9 Κατηγοριοποίηση

Σύνοψη

Το ένατο Κεφάλαιο καλύπτει εν μέρει τη θεματική ενότητα της Κατηγοριοποίησης (Classification). Η κατηγοριοποίηση είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων, με μεγάλο αριθμό εφαρμογών στον χώρο των οικονομικών. Είναι εργασία επιβλεπόμενης μάθησης, που στόχο έχει την ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα στόχο με ονομαστικές τιμές και σε ένα σύνολο άλλων γνωρισμάτων. Μια άλλη εργασία επιβλεπόμενης μάθησης είναι η Παλινδρόμηση, η οποία όμως στοχεύει στην πρόβλεψη αριθμητικών τιμών. Στην κατηγοριοποίηση εφαρμόζεται ένας επαγωγικός αλγόριθμος και κατασκευάζεται ένα μοντέλο. Η διαδικασία της κατηγοριοποίησης περιλαμβάνει τρία στάδια. Στο πρώτο στάδιο ο αλγόριθμος επεξεργάζεται τα δεδομένα του συνόλου εκπαίδευσης και κατασκευάζει ένα μοντέλο. Στο δεύτερο στάδιο ελέγχεται η ικανότητα του μοντέλου να προβλέπει την κλάση άγνωστων παρατηρήσεων. Εάν η επίδοση του μοντέλου κριθεί ικανοποιητική, τότε ακολουθεί το τρίτο στάδιο, το οποίο συνίσταται στη χρήση του μοντέλου για τη διατύπωση προβλέψεων. Κατά την εκπαίδευση πρέπει να αποφευχθεί η υπερπροσαρμογή του μοντέλου, η απομνημόνευση δηλαδή του συγκεκριμένου συνόλου εκπαίδευσης. Αποτέλεσμα της υπερπροσαρμογής είναι η πτώση της επίδοσης έναντι άγνωστων παρατηρήσεων. Κριτήρια για την αξιολόγηση των μεθόδων κατηγοριοποίησης είναι η ακρίβεια πρόβλεψης, η ταχύτητα, η ερμηνευσιμότητα, η επεκτασιμότητα και η ανθεκτικότητα.
Στα πλαίσια του παρόντος κεφαλαίου γίνεται παρουσίαση τριών, πολύ γνωστών μεθόδων κατηγοριοποίησης. Οι μέθοδοι αυτές είναι τα Δένδρα Αποφάσεων, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron και τα Μπαϋεσιανά Δίκτυα. Τα Δένδρα Αποφάσεων βασίζονται στη διαδοχική διάσπαση του συνόλου δεδομένων σε υποσύνολα. Αναπαριστώνται με μια ανεστραμμένη δενδρική δομή, όπου κάθε κόμβος αντιπροσωπεύει έναν έλεγχο στα δεδομένα, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα του ελέγχου και κάθε φύλο αντιπροσωπεύει μια απόφαση κατηγοριοποίησης. Αφού κατασκευαστεί το μοντέλο, μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Μια νέα παρατήρηση κατηγοριοποιείται ακολουθώντας μια διαδρομή από τη ρίζα μέχρι ένα φύλο, σύμφωνα με τους ελέγχους των κόμβων. Έχουν προταθεί διάφοροι αλγόριθμοι για τη δημιουργία Δένδρων Αποφάσεων. Στο παρόν κεφάλαιο παρουσιάζονται αρχικά τα δένδρα τύπου ID3. Τα δένδρα ID3 χρησιμοποιούν ως κριτήριο για τον διαχωρισμό των παρατηρήσεων το Κέρδος Πληροφορίας, δηλαδή τη μείωση της στατιστικής εντροπίας. Τα δένδρα τύπου C4.5 αποτελούν επέκταση – βελτίωση των ID3, χρησιμοποιούν ως κριτήριο διαχωρισμού τον Λόγο Κέρδους και είναι ικανά να χειρίζονται αριθμητικές μεταβλητές εισόδου. Τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron είναι ένα πλέγμα συνδεδεμένων νευρώνων. Κάθε σύνδεση συνοδεύεται από μία αριθμητική τιμή που ονομάζεται βάρος. Ένας νευρώνας μετασχηματίζει το σήμα εισόδου και το μεταβιβάζει σε επόμενους νευρώνες. Οι νευρώνες είναι οργανωμένοι σε επίπεδα, και υπάρχουν ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και ένα ή περισσότερα κρυφά επίπεδα. Ο χρήστης προκαθορίζει τη δομή του δικτύου. Στη συνέχεια ακολουθεί η εκπαίδευση του δικτύου, η οποία συνίσταται στη ρύθμιση των βαρών των συνδέσεων. Ένας πολύ επιτυχημένος αλγόριθμος για την εκπαίδευση του δικτύου είναι η Αντίστροφη Μετάδοση Σφάλματος (Backpropagation). Τα Μπαϋεσιανά Δίκτυα αποτελούνται από έναν κατευθυνόμενο ακυκλικό γράφο και έναν πίνακα κατανομής πιθανοτήτων. Κάθε κόμβος του γράφου συμβολίζει μια στοχαστική μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης ανάμεσα σε δύο μεταβλητές. Οι Αφελείς Μπαϋεσιανοί κατηγοριοποιητές υποθέτουν την υπό συνθήκη ανεξαρτησία των μεταβλητών εισόδου. Αντιθέτως, τα Μπαϋεσιανά Δίκτυα επιτρέπουν την ανεξαρτησία υποσυνόλων των μεταβλητών εισόδου. Η εκπαίδευση ενός Μπαϋεσιανού Δικτύου περιλαμβάνει τον σχεδιασμό του γράφου και τον υπολογισμό του πίνακα πιθανοτήτων. Στο τέλος του κεφαλαίου παρουσιάζεται μια μελέτη περίπτωσης, όπου οι τρεις προαναφερθείσες τεχνικές κατηγοριοποίησης εφαρμόζονται για τον εντοπισμό περιπτώσεων παραποίησης των χρηματοοικονομικών καταστάσεων επιχειρήσεων. Τα τρία μοντέλα επιτυγχάνουν υψηλό βαθμό ακρίβειας έναντι άγνωστων παρατηρήσεων και αποκαλύπτουν σημαντικούς παράγοντες που σχετίζονται με τις περιπτώσεις διοικητικής απάτης

Προηγούμενη γνώση

Η θεματική ενότητα του παρόντος κεφαλαίου είναι αυτόνομη και δεν απαιτούνται ιδιαίτερες προηγούμενες γνώσεις. Ωστόσο, για την καλύτερη κατανόηση των περιεχομένων θα συνιστούσαμε την προηγούμενη ανάγνωση του Κεφαλαίου 6, το οποίο εισάγει τον αναγνώστη στην Εξόρυξη Δεδομένων και την ανάγνωση του Κεφαλαίου 7, το οποίο αναφέρεται στην προεπεξεργασία των δεδομένων. Για τον αναγνώστη που ενδιαφέρεται να αναζητήσει περισσότερες πληροφορίες για την Κατηγοριοποίηση και τις τρεις συγκεκριμένες μεθόδους που παρουσιάζονται,

υπάρχει πληθώρα διαθέσιμων συγγραμμάτων. Ενδεικτικά αναφέρουμε τα βιβλία των Han, Kamber and Pei (2011) και των Maimon and Rokach (2010).

9.1 Εισαγωγή

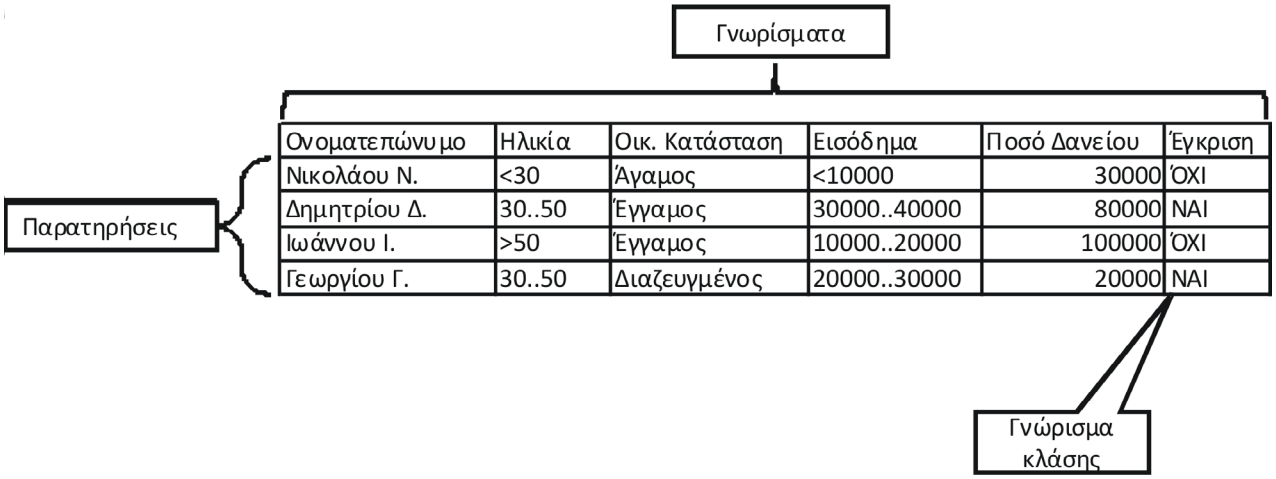
Η κατηγοριοποίηση *(*classification) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων, με μεγάλο αριθμό εφαρμογών στον χώρο των οικονομικών. Η πρόβλεψη χρεοκοπίας, η έγκριση δανείων, η αναγνώριση απάτης είναι τυπικά προβλήματα κατηγοριοποίησης. Η κατηγοριοποίηση είναι εργασία **επιβλεπόμενης μάθησης**. Στόχος της επιβλεπόμενης μάθησης είναι η ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα στόχο και σε ένα σύνολο άλλων γνωρισμάτων. Το γνώρισμα στόχος αναφέρεται και ως εξαρτημένη μεταβλητή, ενώ τα υπόλοιπα γνωρίσματα αναφέρονται και ως ανεξάρτητες μεταβλητές. Με την επιβλεπόμενη μάθηση επιτυγχάνεται η δημιουργία ενός μηχανισμού λήψης αποφάσεων ή υπολογισμών, ο οποίος είναι ικανός να προβλέπει τις τιμές της εξαρτημένης μεταβλητής χρησιμοποιώντας τις ανεξάρτητες μεταβλητές. Ο μηχανισμός λήψης απόφασης καλείται και μοντέλο και μπορεί να έχει διάφορες μορφές, όπως πχ να είναι ένα σύνολο κανόνων ή μια εξίσωση ή το πλέγμα των νευρώνων και συνδέσεων ενός Νευρωνικού Δικτύου.

Στην επιβλεπόμενη μάθηση ανήκουν η **Κατηγοριοποίηση** (Classification) και η **Παλινδρόμηση** (Regression). Η κατηγοριοποίηση και η παλινδρόμηση έχουν πολλές ομοιότητες. Και στις δύο περιπτώσεις στόχος είναι η πρόβλεψη των τιμών ενός γνωρίσματος, με χρήση άλλων γνωρισμάτων. Επίσης, και στις δύο περιπτώσεις χρησιμοποιείται ένα σύνολο δεδομένων εκπαίδευσης, με την επεξεργασία του οποίου κατασκευάζεται το μοντέλο. Η διαφορά ανάμεσα στην κατηγοριοποίηση και στην παλινδρόμηση έχει σχέση με τον τύπο της εξαρτημένης μεταβλητής. Στόχος της παλινδρόμησης είναι η πρόβλεψη μιας εξαρτημένης μεταβλητής, η οποία περιέχει συνεχόμενες (αριθμητικές) τιμές. Αντιθέτως, κατηγοριοποίηση είναι η πρόβλεψη διακριτών ονομαστικών τιμών. Οι τιμές αυτές είναι συγκεκριμένες, γνωστές εκ των προτέρων και ορίζουν την κλάση (κατηγορία) στην οποία ανήκει κάθε αντικείμενο. Για τον λόγο αυτό, η εξαρτημένη μεταβλητή σε προβλήματα κατηγοριοποίησης καλείται και **γνώρισμα κλάσης**.

Με την ένταξη αντικειμένων σε ομάδες ασχολείται και μια άλλη εργασία Εξόρυξης Δεδομένων, η **Ανάλυση Συστάδων** (Clustering). Οι διαφορές ανάμεσα στην Ανάλυση Συστάδων και στην Κατηγοριοποίηση είναι μεγάλες. Η Ανάλυση Συστάδων επιμερίζει τα αντικείμενα σε ομάδες βάσει της ομοιότητας τους. Οι συστάδες και το πλήθος τους δεν είναι εκ των προτέρων γνωστές. Επίσης, δεν υπάρχει στα δεδομένα κάποιο πεδίο που να καθορίζει την ομάδα στην οποία ανήκει το κάθε αντικείμενο. Αντιθέτως, στην κατηγοριοποίηση οι κατηγορίες είναι εκ των προτέρων γνωστές. Οι τιμές του γνωρίσματος κλάσης ορίζουν την κατηγορία στην οποία ανήκει κάθε αντικείμενο.

