές τιμές κλάσης, Ωστόσο, έχουν προταθεί παραλλαγές των ΜΔΥ που τις καθιστούν ικανές να αναπτύσσουν μοντέλα κατηγοριοποίησης για προβλήματα με πολλαπλές τιμές κλάσης. Μια προσέγγιση ονομάζεται one-against-the-rest (Vapnik, 1995). Σύμφωνα με την προσέγγιση αυτή, για ένα πρόβλημα με *k* δυνατές τιμές κλάσης κατασκευάζονται *k* δυαδικοί κατηγοριοποιητές, οι οποίοι προβλέπουν τιμή +1 για τη μια τιμή κλάσης και τιμή -1 για όλες τις υπόλοιπες. Οι άγνωστες παρατηρήσεις εκχωρούνται στην κλάση με τη μεγαλύτερη τιμή απόφασης. Σύμφωνα με μια άλλη προσέγγιση, η οποία ονομάζεται one-against-one (Kreßel, 1999), αναπτύσσονται δυαδικοί κατηγοριοποιητές για κάθε δυνατό ζευγάρι τιμών κλάσης. Για προβλήματα με *k* δυνατές τιμές κλάσης αναπτύσσονται συνολικά *k(k-1)/2* κατηγοριοποιητές. Σχήματα ψηφοφορίας χρησιμοποιούνται για την τελική κατηγοριοποίηση.

Οι Μηχανές Διανυσμάτων Υποστήριξης είναι πολύ δημοφιλείς, χάρη στα ιδιαίτερα χαρακτηριστικά τους και τα πολλά πλεονεκτήματα τους. Τα κύρια **πλεονεκτήματα** τους είναι τα ακόλουθα:

- Η χρήση της συνάρτησης πυρήνα τις καθιστά πολύ αποτελεσματικές σε περιπτώσεις όπου υπάρχουν μη γραμμικές σχέσεις στα δεδομένα.
- Επιτυγχάνουν υψηλές επιδόσεις κατηγοριοποίησης, κυρίως στην περίπτωση δυαδικών κλάσεων.
- Διαθέτουν στιβαρή θεωρητική θεμελίωση.
- Είναι ανθεκτικές στην υπερπροσαρμογή και διαθέτουν πολύ καλή δυνατότητα γενίκευσης με κατάλληλη ρύθμιση της παραμέτρου C.
- Δεν παγιδεύονται σε τοπικά ελάχιστα.
- Είναι αποτελεσματικές σε περιπτώσεις συνόλων δεδομένων με πολλές στήλες και σχετικά λίγες γραμμές.

Τα βασικότερα **μειονεκτήματα** των Μηχανών διανυσμάτων Υποστήριξης είναι τα ακόλουθα:

- Δεν υπάρχει κάποια μεθοδολογία για την επιλογή της συνάρτησης πυρήνα καθώς και των παραμέτρων του πυρήνα.
- Δεν παρέχουν ερμηνεύσιμα μοντέλα. Η συμβολή της εκάστοτε μεταβλητής εισόδου στο τελικό αποτέλεσμα κατηγοριοποίησης είναι αδιαφανής.
- Έχουν σχετικά μεγάλους χρόνους εκπαίδευσης, αν και σημαντικά χαμηλότερους από αυτούς των Νευρωνικών Δικτύων.
- Έχουν μεγάλες απαιτήσεις σε μνήμη υπολογιστή.
- Σε περίπτωση κλάσεων με πολλαπλές τιμές το πρόβλημα διατυπώνεται σαν συνδυασμός προβλημάτων δυαδικών κλάσεων.

## 10.2 k-Πλησιέστεροι Γείτονες

Οι **Κατηγοριοποιητές Βασισμένοι σε Παραδείγματα** (Instance Based Classifiers (IBC)) είναι μια οικογένεια κατηγοριοποιητών, όπου η μάθηση βασίζεται στην αναλογία. Οι κατηγοριοποιητές IBC δεν παράγουν κάποιο μοντέλο γενίκευσης. Μέθοδοι κατηγοριοποίησης όπως τα Νευρωνικά Δίκτυα, τα Δένδρα Αποφάσεων ή τα Μπαϋεσιανά Δίκτυα Πίστης ολοκληρώνουν την εκπαίδευση με τη δημιουργία κάποιου μοντέλου. Ακολούθως, το μοντέλο χρησιμοποιείται για την κατηγοριοποίηση νέων παρατηρήσεων. Σε αντίθεση με αυτές τις μεθόδους, στους κατηγοριοποιητές IBC δεν υπάρχει κάποιο στάδιο εκπαίδευσης και δεν παράγεται κάποιο μοντέλο, μέχρι να χρειαστεί να κατηγοριοποιηθεί μια νέα παρατήρηση. Για τον λόγο αυτό, οι κατηγοριοποιητές IBC καλούνται και «οκνηροί» (lazy classifiers). Όταν χρειαστεί να κατηγοριοποιήσουν μια νέα παρατήρηση, τη συγκρίνουν με γνωστές παρατηρήσεις του συνόλου εκπαίδευσης. Αυτό απαιτεί την αποθήκευση όλων ή τουλάχιστον ενός μέρους των παρατηρήσεων εκπαίδευσης. Αντιθέτως, σε άλλες τεχνικές, όπως τα SVM, μπορούν να απορριφθούν όλες οι παρατηρήσεις εκπαίδευσης που δεν είναι διανύσματα υποστήριξης

Η μέθοδος των **k-Πλησιέστερων Γειτόνων** (k-Nearest Neighbors - kNN) είναι ένας αλγόριθμος της οικογένειας των Κατηγοριοποιητών Βασισμένων σε Παραδείγματα. Για λόγους απλότητας θεωρούμε αρχικά ένα πρόβλημα κατηγοριοποίησης, όπου οι παρατηρήσεις αποτελούνται από δύο αριθμητικά πεδία και το γνώρισμα της κλάσης. Κάθε παρατήρηση μπορεί να θεωρηθεί ως ένα σημείο στον χώρο των δύο διαστάσεων. Μια παρατήρηση X απέχει από μια άλλη παρατήρηση Y, απόσταση d(X,Y) μέσα στον δισδιάστατο χώρο. Η απόσταση d(X,Y) μπορεί να υπολογιστεί ως η Ευκλείδεια απόσταση σύμφωνα με την Εξίσωση 10.14:

$$d(X,Y) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2)}$$

(10.14)

όπου $x_1$, $y_1$ οι τιμές των $X$ και $Y$ για την πρώτη διάσταση και $x_2$, $y_2$ οι τιμές των $X$ και $Y$ για τη δεύτερη διάσταση. Σύμφωνα με τον αλγόριθμο k-NN, ο χρήστης προκαθορίζει την τιμή της σταθερής παραμέτρου $k$. Ο αλγόριθμος αναζητά μέσα στον δισδιάστατο χώρο τα k σημεία-παρατηρήσεις που βρίσκονται πλησιέστερα στη νέα παρατήρηση. Ο κατηγοριοποιητής εκχωρεί τη νέα παρατήρηση στην κλάση που πλειοψηφεί μεταξύ των $k$ πλησιέστερων γειτόνων. Εάν οριστεί ότι $k=1$, τότε η νέα παρατήρηση εκχωρείται στην κλάση της πιο όμοιας παρατήρησης εκπαίδευσης

Τα παραπάνω παρουσιάζονται διαγραμματικά στο Σχήμα 10.3. Στο παράδειγμα υπάρχουν δύο δυνατές τιμές κλάσης, οι οποίες συμβολίζονται με το χρώμα των σημείων. Το κίτρινο σημείο συμβολίζει τη νέα πα-

ρατήρηση που θα κατηγοριοποιηθεί. Στο παράδειγμα η τιμή του *k* έχει οριστεί να είναι 5. Εντοπίζονται τα 5 πλησιέστερα σημεία. Παρατηρούμε ότι τρία από αυτά είναι κόκκινα και δύο είναι μπλε. Η νέα παρατήρηση εκχωρείται στην «κόκκινη» κλάση.

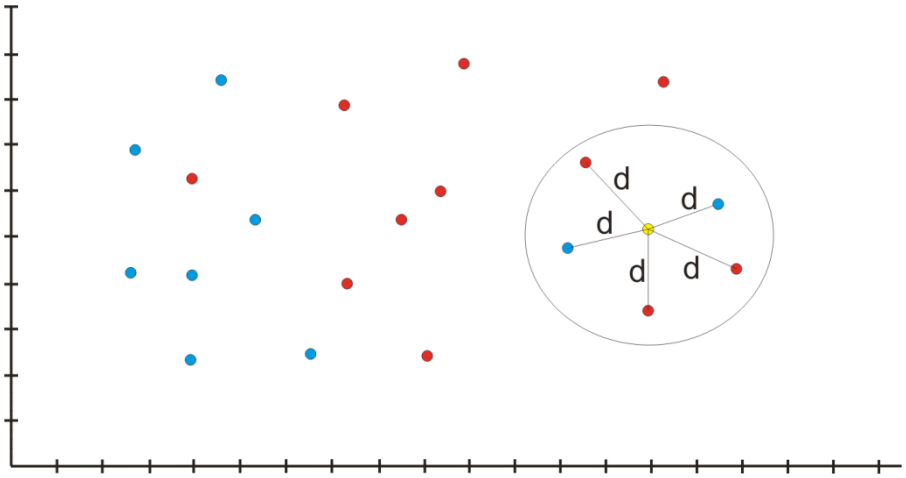
Κατ' αντιστοιχία, ο ίδιος αλγόριθμος ισχύει για παρατηρήσεις με *n* αριθμητικές διαστάσεις. Οι παρατηρήσεις θεωρούνται σημεία στον n-διάστατο χώρο και η Ευκλείδεια απόσταση υπολογίζεται σύμφωνα με την Εξίσωση 10.15.

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

(10.15)

Μια βελτίωση στον παραπάνω αλγόριθμο είναι να μην λαμβάνεται η απόφαση κατηγοριοποίησης με ισότιμη ψηφοφορία μεταξύ των επιλεγμένων γειτόνων, αλλά να συνεισφέρουν περισσότερο τα σημεία τα οποία είναι πλησιέστερα στη νέα παρατήρηση. Ένας απλός τρόπος για να επιτευχθεί αυτό είναι να εκχωρηθούν συντελεστές βαρύτητας ψήφου στα επιλεγμένα σημεία. Οι συντελεστές θα μπορούσαν να είναι ίσοι με *1/d*, όπου *d* η απόσταση του εκάστοτε σημείου από τη νέα παρατήρηση.

Μία αδυναμία στον υπολογισμό της ομοιότητας με βάση την Ευκλείδεια απόσταση είναι το γεγονός ότι οι μεταβλητές με μεγάλο εύρος τιμών επηρεάζουν περισσότερο το αποτέλεσμα από τις μεταβλητές με μικρό εύρος τιμών. Εάν πχ οι παρατηρήσεις έχουν δύο γνωρίσματα Α και Β και το Α παίρνει τιμές από 1 έως 1000, ενώ το Β παίρνει τιμές από 1 έως 10, τότε το γνώρισμα Α επηρεάζει δυσανάλογα την απόσταση σε σχέση με το γνώρισμα Β. Το πρόβλημα αυτό αντιμετωπίζεται με κανονικοποίηση των αριθμητικών τιμών. Αυτό μπορεί να επιτευχθεί διαιρώντας τις τιμές των γνωρισμάτων με την περιοχή τιμών των γνωρισμάτων.




**Σχήμα 10.3** *Κατηγοριοποιητής k-NN με k=5*

Ένα άλλο συγγενές πρόβλημα είναι το γεγονός ότι ο υπολογισμός της ομοιότητας με βάση την Ευκλείδεια απόσταση υποθέτει την ισότιμη συμμετοχή όλων των γνωρισμάτων, κάτι που γενικώς δεν ισχύει. Το πρόβλημα αυτό αντιμετωπίζεται με τον καθορισμό «βαρών» για την κάθε διάσταση. Ο καθορισμός των βαρών επιτρέπει την αναδιατύπωση του υπολογισμού της απόστασης σύμφωνα με την Εξίσωση 10.16:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} w_i * (x_i - y_i)^2}$$

(10.16)

όπου $w_i$ είναι το βάρος που αντιστοιχεί στην *i*-οστή διάσταση. Ο καθορισμός των βαρών αποτελεί ένα ενεργό πεδίο έρευνας που έχει αποδώσει διάφορες μεθόδους υπολογισμού των βαρών.

Ένας άλλος περιορισμός του υπολογισμού της ομοιότητας με βάση την Ευκλείδεια απόσταση ή κάποια παραλλαγή της είναι το γεγονός ότι αυτές οι προσεγγίσεις προϋποθέτουν γνωρίσματα αριθμητικών τιμών. Για να αντιμετωπιστεί αυτό το πρόβλημα, έχουν προταθεί συναρτήσεις που υπολογίζουν την απόσταση παρατηρήσεων που αποτελούνται από ονομαστικές τιμές. Στην απλούστερη εκδοχή τους οι συναρτήσεις αυτές επιστρέφουν την τιμή 0 εάν οι τιμές του ίδιου ονομαστικού γνωρίσματος δύο διαφορετικών παρατηρήσεων είναι ίδιες, αλλιώς επιστρέφουν την τιμή 1. Επίσης, έχουν προταθεί αλγόριθμοι που υπολογίζουν την απόσταση αντικειμένων που αποτελούνται και από αριθμητικές και από ονομαστικές τιμές. (Wilson & Martinez, 1997).

Η μέθοδος k-NN εκτός από κατηγοριοποίηση μπορεί να χρησιμοποιηθεί και για παλινδρόμηση, δηλαδή για πρόβλεψη αριθμητικών τιμών. Για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής μιας νέας παρατήρησης, ο αλγόριθμος εντοπίζει τις *k* πλησιέστερες παρατηρήσεις και επιστρέφει ως πρόβλεψη τη μέση τιμή των εξαρτημένων μεταβλητών των επιλεγμένων παρατηρήσεων.

Οι κατηγοριοποιητές k-NN διαθέτουν αξιόλογα **πλεονεκτήματα**:

- Είναι αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών.
- Διαθέτουν απλό αλγόριθμο.
- Σε πολλές περιπτώσεις επέτυχαν υψηλές επιδόσεις κατηγοριοποίησης.

Ορισμένα από τα βασικά **μειονεκτήματα** τους είναι τα ακόλουθα:

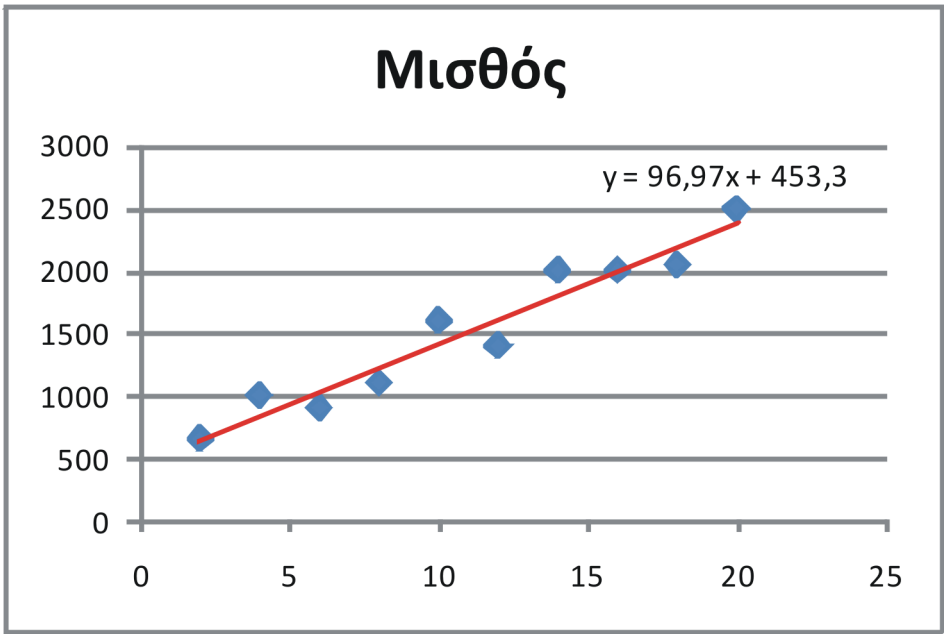
- Το γεγονός ότι γίνονται πολλές συγκρίσεις μεταξύ παρατηρήσεων απαιτεί πολύ αποτελεσματικές τεχνικές καταλογοποίησης (indexing).
- Η κατηγοριοποίηση νέων παρατηρήσεων διαρκεί πολύ περισσότερο χρόνο, ειδικά στις περιπτώσεις όπου ο αριθμός των εν δυνάμει «γειτόνων» είναι μεγάλος.
- Τα αποτελέσματα τους μπορούν να επηρεαστούν σε σημαντικό βαθμό από το πλήθος των γειτόνων k.
- Είναι ευαίσθητοι σε τοπικά χαρακτηριστικά των δεδομένων.
- Είναι ευαίσθητοι στην ύπαρξη μη σημαντικών μεταβλητών εισόδου.