Ένα παράδειγμα προβλήματος κατηγοριοποίησης είναι η έγκριση των τραπεζικών δανείων. Το σύνολο δεδομένων περιλαμβάνει στοιχεία για τον υποψήφιο δανειολήπτη, στοιχεία σχετικά με το δάνειο, καθώς επίσης και την τελική απόφαση για την έγκριση ή την απόρριψη του δανείου. Κάθε γραμμή του συνόλου δεδομένων αντιστοιχεί σε μια αίτηση. Οι γραμμές καλούνται και αντικείμενα, παραδείγματα ή παρατηρήσεις. Οι στήλες αναφέρονται σε μια ιδιότητα των αντικειμένων, όπως πχ το επάγγελμα του δανειολήπτη ή το είδος του δανείου (στεγαστικό, καταναλωτικό κλπ.). Οι στήλες καλούνται και πεδία (fields), μεταβλητές (variables), γνωρίσματα (attributes) ή χαρακτηριστικά (features). Το γνώρισμα το οποίο περιέχει την απόφαση της έγκρισης ή απόρριψης του δανείου είναι το γνώρισμα της κλάσης. Η έγκριση του δανείου εξαρτάται από τα στοιχεία της αίτησης, όπως η ηλικία, το επάγγελμα και η οικονομική κατάσταση του δανειολήπτη, το ποσό και ο τύπος του δανείου κλπ. Η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να προβλέπει την έγκριση ή απόρριψη του δανείου χρησιμοποιώντας τα υπόλοιπα στοιχεία της αίτησης, είναι ένα πρόβλημα κατηγοριοποίησης. Στο Σχήμα 9.1 παρουσιάζεται το σύνολο δεδομένων αυτού του παραδείγματος

Ένας Επαγωγικός Αλγόριθμος είναι μια οντότητα, η οποία επεξεργάζεται ένα σύνολο δεδομένων και κατασκευάζει ένα μοντέλο. Το μοντέλο είναι μια τυποποίηση, η οποία περιγράφει τη γενίκευση της σχέσης ανάμεσα σε μια εξαρτημένη μεταβλητή και σε ένα σύνολο ανεξάρτητων μεταβλητών. Με άλλα λόγια, το μοντέλο μπορεί και δέχεται ως είσοδο τις τιμές των ανεξάρτητων μεταβλητών και παράγει ως έξοδο μια τιμή για την εξαρτημένη μεταβλητή. Το μοντέλο, αφού κατασκευαστεί, μπορεί να χρησιμοποιηθεί για την πρόβλεψη της κλάσης νέων παρατηρήσεων.



Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Νικολάου Ν.	<30	Άγαμος	<10000	30000	ΟΧΙ
Δημητρίου Δ.	30..50	Έγγαμος	30000..40000	80000	ΝΑΙ
Ιωάννου Ι.	>50	Έγγαμος	10000..20000	100000	ΟΧΙ
Γεωργίου Γ.	30..50	Διαζευγμένος	20000..30000	20000	ΝΑΙ

Σχήμα 9.1 *Έγκριση τραπεζικών δανείων*

9.2 Επαγωγικοί Αλγόριθμοι και Μοντέλα

Επαγωγικοί αλγόριθμοι ή μέθοδοι κατηγοριοποίησης υπάρχουν πολλές, όπως πχ τα Δένδρα Αποφάσεων, τα Μπαϋεσιανά Δίκτυα, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron κλπ. Τα μοντέλα τα οποία κατασκευάζει η κάθε μέθοδος είναι τελείως διαφορετικά. Για παράδειγμα, ένα μοντέλο Δένδρου Αποφάσεων είναι μια δενδρική δομή, όπου κάθε κόμβος είναι ένας έλεγχος σε κάποιο γνώρισμα, κάθε κλάδος είναι ένα αποτέλεσμα του ελέγχου και κάθε φύλο είναι μια απόφαση κατηγοριοποίησης. Ένα μοντέλο Μπαϋεσιανού Δικτύου είναι ένας κατευθυνόμενος ακυκλικός γράφος και μια κατανομή πιθανοτήτων συσχέτισης μεταξύ των μεταβλητών. Στον γράφο κάθε κόμβος αντιστοιχεί σε μια μεταβλητή.

Όπως αναφέρθηκε και προηγουμένως, ένας επαγωγικός αλγόριθμος επεξεργάζεται ένα σύνολο δεδομένων και παράγει ένα μοντέλο. Αν συμβολίσουμε με το *I* ένα επαγωγικό αλγόριθμο και με *D* ένα σύνολο δεδομένων, τότε το *I(D)* συμβολίζει το μοντέλο που θα παραχθεί από την επεξεργασία του *D* από τον *I*. Αν ο ίδιος αλγόριθμος επεξεργαστεί ένα διαφορετικό σύνολο δεδομένων *D'*, τότε θα παραχθεί ένα διαφορετικό μοντέλο *I(D')*. Εάν το *D'* περιέχει διαφορετικά πεδία από το *D,* είναι προφανές ότι τα δύο μοντέλα θα είναι διαφορετικά. Αν τα δύο σύνολα δεδομένων περιέχουν τα ίδια πεδία, αλλά διαφορετικές παρατηρήσεις, τότε και πάλι θα προκύψουν δύο διαφορετικά μοντέλα. Υπάρχουν μάλιστα μέθοδοι, όπως τα Δένδρα Αποφάσεων, όπου μικρές διαφορές στα σύνολα δεδομένων έχουν σαν αποτέλεσμα τη δημιουργία σημαντικά διαφορετικών μοντέλων.

Μια μέθοδος κατηγοριοποίησης μπορεί να είναι αιτιοκρατική (deterministic) ή στοχαστική (stochastic). Οι στοχαστικές μέθοδοι καλούνται και πιθανολογικές (probabilistic). Μια αιτιοκρατική μέθοδος δημιουργεί μοντέλα, τα οποία εκχωρούν μια παρατήρηση σε μια κλάση. Τα στοχαστικά μοντέλα υπολογίζουν την πιθανότητα να ανήκει η παρατήρηση σε κάθε μια από τις δυνατές κλάσεις. Παράδειγμα αιτιοκρατικής μεθόδου είναι τα Δένδρα αποφάσεων τύπου C4.5, ενώ παράδειγμα στοχαστικής μεθόδου είναι τα Μπαϋεσιανά Δίκτυα.

9.3 Στάδια κατηγοριοποίησης

Η κατηγοριοποίηση περιλαμβάνει τρία στάδια, το στάδιο της επιβλεπόμενης μάθησης, το στάδιο της επικύρωσης του μοντέλου και το στάδιο της χρήσης του μοντέλου. Αναλυτικότερα, οι εργασίες που λαμβάνουν χώρα σε κάθε στάδιο είναι οι ακόλουθες:

- **Επιβλεπόμενη μάθηση**. Στο στάδιο αυτό, μια μέθοδος κατηγοριοποίησης αναλύει ένα σύνολο δεδομένων. Η μέθοδος θα ανακαλύψει σχέσεις μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Το αποτέλεσμα αυτής της επεξεργασίας είναι η κατασκευή ενός μοντέλου. Η κατασκευή ή εκπαίδευση του μοντέλου καθοδηγείται από τις τιμές του γνωρίσματος της κλάσης και για τον λόγο αυτό η διαδικασία ονομάζεται επιβλεπόμενη μάθηση. Το σύνολο δεδομένων, το οποίο χρησιμοποιείται για την εκπαίδευση του μοντέλου, ονομάζεται σύνολο εκπαίδευσης (training data set). Η επιλογή του συνόλου εκπαίδευσης είναι καθοριστικής σημασίας, γιατί το μοντέλο που θα προκύψει θα αποτυπώνει σχέσεις που υπάρχουν στο σύνολο εκπαίδευσης. Μεροληπτικά **σύνολα**

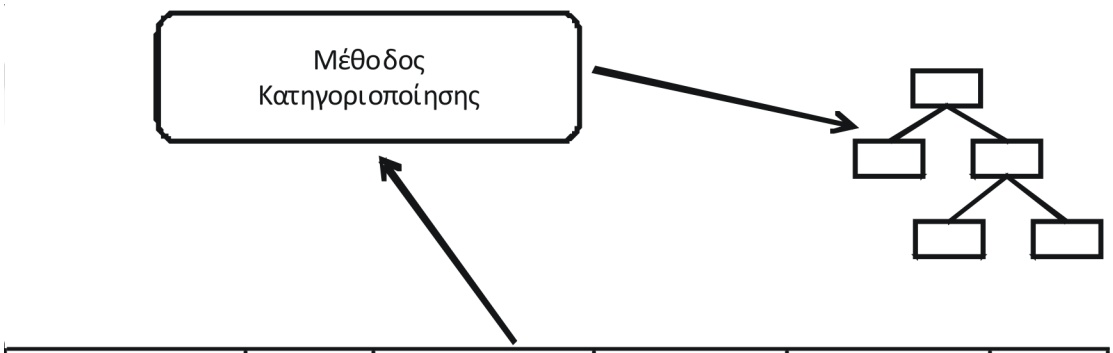
εκπαίδευσης θα οδηγήσουν στην κατασκευή μεροληπτικών μοντέλων.

- **Επικύρωση μοντέλου.** Στο στάδιο αυτό δοκιμάζεται η ακρίβεια του μοντέλου, η ικανότητα του δηλαδή να προβλέπει σωστά την κλάση των παρατηρήσεων. Το μοντέλο τροφοδοτείται με παρατηρήσεις, των οποίων η κλάση είναι γνωστή. Αναλύοντας τα στοιχεία των ανεξάρτητων μεταβλητών κάθε παρατήρησης, το μοντέλο προβλέπει την κλάση της παρατήρησης και στη συνέχεια συγκρίνεται η πρόβλεψη του μοντέλου με την πραγματική τιμή της κλάσης. Αν το μοντέλο επιδείξει ικανοποιητική ακρίβεια προβλέψεων, εάν δηλαδή προβλέψει σωστά την κλάση ενός ικανοποιητικού ποσοστού παρατηρήσεων, τότε θεωρείται επιτυχημένο και μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Η διαδικασία δοκιμής του μοντέλου καλείται **επικύρωση** (validation) και το σύνολο δεδομένων που χρησιμοποιείται για τη δοκιμή καλείται **σύνολο επικύρωσης** (validation set). Σκοπός ενός μοντέλου είναι να χρησιμοποιηθεί για τη διατύπωση προβλέψεων στην πραγματική ζωή και όχι να αναλύσει ένα συγκεκριμένο σύνολο δεδομένων. Το μοντέλο πρέπει να αποδείξει την ικανότητα του να προβλέπει την κλάση άγνωστων παρατηρήσεων, παρατηρήσεων δηλαδή διαφορετικών από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση του. Για τον λόγο αυτό, το σύνολο εκπαίδευσης και το σύνολο επικύρωσης πρέπει να περιέχουν διαφορετικές παρατηρήσεις.
- **Χρήση του μοντέλο**υ. Το μοντέλο, αφού εκπαιδευτεί και επικυρωθεί, χρησιμοποιείται για τη διατύπωση προβλέψεων. Μια νέα παρατήρηση, της οποίας η κλάση είναι άγνωστη, εισάγεται στο μοντέλο. Το μοντέλο χρησιμοποιώντας τις τιμές των ανεξάρτητων μεταβλητών υπολογίζει την τιμή της κλάσης.

Το Σχήμα 9.2 παρουσιάζει τα στάδια της κατηγοριοποίησης με τη βοήθεια ενός παραδείγματος. Στο τμήμα Α) απεικονίζεται η επιβλεπόμενη μάθηση. Μια μέθοδος κατηγοριοποίησης επεξεργάζεται ένα σύνολο εκπαίδευσης, το οποίο περιέχει στοιχεία δανείων και κατασκευάζεται ένα μοντέλο. Το μοντέλο μπορεί να προβλέψει την έγκριση ή απόρριψη του δανείου από τα υπόλοιπα στοιχεία της εκάστοτε αίτησης. Στο τμήμα Β) απεικονίζεται η επικύρωση του μοντέλου. Το μοντέλο τροφοδοτείται με περιπτώσεις δανείων διαφορετικές από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση. Για κάθε δάνειο, το μοντέλο πραγματοποιεί μια πρόβλεψη και η πρόβλεψη αυτή συγκρίνεται με την πραγματική απόφαση έγκρισης ή απόρριψης του δανείου. Υπολογίζεται η ακρίβεια του μοντέλου. Στο τμήμα Γ) το μοντέλο χρησιμοποιείται για την πρόβλεψη έγκρισης νέων δανείων.

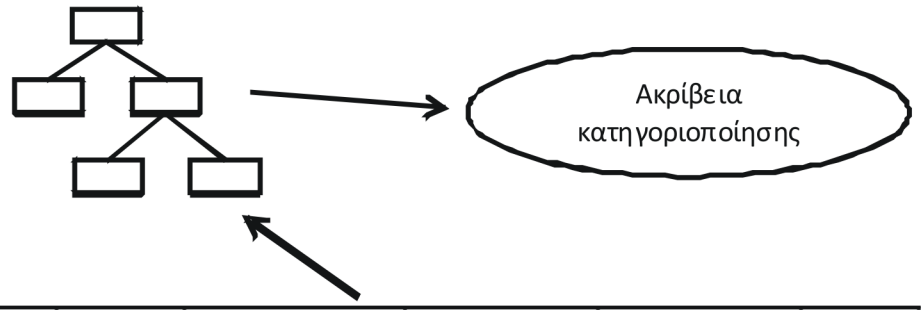
9.4 Υπερπροσαρμογή μοντέλων

Αναφέρθηκε προηγουμένως ότι για την εκτίμηση της ακρίβειας του μοντέλου πρέπει να χρησιμοποιηθούν παρατηρήσεις διαφορετικές από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση. «Γιατί το σύνολο εκπαίδευσης δεν είναι αρκετά καλό για τον έλεγχο της ακρίβειας;», θα μπορούσε να αναρωτηθεί κανείς. Η απάντηση έχει σχέση με το πρόβλημα της υπερπροσαρμογής των μοντέλων στο σύνολο δεδομένων εκπαίδευσης. Με τον όρο **υπερπροσαρμογή** στα δεδομένα εκπαίδευσης (data overfitting) ορίζουμε το φαινόμενο όπου το μοντέλο «απομνημονεύει» τις περιπτώσεις οι οποίες υπάρχουν στο σύνολο εκπαίδευσης, αντί να εκπαιδεύεται ουσιαστικά, ενσωματώνοντας «κανόνες» γενικότερης ισχύος. Ένα υπερβολικά προσαρμοσμένο μοντέλο ενσωματώνει και τον θόρυβο των δεδομένων. Ακόμα όμως και όταν δεν υπάρχει θόρυβος, η υπερβολική προσαρμογή του μοντέλου στα συγκεκριμένα δεδομένα θα το εμποδίσει να προβλέψει σωστά την κλάση νέων παρατηρήσεων. Η υπερπροσαρμογή παρουσιάζεται όταν ένα μοντέλο είναι υπερβολικά περίπλοκο. Το μοντέλο αυτό είναι ικανό να αφομοιώσει τις ιδιαιτερότητες των δεδομένων εκπαίδευσης, αντί να καταγράψει σχέσεις γενικότερης ισχύος. Στο Σχήμα 9.3 στο τμήμα Α) απεικονίζεται ένα σύνολο δεδομένων εκπαίδευσης με δύο μεταβλητές. Επίσης παρουσιάζονται δύο μοντέλα, το ένα από τα οποία συμβολίζεται με τη συνεχόμενη γραμμή, ενώ το δεύτερο με τη διακεκομμένη γραμμή. Το πρώτο μοντέλο επιτυγχάνει ικανοποιητικό βαθμό γενίκευσης, ενώ το δεύτερο είναι πολύ σύνθετο και υπερπροσαρμοσμένο στα δεδομένα.



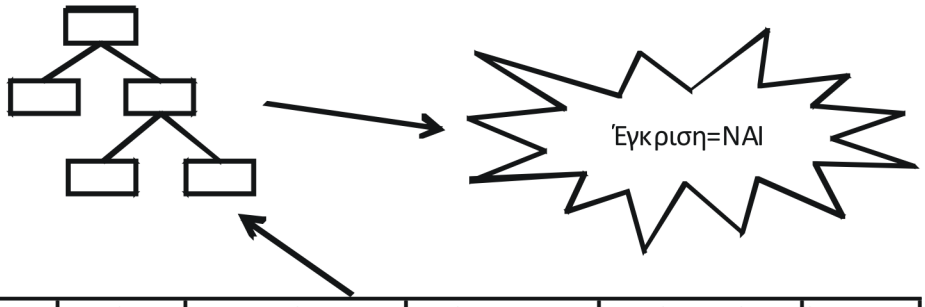
Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Νικολάου Ν.	<30	Άγαμος	<10000	30000	ΟΧΙ
Δημητρίου Δ.	30..50	Έγγαμος	30000..40000	80000	ΝΑΙ
Ιωάννου Ι.	>50	Έγγαμος	10000..20000	100000	ΟΧΙ
Γεωργίου Γ.	30..50	Διαζευγμένος	20000..30000	20000	ΝΑΙ

Α)



Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Κωνσταντίνου Κ.	<30	Άγαμος	<10000	40000	ΟΧΙ
Παναγιώτου Π.	30..50	Έγγαμος	30000..40000	100000	ΝΑΙ

Β)

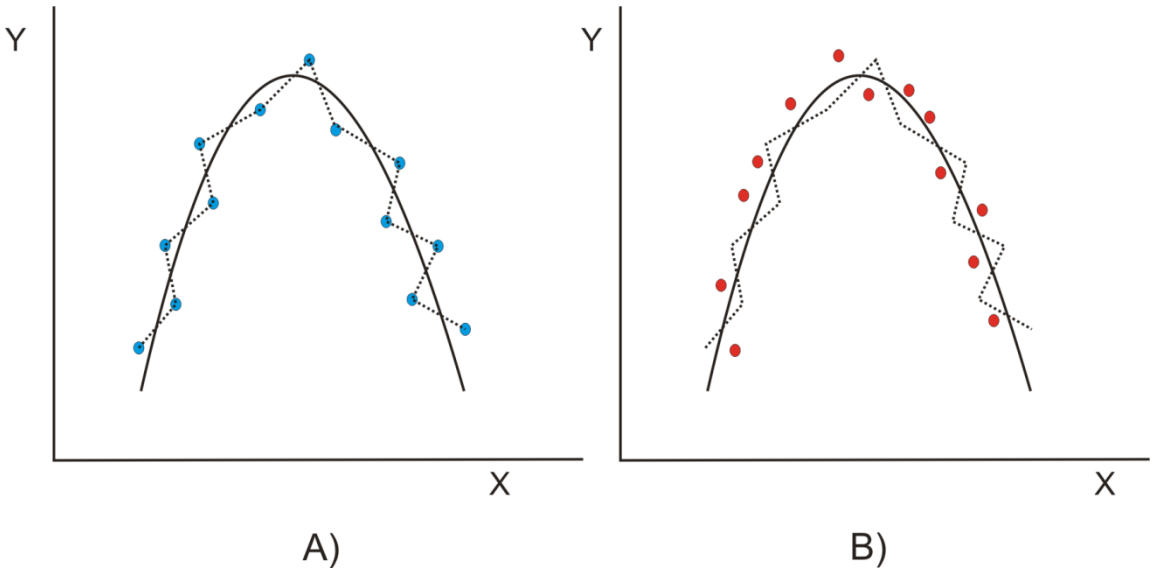


Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Χρήστου Χ.	30..50	Έγγαμος	30000..40000	90000	

Γ)

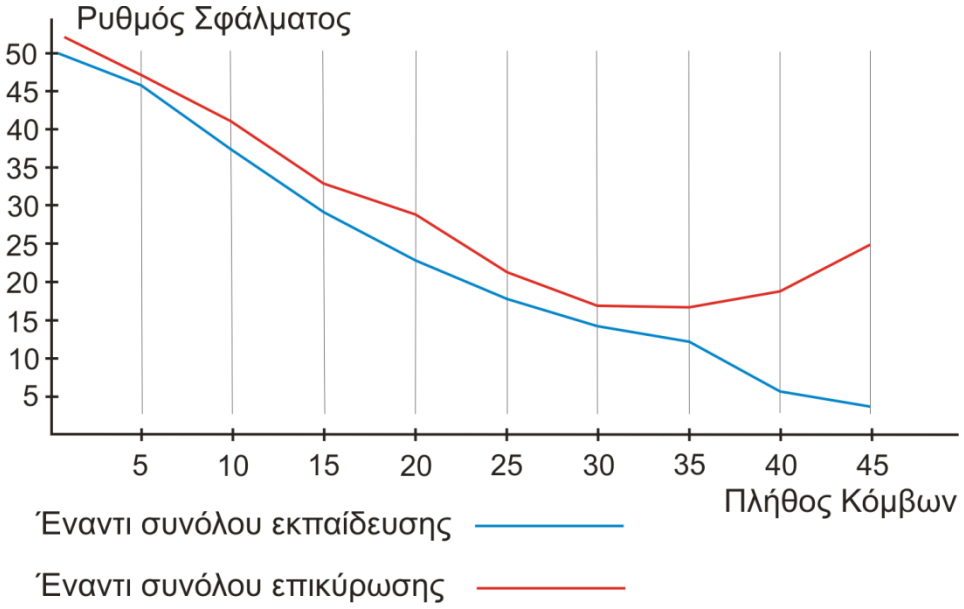
Σχήμα 9.2 *Στάδια Κατηγοριοποίησης*

Ένα υπερπροσαρμοσμένο μοντέλο επιτυγχάνει εξαιρετικά υψηλές επιδόσεις έναντι των δεδομένων εκπαίδευσης, οι επιδόσεις του όμως έναντι άγνωστων παρατηρήσεων δεν είναι ικανοποιητικές. Για τον λόγο αυτό, εξαιρετικά υψηλός ρυθμός ακρίβειας έναντι του συνόλου εκπαίδευσης, όχι μόνον δεν είναι ασφαλές μέτρο της επιτυχίας του μοντέλου, αλλά αποτελεί ένδειξη πιθανής υπερπροσαρμογής του. Στο Σχήμα 9.3 παρουσιάζονται τα δύο μοντέλα και ένα σύνολο νέων παρατηρήσεων, οι οποίες συμβολίζονται με κόκκινες τελείες. Είναι εμφανές ότι το γενικευμένο μοντέλο που συμβολίζεται με τη συνεχόμενη γραμμή προβλέπει καλύτερα τις τιμές του Y από τις τιμές του X.



Σχήμα 9.3 *Υπερπροσαρμογή κατηγοριοποιητή*

Αντίστροφο πρόβλημα της υπερπροσαρμογής είναι η υποπροσαρμογή. Στην περίπτωση της **υποπροσαρμογής**, το μοντέλο είναι υπερβολικά απλό για να ενσωματώσει τις ουσιαστικές σχέσεις, οι οποίες υπάρχουν στα δεδομένα εκπαίδευσης. Αποτέλεσμα της υποπροσαρμογής είναι η χαμηλή ακρίβεια έναντι και των δεδομένων εκπαίδευσης και των άγνωστων παρατηρήσεων. Στο Σχήμα 9.4 παρουσιάζεται η πτώση του ρυθμού σφάλματος σε σχέση με την πολυπλοκότητα του μοντέλου, η οποία εκφράζεται ως πλήθος κόμβων ενός Δένδρου Αποφάσεων. Αρχικά το μοντέλο είναι εξαιρετικά απλό και ο ρυθμός σφάλματος είναι περίπου 50%. Ο ρυθμός σφάλματος μειώνεται με την αύξηση της πολυπλοκότητας, μέχρι τους 30 κόμβους. Πέραν των 30 κόμβων, ο ρυθμός σφάλματος έναντι του συνόλου εκπαίδευσης εξακολουθεί να μειώνεται. Αντιθέτως, ο ρυθμός σφάλματος έναντι του συνόλου επικύρωσης, το οποίο αποτελείται από άγνωστες παρατηρήσεις, αρχικά διατηρείται σταθερός μέχρι τους 35 κόμβους και στη συνέχεια αυξάνεται.



Σχήμα 9.4 *Πτώση Ρυθμού Σφάλματος και πολυπλοκότητα μοντέλου*

9.5 Κριτήρια αξιολόγησης μεθόδων κατηγοριοποίησης

Η έρευνα σχετικά με την κατηγοριοποίηση έχει αποδώσει πλούσιους καρπούς και σήμερα υπάρχουν διαθέσιμες αρκετές και πολύ διαφορετικές μέθοδοι κατηγοριοποίησης. Ορισμένες από αυτές, όπως πχ τα Νευρωνικά

Δίκτυα, θεωρούνται ιδιαίτερα ικανές να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Οι μέθοδοι αυτές μπορούν να θεωρηθούν «καλύτερες» από άλλες, όμως η ακρίβεια δεν είναι το μοναδικό κριτήριο αξιολόγησης των μεθόδων κατηγοριοποίησης. Αναλυτικότερα, οι μέθοδοι κατηγοριοποίησης μπορούν να αξιολογηθούν με βάση τα παρακάτω κριτήρια:

- **Ακρίβεια πρόβλεψης** (accuracy). Είναι η ικανότητα των μοντέλων να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Προφανώς πρόκειται για ένα πολύ σημαντικό κριτήριο και μεγάλο μέρος της έρευνας προσανατολίζεται στην ανακάλυψη μεθόδων υψηλών επιδόσεων.
- **Ταχύτητα** (speed). Σχετίζεται με την πολυπλοκότητα της μεθόδου και το υπολογιστικό κόστος που αυτή συνεπάγεται. Η εκτέλεση περίπλοκων αλγορίθμων, οι οποίοι απαιτούν εκτεταμένους υπολογισμούς, προκαλούν καθυστερήσεις. Καθυστερήσεις μπορεί να υπάρχουν στη διαδικασία κατασκευής, αλλά και στη χρήση των μοντέλων, στην εφαρμογή τους δηλαδή για την κατηγοριοποίηση μιας νέας παρατήρησης. Ορισμένες μέθοδοι, όπως τα Δένδρα Αποφάσεων, διαθέτουν γρήγορους αλγορίθμους και ο χρόνος κατασκευής των μοντέλων είναι μικρός. Άλλες μέθοδοι, όπως τα Νευρωνικά Δίκτυα, χρειάζονται πολύ περισσότερο χρόνο για την εκπαίδευση των μοντέλων. Κατά κανόνα ο χρόνος χρήσης των μοντέλων είναι πολύ μικρός. Ωστόσο, υπάρχουν μέθοδοι, όπως οι k-Πλησιέστεροι Γείτονες, οι οποίες δεν εκπαιδεύουν κάποιο μοντέλο, όμως ο χρόνος για την κατηγοριοποίηση νέων παρατηρήσεων είναι μεγάλος.
- **Ερμηνευσιμότητα** (interpretability). Είναι η ικανότητα της μεθόδου να παράγει μοντέλα, τα οποία είναι κατανοητά από τον άνθρωπο. Για παράδειγμα, στα Δένδρα Αποφάσεων ο τρόπος λήψης της απόφασης κατηγοριοποίησης είναι απολύτως κατανοητός και το μοντέλο μπορεί εύκολα νε μετατραπεί σε ένα σύνολο κανόνων της μορφής ΕΑΝ-ΤΟΤΕ. Αντιθέτως, τα μοντέλα άλλων μεθόδων, όπως τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης, λειτουργούν ως «μαύρα κουτιά». Στα μοντέλα αυτά παρέχονται οι τιμές των μεταβλητών εισόδου και υπολογίζεται η απόφαση κατηγοριοποίησης στην έξοδο. Ο τρόπος λήψης της απόφασης όμως δεν είναι κατανοητός στον άνθρωπο. Η ερμηνευσιμότητα είναι μια σημαντική ιδιότητα των μεθόδων κατηγοριοποίησης. Σε πολλές περιπτώσεις οι χρήστες των μοντέλων επιθυμούν να γνωρίζουν τον τρόπο λήψης της απόφασης, ώστε να είναι πιο σίγουροι για το αποτέλεσμα. Επίσης, στο μοντέλο καταγράφονται σχέσεις μεταξύ των δεδομένων. Ορισμένες από τις σχέσεις αυτές μπορεί να είναι νέες και άγνωστες. Αν το μοντέλο είναι ερμηνεύσιμο θα αποκαλυφθούν οι νέες σχέσεις και η μέθοδος κατηγοριοποίησης θα χρησιμοποιηθεί ως εργαλείο ανάλυσης, ικανό να προσφέρει καινοτόμα γνώση.
- **Επεκτασιμότητα** (scalability). Αναφέρεται στην ικανότητα των μεθόδων να χειριστούν πολύ μεγάλα σύνολα δεδομένων. Η Μηχανική Μάθηση και η Στατιστική προσφέρουν μεθόδους κατηγοριοποίησης. Ωστόσο, η εφαρμογή αυτών των μεθόδων για την επεξεργασία δεδομένων μεγάλου όγκου δεν είναι πάντα εύκολη. Σε αρκετές περιπτώσεις η υπολογιστική πολυπλοκότητα των μεθόδων είναι συνάρτηση του πλήθους των παρατηρήσεων και μάλιστα με σχέση περισσότερο από γραμμική. Επίσης, οι περισσότερες μέθοδοι απαιτούν την εγκατάσταση του συνόλου εκπαίδευσης στην κύρια μνήμη του υπολογιστή. Τα ζητήματα αυτά θέτουν όρια στη δυνατότητα εφαρμογής των μεθόδων. Όμως αντικείμενο της Εξόρυξης Δεδομένων είναι η ανακάλυψη γνώσης από δεδομένα μεγάλου όγκου. Ειδικά στη σημερινή εποχή, η παραγωγή και καταγραφή δεδομένων είναι μαζικότατη. Σε ότι αφορά την εφαρμογή των μεθόδων αυτών για επιχειρηματικούς σκοπούς, η τάση που παρουσιάστηκε στα τέλη της δεκαετίας του 90’ για δημιουργία Αποθηκών Δεδομένων, έχει οδηγήσει στην αποθήκευση δεδομένων, που ο όγκος τους είναι της τάξης μεγέθους terabyte. Για να έχουν πρακτική χρησιμότητα, οι μέθοδοι Εξόρυξης Δεδομένων πρέπει να είναι ικανές να χειριστούν αυτά τα πολύ μεγάλου όγκου δεδομένα. Όπως χαρακτηριστικά επισημαίνουν οι Fayyad, Piatetsky-Shapiro and Smyth (1996), η πρόκληση για την κοινότητα των ερευνητών Εξόρυξης Δεδομένων είναι η κατασκευή μεθόδων που διευκολύνουν τη χρήση αλγορίθμων εξόρυξης δεδομένων σε βάσεις δεδομένων του πραγματικού κόσμου.
- **Ανθεκτικότητα** (robustness). Αναφέρεται στην ικανότητα των μεθόδων να πραγματοποιήσουν ορθές προβλέψεις, όταν τα δεδομένα χαρακτηρίζονται από προβλήματα, όπως ο θόρυβος και οι χαμένες τιμές.