## 10.3 Παλινδρόμηση

Ο όρος **Ανάλυση Παλινδρόμησης** (Regression Analysis) αναφέρεται σε μια οικογένεια στατιστικών τεχνικών, που στοχεύουν στη διερεύνηση σχέσεων μεταξύ μεταβλητών. Πιο συγκεκριμένα, η παλινδρόμηση μοντελοποιεί τη σχέση επίδρασης μιας ή περισσότερων μεταβλητών σε μια άλλη μεταβλητή. Η μεταβλητή της οποίας η τιμή υπολογίζεται καλείται **εξαρτημένη μεταβλητ**ή (dependent variable) ή **μεταβλητή απόκρισης** (respond variable). Οι μεταβλητές οι οποίες χρησιμοποιούνται για τον υπολογισμό της εξαρτημένης μεταβλητής ονομάζονται **ανεξάρτητες** (independent) ή **επεξηγηματικές** (explanatory). Με την Παλινδρόμηση ο χρήστης κατανοεί τον τρόπο με τον οποίο επηρεάζουν οι μεταβολές των ανεξάρτητων μεταβλητών την εξαρτημένη μεταβλητή. Επίσης, η Παλινδρόμηση μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων της τιμής της εξαρτημένης μεταβλητής. Τεχνικές Ανάλυσης Παλινδρόμησης έχουν χρησιμοποιηθεί κατά κόρον σε οικονομικές μελέτες πάσης φύσεως και αποτελούν ένα από τα βασικά και παραδοσιακά εργαλεία διεξαγωγής οικονομικών αναλύσεων.

### 10.3.1 Απλή Γραμμική Παλινδρόμηση

Η απλούστερη εκδοχή παλινδρόμησης είναι όταν η μεταβλητή απόκρισης εξαρτάται από μια μόνο επεξηγηματική μεταβλητή. Ως παράδειγμα, ας υποθέσουμε ότι ο μισθός εξαρτάται μόνο από την εκπαίδευση του εργαζομένου. Για τις ανάγκες του παραδείγματος μετρούμε τον βαθμό εκπαίδευσης με βάση τον χρόνο εκπαίδευσης. Αφού συγκεντρωθούν στοιχεία για ένα σύνολο εργαζομένων, τα στοιχεία αυτά καταγράφονται σε ένα διάγραμμα διασποράς. Κάθε σημείο αντιστοιχεί σε έναν εργαζόμενο, ενώ οι άξονες αντιστοιχούν στον μισθό και τον χρόνο εκπαίδευσης. Τα στοιχεία αυτά παρουσιάζονται στο Σχήμα 10.4. Είναι προφανές ότι ο μισθός αυξάνεται με τα χρόνια εκπαίδευσης. Αυτό που δεν είναι προφανές είναι η ακριβής σχέση ανάμεσα σε αυτές τις δύο μεταβλητές.




**Σχήμα 10.4** *Γραμμική Παλινδρόμηση*

Ο αναλυτής διατυπώνει μια υπόθεση σχετικά με τη σχέση ανάμεσα στις μεταβλητές, τις οποίες μελετά. Υποθέτουμε ότι η σχέση μεταξύ του μισθού και του χρόνου εκπαίδευσης είναι γραμμική. Στην περίπτωση αυτή, η σχέση ανάμεσα στις δύο μεταβλητές μπορεί να αναπαρασταθεί με μια ευθεία γραμμή. Μια γραμμική σχέση ανάμεσα στη μεταβλητή Y και X αποδίδεται με την Εξίσωση 10.17.

$$Y = a + b * X$$

(10.17)

Η παράμετρος α είναι η τιμή του *Y* όταν το *X* ισούται με 0. Στο παράδειγμα μας είναι η αμοιβή του εργαζόμενου με μηδενικό χρόνο εκπαίδευσης. Η παράμετρος *β* καθορίζει την κλίση της ευθείας. Οι δύο παράμετροι, *α* και *β*, είναι άγνωστες και πρέπει να υπολογιστούν με βάση την πληροφορία, η οποία βρίσκεται στο σύνολο δεδομένων. Για τον υπολογισμό τους εφαρμόζεται η μέθοδος των ελαχίστων τετραγώνων. Εάν έχουμε *n* παρατηρήσεις $(x_i, y_i)$ τότε οι συντελεστές α και β υπολογίζονται σύμφωνα με τις ακόλουθες εξισώσεις.

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(10.18)

$$a = \bar{y} - \beta\bar{x}$$

(10.19)

όπου $\bar{x}$ η μέση τιμή των $x_i$ και $\bar{y}$ η μέση τιμή των $y_i$. Για τα δεδομένα τα οποία παρουσιάζονται στο Σχήμα 10.4, η σχέση ανάμεσα στον μισθό και τα χρόνια εκπαίδευσης αποδίδεται από την εξίσωση μισθός=453,3+96,97*χρόνια_εκπαίδευσης. Το Τετραγωνικό Σφάλμα μας δίνει μια εκτίμηση του βαθμού προσέγγισης των πραγματικών τιμών από τη συνάρτηση. Αν $y_i$ είναι οι πραγματικές τιμές και $\hat{y}_i$ είναι οι υπολογισμένες τιμές, τότε το Τετραγωνικό Σφάλμα δίνεται από τη Σχέση 10.29.

$$TΣ = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(10.20)

Ο **συντελεστής προσδιορισμού** r2 (coefficient of determination) είναι ένα μέτρο του βαθμού της συνολικής μεταβλητότητας της εξαρτώμενης μεταβλητής, που εξηγείται από την παλινδρόμηση. Ο συντελεστής προσδιορισμού ορίζεται από την Εξίσωση 10.21.

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(10.21)

Ο συντελεστής προσδιορισμού παίρνει τιμές μεταξύ 0 και 1. Μεγάλη τιμή του σημαίνει ότι η παλινδρόμηση εξηγεί τη μεταβλητότητα της εξαρτημένης μεταβλητής. Αυτό είναι σημαντικό, εάν ο αναλυτής επιθυμεί να χρησιμοποιήσει το μοντέλο για τη διατύπωση προβλέψεων σχετικά με τις τιμές της *Y*.

Σημειώνεται ότι στην Απλή Γραμμική Παλινδρόμηση γίνονται οι παρακάτω παραδοχές:

- **Γραμμικότητα**. Υποθέτουμε ότι οι μέσες τιμές της Y, για τα διάφορα επίπεδα της X, είναι γραμμικές συναρτήσεις της X.
- **Ομοσκεδαστικότητα-Σταθερότητα Διασποράς**. Οι κατανομές της Y έχουν ίδια διασπορά για όλα τα επίπεδα της X.
- **Ανεξαρτησία**. Οι τιμές της Y, που αντιστοιχούν στα διάφορα επίπεδα της X, είναι μεταξύ τους ανεξάρτητες.
- **Κανονικότητα**. Η κατανομή της Y για όλα τα επίπεδα της X είναι κανονική.

### 10.3.2 Πολλαπλή Γραμμική Παλινδρόμηση

Σε πολλά προβλήματα η εξαρτημένη μεταβλητή εξαρτάται όχι από μια, αλλά από περισσότερες μεταβλητές. Η Πολλαπλή Παλινδρόμηση επιτρέπει την προσθήκη πρόσθετων παραγόντων και ποσοτικοποιεί την επίδραση τους στην εξαρτημένη μεταβλητή. Η σχέση ανάμεσα στην εξαρτημένη μεταβλητή Y και στις *n* ανεξάρτητες μεταβλητές $X_i$ δίνεται από την Εξίσωση 10.22

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

(10.22)

Κατ' αντιστοιχία με την απλή γραμμική παλινδρόμηση, οι παράμετροι $a, b_1, b_2, \ldots b_n$ πρέπει να υπολογιστούν. Στην απλή γραμμική παλινδρόμηση, η Εξίσωση 10.17 περιγράφει μια ευθεία γραμμή. Στην πολλαπλή γραμμική παλινδρόμηση με δύο ανεξάρτητες μεταβλητές, η Εξίσωση 10.22 περιγράφει ένα επίπεδο. Η μέθοδος των ελαχίστων τετραγώνων εφαρμόζεται για τον υπολογισμό των παραμέτρων $a, b_1, b_2$, δηλαδή του επιπέδου. Το επίπεδο ορίζεται με τέτοιο τρόπο, ώστε το άθροισμα των τετραγωνικών λαθών ανάμεσα στις προβλεπόμενες και τις πραγματικές τιμές του *Y* να ελαχιστοποιείται. Το α είναι το σημείο τομής του επιπέδου με τον άξονα του *Y*. Το $b_1$ και το $b_2$ είναι οι κλίσεις του επιπέδου ως προς τους άξονες των $X_1$ και $X_2$ αντίστοιχα. Στην πολλαπλή παλινδρόμηση μπορούμε να έχουμε *n* ανεξάρτητες μεταβλητές $X_i$ και στην περίπτωση αυτή κατασκευάζεται ένα υπερεπίπεδο στον αντίστοιχο χώρο. Σημειώνεται ότι για τη σωστή εκτίμηση των τιμών των παραμέτρων απαιτείται μεγάλος αριθμός παρατηρήσεων.

Ένα μοντέλο πολλαπλής παλινδρόμησης μπορεί να χρησιμοποιηθεί για την πρόβλεψη των τιμών της εξαρτημένης μεταβλητής. Σε αυτήν την περίπτωση η τιμή του συντελεστή προσδιορισμού $r^2$ είναι σημαντική. Επίσης, το μοντέλο μπορεί να χρησιμοποιηθεί για την εκτίμηση της σημαντικότητας των ανεξάρτητων μεταβλητών. Οι συντελεστές $b_i$ αποτελούν μια εκτίμηση της επίδρασης της εκάστοτε μεταβλητής $X_i$ στην *Y*. Με τη χρήση στατιστικών τεχνικών, η αναλυτική περιγραφή των οποίων βρίσκεται έξω από τα όρια του παρόντος συγγράμματος, ο αναλυτής μπορεί να ελέγξει και να αποδεχτεί ή να απορρίψει τη μηδενική υπόθεση, ότι η πραγματική τιμή ενός συντελεστή είναι μηδενική, και να εκτιμήσει το κατά πόσο ο συντελεστής είναι στατιστικά σημαντικός.

Ένα ενδεχόμενο πρόβλημα στην πολλαπλή παλινδρόμηση είναι η παράλειψη σημαντικών μεταβλητών, η μη συμμετοχή δηλαδή στο μοντέλο της Εξίσωσης 10.22 ανεξάρτητων μεταβλητών, οι οποίες επηρεάζουν ουσιαστικά την εξαρτημένη μεταβλητή Y. Η παράλειψη σημαντικών μεταβλητών έχει ουσιαστικές επιπτώσεις στην παλινδρόμηση. Ο συντελεστής προσδιορισμού $r^2$ μειώνεται. Επίσης, προκαλούνται μεταβολές στην τιμή

του σταθερού συντελεστή *α*. Αν η μεταβλητή η οποία παραλήφθηκε έχει θετική επίπτωση στην ανεξάρτητη μεταβλητή, τότε η τιμή του *α* αυξάνεται, ενώ αν έχει αρνητική επίπτωση, τότε η τιμή του *α* μειώνεται. Σε περίπτωση όπου η μεταβλητή η οποία παραλήφθηκε συσχετίζεται με κάποια άλλη μεταβλητή, τότε ο συντελεστής αυτής της μεταβλητής τροποποιείται. Ο αναλυτής πρέπει να προσπαθεί να συμπεριλάβει όλες τις σημαντικές μεταβλητές, αν και αυτό δεν είναι πάντα δυνατόν, καθώς μερικές σημαντικές μεταβλητές μπορεί να μην είναι παρατηρήσιμες.

Ένα άλλο γνωστό πρόβλημα στην πολλαπλή παλινδρόμηση είναι το πρόβλημα της πολυσυγγραμμικότητας (multicollinearity). Πολυσυγγραμμικότητα υπάρχει όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες μεταξύ τους, και οι τιμές της μιας μπορούν να υπολογιστούν από την άλλη. Η πολυσυγγραμμικότητα δεν έχει επιπτώσεις στην ικανότητα του μοντέλου να προβλέπει τις τιμές της εξαρτημένης μεταβλητής, έχει όμως επιπτώσεις στους συντελεστές των ανεξάρτητων μεταβλητών. Εάν ο χρήστης χρησιμοποιεί το μοντέλο για την εκτίμηση της σημαντικότητας των ανεξάρτητων μεταβλητών και υπάρχει πρόβλημα πολυσυγγραμμικότητας, τότε τα αποτελέσματα δεν είναι ασφαλή. Οι τιμές των συντελεστών μπορεί να αλλάξουν πολύ, αν προστεθεί ή αφαιρεθεί μια νέα μεταβλητή ή εάν συμβούν μικρές μεταβολές στα δεδομένα. Ένας απλός τρόπος αντιμετώπισης του προβλήματος είναι η απομάκρυνση μεταβλητών από το μοντέλο.

### 10.3.3 Πολυωνυμική Παλινδρόμηση

Υπάρχουν προβλήματα όπου η σχέση ανάμεσα στην εξαρτημένη και την ανεξάρτητη μεταβλητή δεν είναι γραμμική και δεν μπορεί να αποδοθεί από μια συνάρτηση της μορφής Y=α+β*X. Η **Μη Γραμμική Παλινδρόμηση** είναι μια παλινδρόμηση όπου ανάμεσα στην εξαρτημένη και στην ανεξάρτητη μεταβλητή υπάρχει μη γραμμική σχέση.

Η **Πολυωνυμική Παλινδρόμηση** είναι μια περίπτωση Μη Γραμμικής Παλινδρόμησης, όπου η σχέση ανάμεσα στα Y και X περιγράφεται με τη χρήση πολυωνύμου:

$$Y = a + b_1 X + b_2 X^2 + \cdots + b_k X^k$$

(10.23)

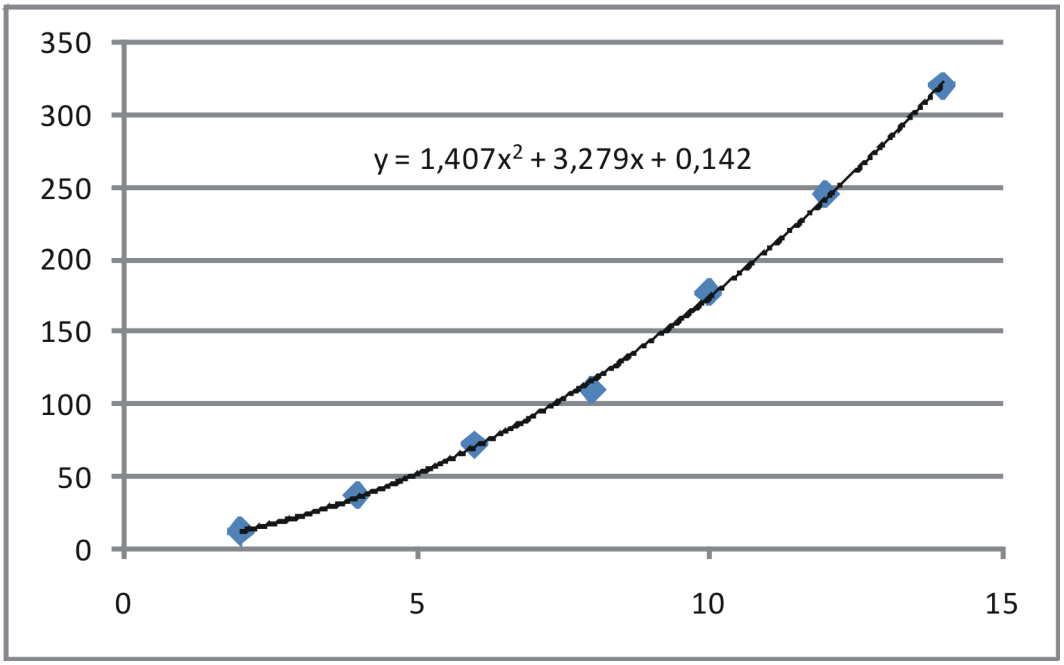
Μια πολυωνυμική συνάρτηση μπορεί να προσεγγίσει την πραγματική παλινδρόμηση, εάν από το διάγραμμα διασποράς ή από τις θεωρητικές γνώσεις για το πρόβλημα προκύπτει ότι η συνάρτηση είναι καμπυλόγραμμη. Το πολυώνυμο της Εξίσωσης 10.23 είναι βαθμού $k$. Στην πράξη πολυώνυμα βαθμού μεγαλύτερου από 2 αποφεύγονται, εκτός εάν δικαιολογούνται θεωρητικά. Παράδειγμα πολυωνυμικής παλινδρόμησης βαθμού 2 παρουσιάζεται στο Σχήμα 10.5

Για τον υπολογισμό των συντελεστών $a$, $b_1$, $b_2$ μπορεί να εφαρμοστεί η μέθοδος των ελαχίστων τετραγώνων. Μια συνάρτηση παλινδρόμησης με πολυώνυμο βαθμού 2 μπορεί να γραφεί ως

$$Y = a + b_1 X_1 + b_2 X_2$$

(10.24)

όπου $X_1$=X και $X_2$=$X^2$. Με τον τρόπο αυτό, η πολυωνυμική παλινδρόμηση μετατρέπεται σε πολλαπλή γραμμική παλινδρόμηση, και ο υπολογισμός των συντελεστών γίνεται με τη μέθοδο των ελαχίστων τετραγώνων.




**Σχήμα 10.5** *Πολυωνυμική Παλινδρόμηση*

### 10.3.4 Λογιστική (ή Λογαριθμική) Παλινδρόμηση

Όλα τα είδη παλινδρόμησης, τα οποία περιγράψαμε μέχρι αυτό το σημείο, μοντελοποιούν τη σχέση ανάμεσα σε ένα σύνολο ανεξάρτητων μεταβλητών και σε μια εξαρτημένη μεταβλητή, η οποία παίρνει αριθμητικές τιμές. Η παλινδρόμηση όμως μπορεί να χρησιμοποιηθεί και για την πρόβλεψη τιμών ονομαστικών πεδίων, μπορεί δηλαδή να εφαρμοστεί σε προβλήματα κατηγοριοποίησης. Ένας πολύ συνηθισμένος τύπος παλινδρόμησης, που χρησιμοποιείται για κατηγοριοποίηση, είναι **Λογιστική Παλινδρόμηση** (Logistic Regression). Στην ελληνική γλώσσα θα τη συναντήσουμε και με το όνομα **Λογαριθμική Παλινδρόμηση**.