9.6 Προεπεξεργασία δεδομένων για κατηγοριοποίηση

Όπως εξηγείται αναλυτικά στο Κεφάλαιο 6, η προεπεξεργασία των δεδομένων είναι ένα απαραίτητο στάδιο,

το οποίο προηγείται της καθαυτό εξόρυξης δεδομένων. Για την περίπτωση της κατηγοριοποίησης, η προεπεξεργασία μπορεί να βελτιώσει την αποτελεσματικότητα, την αποδοτικότητα και την επεκτασιμότητα των μεθόδων. Στο στάδιο της προεπεξεργασίας αντιμετωπίζεται το πρόβλημα του θορύβου και των χαμένων τιμών. Επίσης, τα δεδομένα μπορούν να αναχθούν σε υψηλότερα επίπεδα γενίκευσης, να διακριτοποιηθούν ώστε να μετατραπούν τα αριθμητικά πεδία σε ονομαστικά και τέλος, να κανονικοποιηθούν, να αντικατασταθούν δηλαδή οι αριθμητικές τιμές με άλλες, πιο «κατάλληλες», αριθμητικές τιμές. Σε πολλές περιπτώσεις η διακριτοποίηση και η κανονικοποίηση είναι απαραίτητες, ώστε να προσαρμοστούν τα δεδομένα σε ιδιαιτερότητες των μεθόδων κατηγοριοποίησης. Για παράδειγμα, η μέθοδος των k-Πλησιέστερων Γειτόνων είναι ιδιαίτερα ευπαθής σε δεδομένα που περιέχουν πεδία με πολύ μεγάλες τιμές και πεδία με πολύ μικρές τιμές.

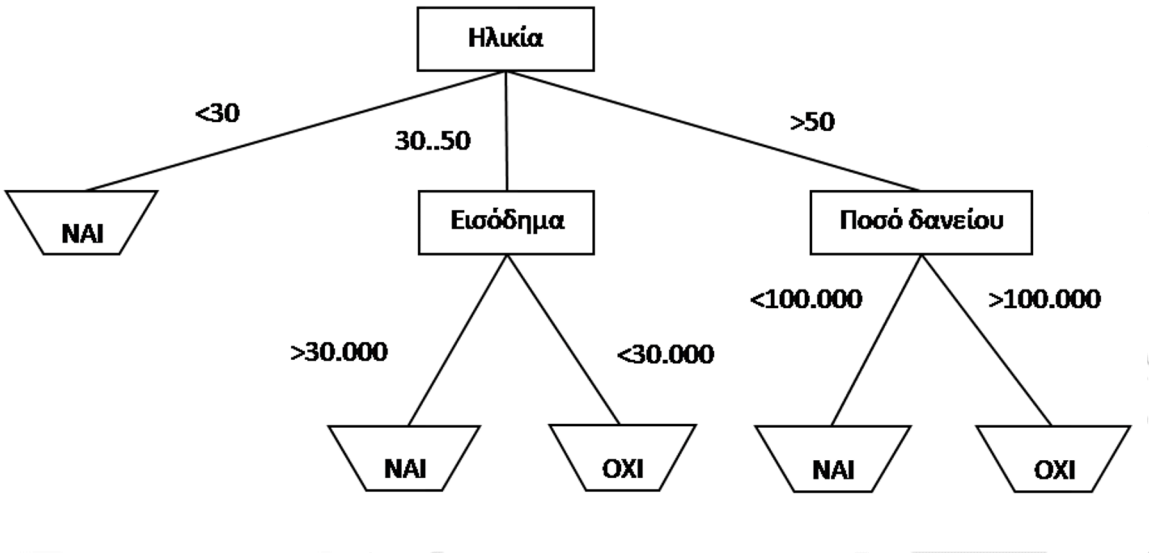
Ιδιαίτερης σημασίας είναι το ζήτημα του πλήθους των διαστάσεων και της επιλογής χαρακτηριστικών. Ερευνητικές εργασίες έχουν αποδείξει ότι το πλήθος των διαστάσεων είναι άμεσα συναρτημένο με το πλήθος των παρατηρήσεων, οι οποίες είναι απαραίτητες για την κατασκευή των μοντέλων. Ανάλογα με το είδος του κατηγοριοποιητή, το πλήθος των παρατηρήσεων μπορεί να είναι γραμμική ή και εκθετική συνάρτηση του πλήθους των διαστάσεων (Fukunaga, 1990; Hwang, Lay & Lippman, 1994). Επίσης, ορισμένες μέθοδοι όπως τα Δένδρα Αποφάσεων είναι ιδιαίτερα ευπαθείς στην ύπαρξη πολλών μεταβλητών εισόδου. Για την αντιμετώπιση του προβλήματος των πολλών διαστάσεων εφαρμόζονται μέθοδοι επιλογής χαρακτηριστικών (feature selection). Αναλυτική παρουσίαση των μεθόδων επιλογής χαρακτηριστικών γίνεται στα πλαίσια του Κεφαλαίου 7. Ωστόσο, η επιλογή χαρακτηριστικών δεν αποτελεί πανάκεια. Σε ορισμένες περιπτώσεις, είναι πιθανόν να επιλεχθεί ένας μεγάλος αριθμός μεταβλητών εισόδου. Αυτό συμβαίνει όταν το γνώρισμα της κλάσης εξαρτάται ουσιαστικά από πολλά άλλα γνωρίσματα. Επίσης, ορισμένες μέθοδοι συναρτούν το πλήθος των επιλεγμένων γνωρισμάτων με το πλήθος των παρατηρήσεων που χρησιμοποιούνται. Αν οι παρατηρήσεις που θα χρησιμοποιηθούν για την επιλογή χαρακτηριστικών είναι λίγες, τότε και τα επιλεγμένα χαρακτηριστικά θα είναι λίγα. Το αποτέλεσμα είναι η απόρριψη σημαντικών χαρακτηριστικών και ο αποκλεισμός τους από τη διαδικασία της κατηγοριοποίησης.

Σε κάθε περίπτωση, ο αναλυτής θα πρέπει να έχει επίγνωση του προβλήματος των διαστάσεων και της επιλογής σημαντικών χαρακτηριστικών σε εργασίες κατηγοριοποίησης. Ο αναλυτής θα πρέπει πιθανώς να πειραματιστεί με διαφορετικές μεθόδους επιλογής χαρακτηριστικών και να μην επαφίεται άκριτα σε μία μόνο μέθοδο.

9.7 Δένδρα Αποφάσεων

9.7.1 Εισαγωγή στα Δένδρα Αποφάσεων

Τα Δένδρα Αποφάσεων είναι μια από τις βασικότερες και πιο δημοφιλείς μεθόδους κατηγοριοποίησης. Βασική λογική της κατασκευής τους είναι η διαδοχική διάσπαση του συνόλου των παρατηρήσεων σε υποσύνολα. Κριτήριο για τη διάσπαση είναι οι τιμές των μεταβλητών. Η διαδικασία των διαδοχικών διασπάσεων αναπαρίσταται με μια ανεστραμμένη δενδρική δομή. Στην κορυφή βρίσκεται ο κόμβος-ρίζα του δένδρου. Σε κατώτερα επίπεδα βρίσκονται επιπλέον κόμβοι, οι οποίοι συνδέονται με ακμές με άλλα στοιχεία του δένδρου. Στο κατώτερο επίπεδο κάθε κλάδου βρίσκονται τα φύλλα του δένδρου. Ο κόμβος - ρίζα έχει μόνο εξερχόμενες ακμές που τον συνδέουν με στοιχεία του κατώτερου επιπέδου. Οι υπόλοιποι κόμβοι έχουν εισερχόμενες ακμές που τους συνδέουν με τους κόμβους του ανώτερου επιπέδου και εξερχόμενες ακμές που τους συνδέουν με στοιχεία του κατώτερου επιπέδου. Τέλος, τα φύλα έχουν μόνο εισερχόμενες ακμές, οι οποίες τα συνδέουν με τους κόμβους του ανώτερου επιπέδου. Κάθε κόμβος αντιπροσωπεύει έναν έλεγχο στα δεδομένα και αντίστοιχη διάσπαση τους σε δύο ή περισσότερα υποσύνολα, ανάλογα με το αποτέλεσμα του ελέγχου. Η συνηθέστερη εκδοχή είναι ο έλεγχος να περιλαμβάνει μία μόνο μεταβλητή, έχουν προταθεί ωστόσο αλγόριθμοι όπου σε έναν κόμβο ελέγχονται περισσότερες μεταβλητές. Κάθε ακμή αντιπροσωπεύει ένα αποτέλεσμα του ελέγχου και το αντίστοιχο υποσύνολο των δεδομένων. Τέλος, κάθε φύλο αντιπροσωπεύει μια απόφαση κατηγοριοποίησης.



Σχήμα 9.5 *Δένδρο Αποφάσεων*

Στο Σχήμα 9.5 απεικονίζεται ένα Δένδρο Αποφάσεων για την έγκριση τραπεζικών δανείων. Ο κόμβος-ρίζα αναφέρεται στο σύνολο των δεδομένων. Στο επίπεδο αυτό οι υποψήφιοι δανειολήπτες χωρίζονται σε τρία υποσύνολα ανάλογα με την ηλικία τους. Στο πρώτο υποσύνολο ανήκουν όσοι έχουν ηλικία μικρότερη των 30 ετών, στο δεύτερο όσοι είναι μεταξύ 30 και 50, ενώ στο τρίτο υποσύνολο ανήκουν όσοι έχουν ηλικία μεγαλύτερη των 50 ετών. Τα τρία υποσύνολα συμβολίζονται με αντίστοιχους κλάδους. Ο πρώτος κλάδος, ο οποίος αντιστοιχεί σε όσους είναι λιγότερο από 30, καταλήγει σε ένα φύλο, δηλαδή σε μια απόφαση κατηγοριοποίησης. Η απόφαση είναι θετική, και αυτό σημαίνει ότι γι' αυτήν την κατηγορία τα δάνεια εγκρίνονται χωρίς περαιτέρω ελέγχους. Ο δεύτερος κλάδος αντιστοιχεί σε όσους είναι μεταξύ 30 και 50 και καταλήγει σε έναν εσωτερικό κόμβο. Στον κόμβο αυτό γίνεται ένας δεύτερος έλεγχος που αφορά το εισόδημα. Αν το εισόδημα είναι μεγαλύτερο των 30.000 τότε το δάνειο εγκρίνεται, διαφορετικά η αίτηση απορρίπτεται. Με τον ίδιο τρόπο, οι υποψήφιοι δανειολήπτες που είναι περισσότερο από 50 χρονών ελέγχονται ως προς το ύψος του δανείου.

Το μοντέλο κατασκευάζεται από έναν αλγόριθμο με επεξεργασία ενός συνόλου δεδομένων εκπαίδευσης. Το μοντέλο, αφού κατασκευαστεί, μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση νέων παρατηρήσεων. Για κάθε νέα παρατήρηση πραγματοποιούνται έλεγχοι τιμών των μεταβλητών της, σύμφωνα με τους κόμβους του δένδρου, και ακολουθείται μια διαδρομή από τη ρίζα μέχρι κάποιο φύλο, όπου λαμβάνεται και η απόφαση κατηγοριοποίησης. Στο παράδειγμα του Σχήματος 9.5, ένας υποψήφιος δανειολήπτης θα ελεγχθεί πρώτα ως προς την ηλικία του. Εάν η ηλικία του είναι από 30 έως 50 χρονών, θα ελεγχθεί το εισόδημα του. Αν το εισόδημα του είναι μεγαλύτερο από 30.000 ευρώ το δάνειο θα εγκριθεί.

9.7.2 Δένδρα Αποφάσεων ID3

Έχουν προταθεί διάφοροι αλγόριθμοι για τη δημιουργία Δένδρων Αποφάσεων. Ένας από τους πιο διαδεδομένους είναι ο ID3, καθώς και οι μετεξελίξεις του, ο C4.5 και η εμπορική του εκδοχή C5.0. Ο ID3 προτάθηκε από τον Quinlan (1986) και υλοποιεί μια καθοδική (top-down) στρατηγική διαίρεσης. Ένα χαρακτηριστικό του είναι ότι απαιτεί την ύπαρξη μόνο ονομαστικών πεδίων. Η βασική ακολουθία βημάτων του ID3 είναι η εξής:

- Δημιουργείται ένας αρχικός κόμβος που αντιπροσωπεύει ολόκληρο το δείγμα.
- Εάν όλες οι παρατηρήσεις του δείγματος ανήκουν στην ίδια κλάση, τότε ο κόμβος μετατρέπεται σε φύλο.
- Διαφορετικά επιλέγεται το γνώρισμα που βέλτιστα διαχωρίζει τις παρατηρήσεις του δείγματος ανάλογα με την κλάση στην οποία ανήκουν.
- Δημιουργούνται κλάδοι που διαχωρίζουν τις παρατηρήσεις του δείγματος. Για κάθε δυνατή τιμή του γνωρίσματος δημιουργείται ένας κλάδος.
- Η διαδικασία επαναλαμβάνεται για κάθε ένα από τα υποσύνολα του δείγματος που δημιουργήθηκαν από τους κλάδους του προηγούμενου βήματος. Η επανάληψη τερματίζεται όταν ικανοποιηθεί του-

λάχιστον μία από τις επόμενες συνθήκες εξόδου:

- Όλες οι παρατηρήσεις ενός κόμβου ανήκουν στην ίδια κλάση.
- Δεν υπάρχουν άλλα γνωρίσματα για τον διαχωρισμό του δείγματος. Σε αυτήν την περίπτωση ο κόμβος μετατρέπεται σε φύλο. Η απόφαση κατηγοριοποίησης στο φύλο είναι η κλάση που πλειοψηφεί στο συγκεκριμένο υποσύνολο παρατηρήσεων.
- Δεν υπάρχουν παρατηρήσεις που να ανήκουν στο υποσύνολο του δείγματος που ορίζει ο κλάδος.

Το βασικότερο πρόβλημα στα Δένδρα Αποφάσεων είναι ο καθορισμός του κριτηρίου, βάση του οποίου θα γίνει ο διαχωρισμός των παρατηρήσεων. Έχουν προταθεί αλγόριθμοι οι οποίοι πραγματοποιούν ελέγχους σε συνδυασμούς μεταβλητών (multivariate). Όμως στους περισσότερους αλγορίθμους, συμπεριλαμβανομένου του ID3, ο έλεγχος πραγματοποιείται σε μία μόνο μεταβλητή (univariate). Διάφοροι ερευνητές έχουν επινοήσει και προτείνει πλήθος μονομεταβλητών κριτηρίων. Σε ένα δεδομένο σημείο του δένδρου, το ερώτημα που τίθεται είναι ποιο γνώρισμα πρέπει να χρησιμοποιηθεί για τον διαχωρισμό των παρατηρήσεων. Στον ID3 το κριτήριο που χρησιμοποιείται ονομάζεται Κέρδος Πληροφορίας (ΚΠ). Για κάθε διαθέσιμο γνώρισμα υπολογίζεται το Κέρδος Πληροφορίας και επιλέγεται το γνώρισμα με τη μεγαλύτερη τιμή ΚΠ.

Το ***Κέρδος Πληροφορίας(S,A)*** *(Information Gain(S,A))* εκφράζει τη μείωση της εντροπίας που θα προκύψει, εάν ένα σύνολο παρατηρήσεων *S* διαχωριστεί σε υποσύνολα με βάση τις τιμές του γνωρίσματος *A*. Η εντροπία μετρά την ανομοιογένεια του συνόλου *S,* ανάλογα με τη διασπορά των παρατηρήσεων ως προς την κλάση στην οποία ανήκουν. Ας θεωρήσουμε ένα σύνολο *S* το οποίο περιέχει s παρατηρήσεις. Εάν η κλάση είναι δυαδική, εάν δηλαδή υπάρχουν δύο δυνατές τιμές για το γνώρισμα της κλάσης, τότε οι παρατηρήσεις της μίας τιμής κλάσης μπορούν να χαρακτηριστούν θετικές, ενώ οι υπόλοιπες μπορούν να χαρακτηριστούν αρνητικές. Το πλήθος των θετικών παρατηρήσεων είναι s_p και το πλήθος των αρνητικών παρατηρήσεων είναι s_n. Η εντροπία του συνόλου *S* ορίζεται από την Εξίσωση 9.1

$$E(S) = -p_p * log_2(p_p) - p_n * log_2(p_n) \tag{9.1}$$

όπου p_p είναι το ποσοστό των θετικών παρατηρήσεων ($p_p=s_p/s$) και p_n είναι το ποσοστό των αρνητικών παρατηρήσεων ($p_n=s_n/s$).

Εάν το γνώρισμα της κλάσης μπορεί να πάρει c διαφορετικές τιμές και το πλήθος των παρατηρήσεων με τιμή κλάσης *i* είναι s_i, τότε η εντροπία του *S* ορίζεται από την Εξίσωση 9.2

$$E(S) = -\sum_{i=1}^{c} p_i * log_2(p_i) \tag{9.2}$$

όπου p_i είναι το ποσοστό των παρατηρήσεων που ανήκουν στην κλάση *i* ($p_i=s_i/s$)

Ας θεωρήσουμε ότι το γνώρισμα *A* μπορεί να πάρει u δυνατές διακριτές τιμές ($a_1,a_2,\ldots,a_u$). Το σύνολο *S* μπορεί να χωριστεί στα υποσύνολα ($S_1,S_2,\ldots,S_u$). Το S_1 αποτελείται από τις παρατηρήσεις οι οποίες έχουν τιμή a_1 στο γνώρισμα *A*. Αντιστοίχως, τα υπόλοιπα υποσύνολα S_j απαρτίζονται από τις παρατηρήσεις που έχουν την εκάστοτε τιμή a_j στο γνώρισμα *A*. Εάν επιλεχθεί ως μεταβλητή διαχωρισμού το γνώρισμα *A*, τότε η εντροπία του διαχωρισμού του συνόλου *S* σε υποσύνολα ανάλογα με τις τιμές του *A* δίνεται από την Εξίσωση 9.3

$$E(S,A) = \sum_{j=1}^{u} \frac{s_j}{s} * E(S_j) \tag{9.3}$$

όπου *u* το πλήθος των δυνατών τιμών του γνωρίσματος *A*, S_j το υποσύνολο των παρατηρήσεων οι οποίες έχουν την τιμή a_j στο γνώρισμα *A*, s_j το πλήθος των μελών του S_j, *s* είναι το πλήθος των μελών του *S* και $E(S_j)$ είναι η εντροπία του S_j, η οποία υπολογίζεται σύμφωνα με την Εξίσωση 9.2 και με το S_j στη θέση του *S*. Ου-

σιαστικά η εντροπία που προκύπτει από τον διαχωρισμό του S ισούται με το άθροισμα των εντροπιών των S_j πολλαπλασιασμένες με έναν συντελεστή βαρύτητας, ο οποίος σχετίζεται με το πλήθος των μελών τους. Όσο μικρότερη είναι η εντροπία τόσο αυξάνει ο βαθμός ομοιογένειας των υποσυνόλων.

Το Κέρδος Πληροφορίας είναι η μείωση της εντροπίας, η οποία προκύπτει από τον διαχωρισμό και ορίζεται από την Εξίσωση 9.4

$$IG(S,A) = E(S) - E(S,A)$$

(9.4)

Ο ID3 υπολογίζει για κάθε γνώρισμα το Κέρδος Πληροφορίας. Το γνώρισμα με το μεγαλύτερο Κέρδος Πληροφορίας επιλέγεται και ο διαχωρισμός των παρατηρήσεων γίνεται με βάση τις τιμές αυτού του γνωρίσματος. Με τον τρόπο αυτόν μεταβαίνουμε σε υποσύνολα μεγαλύτερης ομοιογένειας.

9.7.3 Δένδρα Αποφάσεων C4.5

Ο αλγόριθμος C4.5 αποτελεί επέκταση του ID3 και προτάθηκε από τον ίδιο ερευνητή (Quinlan, 1993). Μια από τις βασικές βελτιώσεις αφορά το κριτήριο διαχωρισμού. Σύμφωνα με τον Quinlan το Κέρδος Πληροφορίας τείνει να ευνοεί γνωρίσματα με μεγάλο πλήθος τιμών. Τα γνωρίσματα αυτά οδηγούν σε μεγάλο αριθμό μικρών και πολύ ομοιογενών υποσυνόλων. Σε πολλές περιπτώσεις όμως, τα γνωρίσματα αυτά δεν περιέχουν ουσιαστική πληροφορία. Αν για παράδειγμα τα δεδομένα περιέχουν πεδίο για κάποιον κωδικό, όπως ο αριθμός ταυτότητας, τότε το πεδίο αυτό θα έχει μεγάλο κέρδος πληροφορίας και θα επιλεγεί. Ωστόσο, δεν περιέχει πληροφορία χρήσιμη για την κατηγοριοποίηση. Για την αντιμετώπιση αυτού του προβλήματος, στον C4.5 χρησιμοποιείται το κριτήριο Λόγος Κέρδους (Gain Ratio), το οποίο ορίζεται με την Εξίσωση 9.5.

$$GainRatio(S,A) = \frac{Information\ Gain\ (S,A)}{Entropy(S,A)}$$

(9.5)

Ο Λόγος Κέρδους κανονικοποιεί το κέρδος πληροφορίας ως προς την εντροπία. Μελέτες έχουν δείξει ότι ο Λόγος Κέρδους βελτιώνει την ακρίβεια και μειώνει την πολυπλοκότητα των δένδρων.

Μια άλλη σημαντική βελτίωση στον C4.5 είναι ότι, σε αντίθεση με τον ID3, μπορεί και χειρίζεται πεδία αριθμητικών τιμών. Για κάθε αριθμητικό πεδίο, ο αλγόριθμος ταξινομεί τις τιμές του, το πλήθος των οποίων είναι πεπερασμένο, σε αύξουσα σειρά, και ορίζει μια τιμή κατωφλιού. Με τον τρόπο αυτόν οι παρατηρήσεις χωρίζονται σε εκείνες των οποίων η τιμή στο συγκεκριμένο πεδίο είναι μικρότερη ή ίση με την τιμή κατωφλιού και σε εκείνες που η τιμή τους είναι μεγαλύτερη. Ακολούθως, το γνώρισμα αντιμετωπίζεται σαν να έχει διακριτές τιμές, όπου οι δύο διακριτές τιμές είναι οι δύο καθορισμένες περιοχές συνεχών τιμών. Επίσης ο C4.5 μπορεί και χειρίζεται δεδομένα με χαμένες τιμές.

9.7.4 Δένδρα CART

Τα δένδρα τύπου CART (Classification And Regression Trees) προτάθηκαν από τους Breiman, Friedman, Olshen and Stone (1984). Ο πρωτότυπος αλγόριθμος τους είναι ενσωματωμένος σε λογισμικά της Salford Systems. Τα δένδρα CART παρουσιάζουν αρκετά ενδιαφέροντα χαρακτηριστικά. Ένα από τα σημαντικότερα χαρακτηριστικά τους είναι ότι μπορούν να χρησιμοποιηθούν και για κατηγοριοποίηση και για παλινδρόμηση, μπορούν δηλαδή να προβλέψουν και ονομαστικές τιμές κλάσης και τιμές αριθμητικών πεδίων. Τα δένδρα CART είναι δυαδικά και κάθε κόμβος μπορεί να έχει μόνο δύο κλάδους. Για τον διαχωρισμό των παρατηρήσεων χρησιμοποιείται το κριτήριο Twoing. Στα πλεονεκτήματα τους περιλαμβάνονται οι υψηλές επιδόσεις σε ταχύτητα και ακρίβεια, καθώς και η ικανότητα τους να χειρίζονται δεδομένα με χαμένες τιμές. Επιπλέον, τα δένδρα CART είναι ικανά να εκτελέσουν κατηγοριοποίηση που λαμβάνει υπόψη το διαφορετικό κόστος σφάλματος (cost sensitive classification). Σε αυτήν την περίπτωση ο αλγόριθμος επιδιώκει να μειώσει τις εσφαλμένες προβλέψεις της πιο ακριβής κλάσης. Το αντικείμενο του διαφορετικού κόστους σφάλματος καλύπτεται στο Κεφάλαιο 10.

Πρόσθετοι αλγόριθμοι δημιουργίας δένδρων αποφάσεων έχουν προταθεί από διάφορους ερευνητές. Ορι-

σμένα δένδρα αποφάσεων είναι τα AID (Sonquist, Baker & Morgan, 1971), CHAID (Kass, 1980) και QUEST (Loh & Shih, 1997).

9.7.5 Κλάδεμα

Κατά τη διαδικασία ανάπτυξης του δένδρου, ο αλγόριθμος έρχεται αντιμέτωπος με παρατηρήσεις που περιέχουν εσφαλμένες ή ακραίες τιμές. Η δημιουργία κλάδων που να αντιστοιχούν σε τέτοιου είδους τιμές, καταγράφει ανωμαλίες των δεδομένων και περιορίζει την ικανότητα του δένδρου να προβλέψει την κλάση νέων παρατηρήσεων. Ένα πρόσθετο ζήτημα είναι οι συνθήκες τερματισμού του βρόχου ανάπτυξης του δένδρου. Αν τα κριτήρια είναι πολύ περιοριστικά, θα δημιουργηθούν μικρά και υποπροσαρμοσμένα δένδρα, ενώ αν τα κριτήρια είναι πολύ χαλαρά, θα δημιουργηθούν υπερβολικά μεγάλα και υπερπροσαρμοσμένα δένδρα. Για την αντιμετώπιση αυτού του προβλήματος προτάθηκε από τους Breiman et al. (1984) μια τεχνική, η οποία προβλέπει την εφαρμογή χαλαρών κριτηρίων, τη δημιουργία υπερπροσαρμοσμένων δένδρων και την ακόλουθη απομάκρυνση των περιττών κλάδων. Η διαδικασία διαγραφής των περιττών κλάδων καλείται **κλάδεμα** (pruning).

Έχουν προταθεί πολλές τεχνικές κλαδέματος. Μια από τις πιο γνωστές είναι η τεχνική Cost-Complexity Pruning. Για κάθε κόμβο του δένδρου υπολογίζεται το αναμενόμενο σφάλμα που θα προκύψει εάν κλαδευτεί το υποδένδρο του κόμβου, καθώς και το αναμενόμενο σφάλμα εάν τα υποδένδρο δεν κλαδευτεί. Εάν το κλάδεμα οδηγήσει σε μεγαλύτερο αναμενόμενο σφάλμα, τότε το υποδένδρο διατηρείται, διαφορετικά κλαδεύεται. Προβλέπεται η δημιουργία πολλών εναλλακτικών δένδρων με κλάδεμα που αυξάνεται προοδευτικά. Τα εναλλακτικά αυτά δένδρα δοκιμάζονται έναντι ενός συνόλου άγνωστων παρατηρήσεων και επιλέγεται το δένδρο το οποίο ελαχιστοποιεί τον αναμενόμενο ρυθμό σφάλματος.