Θεωρούμε την περίπτωση δυαδικής κλάσης. Μπορούμε να χρησιμοποιήσουμε την τιμή 1 για τη μια τιμή της κλάσης (πχ χρεοκοπία) και την τιμή 0 για την άλλη τιμή της κλάσης (μη χρεοκοπία). Εάν $p$ είναι η πιθανότητα να πάρει η εξαρτημένη μεταβλητή $Y$ την τιμή 1, τότε η πιθανότητα να πάρει το $Y$ την τιμή 0 είναι $1-p$.

$$P(Y = 1) = p; \qquad P(Y = 0) = 1 - p$$

(10.25)

Μπορούμε να συνδέσουμε την πιθανότητα $p$ με μια γραμμική έκφραση $\alpha+\Sigma bx$ με τη βοήθεια μιας συνάρτησης, η οποία επιστρέφει τιμές στο διάστημα [0..1]. Η συνάρτηση Logit έχει αυτήν την ιδιότητα. Η Λογιστική Παλινδρόμηση περιγράφεται από τη Σχέση 10.26

$$Logit(Y = 1) = ln\left(\frac{P(Y = 1)}{P(Y = 0)}\right) = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

(10.26)

Η πιθανότητα να παίρνει το Y την τιμή 1 υπολογίζεται από την Εξίσωση 10.27.

$$P(Y = 1) = \frac{1}{1 + e^{-(a+b_1X_1+\cdots+b_nX_n)}}$$

(10.27)

Για την εκτίμηση της σημαντικότητας των ανεξάρτητων μεταβλητών, ο αναλυτής μπορεί να χρησιμοποιή-

σει ειδικούς ελέγχους, όπως η στατιστική Wald ή το Likelihood-ratio test. Η Λογιστική Παλινδρόμηση μπορεί να χρησιμοποιηθεί και για την πρόβλεψη κλάσεων με περισσότερες από δύο τιμές.

Η Λογιστική Παλινδρόμηση είναι η παραδοσιακή μέθοδος που χρησιμοποιούν οι οικονομολόγοι για να αντιμετωπίσουν προβλήματα κατηγοριοποίησης. Ο κυριότερος λόγος γι' αυτό είναι το γεγονός ότι κατά κανόνα οι οικονομολόγοι δεν είναι εξοικειωμένοι με μεθόδους κατηγοριοποίησης, οι οποίες προέρχονται από τη Μηχανική Μάθηση

Η Λογιστική Παλινδρόμηση συγκεντρώνει αρκετά **πλεονεκτήματα:**

- Είναι μια μέθοδος αρκετά απλή, δοκιμασμένη και ευρύτατα χρησιμοποιούμενη.
- Ο υπολογισμός των συντελεστών $b_1 \ldots b_n$ είναι ένα μέτρο της σημαντικότητας των ανεξάρτητων μεταβλητών. Υπό τη έννοια αυτή, η Λογιστική Παλινδρόμηση παρέχει μοντέλα ερμηνεύσιμα.
- Η Λογιστική Παλινδρόμηση επιτυγχάνει ικανοποιητικές επιδόσεις κατηγοριοποίησης.

**Μειονεκτήματα** της Λογιστικής Παλινδρόμησης είναι τα εξής:

- Το βασικό μειονέκτημα της Λογιστικής Παλινδρόμησης είναι η διατύπωση αυθαίρετων υποθέσεων, όπως η ύπαρξη γραμμικής σχέσης με τον λογάριθμο του κλάσματος των πιθανοτήτων.
- Σύμφωνα με τα πολλά ερευνητικά αποτελέσματα, άλλες μέθοδοι, όπως τα Νευρωνικά Δίκτυα ή οι Μηχανές διανυσμάτων Υποστήριξης, επιτυγχάνουν τουλάχιστον εφάμιλλες ή και καλύτερες επιδόσεις κατηγοριοποίησης.

## 10.4 Σύνθετοι Κατηγοριοποιητές

Η κατηγοριοποίηση είναι ένα από τα βασικότερα αντικείμενα της Μηχανικής Μάθησης και της Εξόρυξης Δεδομένων. Σήμερα υπάρχουν διαθέσιμες αρκετές μέθοδοι κατηγοριοποίησης, ορισμένες από τις οποίες παρουσιάστηκαν στα πλαίσια του παρόντος και του προηγούμενου κεφαλαίου. Εκτός όμως από αυτήν την ποικιλία «ατομικών» και ριζικά διαφορετικών μεθόδων, υπάρχουν και οι λεγόμενες σύνθετες τεχνικές. Στους σύνθετους κατηγοριοποιητές γίνεται συνδυασμός μοντέλων ή μεθόδων. Αποτελέσματα ερευνητικών εργασιών παρέχουν ισχυρές ενδείξεις ότι οι σύνθετοι κατηγοριοποιητές μπορούν να επιτύχουν υψηλότερες επιδόσεις από τις ατομικές τεχνικές. Τα τελευταία χρόνια μάλιστα, οι σύνθετοι κατηγοριοποιητές γνωρίζουν μεγάλη άνθηση. Ενδεικτικά αναφέρουμε ότι ο Kirkos (2015), σε μια εργασία επισκόπησης βιβλιογραφίας σχετικά με την πρόβλεψη χρεοκοπίας με χρήση ευφυών τεχνικών, διαπιστώνει ότι στις είκοσι από τις συνολικά σαράντα δύο εργασίες εφαρμόστηκαν σύνθετοι κατηγοριοποιητές.

Ο συνδυασμός ατομικών τεχνικών είναι μια απαιτητική εργασία. Οι δυνατότητες συνδυασμού των βασικών μεθόδων είναι πάρα πολλές, και στη σχετική βιβλιογραφία προτείνονται συνεχώς νέες τεχνικές. Παρά τις πολλές διαφορές τους, οι σύνθετοι κατηγοριοποιητές μπορούν να χωριστούν σε δύο βασικές κατηγορίες:

- στους συνδυασμούς κατηγοριοποιητών,
- στους υβριδικούς κατηγοριοποιητές.

Στους **Συνδυασμούς Κατηγοριοποιητών** δημιουργείται ένας σχετικά μεγάλος αριθμός επιμέρους κατηγοριοποιητών. Όλοι αυτοί οι κατηγοριοποιητές εκτελούν την ίδια εργασία, δηλαδή δίνουν απαντήσεις στο ίδιο πρόβλημα κατηγοριοποίησης. Οι αποφάσεις των επιμέρους μοντέλων συναθροίζονται και προκύπτει η τελική απόφαση. Ένα παράδειγμα, το οποίο αναφέρεται συχνά για την καλύτερη κατανόηση των συνδυασμών κατηγοριοποιητών, είναι αυτό του ασθενή, ο οποίος επισκέπτεται πολλούς γιατρούς, συγκρίνει τις γνωματεύσεις τους και αποδέχεται την πλειοψηφούσα γνωμάτευση.

Προφανώς δεν έχει νόημα να αναπαραχθεί πολλές φορές ο ίδιος κατηγοριοποιητής. Κατά συνέπεια οι επιμέρους κατηγοριοποιητές πρέπει να διαφέρουν μεταξύ τους ουσιαστικά. Η επιθυμητή διαφοροποίηση μπορεί να επιτευχθεί με πολλούς τρόπους:

- **Εφαρμογή διαφορετικών ατομικών μεθόδων** και ανάπτυξη αντίστοιχων μοντέλων. Μπορούν να αναπτυχθούν μοντέλα Νευρωνικών Δικτύων, Δένδρων Αποφάσεων, Μηχανών Διανυσμάτων Υποστήριξης και άλλων μεθόδων, και να συνδυαστούν οι προβλέψεις τους. Κατά κανόνα όλα τα μοντέλα εκπαιδεύονται χρησιμοποιώντας τα ίδια δεδομένα.
- **Χρήση διαφορετικών δεδομένων εκπαίδευσης**. Εφαρμόζεται μόνο μια μέθοδος. Από το αρχικό

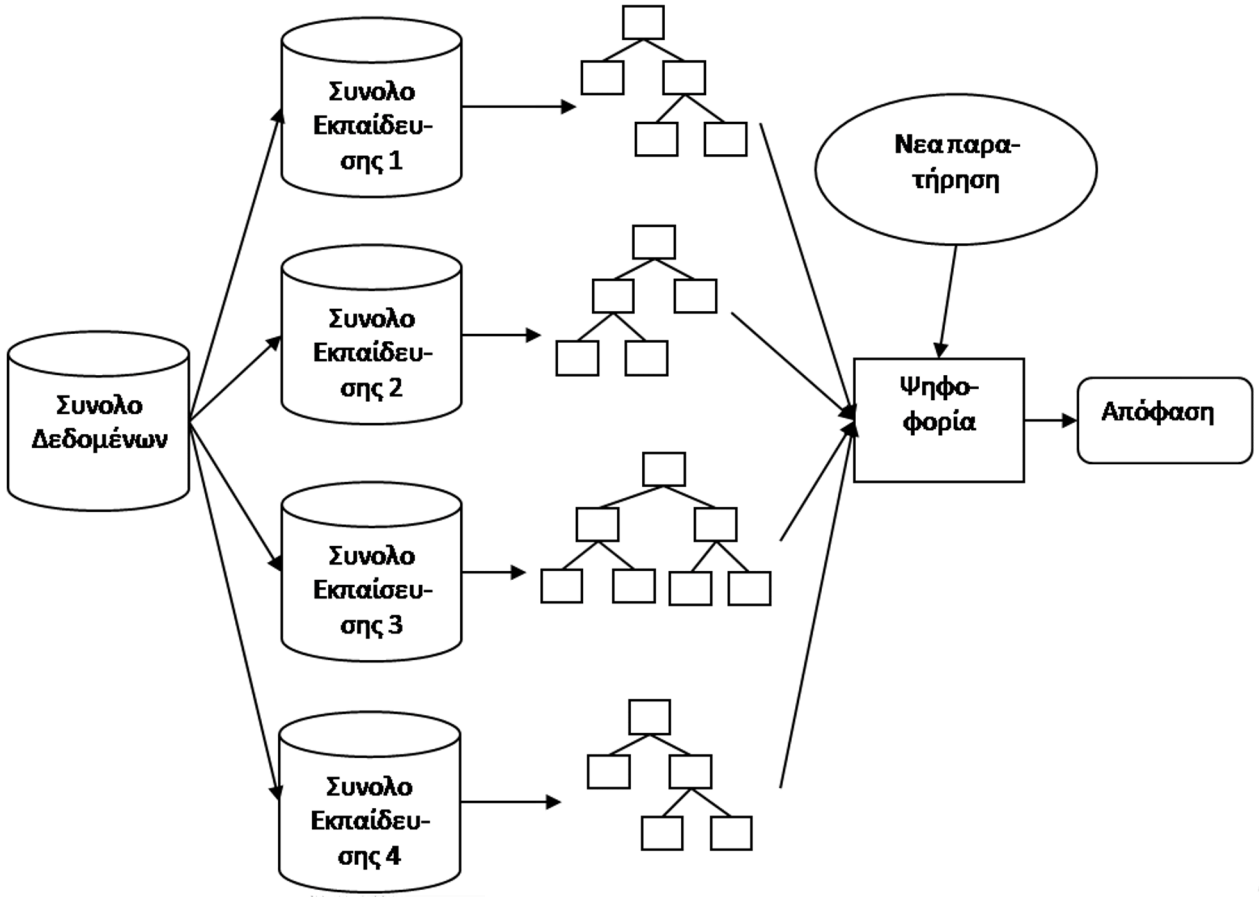
σύνολο δεδομένων δημιουργούνται πολλά σύνολα εκπαίδευσης. Η διαφοροποίηση των μοντέλων πηγάζει από τη χρήση διαφορετικών δεδομένων για την εκπαίδευση.

- **Χρήση διαφορετικών χώρων χαρακτηριστικών** (feature space). Επιλέγονται διαφορετικά σύνολα μεταβλητών εισόδου και εκπαιδεύονται αντίστοιχα μοντέλα. Τα διαφορετικά σύνολα μεταβλητών εισόδου μπορούν να προκύψουν από την εφαρμογή διαφορετικών μεθόδων επιλογής χαρακτηριστικών (feature selection).
- **Διαφορετική ρύθμιση παραμέτρων της ίδιας βασικής μεθόδου**. Εφαρμόζεται μια βασική μέθοδος και χρησιμοποιείται το ίδιο σύνολο δεδομένων για την εκπαίδευση των μοντέλων. Η διαφοροποίηση των μοντέλων προέρχεται από τη διαφορετική ρύθμιση των παραμέτρων. Για παράδειγμα, αν χρησιμοποιηθεί η μέθοδος των Νευρωνικών Δικτύων μπορούν να εκπαιδευτούν μοντέλα με διαφορετικές αρχιτεκτονικές, ρυθμούς εκπαίδευσης, εποχές κλπ.

Ένα βασικό ζήτημα στον σχεδιασμό συνδυασμού κατηγοριοποιητών είναι ο καθορισμός του τρόπου με τον οποίο ένας κατηγοριοποιητής επηρεάζει τους άλλους κατηγοριοποιητές. Οι κατηγοριοποιητές μπορούν να συνδυαστούν με σειριακό ή με παράλληλο τρόπο. Στον σειριακό τρόπο διεξάγεται μια επαναληπτική διαδικασία. Σε κάθε επανάληψη χρησιμοποιείται γνώση, η οποία αποκτήθηκε στο προηγούμενο στάδιο, και επηρεάζει την εκπαίδευση στο τρέχων στάδιο. Η αποκτηθείσα γνώση μπορεί να εκφραστεί σαν χειρισμός και τροποποίηση των δεδομένων εκπαίδευσης. Μια διαφορετική προσέγγιση είναι να χρησιμοποιηθεί ο κατηγοριοποιητής που δημιουργήθηκε σε ένα στάδιο για τη δημιουργία του νέου κατηγοριοποιητή στο επόμενο στάδιο.

Το πιο γνωστό **παράδειγμα σειριακού συνδυασμού** με ταυτόχρονο χειρισμό των δεδομένων εκπαίδευσης είναι η τεχνική **boosting.** Στον αλγόριθμο AdaBoost (Freund and Schapire, 1996), από ένα αρχικό σύνολο δεδομένων δημιουργείται ένα νέο σύνολο εκπαίδευσης, ίδιου μεγέθους με το αρχικό. Για τη δημιουργία του νέου συνόλου εκπαίδευσης εφαρμόζεται δειγματοληψία με επανατοποθέτηση. Αυτό σημαίνει ότι όταν επιλέγεται μια παρατήρηση και τοποθετείται στο νέο σύνολο εκπαίδευσης δεν απομακρύνεται από το αρχικό σύνολο δεδομένων. Κατά συνέπεια, στην επόμενη επιλογή είναι πιθανό να επιλεγεί πάλι η ίδια παρατήρηση. Με τον τρόπο αυτό, μια παρατήρηση μπορεί να συμμετέχει στο νέο σύνολο εκπαίδευσης πολλές φορές. Από το νέο σύνολο δεδομένων κατασκευάζεται ένας κατηγοριοποιητής. Σε κάθε παρατήρηση εκχωρείται ένας συντελεστής βαρύτητας. Εάν το μοντέλο κατηγοριοποιεί εσφαλμένα μια παρατήρηση τότε ο συντελεστής βαρύτητας αυξάνεται, διαφορετικά μειώνεται. Οι συντελεστές βαρύτητας χρησιμοποιούνται στο επόμενο στάδιο εκπαίδευσης, έτσι ώστε ο επόμενος κατηγοριοποιητής να δώσει περισσότερη προσοχή σε αυτές τις παρατηρήσεις. Με τον τρόπο αυτό, δημιουργείται μια σειρά διαδοχικών κατηγοριοποιητών. Κάθε κατηγοριοποιητής σχετίζεται με έναν συντελεστή βαρύτητας, ο οποίος είναι συνάρτηση της ακρίβειας του. Η τελική απόφαση λαμβάνεται από τις αποφάσεις των επιμέρους μοντέλων με ψηφοφορία και με χρήση των βαρών. Ο αλγόριθμος δίνει υψηλά ποσοστά ακρίβειας, υπάρχει όμως κίνδυνος υπερπροσαρμογής στις παρατηρήσεις που κατηγοριοποιούνται λανθασμένα.

Σύμφωνα με την προσέγγιση του **παράλληλου συνδυασμού κατηγοριοποιητών**, δημιουργούνται πολλαπλά σύνολα εκπαίδευσης από το αρχικό σύνολο δεδομένων. Για κάθε σύνολο εκπαίδευσης δημιουργείται ένα διαφορετικό μοντέλο. Η τελική απόφαση κατηγοριοποίησης μιας νέας παρατήρησης λαμβάνεται με κάποια μορφή συνάθροισης των προβλέψεων των επιμέρους μοντέλων. Ένας πολύ δημοφιλής αλγόριθμος παράλληλου συνδυασμού είναι ο **Bagging (**Breiman, 1996). Στον αλγόριθμο Bagging, τα σύνολα εκπαίδευσης δημιουργούνται με δειγματοληψία με επανατοποθέτηση. Για την κατηγοριοποίηση μιας νέας παρατήρησης γίνεται ψηφοφορία μεταξύ των μοντέλων, και η παρατήρηση εκχωρείται στην κλάση που συγκέντρωσε τις περισσότερες ψήφους. Σύμφωνα με τον Breiman, το Bagging λειτουργεί πολύ αποτελεσματικά με «ασταθείς» μεθόδους, όπου μικρές αλλαγές στο σύνολο εκπαίδευσης οδηγούν σε σημαντικά διαφορετικά μοντέλα. Παράδειγμα τέτοιας μεθόδου είναι τα Δένδρα Αποφάσεων. Η τεχνική Bagging παρουσιάζεται διαγραμματικά στο Σχήμα 10.6.