Η προσέγγιση της ανάπτυξης ενός λεπτομερούς δένδρου και του ακόλουθου κλαδέματος του επισύρει πρόσθετο υπολογιστικό κόστος. Εναλλακτικά, μπορεί κατά τη διάρκεια της δημιουργίας του δένδρου να διακοπεί η ανάπτυξη περιττών κλάδων με τον καθορισμό κατάλληλων συνθηκών εξόδου. Η προσέγγιση αυτή έχει το πλεονέκτημα ότι αποφεύγει την άσκοπη εργασία δημιουργίας άχρηστων κλάδων. Ωστόσο, η τεχνική του κλαδέματος αποδίδει καλύτερα αποτελέσματα, ειδικά στην περίπτωση όπου ένας κλάδος δεν είναι σημαντικός, αλλά δύο συνεχόμενοι κλάδοι επιτυγχάνουν ισχυρή κατηγοριοποίηση. Άλλες τεχνικές κλαδέματος είναι η Minimum Description Length Pruning (Quinlan & Rivest, 1989), η Minimum Error Pruning (Olaru & Wehenkel, 2003) και η Pessimistic Pruning (Quinlan, 1993).

9.7.6 Δημιουργία κανόνων από Δένδρα Αποφάσεων

Στα Δένδρα Αποφάσεων η ανακαλυφθείσα γνώση αναπαρίσταται με τέτοιον τρόπο, ώστε είναι εύκολη η εξαγωγή επαγωγικών κανόνων της μορφής IF-THEN. Για κάθε φύλο δημιουργείται ένας κανόνας, ο οποίος περιλαμβάνει τις λογικές συνθήκες όλων των ελέγχων από τη ρίζα έως το φύλο. Οι λογικές αυτές συνθήκες συνδέονται με τον λογικό τελεστή AND. Ωστόσο, μια τέτοιου τύπου εξαγωγή κανόνων, οδηγεί συχνά σε κανόνες περιττά περίπλοκους. Για την απλοποίηση των κανόνων μπορεί να υιοθετηθεί μια προσέγγιση κλαδέματος, η οποία συνίσταται στην απομάκρυνση των μη σημαντικών συνθηκών. Ο καθορισμός των μη σημαντικών συνθηκών επιτυγχάνεται με τη σύγκριση του ρυθμού σφάλματος του απλουστευμένου κανόνα με τον ρυθμό σφάλματος του πλήρους κανόνα.

9.7.7 Πλεονεκτήματα και Μειονεκτήματα των Δένδρων Αποφάσεων

Τα Δένδρα Αποφάσεων προσφέρουν πολλά και σημαντικά **πλεονεκτήματα**:

- Σε αντίθεση με άλλες μεθόδους δεν κάνουν αυθαίρετες υποθέσεις για τη γραμμικότητα της σχέσης μεταξύ των μεταβλητών εισόδου και εξόδου ή για την ανεξαρτησία των μεταβλητών εισόδου.
- Τα Δένδρα Αποφάσεων είναι μη παραμετρικά. Η δημιουργία του δένδρου δεν καθορίζεται από τον καθορισμό πολλών και σύνθετων παραμέτρων.
- Η αναπαράσταση της γνώσης γίνεται με κατανοητό τρόπο και είναι εύκολη η εξαγωγή κατανοητών κανόνων.
- Δέχονται ως μεταβλητές εισόδου και ονομαστικά γνωρίσματα και γνωρίσματα με αριθμητικές τιμές.
- Μπορούν να χειριστούν δεδομένα με χαμένες τιμές.
- Διαθέτουν ένα εξαιρετικά γρήγορο αλγόριθμο εκπαίδευσης.

Μειονεκτήματα των Δένδρων Απόφασης είναι τα εξής:

- Αρκετοί αλγόριθμοι, όπως ο ID3 και ο C4.5, λειτουργούν μόνο για διακριτές και όχι για συνεχόμενες τιμές κλάσης.
- Είναι ιδιαίτερα ευπαθή σε μεταβολές του δείγματος εκπαίδευσης. Οριακές μεταβολές του δείγματος εκπαίδευσης μπορεί να οδηγήσουν στη δημιουργία σημαντικά διαφορετικών δένδρων.
- Αρκετοί αλγόριθμοι Δένδρων Απόφασης, όπως ο ID3 ή ο C4.5, απαιτούν την εγκατάσταση ολόκληρου του δείγματος εκπαίδευσης στην κύρια μνήμη του υπολογιστή, κάτι που προκαλεί προβλήματα στον χειρισμό εξαιρετικά μεγάλων δειγμάτων.

9.8 Κατηγοριοποίηση με Νευρωνικά Δίκτυα

Τα **Νευρωνικά Δίκτυα** (Neural Networks) αποτελούν ένα από τα σημαντικότερα επιτεύγματα της Τεχνητής Νοημοσύνης. Εμπνευσμένα από το βιολογικό νευρικό σύστημα, και ειδικότερα από τον ανθρώπινο εγκέφαλο, διαθέτουν αξιοσημείωτα χαρακτηριστικά, όπως τη δυνατότητα τους να αναπαριστούν σύνθετες εξαρτήσεις ή την ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Χάρη στη στιβαρή θεωρητική τους θεμελίωση και στις ιδιαίτερες δυνατότητες τους έχουν καταστεί ιδιαίτερα δημοφιλή και έχουν εφαρμοστεί σε πολλούς τομείς, όπως η ιατρική, η οικονομία, η άμυνα κλπ. Τα Νευρωνικά Δίκτυα είναι μια τεχνική ισχυρά καθοδηγούμενη από τα δεδομένα. Αυτό σημαίνει ότι δεν επιβάλλουν αυθαίρετες υποθέσεις και ότι τα μοντέλα τους πηγάζουν από την επεξεργασία των δεδομένων. Έχουν προταθεί τύποι Νευρωνικών Δικτύων κατάλληλοι για επιβλεπόμενη, αλλά και για μη επιβλεπόμενη μάθηση. Στο παρόν κεφάλαιο καλύπτονται θέματα Νευρωνικών Δικτύων επιβλεπόμενης μάθησης. Αναφορά στα Νευρωνικά Δίκτυα μη επιβλεπόμενης μάθησης, και ειδικά στους Αυτοοργανούμενους Χάρτες, γίνεται στο Κεφάλαιο 11.

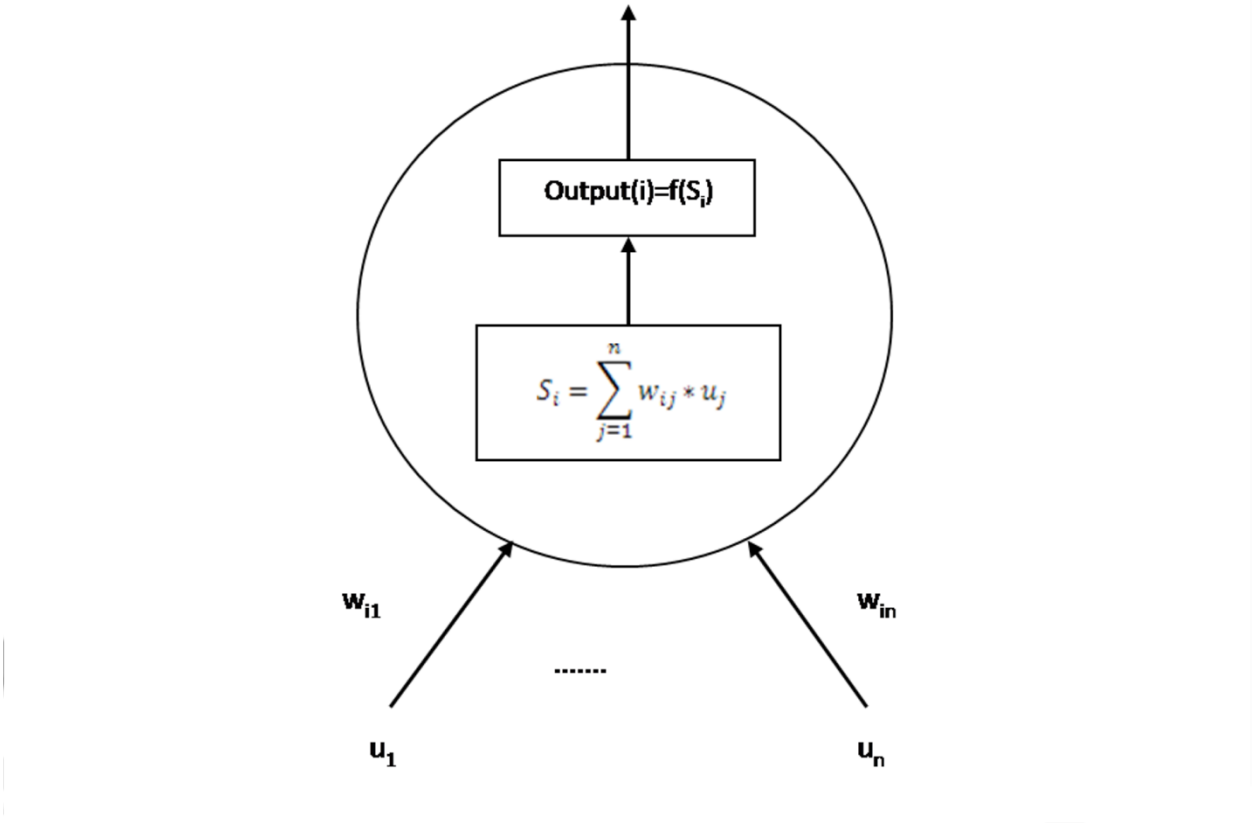
9.8.1 Νευρώνες και Συνδέσεις

Βασική δομική μονάδα των Νευρωνικών Δικτύων είναι οι νευρώνες. Οι νευρώνες ονομάζονται επίσης κόμβοι ή κελιά. Ένας **νευρώνας** είναι μια στοιχειώδης υπολογιστική μονάδα, η οποία δέχεται τιμές εισόδου και υπολογίζει μια τιμή εξόδου. Οι νευρώνες συνδέονται μεταξύ τους με κατευθυνόμενα βέλη ή συνδέσεις. Μέσω των συνδέσεων ένας νευρώνας δέχεται τιμές εισόδου από άλλους νευρώνες. Επίσης, μέσω των συνδέσεων μεταβιβάζει την τιμή εξόδου του σε άλλους νευρώνες. Κάθε σύνδεση συνοδεύεται από μία αριθμητική τιμή που ονομάζεται **βάρος** (weight) w. Το βάρος επηρεάζει την επίδραση μεταξύ των συνδεδεμένων νευρώνων. Εάν u_j είναι η τιμή εξόδου του νευρώνα j, και η τιμή αυτή μεταβιβάζεται στον νευρώνα i, τότε το u_j πολλαπλασιάζεται με το βάρος της σύνδεσης των δύο νευρώνων w_{ij}.

Η επεξεργασία που διενεργεί ένας νευρώνας i ολοκληρώνεται σε δύο στάδια. Στο πρώτο στάδιο αθροίζονται οι τιμές εισόδου. Οι τιμές εισόδου ισούνται με τις τιμές εξόδου των συνδεδεμένων νευρώνων, πολλαπλασιασμένες με τα βάρη των αντίστοιχων συνδέσεων. Για έναν νευρώνα i ο οποίος δέχεται τιμές εισόδου u_j από n νευρώνες, το συνολικό σήμα εισόδου S_i υπολογίζεται σύμφωνα με την Εξίσωση 9.6.

$$S_i = \sum_{j=1}^{n} w_{ij} * u_j \tag{9.6}$$

Στο δεύτερο στάδιο, μετασχηματίζεται το άθροισμα των τιμών εισόδου, με χρήση μιας συνάρτησης γνωστής ως **συνάρτηση ενεργοποίησης** (activation function) ή **συνάρτηση μετασχηματισμού**. Η τιμή που υπολογίζεται είναι η τιμή εξόδου του νευρώνα. Τα παραπάνω απεικονίζονται στο Σχήμα 9.6.



Σχήμα 9.6 *Ενεργοποίηση Νευρώνα*

Διάφορες συναρτήσεις μπορούν να χρησιμοποιηθούν ως συναρτήσεις ενεργοποίησης. Τέτοιες συναρτήσεις είναι η συνάρτηση ημιτόνου, η συνάρτηση συνημίτονου, η συνάρτηση υπερβολικής εφαπτομένης κλπ. Συνήθως όμως χρησιμοποιείται η Σιγμοειδής συνάρτηση, επειδή είναι απλή και μη γραμμική και επειδή μοιάζει με τη συμπεριφορά των πραγματικών νευρώνων. Η Σιγμοειδής συνάρτηση ορίζεται από την Εξίσωση 9.7.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{9.7}$$

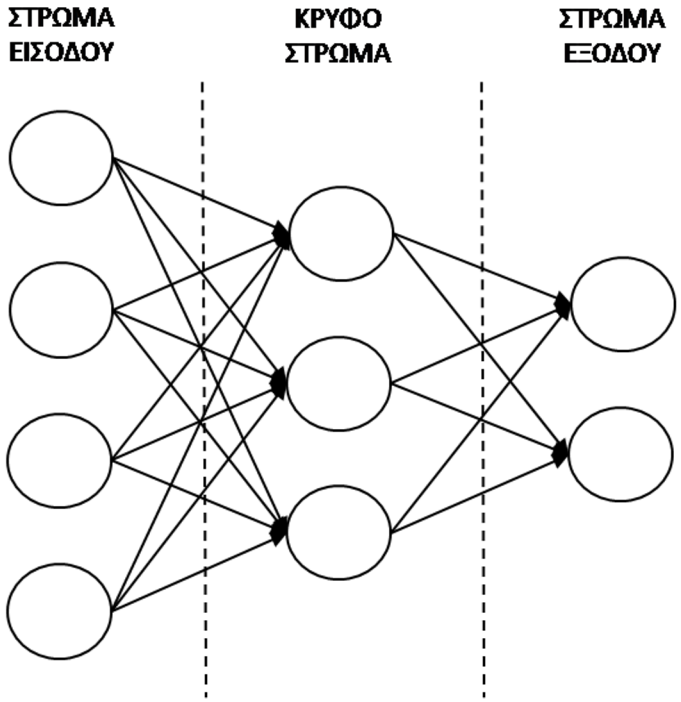
Οι συνδέσεις μπορεί να είναι μονόδρομες ή αμφίδρομες. Όταν ένα δίκτυο δεν περιέχει αμφίδρομες συνδέσεις χαρακτηρίζεται δίκτυο **απλής προώθησης** (feed forward), ενώ όταν περιέχει και αμφίδρομες συνδέσεις χαρακτηρίζεται **αναδρομικό** (recurrent). Τα δίκτυα απλής προώθησης και πολλών επιπέδων είναι ιδιαίτερα αποτελεσματικά για τη μοντελοποίηση σύνθετων, μη γραμμικών σχέσεων ανάμεσα σε μια εξαρτημένη μεταβλητή και πολλές ανεξάρτητες μεταβλητές. Για τον λόγο αυτό, χρησιμοποιούνται συχνά σε προβλήματα κατηγοριοποίησης. Τα νευρωνικά δίκτυα τύπου **Multilayer Perceptron** (**MLP**) είναι δίκτυα απλής προώθησης και ιδιαίτερα δημοφιλή. Σύμφωνα με τους Wong, Bodnovich and Selvi (1997), στο 95% των περιπτώσεων επιχειρηματικών εφαρμογών χρησιμοποιούνται δίκτυα αυτού του τύπου.

9.8.2 Δομή MLP

Σε ένα δίκτυο MLP οι νευρώνες είναι οργανωμένοι σε **στρώματα** (layers) ή επίπεδα. Παράδειγμα Νευρωνικού Δικτύου με επίπεδα παρουσιάζεται στο Σχήμα 9.7. Το πρώτο στρώμα ονομάζεται **στρώμα εισόδου** (input layer). Υπάρχει ένας νευρώνας εισόδου για κάθε ανεξάρτητη μεταβλητή. Χαρακτηριστικό των νευρώνων εισόδου είναι ότι δεν μετασχηματίζουν την τιμή. Απλά δέχονται την τιμή της ανεξάρτητης μεταβλητής και την μεταβιβάζουν στους επόμενους νευρώνες. Το δεύτερο στρώμα ονομάζεται **κρυφό στρώμα** (hidden layer). Οι νευρώνες του κρυφού στρώματος δέχονται τις τιμές των νευρώνων εισόδου πολλαπλασιασμένες με τα βάρη των συνδέσεων, τις αθροίζουν και μετασχηματίζουν το άθροισμα σύμφωνα με τη συνάρτηση μετα-

σχηματισμού. Οι κρυφοί νευρώνες είναι καθοριστικής σημασίας για την καταγραφή των σύνθετων σχέσεων των δεδομένων. Οι τιμές εξόδου των κρυφών νευρώνων, πολλαπλασιασμένες με τα βάρη των συνδέσεων, διαβιβάζονται στους νευρώνες του **στρώματος εξόδου** (output layer). Στους νευρώνες εξόδου υπολογίζεται η τελική πρόβλεψη του δικτύου. Είναι δυνατόν να υπάρχουν περισσότερα κρυφά στρώματα, συνήθως όμως χρησιμοποιείται μόνο ένα κρυφό στρώμα. Για διχότομα προβλήματα κατηγοριοποίησης ένας νευρώνας εξόδου είναι αρκετός. Για προβλήματα με περισσότερες τιμές κλάσης χρειάζεται ένας νευρώνας εξόδου για κάθε δυνατή τιμή της κλάσης. Είναι δυνατόν να υπάρχουν διαφορετικές συναρτήσεις μετασχηματισμού σε νευρώνες διαφορετικών επιπέδων. Οι νευρώνες είναι **πλήρως συνδεδεμένοι** (fully connected), και κάθε νευρώνας διαβιβάζει τιμές σε όλους τους νευρώνες του επόμενου στρώματος και μόνο σε αυτούς. Επίσης, μπορεί να υπάρχει ένας κατά σύμβαση νευρώνας πόλωσης *0*, ο οποίος είναι συνδεδεμένος με όλους τους υπόλοιπους νευρώνες και του οποίου η έξοδος u_0 είναι σταθερά *+1*. Τα βάρη w_{i0} ονομάζονται *πόλωση* (bias).

Ο όρος **αρχιτεκτονική του δικτύου** (network architecture) ή **τοπολογία του δικτύου** (network topology) αναφέρεται στη δομή του δικτύου και περιλαμβάνει ζητήματα όπως το πλήθος των κρυφών στρωμάτων και το πλήθος των νευρώνων σε κάθε στρώμα. Ο χρήστης του νευρωνικού δικτύου οφείλει να προκαθορίσει την αρχιτεκτονική του δικτύου πριν από την εκπαίδευση του. Επίσης, προκαθορίζει τη συνάρτηση μετασχηματισμού των νευρώνων. Η αρχιτεκτονική του δικτύου είναι σημαντική και επηρεάζει την αποτελεσματικότητα του.



Σχήμα 9.7 *Δίκτυο τριών επιπέδων*

9.8.3 Εκπαίδευση Δικτύου

Η εκπαίδευση ενός δικτύου συνίσταται στη ρύθμιση των βαρών των συνδέσεων. Για την εκπαίδευση ενός Νευρωνικού Δικτύου τυπικά απαιτείται ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Το σύνολο εκπαίδευσης χρησιμοποιείται για τον καθορισμό των βαρών των συνδέσεων. Το σύνολο ελέγχου χρησιμοποιείται για την εκτίμηση της επίδοσης του μοντέλου. Στην επιβλεπόμενη μάθηση, κάθε παρατήρηση του συνόλου εκπαίδευσης ή ελέγχου περιλαμβάνει και την τιμή της κλάσης της συγκεκριμένης παρατήρησης.

Υπάρχουν διάφοροι αλγόριθμοι για την εκπαίδευση του δικτύου. Ο πιο σημαντικός και διαδεδομένος αλγόριθμος, λόγω της καταλληλότητας του για επιβλεπόμενη μάθηση και για καθήκοντα κατηγοριοποίησης και πρόβλεψης, είναι ο αλγόριθμος **Αντίστροφης Μετάδοσης Σφάλματος** (**Backpropagation**). Συνεισφορά στη διατύπωση του Backpropagation είχαν οι Parker (1985) και Rumelhart, Hinton and Williams (1896).

Για να εκπαιδεύσει το δίκτυο, ο αλγόριθμος Backpropagation εφαρμόζει μια επαναλαμβανόμενη διαδικα-

σία, όπου μια παρατήρηση εκπαίδευσης εφαρμόζεται στο δίκτυο και ακολούθως υπολογίζεται στην έξοδο μια πρόβλεψη για την κλάση της παρατήρησης. Η πρόβλεψη συγκρίνεται με την πραγματική κλάση. Στη συνέχεια τροποποιούνται τα βάρη των συνδέσεων, έτσι ώστε να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα μεταξύ της πρόβλεψης και της πραγματικής κλάσης. Η τροποποίηση των βαρών γίνεται αρχίζοντας από το επίπεδα εξόδου και συνεχίζοντας προς τα προηγούμενα επίπεδα. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να ικανοποιηθούν οι συνθήκες τερματισμού. Πιο αναλυτικά, τα βήματα του αλγόριθμου Backpropagation είναι τα εξής.

- **Εκχώρηση αρχικών τιμών στα βάρη**. Τα βάρη των συνδέσεων αρχικοποιούνται με τυχαίες τιμές. Οι τιμές αρχικοποίησης είναι μικρές και μπορεί να κυμαίνονται από -1,0 έως 1,0.
- **Διάδοση της εισόδου**. Μια παρατήρηση εκπαίδευσης εφαρμόζεται στην είσοδο του δικτύου. Τα δεδομένα εισόδου διαδίδονται στους νευρώνες των επόμενων στρωμάτων και μετασχηματίζονται χρησιμοποιώντας τα βάρη των συνδέσεων και τις συναρτήσεις μετασχηματισμού. Στο επίπεδο εξόδου υπολογίζεται μια πρόβλεψη.
- **Διάδοση του σφάλματος προς τα πίσω**. Η τιμή κλάσης που υπολόγισε το δίκτυο συγκρίνεται με την πραγματική τιμή. Αν η πρόβλεψη είναι εσφαλμένη, το σφάλμα διαδίδεται προς τα πίσω και επαναϋπολογίζονται τα βάρη των συνδέσεων. Για έναν νευρώνα i που βρίσκεται στο επίπεδο εξόδου το σφάλμα υπολογίζεται ως:

$$Err_i = f(i) * \big(1 - f(i)\big) * \big(T_i - f(i)\big)$$

(9.8)

όπου T_i είναι η πραγματική τιμή που θα έπρεπε να υπολογιστεί με βάση την πραγματική κλάση της συγκεκριμένης παρατήρησης εκπαίδευσης.

Για έναν νευρώνα *i* που βρίσκεται σε κρυφό επίπεδο, το σφάλμα υπολογίζεται σύμφωνα με τη Σχέση 9.9

$$Err_i = f(i) * \big(1 - f(i)\big) * \sum_j Err_j * w_{ij}$$

(9.9)

όπου *j* είναι οι νευρώνες του επόμενου επιπέδου με τους οποίους είναι συνδεδεμένος ο *i*, w_{ij} είναι τα αντίστοιχα βάρη συνδέσεων και Err_j είναι τα σφάλματα των νευρώνων *j*. Τα βάρη των συνδέσεων τροποποιούνται σύμφωνα με την Εξίσωση 9.10

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

(9.10)

όπου το Δw_{ij} ορίζεται από την Εξίσωση 9.11.

$$\Delta w_{ij} = (I) * Err_j * f(i)$$

(9.11)

Η σταθερά *I* στην Εξίσωση 9.12 συμβολίζει τον **ρυθμό εκπαίδευσης** (learning rate) και τυπικά παίρνει τιμές από 0,0 έως και 1,0. Ο ρυθμός εκπαίδευσης επιτρέπει να ρυθμίσουμε τον βαθμό μεταβολής των βαρών σε κάθε επανάληψη. Κατ' επέκταση, ο ρυθμός εκπαίδευσης επηρεάζει την ταχύτητα εκπαίδευσης του δικτύου. Ένας άλλος τρόπος ενίσχυσης του ρυθμού μεταβολής των βαρών είναι με τη χρήση της ορμής (momentum) Η χρήση του momentum συνίσταται στην αύξηση της μεταβολής των βαρών με την πρόσθεση ενός ποσοστού της προηγούμενης μεταβολής. Η προσθήκη του νέου προσθετέου τείνει να διατηρήσει την κατεύθυνση μεταβολής των βαρών. Ο ορμή βοηθά το μοντέλο να μην παγιδευτεί σε τοπικά ελάχιστα.

- **Επανάληψη** της διάδοσης εισόδου με τροφοδοσία μιας νέας παρατήρησης στο δίκτυο **μέχρι την ικανοποίηση της συνθήκης εξόδου**. Συνθήκη εξόδου μπορεί να είναι μια από τις παρακάτω:

    - όλα τα Δw_{ij} έχουν τιμή κάτω από ένα όριο, ή
    - το ποσοστό των εσφαλμένων προβλέψεων είναι κάτω από ένα όριο, ή
    - συμπληρώθηκε ο προκαθορισμένος αριθμός εποχών.

Σύμφωνα με τον αλγόριθμο που περιγράφτηκε προηγουμένως, ο επαναϋπολογισμός των βαρών γίνεται για κάθε παρατήρηση εκπαίδευσης. Σε μια άλλη παραλλαγή του αλγόριθμου οι μεταβολές των βαρών συσσωρεύονται και τα βάρη των συνδέσεων αλλάζουν τιμή όταν ολοκληρωθεί η ανάγνωση όλου του συνόλου εκπαίδευσης. Κάθε επανάληψη του συνόλου εκπαίδευσης καλείται **εποχή**.

9.8.4 Θέματα μοντελοποίησης με νευρωνικά δίκτυα

Η δημιουργία και η εκπαίδευση ενός επιτυχημένου νευρωνικού δικτύου είναι μια απαιτητική και δύσκολη εργασία. Αρχικά πρέπει να καθοριστούν το πλήθος των κρυφών στρωμάτων και το πλήθος των νευρώνων σε κάθε στρώμα. Δυστυχώς, δεν υπάρχουν μαθηματικά θεμελιωμένοι κανόνες γι' αυτά τα ζητήματα. Αξιοποιώντας κυρίως την εμπειρία, έχουν προταθεί ορισμένοι πρακτικοί κανόνες, όπως το πλήθος των κρυφών νευρώνων να είναι το μισό του πλήθους των νευρώνων εισόδου ή να είναι το μισό του αθροίσματος των νευρώνων εισόδου και των δυνατών τιμών της κλάσης. Φυσικά, τέτοιοι πρακτικοί κανόνες δεν έχουν απόλυτη ισχύ και συχνά ο χρήστης είναι υποχρεωμένος να πειραματίζεται με διάφορα μοντέλα, μέχρι να επιλέξει μια αρχιτεκτονική. Επιπλέον, ο χρήστης πρέπει να ρυθμίσει μια σειρά από πρόσθετες παραμέτρους. Ειδικότερα πρέπει να επιλέξει τις συναρτήσεις μετασχηματισμού για κάθε επίπεδο και να ορίσει τιμές για τον ρυθμό εκπαίδευσης, το πλήθος των εποχών και τη ροπή. Οι παράμετροι έχουν επιπτώσεις στην εκπαίδευση του μοντέλου. Μικρός ρυθμός εκπαίδευσης προκαλεί μικρές μεταβολές βαρών και κατ' επέκταση αργή εκπαίδευση του δικτύου. Μεγάλος ρυθμός εκπαίδευσης προκαλεί την ταχεία εκπαίδευση του δικτύου και κίνδυνο υπερπροσαρμογής του μοντέλου. Για τη ρύθμιση όλων αυτών των παραμέτρων δεν υπάρχουν καθορισμένοι κανόνες και οδηγός του χρήστη είναι η εμπειρία.