**Σχήμα 10.6** *Bagging*

Ένα πολύ σημαντικό ζήτημα στην κατασκευή συνδυασμού κατηγοριοποιητών είναι ο καθορισμός του τρόπου με τον οποίο συνδυάζονται οι αποφάσεις των επιμέρους μοντέλων. Υπάρχουν δύο διαθέσιμες προσεγγίσεις, οι **απλές συνδυαστικές μέθοδοι** και οι **μετα-συνδυαστικές μέθοδοι**. Στις απλές συνδυαστικές μεθόδους, οι επιμέρους αποφάσεις συνδυάζονται σύμφωνα με κάποια συνάρτηση. Ο απλούστερος τρόπος είναι η απλή ψηφοφορία και η εκχώρηση της παρατήρησης στην πλειοψηφούσα κλάση. Ωστόσο, έχουν προταθεί και πολλά άλλα συνδυαστικά σχήματα, όπως η Άθροιση Κατανομής (Distribution Summation), η Καταμέτρηση Borda (Borda Count) και η Θεωρία Dempster-Shafer (Dempster Shafer Theory (DST)). Η Άθροιση Κατανομής συναθροίζει τις πιθανότητες ένταξης σε κάθε κλάση, τις οποίες υπολογίζει το εκάστοτε μοντέλο. Σύμφωνα με την καταμέτρηση Borda, οι επιμέρους κατηγοριοποιητές ταξινομούν τις υποψήφιες τιμές κλάσης σε σειρά προτεραιότητας. Κάθε θέση στην κλίμακα ταξινόμησης συσχετίζεται με έναν αριθμό «πόντων». Οι πόντοι της κάθε κλάσης συναθροίζονται, και η παρατήρηση εκχωρείται στην κλάση που συγκεντρώνει τους περισσότερους πόντους. Στην τεχνική DST επικρατεί η κλάση για την οποία μεγιστοποιείται η τιμή μιας συνάρτησης, η οποία χρησιμοποιεί τις πιθανότητες που υπολογίζουν οι βασικοί κατηγοριοποιητές. Οι μετα-συνδυαστικές μέθοδοι χρησιμοποιούν τους βασικούς κατηγοριοποιητές και τις προβλέψεις τους για περαιτέρω μάθηση. Στη μέθοδο Stack Generalization (Wolpert, 1992) τα αποτελέσματα των βασικών κατηγοριοποιητών χρησιμοποιούνται ως είσοδος από τους κατηγοριοποιητές του επόμενου επιπέδου.

Οι **Υβριδικοί Κατηγοριοποιητές** είναι η δεύτερη μεγάλη κατηγορία των σύνθετων κατηγοριοποιητών. Στα υβριδικά συστήματα, όπως και στους συνδυασμούς κατηγοριοποιητών, χρησιμοποιούνται διάφορες τεχνικές. Ωστόσο υπάρχουν σημαντικές διαφορές σε σχέση με τις μεθόδους συνδυασμού κατηγοριοποιητών. Η πρώτη διαφορά είναι ότι εφαρμόζονται ετερογενείς τεχνικές, κάθε μια από τις οποίες επιλύει ένα διαφορετικό πρόβλημα. Η δεύτερη διαφορά είναι ότι η τελική απόφαση κατηγοριοποίησης λαμβάνεται από έναν μόνο κατηγοριοποιητή, ενώ στους συνδυασμούς κατηγοριοποιητών γίνεται συνδυασμός των αποφάσεων πολλών κατηγοριοποιητών.

Οι Lin, Hu and Tsai (2012) ορίζουν τους ακόλουθους τρεις τύπους υβριδικών κατηγοριοποιητών:

- διαδοχικές τεχνικές,

- συνδυασμός κατηγοριοποίησης και ανάλυσης συστάδων,
- ολοκλήρωση δύο τεχνικών με συμπληρωματικό τρόπο.

Στις διαδοχικές υβριδικές τεχνικές πραγματοποιείται μια επεξεργασία σε ένα πρώτο στάδιο και στη συνέχεια, τα αποτελέσματα αυτής της επεξεργασίας χρησιμοποιούνται ως είσοδος στο επόμενο στάδιο. Για παράδειγμα, στο πρώτο στάδιο μπορεί να υπολογίζονται κάποιες τιμές, οι οποίες θα χρησιμοποιηθούν στο επόμενο στάδιο ως πρόσθετα δεδομένα εισόδου. Μια διασταλτική ερμηνεία του όρου των διαδοχικών υβριδικών τεχνικών θα μπορούσε να αποδεχτεί τη μείωση της διαστασιμότητας ή του πλήθους των παρατηρήσεων εκπαίδευσης με την εφαρμογή μεθόδων soft computing, και την ακόλουθη ανάπτυξη ενός κατηγοριοποιητή, ως περίπτωση υβριδικών μοντέλων.

Ο δεύτερος τύπος υβριδικών κατηγοριοποιητών είναι ο συνδυασμός τεχνικών κατηγοριοποίησης με τεχνικές ανάλυσης συστάδων. Οι τεχνικές ανάλυσης συστάδων μπορούν να εφαρμοστούν για τον εντοπισμό και την απομάκρυνση εξαιρέσεων και παρατηρήσεων με ακραίες τιμές. Επίσης, η ανάλυση συστάδων μπορεί να εντοπίσει ομάδες ομοειδών παρατηρήσεων, οι οποίες θα θεωρηθούν κλάσεις και θα χρησιμοποιηθούν για περαιτέρω κατηγοριοποίηση.

Πιθανώς, η πιο γνήσια εκδοχή υβριδικών συστημάτων είναι αυτά τα οποία ολοκληρώνουν δύο ετερογενείς μεθόδους σε μια ενιαία διαδικασία εκπαίδευσης. Μια διαδεδομένη τεχνική, για την ανάπτυξη υβριδικών κατηγοριοποιητών αυτού του τύπου, είναι ο συνδυασμός Εξελικτικών Αλγορίθμων με μεθόδους κατηγοριοποίησης. Εξελικτικοί αλγόριθμοι που εφαρμόζονται συχνά για τη δημιουργία υβριδικών κατηγοριοποιητών είναι οι Γενετικοί Αλγόριθμοι (Genetic Algorithms) και η Βελτιστοποίηση Σμήνους Σημείων (Particle Swarm Optimization). Οι Γενετικοί Αλγόριθμοι παρουσιάζονται στο Κεφάλαιο 3. Η Βελτιστοποίηση Σμήνους Σημείων είναι μια τεχνική, που προσομοιάζει την κίνηση σμήνους πτηνών ή κοπαδιού ψαριών, με στόχο την εύρεση της βέλτιστης θέσης μέσα στο σμήνος.

Οι εξελικτικοί αλγόριθμοι μπορούν να εφαρμοστούν με διάφορους τρόπους. Η απλούστερη τεχνική είναι να χρησιμοποιηθούν για τη ρύθμιση των παραμέτρων άλλων μεθόδων. Για παράδειγμα, μπορεί να δημιουργηθεί ένας πληθυσμός Νευρωνικών Δικτύων με διαφορετικές αρχιτεκτονικές, ρυθμούς μάθησης, εποχές εκπαίδευσης κλπ. και με τη χρήση των εξελικτικών αλγορίθμων να επιλεγεί το δίκτυο με την καλύτερη ρύθμιση παραμέτρων. Άλλη εκδοχή είναι να ενσωματωθούν οι εξελικτικοί αλγόριθμοι στη διαδικασία εκπαίδευσης του κατηγοριοποιητή. Μια τρίτη εκδοχή είναι να γίνει επιλογή σημαντικών χαρακτηριστικών και ταυτόχρονη ρύθμιση παραμέτρων. Με τον τρόπο αυτό, επιτυγχάνεται κατάλληλη ρύθμιση των παραμέτρων για τις συγκεκριμένες μεταβλητές εισόδου. Τέλος, σε αυτήν την κατηγορία των υβριδικών κατηγοριοποιητών ανήκουν και τα Νεύρω-Ασαφή συστήματα, τα οποία συνδυάζουν την Ασαφή Λογική με τα Νευρωνικά Δίκτυα.

Στην πρόσφατη έρευνα έχουν προταθεί σχήματα, τα οποία επιτρέπουν τη συναρμογή συνδυασμών κατηγοριοποιητών και υβριδικών τεχνικών. Εξελικτικοί αλγόριθμοι έχουν εφαρμοστεί για τη βελτιστοποίηση συνδυασμών κατηγοριοποιητών. Στους συνδυασμούς κατηγοριοποιητών μια σημαντική ιδιότητα, η οποία επηρεάζει και την επίδοση, είναι η ουσιαστική διαφοροποίηση των επιμέρους μοντέλων. Οι Γενετικοί Αλγόριθμοι μπορούν να χρησιμοποιηθούν για να επιλέξουν από μια δεξαμενή διαθέσιμων βασικών κατηγοριοποιητών εκείνους τους κατηγοριοποιητές που παρουσιάζουν αυξημένη διαφοροποίηση.

## 10.5 Επικύρωση Κατηγοριοποιητών

Σύμφωνα με ότι έχει αναφερθεί μέχρι τώρα, οι μέθοδοι κατηγοριοποίησης επεξεργάζονται ένα σύνολο παρατηρήσεων και εκπαιδεύουν ένα μοντέλο. Το μοντέλο συνίσταται σε έναν μηχανισμό λήψης απόφασης, για το εάν οι παρατηρήσεις του δείγματος ανήκουν σε μια κλάση. Η ακρίβεια του μοντέλου μέχρι στιγμής έχει οριστεί σε σχέση με την ικανότητα του να κατατάσσει σωστά τις παρατηρήσεις του συνόλου εκπαίδευσης. **Το ερώτημα που προκύπτει είναι τι θα συμβεί εάν το μοντέλο αντιμετωπίσει «άγνωστες» παρατηρήσεις**, παρατηρήσεις δηλαδή που δεν ανήκουν στο σύνολο εκπαίδευσης. Ουσιαστικά, το ερώτημα που τίθεται είναι εάν το μοντέλο ενσωματώνει γενικευμένους κανόνες ευρύτερης ισχύος, οι οποίοι καθορίζουν τον προσδιορισμό της κλάσης μιας παρατήρησης στον πραγματικό κόσμο, ή εάν το μοντέλο ενσωματώνει εξειδικευμένους κανόνες, που καθορίζουν τον προσδιορισμό της κλάσης των παρατηρήσεων του συγκεκριμένου συνόλου. Προφανώς η πραγματική αξία ενός μοντέλου βρίσκεται στην ικανότητα του να προβλέπει την κλάση άγνωστων παρατηρήσεων του πραγματικού κόσμου.

Οι αυξημένες επιδόσεις ενός μοντέλου έναντι του συνόλου εκπαίδευσης δεν συνεπάγονται και αυξημένη ικανότητα κατηγοριοποίησης άγνωστων παρατηρήσεων. Στο Κεφάλαιο 9 έγινε αναφορά στο πρόβλημα της υπερπροσαρμογής των μοντέλων (data overfitting). Η υπερπροσαρμογή παρουσιάζεται όταν ένα μοντέλο εί-

ναι υπερβολικά περίπλοκο. Το μοντέλο αυτό είναι ικανό να αφομοιώσει τις ιδιαιτερότητες των δεδομένων εκπαίδευσης, αντί να καταγράψει σχέσεις γενικότερης ισχύος. Το αποτέλεσμα της υπερπροσαρμογής είναι ιδιαίτερα ψηλές επιδόσεις έναντι του συνόλου εκπαίδευσης, αλλά δυσανάλογα χαμηλές επιδόσεις έναντι άγνωστων παρατηρήσεων.

Για τους παραπάνω λόγους, η ακρίβεια ενός μοντέλου πρέπει να εκτιμάται έναντι άγνωστων παρατηρήσεων. Ο καθορισμός της ακρίβειας ενός μοντέλου είναι ιδιαίτερα σημαντικός, γιατί μας επιτρέπει να αποφανθούμε εάν το μοντέλο μπορεί να χρησιμοποιηθεί για τη λήψη αποφάσεων στον πραγματικό κόσμο. Επίσης, μας επιτρέπει να συγκρίνουμε διαφορετικά μοντέλα, ώστε να επιλέξουμε το καλύτερο. Για την εκτίμηση τις ικανότητας ενός μοντέλου να προβλέπει άγνωστες παρατηρήσεις έχουν προταθεί διάφορες μέθοδοι, όπως η διάσπαση του δείγματος σε δείγμα εκπαίδευσης και δείγμα επικύρωσης (holdout method), η επικύρωση 10 τμημάτων (10-fold cross validation), η μέθοδος «άφησε ένα έξω» (leave one out) και τέλος, η μέθοδος bootstrapping.

**Μέθοδος Holdout.** Κατά τη μέθοδο holdout, το σύνολο δεδομένων διασπάται σε δύο υποσύνολα, κάθε ένα από τα οποία περιέχει διαφορετικές παρατηρήσεις. Το ένα υποσύνολο χρησιμοποιείται για την εκπαίδευση του μοντέλου και ονομάζεται **σύνολο εκπαίδευσης** (training set). Αφού ολοκληρωθεί η εκπαίδευση, το μοντέλο αποπειράται να προβλέψει την κλάση των παρατηρήσεων του δεύτερου υποσυνόλου, και ακολούθως συγκρίνονται οι προβλέψεις του μοντέλου με την πραγματική κλάση των παρατηρήσεων. Το δεύτερο υποσύνολο είναι γνωστό ως **σύνολο επικύρωσης** (validation set ή holdout set). Μια ενδεδειγμένη πρακτική είναι να χρησιμοποιούνται τα δύο τρίτα του αρχικού συνόλου ως σύνολο εκπαίδευσης και το ένα τρίτο ως σύνολο επικύρωσης. Η επίδοση του μοντέλου είναι το ποσοστό των ορθών προβλέψεων. Μια παραλλαγή της μεθόδου holdout είναι η μέθοδος της τυχαίας υποδειγματοληψίας (random subsampling). Σύμφωνα με τη μέθοδο αυτή, γίνεται επανάληψη της μεθόδου holdout πολλές φορές. Σε κάθε επανάληψη δημιουργούνται νέα σύνολα εκπαίδευσης και επικύρωσης, εφαρμόζοντας τυχαία δειγματοληψία.

**Διασταυρούμενη Επικύρωση 10 τμημάτων.** Στη μέθοδο επικύρωσης 10 τμημάτων (10 fold cross validation) το σύνολο δεδομένων διαιρείται σε 10 υποσύνολα. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις. Η επιλογή των υποσυνόλων είναι τυχαία. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης και τα υπόλοιπα εννέα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και δοκιμάζεται έναντι του συνόλου επικύρωσης. Η διαδικασία επαναλαμβάνεται δέκα φορές, κάθε φορά χρησιμοποιώντας ένα διαφορετικό σύνολο ως σύνολο επικύρωσης και τα υπόλοιπα εννέα ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου. Η μέθοδος μπορεί να διαφοροποιηθεί ως προς το πλήθος των τμημάτων. Γενικότερα ονομάζεται μέθοδος επικύρωσης *k* τμημάτων, όπου *k* συμβολίζει τον αριθμό των δημιουργημένων υποσυνόλων και των επαναλήψεων. Μια άλλη εκδοχή της μεθόδου είναι η **στρωματοποιημένη** επικύρωση 10 τμημάτων (stratified 10 fold cross validation). Σύμφωνα με αυτήν την εκδοχή, κάθε υποσύνολο περιέχει περίπου ίσο αριθμό παρατηρήσεων για την κάθε κλάση.

**Μέθοδος «άφησε ένα έξω»**. Η μέθοδος «άφησε ένα έξω» (leave one out) ουσιαστικά αποτελεί παραλλαγή της μεθόδου επικύρωσης *k* τμημάτων. Στην παραλλαγή αυτή το *k=n*, όπου n είναι το πλήθος των παρατηρήσεων, οι οποίες απαρτίζουν το σύνολο δεδομένων. Για κάθε μια παρατήρηση, το μοντέλο εκπαιδεύεται χρησιμοποιώντας τις υπόλοιπες *n-1* παρατηρήσεις και επικυρώνεται έναντι της επιλεγμένης παρατήρησης. Η διαδικασία επαναλαμβάνεται *n* φορές. Στο τέλος υπολογίζεται το ποσοστό ορθών παρατηρήσεων.

**Μέθοδος bootstrap**. Στη μέθοδο bootstrap δημιουργούνται πάλι πολλαπλά σύνολα επικύρωσης με δειγματοληψία. Η διαφορά έγκειται στο γεγονός ότι η δειγματοληψία γίνεται με επανατοποθέτηση. Κάθε παρατήρηση που επιλέγεται να συμμετάσχει στο δείγμα επικύρωσης δεν αφαιρείται από το αρχικό δείγμα. Κατά τον τρόπο αυτό, μια παρατήρηση μπορεί να επιλεγεί περισσότερες από μία φορές για να συμμετάσχει στο ίδιο σύνολο επικύρωσης.

Ο Kohavi (1995) πραγματοποίησε μια συγκριτική μελέτη σχετικά με τις μεθόδους επικύρωσης κατηγοριοποιητών. Σύμφωνα με τα αποτελέσματα του, η πλέον κατάλληλη μέθοδος είναι η στρωματοποιημένη επικύρωση 10 τμημάτων.

## 10.6 Ανισοκατανομή κλάσεων και κόστος σφάλματος

Το ποσοστό των επιτυχών προβλέψεων ενός κατηγοριοποιητή είναι ένα ισχυρό μέτρο της ικανότητας πρόβλεψης, σε ορισμένες περιπτώσεις όμως δεν είναι επαρκές. Σε πολλά προβλήματα του πραγματικού κόσμου **οι παρατηρήσεις δεν είναι ισομερώς κατανεμημένες στις διάφορες κλάσεις**. Το φαινόμενο είναι πολύ συνηθισμένο σε ιατρικά δεδομένα, όπου η πιθανότητα εμφάνισης μιας ασθένειας είναι μικρή. Σε μια βάση δεδομένων με αποτελέσματα εξετάσεων, ένα μικρό ποσοστό των παρατηρήσεων θα αφορά περιπτώσεις ασθένειας.

Σε οικονομικά δεδομένα, μια από τις χαρακτηριστικότερες περιπτώσεις ανισοκατανομής των κλάσεων είναι η πρόβλεψη χρεοκοπίας. Στις ΗΠΑ το ποσοστό αποτυχίας των επιχειρήσεων είναι περίπου 2%. Αυτό σημαίνει ότι ένας κατηγοριοποιητής, ο οποίος προβλέπει πάντα μη χρεοκοπία, θα είχε ακρίβεια 98%. Η ακρίβεια του μοντέλου είναι εξαιρετικά υψηλή, στην πραγματικότητα όμως το μοντέλο είναι απολύτως αποτυχημένο, αφού αδυνατεί να προβλέψει τις περιπτώσεις χρεοκοπίας, που είναι και το ζητούμενο. Σε σύνολα δεδομένων όπου οι παρατηρήσεις δεν είναι ισομερώς κατανεμημένες στις κλάσεις, το γενικό ποσοστό επιτυχών προβλέψεων δεν επαρκεί για την εκτίμηση των δυνατοτήτων του κατηγοριοποιητή. Επίσης, η ανισοκατανομή των κλάσεων έχει επιπτώσεις στην εκπαίδευση των κατηγοριοποιητών. Μελέτες έχουν δείξει ότι τα παραγόμενα μοντέλα τείνουν να προβλέπουν καλύτερα την πλειοψηφούσα κλάση και σημαντικά χειρότερα τη μειοψηφούσα κλάση (Weiss & Provost, 2003).

Η φυσική κατανομή των παρατηρήσεων σε κλάσεις συχνά δεν είναι η καλύτερη κατανομή για την εκπαίδευση ενός κατηγοριοποιητή. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες τεχνικές επαναδειγματοληψίας (resampling). Η τυχαία υποδειγματοληψία (random undersampling) είναι μια τεχνική, η οποία απομακρύνει με τυχαίο τρόπο παρατηρήσεις της πλειοψηφούσας κλάσης, μέχρι να επιτευχθεί ίσο πλήθος παρατηρήσεων για την κάθε κλάση. Μειονέκτημα αυτής της τεχνικής είναι ότι διαγράφονται παρατηρήσεις, οι οποίες πιθανώς περιείχαν ουσιαστική πληροφορία. Η τυχαία υπερδειγματοληψία (random oversampling) αναπαράγει με τυχαίο τρόπο τις παρατηρήσεις της μειοψηφούσας κλάσης, μέχρι να ισοσταθμιστεί το πλήθος των κλάσεων. Μειονέκτημα της τεχνικής είναι ότι μπορεί να οδηγήσει σε υπερπροσαρμογή. Για την αντιμετώπιση αυτών των προβλημάτων έχουν προταθεί τεχνικές, οι οποίες επιλεκτικά αναπαράγουν παρατηρήσεις της μειοψηφούσας κλάσης ή/και διαγράφουν παρατηρήσεις της πλειοψηφούσας κλάσης (Kubat & Martin, 1997; Japkowicz, 2000; Chawla, Boywer, Hall & Kegelmeyer, 2002).

Μια άλλη περίπτωση όπου το γενικό ποσοστό επιτυχών προβλέψεων δεν επαρκεί, είναι όταν οι **αποτυχίες πρόβλεψης διαφορετικών κλάσεων δεν έχουν το ίδιο κόστος**. Ας επανέρθουμε στο παράδειγμα πρόβλεψης χρεοκοπίας. Οι περιπτώσεις εσφαλμένων προβλέψεων είναι οι εξής δύο: α) Μια επιχείρηση που θα χρεοκοπήσει κατηγοριοποιείται ως βιώσιμη. β) Μια επιχείρηση που δεν θα χρεοκοπήσει κατηγοριοποιείται ως χρεοκοπημένη. Η πρώτη περίπτωση είναι ένα σφάλμα Τύπου Ι. Η δεύτερη περίπτωση είναι ένα σφάλμα Τύπου ΙΙ. Ποιο είναι το κόστος των σφαλμάτων για έναν τραπεζικό οργανισμό; Στην περίπτωση σφάλματος Τύπου ΙΙ, η τράπεζα δεν θα ενέκρινε ένα επιτυχημένο δάνειο και θα έχανε το σχετικό κέρδος. Στην περίπτωση σφάλματος Τύπου Ι, η τράπεζα θα ενέκρινε ένα αποτυχημένο δάνειο και θα έχανε το αντίστοιχο κεφάλαιο. Είναι προφανές ότι τα δύο σφάλματα έχουν διαφορετικό κόστος. Στον πραγματικό κόσμο είναι δύσκολο να βρεθούν προβλήματα όπου το κόστος διαφορετικού τύπου σφαλμάτων είναι το ίδιο (Witten & Frank, 2000).

Οι μέθοδοι κατηγοριοποίησης είναι σχεδιασμένες έτσι ώστε να ελαχιστοποιούν τον ρυθμό σφάλματος, δηλαδή τον συνολικό αριθμό εσφαλμένων προβλέψεων ή την πιθανότητα εσφαλμένων προβλέψεων. Η προσέγγιση αυτή υποθέτει ότι το κόστος διαφορετικών τύπων σφάλματος είναι το ίδιο. Σε ρεαλιστικά προβλήματα όμως αυτό που συχνά έχει σημασία είναι η μείωση του κόστους εσφαλμένων κατηγοριοποιήσεων, και όχι η αύξηση του ρυθμού ακρίβειας. Για μια τράπεζα σημασία έχει η λήψη επικερδών αποφάσεων και όχι η διατύπωση πολλών επιτυχών προβλέψεων.

Η εκπαίδευση μοντέλων με τρόπο τέτοιο ώστε να μειώνεται το συνολικό κόστος εσφαλμένων προβλέψεων ονομάζεται **ευαίσθητη ως προς το κόστος εκπαίδευση** (cost sensitive learning). Για εκπαίδευση ευαίσθητη ως προς το κόστος, απαιτείται καταρχάς ο καθορισμός του συνολικού σφάλματος. Έχουν προταθεί διάφοροι ορισμοί για το συνολικό κόστος σφάλματος. Για περιπτώσεις δυαδικής κλάσης, οι Chen, Huang and Lin (2009) το ορίζουν σύμφωνα με τη Σχέση 10.28

$$MC = (k_1 * T1Err + k_2 * T2Err)$$

(10.28)

όπου *T1Err* είναι το πλήθος σφαλμάτων Τύπου Ι, *T2Err* το πλήθος σφαλμάτων Τύπου ΙΙ, και $k_1$, $k_2$ το κόστος σφάλματος Τύπου Ι και ΙΙ αντίστοιχα. Οι Chen, Ribeiro, Vieira, Duarte and Neves (2011) ορίζουν το συνολικό κόστος σφάλματος σύμφωνα με τη Σχέση 10.29.

$$MC = \frac{k_1}{k_1 + k_2} * T1Err + \frac{k_2}{k_1 + k_2} * T2Err$$

(10.29)

Πρόσθετοι ορισμοί έχουν προταθεί από άλλους ερευνητές. Ένα ανοικτό ζήτημα σε κάθε πρόβλημα εκπαίδευσης ευαίσθητης ως προς το κόστος, είναι ο καθορισμός της αναλογίας του κόστους διαφορετικών σφαλμάτων $k_1/k_2$. Για τον καθορισμό της αναλογίας απαιτείται καλή γνώσης του πεδίου εφαρμογής και η γνώμη ειδικών. Σε πολλά προβλήματα δεν υπάρχει μια αναγνωρισμένη και γενικώς αποδεκτή αναλογία. Για παράδειγμα, σε εργασίες πρόβλεψης χρεοκοπίας έχουν χρησιμοποιηθεί αναλογίες, οι οποίες κυμαίνονται από *1/1* έως και *1/100*.

Στην ευαίσθητη ως προς το κόστος εκπαίδευση, τα μοντέλα εκπαιδεύονται με τέτοιο τρόπο ώστε να μειωθεί το κόστος των εσφαλμένων κατηγοριοποιήσεων. Μια τέτοιου τύπου εκπαίδευση μπορεί να επιτευχθεί με παρεμβάσεις είτε στο επίπεδο των δεδομένων είτε στο επίπεδο του αλγορίθμου. Σύμφωνα με την πρώτη προσέγγιση πραγματοποιείται επαναδειγματοληψία, ώστε να μεταβληθεί κατάλληλα η αναλογία ανάμεσα στις παρατηρήσεις που ανήκουν στην «ακριβή» κλάση και στις παρατηρήσεις που ανήκουν στη «φθηνή» κλάση. Ο Elkan (2001) ασχολείται με την περίπτωση της δυαδικής κλάσης, και αναφέρεται σε τρόπους αλλαγής της αναλογίας θετικών και αρνητικών παρατηρήσεων, έτσι ώστε να επιτευχθεί εκπαίδευση με μείωση του κόστους σε μοντέλα που εκπαιδεύονται με μεθόδους, οι οποίες δεν είναι ευαίσθητες ως προς το κόστος. Σύμφωνα με τη δεύτερη προσέγγιση, η πληροφορία για το κόστος ενσωματώνεται και αξιοποιείται στον αλγόριθμο εκπαίδευσης. Παράδειγμα τέτοιας τεχνικής υπάρχει στην εργασία των Chen et al. (2011). Στην εργασία αυτή οι ερευνητές συνδυάζουν τους Γενετικούς Αλγορίθμους με τη μέθοδο κατηγοριοποίησης Learning Vector Quantization, και ενσωματώνουν το κόστος στη συνάρτηση καταλληλότητας (fitness function).

## 10.7 Επιδόσεις ανά κλάση

Σε περιπτώσεις όπου οι κλάσεις δεν είναι ισομερώς κατανεμημένες ή όπου οι εσφαλμένες κατηγοριοποιήσεις διαφορετικών κλάσεων έχουν διαφορετικό κόστος, είναι σημαντική η εκτίμηση της ικανότητας πρόβλεψης του κατηγοριοποιητή για την κάθε κλάση.

Για να εκτιμήσουμε τις ανά κλάση επιδόσεις ενός κατηγοριοποιητή, εισάγουμε την αναγκαία ορολογία. Για την περίπτωση δυαδικής κλάσης ισχύουν οι ακόλουθοι όροι:

**Θετικές παρατηρήσεις** (positive) ονομάζονται οι παρατηρήσεις, οι οποίες ανήκουν σε μια τιμή της κλάσης (πχ χρεοκοπία)

**Αρνητικές παρατηρήσεις** (negative) ονομάζονται οι παρατηρήσεις, οι οποίες ανήκουν στην άλλη τιμή της κλάσης (πχ μη χρεοκοπία)

**Αληθινές Θετικές Προβλέψεις** (true positive – tp) είναι το πλήθος των επιτυχών προβλέψεων για θετικές παρατηρήσεις (πχ η επιχείρηση είναι χρεοκοπημένη και ο κατηγοριοποιητής προβλέπει σωστά την κλάση).

**Αληθινές Αρνητικές Προβλέψεις** (true negative – tn) είναι το πλήθος των επιτυχημένων προβλέψεων για αρνητικές παρατηρήσεις (πχ η επιχείρηση δεν είναι χρεοκοπημένη και ο κατηγοριοποιητής προβλέπει σωστά την κλάση).

**Ψευδείς Θετικές Προβλέψεις** (false positive – fp) είναι το πλήθος των αποτυχημένων προβλέψεων για αρνητικές παρατηρήσεις (η επιχείρηση δεν είναι χρεοκοπημένη, ο κατηγοριοποιητής όμως την προβλέπει ως χρεοκοπημένη).

**Ψευδείς Αρνητικές Προβλέψεις** (false negative – fn) είναι το πλήθος των αποτυχημένων προβλέψεων για θετικές παρατηρήσεις (η επιχείρηση είναι χρεοκοπημένη, ο κατηγοριοποιητής όμως την προβλέπει ως μη χρεοκοπημένη).

Ένας τρόπος παρουσίασης των επιδόσεων ανά κλάση ενός κατηγοριοποιητή είναι με τη χρήση του **πίνακα σύγχυσης** (confusion matrix). Ο Πίνακας Σύγχυσης είναι ένας δισδιάστατος πίνακας, όπου οι στήλες αντιστοιχούν στις προβλέψεις και οι γραμμές στις πραγματικές τιμές κλάσης. Στα κελιά του πίνακα αναγράφονται οι αληθινές θετικές, οι αληθινές αρνητικές, οι ψευδείς θετικές και οι ψευδείς αρνητικές προβλέψεις. Στο Σχήμα 10.7 απεικονίζεται ένας Πίνακας Σύγχυσης.

| | Πρόβλεψη Αρνητικής Κλάσης | Πρόβλεψη Θετικής Κλάσης |
|---|---|---|
| Πραγματική Αρνητική Κλάση | tn | fp |
| Πραγματική Θετική Κλάση | fn | tp |

tn = true negative
tp = true positive
fn = false negative
fp = false positive

**Σχήμα 10.7** *Πίνακας Σύγχυσης*

Ορισμένα πρόσθετα μέτρα για τις επιδόσεις ενός κατηγοριοποιητή είναι τα ακόλουθα:

$$sensitivity = \frac{tp}{pos} \tag{10.30}$$

$$specificity = \frac{tn}{negat} \tag{10.31}$$

$$precision = \frac{tp}{tp + fp} \tag{10.32}$$

$$accuracy = sensitivity * \frac{pos}{pos + negat} + specificity * \frac{neg}{pos + negat} = \frac{tp + tn}{pos + negat} \tag{10.33}$$

όπου pos είναι το πλήθος των θετικών παρατηρήσεων και negat είναι το πλήθος των αρνητικών παρατηρήσεων. Σύμφωνα με τα παραπάνω, η ακρίβεια (accuracy) ορίζεται ως το ποσοστό των ορθών θετικών προβλέψεων επί το ποσοστό των θετικών παρατηρήσεων συν το ποσοστό των ορθών αρνητικών προβλέψεων επί το ποσοστό των αρνητικών παρατηρήσεων ή ισοδύναμα ως το πλήθος των ορθών προβλέψεων προς το πλήθος των παρατηρήσεων.

## 10.8 Καμπύλες ROC

Ένα ισχυρό μέτρο για την εκτίμηση της ανά κλάση ακρίβειας του κατηγοριοποιητή είναι οι λεγόμενες καμπύλες ROC (Receiver Operating Characteristics). Οι καμπύλες ROC σχεδιάζονται σε έναν δυσδιάστατο επίπεδο χώρο. Ο οριζόντιος άξονας εκφράζει το μέγεθος 1-specificity, το οποίο ονομάζεται και False Positive Rate.

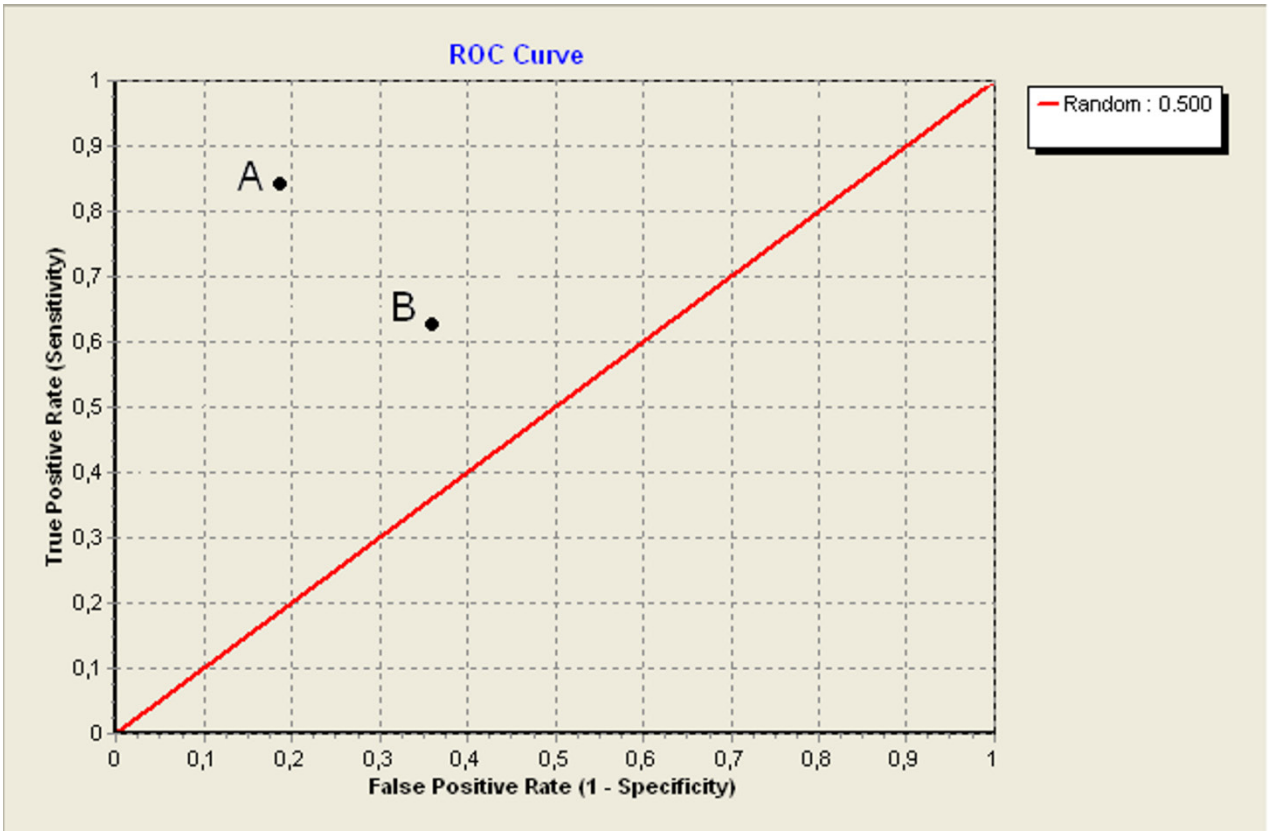
$$False\_Positive\_Rate = 1 - Specificity = \frac{fp}{negat} \tag{10.34}$$

Ο κατακόρυφος άξονας εκφράζει το μέγεθος sensitivity, το οποίο ονομάζεται και True Positive Rate

$$True\_Positive\_Rate = Sensitivity = \frac{tp}{pos}$$

(10.35)

Ουσιαστικά, ο οριζόντιος άξονας εκφράζει το ποσοστό των αρνητικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν λάθος, και ο κατακόρυφος άξονας εκφράζει το ποσοστό των θετικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν σωστά. Το Σχήμα 10.8 απεικονίζει τον δυσδιάστατο χώρο καμπύλων ROC. Κάθε σημείο του χώρου αυτού εκφράζει ένα ισοζύγιο ανάμεσα στο ποσοστό ορθών θετικών προβλέψεων και εσφαλμένων θετικών προβλέψεων. Το σημείο 0,0 είναι ένας κατηγοριοποιητής, που δεν προβλέπει ποτέ θετική παρατήρηση. Το σημείο 1,1 είναι ένας κατηγοριοποιητής, που προβλέπει πάντα θετική παρατήρηση. Η διαγώνια γραμμή, από το σημείο 0,0 στο σημείο 1,1 είναι ένας κατηγοριοποιητής που προβλέπει τυχαία την κλάση. Οι κατηγοριοποιητές που βρίσκονται κάτω από τη διαγώνια γραμμή είναι χειρότεροι από την τυχαία πρόβλεψη. Οι κατηγοριοποιητές που βρίσκονται πάνω από τη διαγώνια γραμμή είναι καλύτεροι από την τυχαία πρόβλεψη. Το σημείο 0,1 είναι ο άριστος κατηγοριοποιητής, οι οποίος προβλέπει σωστά όλες τις θετικές και αρνητικές παρατηρήσεις. Γενικώς, όσο πιο μετατοπισμένο είναι προς τα επάνω και προς τα αριστερά ένα σημείο, τόσο καλύτερη θεωρείται η επίδοση. Στο Σχήμα 10.8 το σημείο A είναι καλύτερο από το σημείο B.

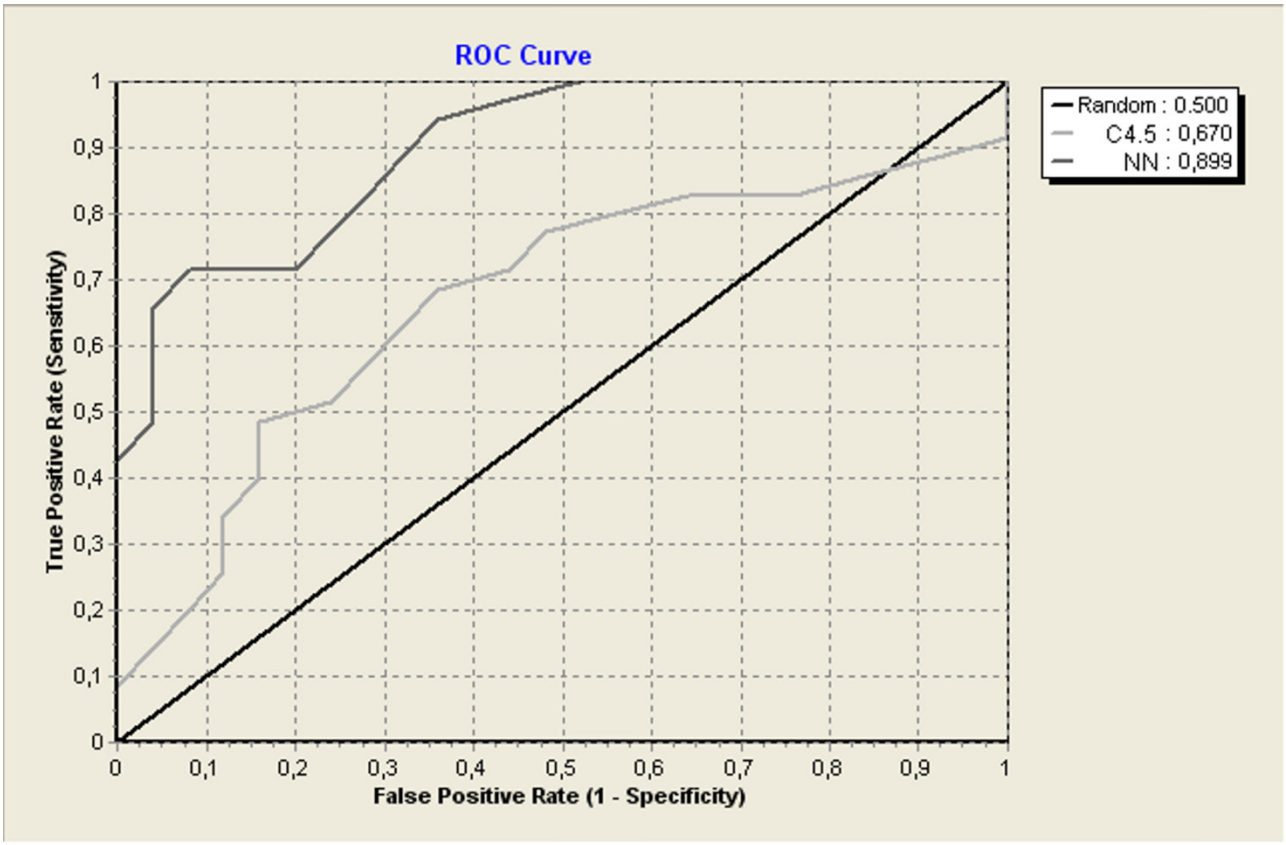



**Σχήμα 10.8** *Σημεία στον χώρο καμπύλων ROC*

Η επίδοση των κατηγοριοποιητών στον χώρο ROC συμβολίζεται με μία καμπύλη. Για να συγκρίνουμε κατηγοριοποιητές χρειαζόμαστε ένα μέτρο σύγκρισης. Τέτοιο μέτρο σύγκρισης είναι η Περιοχή Κάτω από την Καμπύλη ROC (Area Under ROC Curve (AUC)). Η AUC εκφράζει το ποσοστό του χώρου που βρίσκεται κάτω από την καμπύλη, και παίρνει τιμές από 0 έως 1. Η διαγώνια γραμμή τυχαίας πρόβλεψης έχει *AUC = 0,5*. Συνεπώς, κάθε κατηγοριοποιητής καλύτερος της τυχαίας πρόβλεψης έχει *AUC>0,5*. Όσο μεγαλύτερη περιοχή AUC έχει ένας κατηγοριοποιητής τόσο καλύτερος είναι. Στο Σχήμα 10.9, παρουσιάζονται οι καμπύλες ROC ενός Δένδρου Αποφάσεων και ενός Νευρωνικού Δικτύου, τα οποία προβλέπουν περιπτώσεις, όπου οι εξωτερικοί ελεγκτές εκδίδουν δυσμενή σχόλια. Όπως φαίνεται στο σχήμα, η τιμή AUC του Νευρωνικού Δικτύου είναι μεγαλύτερη από την αντίστοιχη του Δένδρου Αποφάσεων, γεγονός που σημαίνει ότι το μοντέλο του Νευρωνικού Δικτύου προβλέπει πιο αποτελεσματικά τις περιπτώσεις έκδοσης δυσμενών σχολίων.

## 10.9 Μελέτη Περίπτωσης – Πρόβλεψη τύπου εξωτερικού ελεγκτή με χρήση μεθόδων κατηγοριοποίησης

Αναφερθήκαμε και στο προηγούμενο κεφάλαιο στη σημασία του εξωτερικού ελέγχου, ιδιαίτερα στη σημερινή εποχή. Ο πρωτεύων σκοπός του εξωτερικού ελέγχου είναι να διασφαλίσει την αντικειμενική παρουσίαση της οικονομικής κατάστασης της επιχείρησης, και έτσι να μειώσει την ασυμμετρία στη ροή πληροφορίας ανάμεσα στα διοικητικά στελέχη, τους μετόχους και τους πιστωτές. Ο έλεγχος μπορεί να έχει μια σειρά από ευεργετικά αποτελέσματα. Ο διαχωρισμός της ιδιοκτησίας από τη διοίκηση στις μοντέρνες επιχειρήσεις δημιουργεί κίνητρα στους μάνατζερς να λειτουργήσουν προς ίδιον όφελος και σε βάρος των μετόχων και πιστωτών (Kane & Velury, 2004). Ο έλεγχος αυξάνει την αξιοπιστία των χρηματοοικονομικών καταστάσεων και μειώνει το ρίσκο σφαλερής πληροφόρησης. Συνακόλουθα μειώνει τον επενδυτικό κίνδυνο. Με αυτόν τον τρόπο ο αξιόπιστος έλεγχος μπορεί επίσης να αυξήσει τις τιμές των μετοχών και να μειώσει το κόστος του χρήματος. Ο έλεγχος μπορεί να βελτιώσει την αποδοτικότητα των επιχειρηματικών διαδικασιών και να βοηθήσει στη συμμόρφωση με τις κανονιστικές διατάξεις (Knechel, Niemi & Sundgren, 2008). Αυτά τα ευεργετικά αποτελέσματα αυξάνονται με τη διεξαγωγή ελέγχου υψηλής ποιότητας (Broye & Weill, 2008).




**Σχήμα 10.9** *Σύγκριση κατηγοριοποιητών με καμπύλες ROC*

Παρά τα σημαντικά του αποτελέσματα, ο έλεγχος πάσχει από μια εσωτερική αντίθεση. Η αντίθεση προέρχεται από το γεγονός ότι ο ελεγκτής πρέπει να παραμείνει ανεξάρτητος και να προστατεύσει τα συμφέροντα των επενδυτών και των πιστωτών, όμως η πρόσληψη του και η αμοιβή του αποφασίζεται από τη διοίκηση της ελεγχόμενης επιχείρησης. Αυτή η αντίφαση μπορεί να θέσει σε κίνδυνο την αντικειμενικότητα του ελεγκτή. Το ερώτημα της ποιότητας του ελέγχου παραμένει ανοικτό. Ωστόσο, είναι γεγονός ότι όταν επιλέγεται έλεγχος υψηλής ποιότητας το φαινόμενο της ασυμμετρίας μειώνεται (Broye & Weill, 2008).

Είναι γενικά αποδεκτό ότι οι ελεγκτικές εταιρείες χωρίζονται σε δύο κατηγορίες Τη μια κατηγορία απαρτίζουν οι τέσσερεις μεγάλες ελεγκτικές εταιρείες (4 Μεγάλοι Ελεγκτές - 4ΜΕ). Αυτές οι εταιρείες είναι η KPMG, η PriceWaterhouseCoopers, η Ernst & Young και η Deloitte & Touche. Όλοι οι υπόλοιποι ελεγκτές εντάσσονται στη δεύτερη κατηγορία (Όχι 4 Μεγάλοι Ελεγκτές – Ο4ΜΕ). Οι 4ΜΕ θεωρούνται ότι διεξάγουν πιο ποιοτικό έλεγχο (DeAngelo, 1981; Palmrose, 1988). Χάρη στο μέγεθος τους, οι 4ΜΕ είναι σε θέση να αντιστέκονται περισσότερο στις πιέσεις των πελατών τους. Επίσης, επενδύουν περισσότερο σε τεχνολογία, εκπαίδευση και υποδομές, και έχουν περισσότερα κίνητρα να διατηρήσουν την επαγγελματική φήμη τους.

Μελέτες έχουν δείξει ότι ψευδής δήλωση αυξημένων εσόδων είναι σπανιότερη σε εταιρείες που ελέγχονται από τους 4ΜΕ (Francis, Maydew & Sparks, 1999). Σε προηγούμενες μελέτες το μέγεθος του ελεγκτή έχει χρησιμοποιηθεί ως μέτρο της ποιότητας του ελέγχου. (Teoh & Wong, 1993).

Η πρόσληψη του εξωτερικού ελεγκτή είναι μια σύνθετη διαδικασία. Οι μέτοχοι επιδιώκουν να προσλάβουν έναν ελεγκτή υψηλής ποιότητας για να περιορίσουν ενδεχόμενο κίνδυνο χειραγώγησης των οικονομικών στοιχείων, και να επιβεβαιώσουν την αξιοπιστία των οικονομικών καταστάσεων. Επίσης, τα διοικητικά στελέχη, τα οποία θέλουν να σηματοδοτήσουν την αξιοπιστία τους και την ευθυγράμμιση τους με τα συμφέροντα των μετόχων, επιθυμούν επίσης να προσλάβουν ελεγκτές υψηλής ποιότητας. Ωστόσο, οι ελεγκτές υψηλής ποιότητας επενδύουν περισσότερο σε τεχνολογία και εκπαίδευση και επομένως έχουν υψηλότερη αμοιβή. Ένα άλλο ζήτημα είναι ότι, σε περίπτωση αποτυχίας της επιχείρησης και συνακόλουθα αποτυχίας του ελέγχου, η πρόσληψη του ελεγκτή πιθανόν να πρέπει να αιτιολογηθεί. Η υιοθέτηση μιας αποτελεσματικής διαδικασίας πρόσληψης του εξωτερικού ελεγκτή αυξάνει την πιθανότητα να προσλάβει η επιχείρηση τον κατάλληλο ελεγκτή στην κατάλληλη τιμή.

Η ερευνητική βιβλιογραφία αποκαλύπτει ότι οι ερευνητές απορρίπτουν τη μηδενική υπόθεση, ότι οι επιχειρήσεις είναι κατανεμημένες τυχαία μεταξύ των 4ΜΕ και των Ο4ΜΕ. Σε ερευνητικές εργασίες έχει μελετηθεί το θέμα της πρόσληψης εξωτερικού ελεγκτή. Ωστόσο, οι μέθοδοι που χρησιμοποιήθηκαν ήταν κλασσικές στατιστικές, όπως η Λογιστική Παλινδρόμηση. Οι Kirkos, Spathis and Manolopoulos (2010) εφαρμόζουν μεθόδους Εξόρυξης Δεδομένων για την ανάπτυξη μοντέλων, τα οποία προβλέπουν την κατηγορία του εξωτερικού ελεγκτή. Η μελέτη αυτή αυξάνει την κατανόηση σχετικά με την επιλογή κατηγορίας εξωτερικών ελεγκτών. Οι ελεγκτικές εταιρείες μπορούν να χρησιμοποιήσουν τα αποτελέσματα και να ανακαλύπτουν τα χαρακτηριστικά των εταιρειών στις οποίες μπορούν να στοχεύσουν.

Τα δεδομένα της έρευνας προέρχονται από τη βάση οικονομικών δεδομένων FAME (Financial Analysis Made Easy), η οποία περιλαμβάνει στοιχεία για Βρετανικές και Ιρλανδικές επιχειρήσεις. Επιλέχθηκαν οι επιχειρήσεις, οι οποίες ήταν εισηγμένες στο χρηματιστήριο και δραστηριοποιούνταν στους τομείς της βιομηχανίας, των κατασκευών, της πληροφορικής και της εξόρυξης μεταλλευμάτων, και οι οποίες άλλαξαν εξωτερικό ελεγκτή τα έτη 2003-2005. Οι επιλεγμένες επιχειρήσεις συνταιριάστηκαν με ίσο αριθμό επιχειρήσεων, οι οποίες δεν άλλαξαν τον εξωτερικό τους ελεγκτή.

Η αρχική επιλογή ανεξάρτητων μεταβλητών στηρίχθηκε στην προηγούμενη έρευνα. Περιλήφθηκαν μεταβλητές που αφορούσαν το μέγεθος της ελεγχόμενης εταιρείας (Krishnan, Krishnan & Stephens, 1996), το πλήθος των θυγατρικών εταιρειών, το ύψος του δανεισμού ( Knechel et al., 2008; Broye & Weil, 2008), την αποθήκη και τους εισπρακτέους λογαριασμούς (Icerman & Hillison, 1991), την κερδοφορία (Citron & Manalis, 2001), την έκδοση δυσμενών σχολίων (Chow & Rice, 1982; Citron & Taffler, 1992; Krishnan et al., 1996), τις τάσεις αύξησης του μεγέθους της εταιρείας (Velury, Reish & O'Reilly, 2003), τις αμοιβές των εξωτερικών ελεγκτών, τη χρηματιστηριακή αξία της επιχείρησης (Kane & Velury, 2004), καθώς και μερικοί ακόμα γνωστοί αριθμοδείκτες, όπως το Quick Ratio και το Z-Score του Altman.

Συνολικά επιλέχθηκαν τριάντα πέντε μεταβλητές. Οι μεταβλητές αυτές υποβλήθηκαν σε έλεγχο σημαντικότητας με εφαρμογή της μεθόδου ANOVA. Δεκαοκτώ μεταβλητές παρουσίαζαν μικρή τιμή p και επιλέχθηκαν να συμμετέχουν στο τελικό άνυσμα εισόδου. Η στατιστική ανάλυση των μεταβλητών αποκάλυψε και ορισμένες πρώτες ενδείξεις συσχέτισης τιμών των μεταβλητών με τον τύπο του εξωτερικού ελεγκτή. Το μέγεθος της ελεγχόμενης επιχείρησης είναι σημαντικό, και οι μεγάλες επιχειρήσεις τείνουν τα προσλάβουν μεγάλους ελεγκτές. Σημαντικό είναι επίσης το ύψος του χρέους. Επιχειρήσεις με μεγαλύτερο ποσοστό χρέους τείνουν να προσλάβουν μεγάλους ελεγκτές. Αξιόλογη διαφοροποίηση παρουσιάζεται και σε μεταβλητές που αναφέρονται στη ρευστότητα. Είναι αξιοσημείωτο ότι ο αριθμοδείκτης Z-Score παρουσίασε υψηλή τιμή *p*, παρέχοντας ισχυρές ενδείξεις ότι η οικονομική ευρωστία ή δυσπραγία δεν επηρεάζει την επιλογή τύπου εξωτερικού ελεγκτή. Επίσης, όλες οι μεταβλητές, οι οποίες αναφέρονταν σε τάσεις οικονομικών μεγεθών, απορρίφθηκαν ως μη σημαντικές.

Τρεις μέθοδοι εξόρυξης δεδομένων, τα Δένδρα Αποφάσεων τύπου C4.5, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron και οι k-Πλησιέστεροι Γείτονες εφαρμόστηκαν για την πρόβλεψη του τύπου του εξωτερικού ελεγκτή. Οι τρεις αυτές μέθοδοι συγκρίθηκαν με τη Λογιστική Παλινδρόμηση, η οποία ήταν και η μοναδική μέθοδος που είχε εφαρμοστεί σε προηγούμενες εργασίες. Το Δένδρο Απόφασης εκπαιδεύτηκε με επίπεδο εμπιστοσύνης 0,25 και περιείχε εικοσιπέντε κόμβους και δεκατρία φύλλα. Ως μεταβλητή διαχωρισμού πρώτου επιπέδου επιλέχθηκε το Σύνολο Χρέους (Total Debt). Σύμφωνα με το κριτήριο του Λόγου Κέρδους (Gain Ratio), η μεταβλητή αυτή διαχωρίζει βέλτιστα τις δυο κατηγορίες. Η μεγάλη πλειοψηφία των επιχειρήσεων με υψηλό επίπεδο χρέους (95 από 98 παρατηρήσεις) επιλέγει ελεγκτή 4ΜΕ. Συμπεραίνουμε ότι οι επιχειρήσεις με υψηλό επίπεδο χρέους επιδιώκουν ποιοτικότερο έλεγχο. Οι αμοιβή των ελεγκτών και οι εισπρακτέοι λογα-

ριασμοί επιλέχθηκαν επίσης ως μεταβλητές διαχωρισμού υψηλού επιπέδου.

Διάφορες εναλλακτικές αρχιτεκτονικές δοκιμάστηκαν για το Νευρωνικό Δίκτυο, και τελικά επιλέχθηκε μια αρχιτεκτονική με ένα κρυφό στρώμα, το οποίο περιείχε ένδεκα κρυφούς νευρώνες. Επειδή το Νευρωνικό Δίκτυο δεν παρέχει κάποια κατανοητή ερμηνεία σχετικά με τη σημαντικότητα των μεταβλητών εισόδου, εφαρμόστηκε ένας έμμεσος έλεγχος, ο οποίος αποτελούνταν από μια επαναληπτική διαδικασία. Σε κάθε στάδιο της επανάληψης αφαιρούνταν μια από τις μεταβλητές εισόδου, και ελέγχονταν η ακρίβεια του μοντέλου. Αν η αφαίρεση μιας μεταβλητής προκαλούσε σημαντική πτώση της ακρίβειας, τότε η μεταβλητή θεωρούνταν σημαντική. Σύμφωνα με τα αποτελέσματα, η σημαντικότερη μεταβλητή ήταν το Σύνολο Χρέους. Επισημαίνεται ότι η ίδια μεταβλητή είχε επιλεγεί ως διαχωριστής πρώτου επιπέδου από το Δένδρο Απόφασης.

Για τη μέθοδο των k-Πλησιέστερων Γειτόνων, το πλήθος των γειτονικών σημείων k ορίστηκε να είναι ίσο με πέντε. Δυστυχώς, για τη συγκεκριμένη μέθοδο δεν υπήρχε η δυνατότητα ελέγχου της σημαντικότητας των μεταβλητών εισόδου. Η τελευταία μέθοδος που εφαρμόστηκε ήταν η Λογιστική Παλινδρόμηση. Σύμφωνα με το κριτήριο Wald, η πιο σημαντική μεταβλητή εισόδου ήταν το Σύνολο Χρέους. Είναι αξιοσημείωτο ότι και οι τρεις μέθοδοι, οι οποίες παρείχαν κριτήρια εκτίμησης της σημαντικότητας των ανεξάρτητων μεταβλητών, συμφωνούν ότι η μεταβλητή, η οποία επηρεάζει σε μεγαλύτερο βαθμό το αποτέλεσμα της κατηγοριοποίησης, είναι το σύνολο χρέους.

Και τα τέσσερα μοντέλα επέτυχαν υψηλούς ρυθμούς ακρίβειας έναντι του συνόλου εκπαίδευσης, οι οποίοι κυμαίνονταν από 79% έως 93%. Ωστόσο, αυτές οι επιδόσεις δεν μπορούν να θεωρηθούν ενδεικτικές των πραγματικών δυνατοτήτων των μοντέλων. Ένας υπαρκτός κίνδυνος είναι η υπερπροσαρμογή των μοντέλων, η απομνημόνευση δηλαδή των παρατηρήσεων του συνόλου εκπαίδευσης. Η πραγματική αξία όμως των μοντέλων βρίσκεται στην εφαρμογή τους στην καθημερινή πράξη, όπου και θα συναντήσουν άγνωστες παρατηρήσεις, διαφορετικές από αυτές του συνόλου εκπαίδευσης. Για τον λόγο αυτό, τα μοντέλα υποβλήθηκαν σε διαδικασία επικύρωσης, ώστε να εκτιμηθεί η ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Εφαρμόστηκαν δύο διαφορετικές τεχνικές επικύρωσης. Η πρώτη τεχνική ήταν η διασταυρούμενη επικύρωση 10 τμημάτων (10 fold cross validation). Η δεύτερη τεχνική ήταν η διάσπαση του συνόλου δεδομένων σε δύο υποσύνολα, όπου το ένα χρησιμοποιήθηκε για εκπαίδευση και το δεύτερο για επικύρωση. Ειδικότερα, επιλέχθηκαν οι επιχειρήσεις οι οποίες άλλαξαν ελεγκτή τα έτη 2003 και 2004 για την εκπαίδευση των μοντέλων (220 παρατηρήσεις), ενώ οι υπόλοιπες επιχειρήσεις (118 παρατηρήσεις) χρησιμοποιήθηκαν για επικύρωση. Σύμφωνα με τα αποτελέσματα των δύο ελέγχων, και τα τέσσερα μοντέλα αποδείχθηκαν ικανά να κατηγοριοποιούν άγνωστες παρατηρήσεις και επέτυχαν ικανοποιητικούς ρυθμούς ακρίβειας. Και στις δύο περιπτώσεις, το Δένδρο Αποφάσεων επέτυχε τις καλύτερες επιδόσεις, ακολουθούμενο από το Νευρωνικό Δίκτυο. Οι k-Πλησιέστεροι Γείτονες και η Λογιστική Παλινδρόμηση εναλλάσσονται στην τρίτη και τέταρτη θέση ανάλογα με την τεχνική επικύρωσης. Τα αποτελέσματα παρέχουν αποδείξεις ότι οι νέες τεχνικές, οι οποίες προέρχονται από τον χώρο της Μηχανικής Μάθησης, υπερβαίνουν σε επιδόσεις την ευρέως χρησιμοποιούμενη Λογιστική Παλινδρόμηση.

Και οι τέσσερεις τεχνικές επέτυχαν αρκετά υψηλά ποσοστά ακρίβειας. Ωστόσο, η περαιτέρω βελτίωση των επιδόσεων αποτελεί μόνιμη επιδίωξη. Όπως αναφέρθηκε και προηγουμένως, οι σύνθετοι κατηγοριοποιητές μπορούν να υπερβούν τις επιδόσεις των ατομικών τεχνικών. Σε μια απόπειρα αύξησης του ρυθμού ακρίβειας εφαρμόστηκε η τεχνική Bagging, και τα νέα μοντέλα δοκιμάστηκαν με τη μέθοδο της διασταυρούμενης επικύρωσης 10 τμημάτων. Σύμφωνα με τα αποτελέσματα, το Δένδρο Αποφάσεων βελτίωσε τις επιδόσεις του κατά 3,5% περίπου, ενώ μικρότερη βελτίωση υπήρχε στο Νευρωνικό Δίκτυο και τη Λογιστική Παλινδρόμηση. Οι ρυθμοί ακρίβειας των κατηγοριοποιητών παρουσιάζονται αναλυτικά στον Πίνακα 10.1

| | **C4.5** | **Νευρωνικό Δίκτυο** | **k-NN** | **Λογ. Παλινδρόμηση** |
|---|---|---|---|---|
| 10 fold cross validation | 82,12 | 77,27 | 69,09 | 76,66 |
| Validation set | 82,09 | 72,88 | 71,19 | 63,56 |
| Bagging + 10 fold cross validation | 85,45 | 79.09 | 69,09 | 77,88 |

**Πίνακας 10.1** *Ρυθμοί ακρίβειας μοντέλων*

# Βιβλιογραφία/Αναφορές


Bergadano, F., Matwin, S., Michalski, R. S., & Zhang, J. (1988). Measuring Quality of Concept Descriptions. In *3rd European Working Session on Learning* (pp. 1-14). Glasgow, SCT: Pittman.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123–140. doi: 10.1007/bf00058655

Broye, G., & Weill, L. (2008). Does Leverage Influences Auditor Choice? A Cross-country Analysis. A*pplied Financial Economics, 18*(9), 715-731. doi: 10.1080/09603100701222325

Chawla, N. V., Boywer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research, 16*(1), 321-357.

Chen, H. J., Huang, S. Y., & Lin, C. S. (2009). Alternative Diagnosis of Corporate Bankruptcy: A Neuro- Fuzzy Approach. *Expert Systems with Applications, 36*(4), 7710-7720. doi: 10.1016/j.eswa.2008.09.023

Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J., & Neves, C. J. (2011). A Genetic Algorithm-Based Approach to Cost-Sensitive Bankruptcy Prediction. *Expert Systems with Applications*, *38*(10), 12939–12945. doi: 10.1016/j.eswa.2011.04.090

Chow, C., & Rice, S. (1982). Qualified Audit Opinions and Auditor Switching. *The Accounting Review, 57*(2), 326-335.

Citron, D., & Taffler, R. (1992). The Audit Report under Going Concern Uncertainties: An Empirical Analysis. *Accounting and Business Research, 22*(88), 337-345. doi: 10.1080/00014788.1992.9729449

Citron, D., & Manalis, G. (2001). The International Firms as new Entrants to the Statutory Audit Market: An Empirical analysis of auditor selection in Greece, 1993 to 1997. *The European Accounting Review, 10*(3), 439–459. doi: 10.2139/ssrn.233635

DeAngelo, L. (1981). Auditor Size and Auditor Quality. *Journal of Accounting and Economics, 3*(3), 183–199. doi: 10.1016/0165-4101(81)90002-1

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2,* 973-978. San Francisco, CA: Morgan Kaufman.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Francis, J., Maydew, E., & Sparks, H. C. (1999). The Role of Big 6 Auditors in the Credible Reporting of Accruals. *Auditing: A Journal of Practice and Theory, 18*(2), 17-34. doi: 10.2308/aud.1999.18.2.17

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the 13th International Conference on Machine Learning,* 148–156. San Francisco, CA: Morgan Kaufmann.

Icerman, R., & Hillison, W. (1991). Disposition of Auditor-Detected Errors: Some Evidence on Evaluative Materiality. *Auditing: A Journal of Practice and Theory, 10*, 22-34.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Truck on Inductive Learning,* 111-117. Las Vegas, NV.

Kane, G., & Velury, U. (2004). The Role of Institutional Ownership in the Market for Auditing Services: An empirical Investigation. *Journal of Business Research, 57*(9), 976-983. doi: 10.1016/s0148-2963(02)00499-x

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2010), Audit-Firm Group Appointment: An Artificial Intelligence Approach. *Intelligent Systems in Accounting, Finance and Management, 17*(1), 1-17. doi: 10.1002/isaf.310

Kirkos, E. (2015). Assessing Methodologies for Intelligent Bankruptcy Prediction. *Artificial Intelligence Review, 43*(1), 83-123. doi: 10.1007/s10462-012-9367-6

Knechel, R., Niemi, L., & Sundgren, S. (2008). Determinants of Auditor Choice: Evidence from a Small Client Market. *International Journal of Auditing, 12*(1), 65-88. doi: 1099-1123.2008.00370.x

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2,* 1137-1143. San Francisco, CA: Morgan Kaufmann.

Kreßel, U. (1999). Pairwise Classification and Support Vector Machines. In B. Schoelkopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods – Support Vector Learning* (pp. 255-268). Cambridge, MA: MIT Press.

Krishnan, J., Krishnan, J., & Stephens, R. (1996). The Simultaneous Relation between Auditor Switching

and Audit Opinion: An empirical analysis. *Accounting and Business Research, 26*(3), 224–236. doi: 10.1080/00014788.1996.9729513

Kubat, M., & Martin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. *Proceedings of the 14th International Conference on Machine Learning,* 179-186. Nashville, TN: Morgan Kaufmann.

Lin, W. Y., Hu, Y. H., & Tsai, C. F. (2012). Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man and Cybernetics, 42*(4), 421-436. doi: 10.1109/tsmcc.2011.2170420

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer + Business Media.

Palmrose, Z. (1988). An Analysis of Auditor Litigation and Audit Service Quality. *The Accounting Review, 63*(1), 55-73.

Simonoff, J. (2003). *Analyzing Categorical Data.* New York, NY: Springer-Verlag.

Smola, A. J., & Schoelkopf, B. (2004). A Tutorial on Support Vector Regression. *Statistics and Computing, 14*(3), 199-222. doi: 10.1023/B:STCO.0000035301.49549.88

Steinwart, I. (2003). On the Optimal Parameter Choice for NU-Support Vector Machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(10), 1274-1284. doi: 10.1109/tpami.2003.1233901

Teoh, S. H., & Wong, T. J. (1993). Perceived Auditor Quality and the Earnings Response Coefficient. *The Accounting Review, 68*(2), 346-366.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer Verlag.

Velury, U., Reish, J., & O'Reilly, D. (2003). Institutional Ownership and the Selection of Industry Specialist Auditors. *Review of Quantitative Finance and Accounting, 21*(1), 35–48. doi: 10.1023/A:1024855605207

Weiss, G. M., & Provost, F. (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research, 19*(1), 315-354.

Wilson, R. D., & Martinez, T. R. (1997). Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research, 6*(1), 1-34.

Witten, I. H., & Frank, E. (2000). *Data Mining Practical Machine Learning Tools and Techniques with JAVA Implementations*. San Francisco, CA: Morgan Kaufman.

Wolpert, D. (1992). Stacked Generalization. *Neural Networks*, 5(2), 241-259. doi: 10.1016/S0893-6080(05)80023-1

# Κριτήρια Αξιολόγησης

## Άσκηση Υπολογισμών 10.1

**Στον Πίνακα 10.2 δίνονται οι τιμές των μεταβλητών X και Y. Υπολογίστε τις τιμές των a και b, ώστε να εκφραστεί η μεταβλητή Y ως γραμμική παλινδρόμηση του X (Y=a+b*X). Υπολογίστε την τιμή του Y, εάν το X πάρει την τιμή 10.**

| X | Y |
|---|---|
| 5 | 11 |
| 4 | 6 |
| 7 | 16 |
| 12 | 20 |
| 3 | 6 |
| 6 | 11 |
| 8 | 18 |
| 8 | 15 |
| 11 | 21 |
| 9 | 20 |

**Πίνακας 10.2** *Δεδομένα Άσκησης 1*

### Λύση

Τα *b* και *a* θα υπολογιστούν σύμφωνα με τις Εξισώσεις 10.18 και 10.19 αντίστοιχα. Αρχικά υπολογίζονται οι μέσες τιμές των *X* και *Y* ($X_m$=*7,3* και $Y_m$=*14,4*). Ακολούθως, υπολογίζονται οι τιμές των *a* και *b* (*a=1,085* και *b=1,824*). Το *Y* εκφράζεται ως γραμμική συνάρτηση του *X* σύμφωνα με τη σχέση *Y=1,085+1,824X*.
Εάν το *X* πάρει την τιμή *0*, το *Y* υπολογίζεται ως *Y=1,085+1,824*10=19,325*.

## Άσκηση Υπολογισμών 10.2

**Ένα μοντέλο κατηγοριοποιητή προβλέπει την πιστοληπτική ικανότητα. Υπάρχουν τρεις δυνατές τιμές κλάσης, η «Υψηλή», η «Μεσαία» και η «Χαμηλή». Τα αποτελέσματα του μοντέλου παρουσιάζονται στον ακόλουθο πίνακα σύγχυσης.**

| | Υψηλή | Μεσαία | Χαμηλή |
|---|---|---|---|
| **Υψηλή** | **90** | **6** | **4** |
| **Μεσαία** | **7** | **85** | **8** |
| **Χαμηλή** | **3** | **5** | **92** |

**Απαντήστε στις παρακάτω περιπτώσεις:**
**Πόσες είναι οι παρατηρήσεις συνολικά;**
**Πόσες περιπτώσεις μεσαίας πιστοληπτικής ικανότητας κατηγοριοποιήθηκαν σωστά;**
**Ποιο ποσοστό παρατηρήσεων κατηγοριοποιήθηκαν σωστά;**
**Πόσες περιπτώσεις μεσαίας πιστοληπτικής ικανότητας κατηγοριοποιήθηκαν ως «Υψηλή»;**
**Πόσες περιπτώσεις χαμηλής πιστοληπτικής ικανότητας κατηγοριοποιήθηκαν λάθος;**

### Λύση

Οι παρατηρήσεις συνολικά είναι (90+6+4+7+85+8+3+5+92)=300.

Οι περιπτώσεις μεσαίας πιστοληπτικής ικανότητας που κατηγοριοποιήθηκαν σωστά είναι 85.

Το ποσοστό των παρατηρήσεων που κατηγοριοποιήθηκαν σωστά είναι (90+85+92)*100/300 = 89%
Οι περιπτώσεις μεσαίας πιστοληπτικής ικανότητας που κατηγοριοποιήθηκαν ως «Υψηλή» είναι 7.
Οι περιπτώσεις χαμηλής πιστοληπτικής ικανότητας που κατηγοριοποιήθηκαν λάθος είναι (3+5)=8.

## Άσκηση Εφαρμογής 10.3

**Χρησιμοποιήστε το αρχείο «analcatdata_japansolvent.arff» ( θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή StatLib). Το σύνολο δεδομένων προέρχεται από το βιβλίο του Simonoff (2003), και σχετίζεται με την κατηγοριοποίηση ιαπωνικών επιχειρήσεων σε φερέγγυες (solvent) και αφερέγγυες (insolvent). Υπάρχουν 52 γραμμές, κάθε μια από τις οποίες αναφέρεται σε μια επιχείρηση. Οι 25 επιχειρήσεις χαρακτηρίζονται αφερέγγυες και οι 27 φερέγγυες. Στο σύνολο δεδομένων υπάρχουν 10 πεδία (στήλες). Το πρώτο πεδίο περιέχει τα ονόματα των επιχειρήσεων, το δεύτερο είναι το πεδίο κλάσης, και ακολουθούν οκτώ αριθμοδείκτες. Οι κλάσεις κωδικοποιούνται με τις τιμές «0» για τις αφερέγγυες επιχειρήσεις και «1» για τις φερέγγυες.**
**Αναπτύξτε μοντέλα πρόβλεψης φερεγγυότητας επιχειρήσεων με χρήση των μεθόδων α)Λογιστικής Παλινδρόμησης, β)Μηχανών Διανυσμάτων Υποστήριξης, γ)k-Πλησιέστερων Γειτόνων. Πειραματιστείτε με τις τιμές των παραμέτρων για τις Μηχανές Διανυσμάτων Υποστήριξης και για τη μέθοδο των k-Πλησιέστερων Γειτόνων, προσπαθώντας να αυξήσετε τις επιδόσεις.**

### Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «analcatdata_japansolvent.arff» πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» εμφανίζονται διάφορες πληροφορίες για τα δεδομένα. Παρατηρήστε τα πεδία (Attributes). Ως πεδίο κλάσης ορίστε το πεδίο «Solvent». Η κατανομή των παρατηρήσεων χρωματίζεται ανάλογα με την τιμή της κλάσης.

Το πεδίο με τα ονόματα των επιχειρήσεων δεν προσφέρει κάτι χρήσιμο στην ανάλυση μας. Το επιλέγετε και το απομακρύνετε πιέζοντας το κουμπί «Remove».

Μελετήστε την κατανομή τιμών στα διάφορα γνωρίσματα, κάνοντας κλικ στο όνομα του γνωρίσματος. Παρατηρήστε ότι για τους περισσότερους αριθμοδείκτες οι αφερέγγυες επιχειρήσεις (μπλε) τείνουν προς το αριστερό άκρο της κατανομής, ενώ οι φερέγγυες (κόκκινες) τείνουν προς το δεξιό άκρο.

Βήμα 2. Μεταβείτε στο tab «Classify».

Βεβαιωθείτε ότι ως πεδίο κλάσης έχει οριστεί το πεδίο «Solvent» και ότι είναι επιλεγμένη η μέθοδος ελέγχου «Cross-validation».

Επιλέξτε μέθοδο κατηγοριοποίησης πιέζοντας το κουμπί «Choose» στο πεδίο «Classifier». Επιλέξτε πρώτα τη μέθοδο weka/classifiers/functions/SimpleLogistic για τη Λογιστική Παλινδρόμηση και πατήστε το κουμπί «Start». Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Το μοντέλο κατηγοριοποιεί σωστά 78.8462% του συνόλου των παρατηρήσεων, 68% της κλάσης «0» και 88.9% της κλάσης «1».

Βήμα 3. Επιλέξτε τη μέθοδο weka/classifiers/functions/SMO για τις Μηχανές Διανυσμάτων Υποστήριξης και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά 76.9231% των συνολικών περιπτώσεων, 60% της κλάσης «0» και 92.6% της κλάσης «1».

Βήμα 4. Επιλέξτε τη μέθοδο weka/classifiers/lazy/IBk για τη μέθοδο των k-Πλησιέστερων Γειτόνων και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά 82.6923% των συνολικών περιπτώσεων, 76% της κλάσης «0» και 88.9% της κλάσης «1».

Βήμα 5. Επιλέξτε ξανά τη μέθοδο SMO και πειραματιστείτε με τις τιμές της παραμέτρου «C». Επίσης, στο πεδίο «kernel», πειραματιστείτε με τον εκθέτη της συνάρτησης πυρήνα. Επιλέξτε ξανά τη μέθοδο IBk και πειραματιστείτε με το πλήθος των γειτόνων (πεδίο «KNN») και με τον τρόπο υπολογισμού της απόστασης (πεδίο «nearestNeighborSearchAlgorithm»). Προσπαθήστε με τη ρύθμιση των παραμέτρων να αυξήσετε τις επιδόσεις. Για παράδειγμα, στη μέθοδο SMO, θέτοντας την τιμή 1,5 στην παράμετρο «C», και ορίζοντας τιμή εκθέτη ίση με τρία για τη συνάρτηση πυρήνα, η ακρίβεια του μοντέλου Μηχανών Διανυσμάτων Υποστήριξης αυξάνεται στο 80.7692%.

## Άσκηση Εφαρμογής 10.4

**Χρησιμοποιήστε το αρχείο «labor.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://**

**www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή UCI repository). Τα δεδομένα αποτελούν αποτέλεσμα συλλογικών διαπραγματεύσεων για ζητήματα εργασίας στον Καναδά, προσφέρθηκαν από τον καθηγητή Stan Matwin και χρησιμοποιήθηκαν στην εργασία των Bergadano et al. (1988). Περιλαμβάνονται συνολικά 17 γνωρίσματα, τα οποία αναφέρονται σε αυξήσεις μισθού τον πρώτο, δεύτερο και τρίτο χρόνο, στις ώρες εργασίας, σε συνταξιοδοτικά και ασφαλιστικά πλάνα, σε ημέρες άδειας κλπ. Το τελευταίο γνώρισμα είναι το γνώρισμα κλάσης, και οι δυνατές τιμές κλάσεις είναι «good» και «bad». Τα δεδομένα περιέχουν 57 παρατηρήσεις, εκ των οποίων οι 37 ανήκουν στην κλάση «good» και οι 20 στην κλάση «bad».**

**Εκτελέστε επιλογή χαρακτηριστικών εφαρμόζοντας τη μέθοδο CFS Subset Evaluator. Αναπτύξτε μοντέλο ικανό να προβλέπει την κλάση των παρατηρήσεων, εφαρμόζοντας τη μέθοδο Δένδρων Αποφάσεων C4.5. Αυξήστε τις επιδόσεις της μεθόδου, αναπτύσσοντας συνδυασμό κατηγοριοποιητών με χρήση της μεθόδου Bagging.**

### Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «labor.arff» πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» εκτελέστε την επιλογή χαρακτηριστικών. Στο πεδίο "Filter" πιέστε το κουμπί "Choose" και επιλέξτε weka/filters/supervised/attribute/AttributeSelection. Αυτομάτως επιλέγεται η μέθοδος CfsSubsetEval. Κάνετε κλικ στο κουμπί «Apply». Μετά την εκτέλεση του αλγορίθμου θα διαπιστώσετε ότι στο πεδίο «Attributes» μειώθηκε το πλήθος των στηλών. Ειδικότερα, παραμένουν επτά στήλες και επιπλέον η στήλη της κλάσης, ενώ οι υπόλοιπες στήλες απομακρύνονται.

Βήμα 2. Μεταβείτε στο tab «Classify» και επιλέξτε κατηγοριοποιητή, κάνοντας κλικ στο κουμπί «Choose» του πεδίου «Classifier». Από τις διαθέσιμες μεθόδους επιλέξτε τη μέθοδο weka/classifiers/trees/J48. Η μέθοδος αυτή δημιουργεί Δένδρα Αποφάσεων C4.5.

Εκπαιδεύστε το μοντέλο και επικυρώστε το με τη μέθοδο «CrossValidation». Για να εκτελέσετε αυτήν την εργασία, βεβαιωθείτε ότι είναι επιλεγμένη η μέθοδος «Cross-validation» στο πεδίο «Test-options» και στη συνέχεια κάντε κλικ στο κουμπί «Start».

Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Μπορείτε να δείτε το Δένδρο Αποφάσεων. Το μοντέλο κατηγοριοποιεί σωστά 77.193% των συνολικών περιπτώσεων, 65% των παρατηρήσεων με κλάση «bad» και 83.8% των παρατηρήσεων με κλάση «good».

Βήμα 3. Μπορείτε να αυξήσετε τις επιδόσεις εφαρμόζοντας τη μέθοδο Bagging. Στο πεδίο «Classifier» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka\classifiers\meta\bagging. Το Bagging είναι μια γενική τεχνική και μπορεί να συνδυαστεί με οποιαδήποτε μέθοδο κατηγοριοποίησης. Για τον λόγο αυτό, πρέπει να ορίσετε τη μέθοδο κατηγοριοποίησης για την οποία θα κατασκευαστούν πολλαπλά μοντέλα. Κάνετε κλικ στα περιεχόμενα του πεδίου «Classifier». Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Στο πεδίο «Classifier» κάντε κλικ στο κουμπί «Choose» και επιλέξτε τη μέθοδο weka/classifiers/trees/J48. Κάντε κλικ στο κουμπί «OK».

Στο σημείο αυτό έχετε ορίσει ότι θέλετε να εφαρμόσετε bagging στη μέθοδο κατηγοριοποίησης C4.5. Κάντε κλικ στο κουμπί «Start» για να εκτελέσετε τον αλγόριθμο.

Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Με τη χρήση της μεθόδου Bagging κατηγοριοποιούνται σωστά 84.2105% των συνολικών περιπτώσεων, 75% των παρατηρήσεων με κλάση «bad» και 89.2% των παρατηρήσεων με κλάση «good». Παρατηρούμε ότι επιτεύχθηκε σημαντική αύξηση των επιδόσεων.

## Άσκηση Εφαρμογής 10.5

**Χρησιμοποιήστε το αρχείο «analcatdata_bankruptcy.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή StatLib). Το σύνολο δεδομένων προέρχεται από το βιβλίο του Simonoff (2003) και σχετίζεται με τη χρεοκοπία επιχειρήσεων. Υπάρχουν 50 γραμμές, κάθε μια από τις οποίες αναφέρεται σε μια επιχείρηση. Οι μισές επιχειρήσεις έχουν χρεοκοπήσει. Στο σύνολο δεδομένων υπάρχουν 7 πεδία (στήλες). Το πρώτο πεδίο περιέχει τα ονόματα των επιχειρήσεων, και ακολουθούν 5 πεδία με αριθμοδείκτες. Το τελευταίο πεδίο είναι το πεδίο της κλάσης, και περιέχει μια ένδειξη («1» ή «0») για το εάν η επιχείρηση χρεοκόπησε ή εξακολουθεί τη λειτουργία της αντίστοιχα.**

**Αναπτύξτε μοντέλο πρόβλεψης χρεοκοπίας με χρήση της μεθόδου Νευρωνικό Δίκτυο τύπου Multilayer**

**Perceptron. Ακολούθως εκτελέστε ευαίσθητη στο κόστος κατηγοριοποίηση, εφαρμόζοντας τη μέθοδο MetaCost, σε συνδυασμό με κατηγοριοποιητή Multilayer Perceptron. Ελέγξτε εάν βελτιώθηκε το ποσοστό ορθών προβλέψεων της ακριβής κλάσης.**

## Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «analcatdata_bankruptcy.arff» πιέζοντας το κουμπί «Open file».

Το πεδίο με τα ονόματα των επιχειρήσεων δεν προσφέρει κάτι χρήσιμο στην ανάλυση μας. Το επιλέγετε και το απομακρύνετε πιέζοντας το κουμπί «Remove».

Μεταβείτε στο tab «Classify».

Επιλέξτε τη μέθοδο κατηγοριοποίησης weka/classifiers/functions/MultilayerPerceptron και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά 90% των συνολικών περιπτώσεων, 88% της κλάσης «0» (μη χρεοκοπία) και 92% της κλάσης «1» (χρεοκοπία).

Βήμα 2. Στο πεδίο «Classifier» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka\classifiers\meta\ MetaCost. Η MetaCost είναι μια γενική μέθοδος, η οποία συνδυάζεται με οποιαδήποτε μέθοδο κατηγοριοποίησης, ώστε να επιτευχθεί ευαίσθητη στο κόστος κατηγοριοποίηση.

Κάντε κλικ στα περιεχόμενα του πεδίου «Classifier» και στο όνομα «MetaCost», ώστε να ανοίξει το παράθυρο ρύθμισης των παραμέτρων.

Σε αυτό το παράθυρο κάντε κλικ στο κουμπί «Choose» του πεδίου «classifier» και επιλέξτε weka/classifiers/functions/MultilayerPerceptron. Με τον τρόπο αυτό ορίζετε ότι θα εφαρμόσετε τη μέθοδο MetaCost σε συνδυασμό με νευρωνικό δίκτυο Multilayer Perceptron.

Στο ίδιο παράθυρο κάντε κλικ στα περιεχόμενα του πεδίου «costMatrix». Ανοίγει ένα νέο παράθυρο όπου καθορίζονται οι τιμές του πίνακα κόστους. Στο πεδίο «Classes» γράψτε την τιμή «2» και κάντε κλικ στο κουμπί «Resize». Στο άνω και αριστερά μέρος του παραθύρου εμφανίζεται ένας πίνακας 2X2.

Στον πίνακα 2X2 καταχωρίστε τις τιμές κόστους όπως φαίνεται κατωτέρω.

| | |
|---|---|
| 0.0 | 1.0 |
| 10.0 | 0.0 |

Με τον τρόπο αυτό, ορίζετε ότι το κόστος εσφαλμένης κατηγοριοποίησης χρεοκοπημένων επιχειρήσεων είναι δεκαπλάσιο από το κόστος εσφαλμένης κατηγοριοποίησης μη χρεοκοπημένων επιχειρήσεων.

Κλείστε το παράθυρο με τον πίνακα κόστους και κάντε κλικ στο κουμπί «OK» του παραθύρου ρύθμισης παραμέτρων. Κάντε κλικ στο κουμπί «Start».

Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Παρατηρούμε ότι το ποσοστό ορθών προβλέψεων έπεσε στο 82%. Επίσης το ποσοστό των ορθών προβλέψεων των παρατηρήσεων κλάσης «0» (μη χρεοκοπία) έπεσε στο 64%. Όμως το ποσοστό των ορθών προβλέψεων των παρατηρήσεων κλάσης «1» (χρεοκοπία) αυξήθηκε στο 100%. Με τη χρήση της μεθόδου MetaCost επιτεύχθηκε αύξηση των ορθών προβλέψεων της ακριβής κλάσης, με αντίτιμο την πτώση των ορθών προβλέψεων των μη χρεοκοπημένων επιχειρήσεων. Με δεδομένο ότι σε πραγματικές συνθήκες το κόστος εσφαλμένης πρόβλεψης μιας χρεοκοπημένης επιχείρησης είναι πολύ μεγαλύτερο, μπορούμε να ισχυριστούμε ότι πρακτικά το μοντέλο βελτιώθηκε.