Άλλα σημαντικά ζητήματα αφορούν τα δεδομένα. Τα νευρωνικά δίκτυα είναι μια μέθοδος ισχυρά καθοδηγούμενη από τα δεδομένα, Για τον λόγο αυτό τα δεδομένα είναι ιδιαίτερα σημαντικά. Στο στάδιο της προεπεξεργασίας πρέπει να έχουν επιλεγεί τα σημαντικά χαρακτηριστικά τα οποία και θα γίνουν οι νευρώνες εισόδου του δικτύου. Τα νευρωνικά δίκτυα λειτουργούν καλύτερα με κανονικοποιημένες τιμές. Ορισμένες υλοποιήσεις επιτρέπουν την αυτοματοποιημένη κανονικοποίηση των τιμών. Στις άλλες περιπτώσεις όμως, ο χρήστης πρέπει να κανονικοποιήσει μόνος του τις τιμές στο στάδιο της προεπεξεργασίας. Το σύνολο δεδομένων του νευρωνικού δικτύου χωρίζεται σε σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set). Το σύνολο εκπαίδευσης χρησιμοποιείται για τη ρύθμιση των βαρών των συνδέσεων. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος δοκιμάζει το μοντέλο με το σύνολο ελέγχου και διακόπτει την εκπαίδευση εάν θεωρήσει ότι το μοντέλο εκπαιδεύτηκε επαρκώς. Ένα τρίτο σύνολο παρατηρήσεων μπορεί να χρησιμοποιηθεί για τον τελικό έλεγχο του μοντέλου αφού ολοκληρωθεί η εκπαίδευση. Επιπλέον, τα Νευρωνικά Δίκτυα είναι αρκετά σύνθετα μοντέλα και μπορούν να ενσωματώσουν σημαντικό όγκο πληροφορίας. Για τους λόγους αυτούς απαιτείται σημαντικός αριθμός παρατηρήσεων για την επιτυχημένη εκπαίδευση του μοντέλου.

9.8.5 Πλεονεκτήματα και μειονεκτήματα των Νευρωνικών Δικτύων

Τα Νευρωνικά Δίκτυα οφείλουν τη μεγάλη δημοφιλία τους στα αδιαμφισβήτητα **πλεονεκτήματα** τους:

- Τα Νευρωνικά Δίκτυα είναι ιδιαιτέρως κατάλληλα αν δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών εισόδου και εξόδου. Η ύπαρξη των κρυφών στρωμάτων επιτρέπει την ικανοποιητική προσέγγιση σύνθετων συναρτήσεων.
- Είναι ιδιαιτέρως ικανά να κατηγοριοποιήσουν αντικείμενα που δεν περιλαμβάνονταν στο σύνολο εκπαίδευσης και επομένως είναι άγνωστα στο δίκτυο.
- Μπορούν να χειριστούν θορυβώδη και ασυνεπή δεδομένα.

Πέρα από τα πλεονεκτήματά τους, τα Νευρωνικά Δίκτυα δεν στερούνται **μειονεκτημάτων**:

- Το σημαντικότερο ίσως μειονέκτημα τους είναι ότι απαιτείται ο εμπειρικός προσδιορισμός πολλών

παραμέτρων όπως η τοπολογία του δικτύου, ο αριθμός των εποχών εκπαίδευσης, ο καθορισμός του ρυθμού εκπαίδευσης. Για όλες αυτές τις παραμέτρους δεν υπάρχει καθιερωμένη δεοντολογία για τον προσδιορισμό τους.

- Άλλο σημαντικό μειονέκτημα είναι η προβληματική ερμηνευσιμότητα. Ο τρόπος λήψης αποφάσεων των νευρωνικών δικτύων είναι ακατανόητος στους ανθρώπους. Ιδιαίτερα στα χρηματοοικονομικά, ο χρήστης επιθυμεί να διασφαλίζει ότι ο τρόπος λήψης αποφάσεων συνάδει ή έστω δεν αντικρούει με καθιερωμένη γνώση. Επίσης, γενικότερα, ο σκοπός της Εξόρυξης Δεδομένων είναι η ανακάλυψη γνώσης, όχι προβλέψεων.
- Τα Νευρωνικά Δίκτυα απαιτούν μεγάλους χρόνους εκπαίδευσης.

9.9 Μπαϋεσιανοί Κατηγοριοποιητές

Τα Μπαϋεσιανά Δίκτυα (Bayesian Networks) είναι ισχυρά εργαλεία για αναπαράσταση σύνθετων σχέσεων μεταξύ μεταβλητών και για εξαγωγή συμπερασμάτων σε συνθήκες αβεβαιότητας. Ανήκουν στην κατηγορία των γραφικών πιθανοτικών μοντέλων, τα οποία αναπαριστούν σχέσεις με μορφή γράφων. Κάθε κόμβος του γράφου συμβολίζει μια στοχαστική μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης ανάμεσα σε δύο μεταβλητές. Τα Μπαϋεσιανά Δίκτυα αρχικά δεν θεωρήθηκαν εργαλεία κατηγοριοποίησης, αργότερα όμως ανακαλύφθηκε ότι οι Αφελείς Μπαϋεσιανοί κατηγοριοποιητές (Naive Bayesian Classifiers), μια απλουστευμένη εκδοχή των Μπαϋεσιανών Δικτύων, έχουν αυξημένες δυνατότητες κατηγοριοποίησης, συγκρίσιμες με αυτές των Νευρωνικών Δικτύων και των Δένδρων Αποφάσεων. Σήμερα τα Μπαϋεσιανά Δίκτυα αποτελούν μια καταξιωμένη μέθοδο Εξόρυξης Δεδομένων, λόγω της στιβαρής θεωρητικής τους θεμελίωσης, της ικανότητας τους να καταγράφουν περίπλοκες σχέσεις αλληλεξάρτησης, του συμβολικού φορμαλισμού τους και της δυνατότητας τους να εφαρμόζονται σε προβλήματα κατηγοριοποίησης (Heckerman, 1997).

Τα Μπαϋεσιανά Δίκτυα έλκουν το θεωρητικό τους υπόβαθρο από τη στατιστική και πιο συγκεκριμένα από το θεώρημα του Bayes, που υπολογίζει την υπό συνθήκη πιθανότητα $P(H|X)$, δηλαδή την πιθανότητα να επαληθευτεί η υπόθεση H με δεδομένο ότι ισχύει το γεγονός X. Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα $P(H|X)$ δίνεται από την Εξίσωση 9.12

$$P(H|X) = \frac{P(H) * P(X|H)}{P(X)} \tag{9.12}$$

όπου $P(H)$ είναι η εκ των προτέρων πιθανότητα να ισχύει η υπόθεση H, $P(X)$ είναι η εκ των προτέρων πιθανότητα να συμβεί το γεγονός X και $P(X|H)$ είναι η πιθανότητα να συμβεί το γεγονός X με δεδομένο ότι ισχύει η υπόθεση H.

9.9.1 Αφελείς Μπαϋεσιανοί Κατηγοριοποιητές

Ο Αφελής Μπαϋεσιανός κατηγοριοποιητής αποτελεί ευθεία εφαρμογή του θεωρήματος Bayes. Υποθέτουμε ότι X είναι μια παρατήρηση του συνόλου δεδομένων και H είναι η υπόθεση ότι παρατήρηση αυτή ανήκει στην κλάση C_i. Πιο συγκεκριμένα, το X θεωρείται ως ένα άνυσμα n τιμών $X=(x_1,x_2,\ldots,x_n)$. Υποθέτουμε ότι υπάρχουν m κλάσεις $C_1, C_2, \ldots, C_m$. Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα να ανήκει η παρατήρηση X στην κλάση C_i υπολογίζεται από την Εξίσωση 9.13.

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)} \tag{9.13}$$

Για να προβλέψει την κλάση μιας άγνωστης παρατήρησης, ο Αφελής Μπαϋεσιανός κατηγοριοποιητής υπολογίζει τις πιθανότητες για την κάθε κλάση και εκχωρεί την παρατήρηση στην κλάση με τη μεγαλύτερη πιθανότητα. Εφόσον το $P(X)$ είναι ίδιο για όλες τις κλάσεις και το $P(C_i)$ μπορεί εύκολα να υπολογιστεί (ως το πλήθος των παρατηρήσεων που ανήκουν στην κλάση C_i προς το πλήθος όλων των παρατηρήσεων), το ζητού-

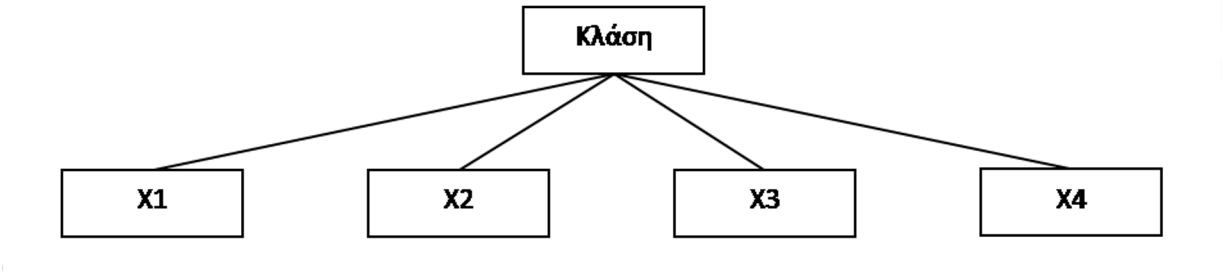
μενο είναι ο υπολογισμός του $P(X|C_i)$. Ο υπολογισμός του $P(X|C_i)$ μπορεί να αποδειχθεί ιδιαίτερα περίπλοκος εάν θεωρηθεί ότι υπάρχει σχέση εξάρτησης μεταξύ των διαστάσεων του ανύσματος Χ, δηλαδή μεταξύ των μεταβλητών εισόδου. Αντιθέτως, αν θεωρηθεί ότι, δοθείσης της κλάσης, οι μεταβλητές εισόδου είναι μεταξύ τους ανεξάρτητες, τότε ο υπολογισμός του $P(X|C_i)$ απλοποιείται και δίνεται από την Εξίσωση 9.14

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

(9.14)

όπου x_k είναι η τιμή της διάστασης k του ανύσματος X.

Ένας Αφελής Μπαϋεσιανός κατηγοριοποιητής με τέσσερεις ανεξάρτητες μεταβλητές παρουσιάζεται με μορφή γραφικού πιθανοτικού μοντέλου στο Σχήμα 9.8.



Σχήμα 9.8 *Αφελής Μπαϋεσιανός Κατηγοριοποιητής*

Ο κατηγοριοποιητής, αφού υπολογίσει τις πιθανότητες $P(C_i|X)$ για όλες τις κλάσεις C_i, εκχωρεί την παρατήρηση στην κλάση με τη μεγαλύτερη πιθανότητα. Εάν ισχύει η υπόθεση ότι δεδομένης της κλάσης είναι ανεξάρτητες οι μεταβλητές εισόδου, ο Αφελής Μπαϋεσιανός κατηγοριοποιητής επιτυγχάνει τους υψηλότερους ρυθμούς ακρίβειας. Ωστόσο, στην πράξη τις περισσότερες φορές η υπόθεση αυτή δεν ισχύει.

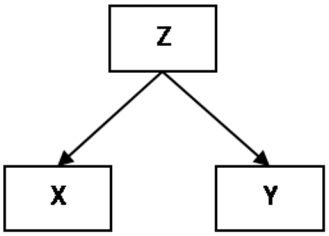
9.9.2 Μπαϋεσιανά Δίκτυα

Τα Μπαϋεσιανά Δίκτυα αποτελούν επέκταση των Αφελών Μπαϋεσιανών κατηγοριοποιητής (ΑΜΚ). Ωστόσο, σε αντίθεση με τους ΑΜΚ δεν υποθέτουν την ανεξαρτησία των μεταβλητών εισόδου. Αντιθέτως, τα Μπαϋεσιανά Δίκτυα επιτρέπουν την ανεξαρτησία υποσυνόλων των μεταβλητών εισόδου. Ένα Μπαϋεσιανό Δίκτυο αναπαριστά τις εξαρτήσεις μεταξύ των μεταβλητών με τη χρήση ενός **Κατευθυνόμενου Ακυκλικού Γράφου** (ΚΑΓ) (Directed Acyclic Graph (DAG)). Κάθε κόμβος του γράφου συμβολίζει μια μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης. Ένα βέλος, το οποίο κατευθύνεται από τη μεταβλητή X προς τη μεταβλητή Y, δηλώνει ότι η Y εξαρτάται από τη X. Η μεταβλητή X καλείται γονέας της Y και η Y καλείται τέκνο της X. Στα Μπαϋεσιανά Δίκτυα μια σημαντική έννοια είναι αυτή της **υπό συνθήκη ανεξαρτησίας** (conditional independence) δύο μεταβλητών. Θεωρούμε τρεις μεταβλητές X,Y,Z οι οποίες συγκροτούν ένα Μπαϋεσιανό Δίκτυο, όπως απεικονίζεται στο Σχήμα 9.9.

Οι μεταβλητές X και Y είναι υπό συνθήκη ανεξάρτητες, εάν οι τιμές της X, με δεδομένες τις τιμές των Y και Z, εξαρτώνται μόνο από τις τιμές της Z. Με μαθηματικό τρόπο η ιδιότητα αυτή αποδίδεται από την Εξίσωση 9.15.

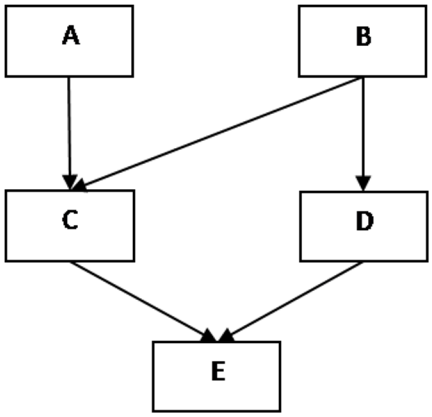
$$P(X|Z,Y) = P(X|Z)$$

(9.15)



Σχήμα 9.9 *Υπό συνθήκη ανεξαρτησία μεταβλητών*

Ο Αφελής Μπαϋεσιανός Κατηγοριοποιητής υποθέτει την υπό συνθήκη ανεξαρτησία των μεταβλητών εισόδου, όταν είναι δεδομένη η τιμή της κλάσης. Στα Μπαϋεσιανά Δίκτυα ισχύει η **τοπική ιδιότητα Markov**, σύμφωνα με την οποία κάθε μεταβλητή είναι υπό συνθήκη ανεξάρτητη από τους μη απογόνους της όταν είναι δεδομένοι οι γονείς της. Το Σχήμα 9.10 απεικονίζει ένα Μπαϋεσιανό Δίκτυο με πέντε μεταβλητές. Η μεταβλητή C είναι ανεξάρτητη από την D εάν είναι γνωστές οι μεταβλητές A και B. Αυτό σημαίνει ότι εάν οι τιμές των μεταβλητών A και B είναι γνωστές, τότε η μεταβλητή D δεν προφέρει πρόσθετη πληροφορία σχετικά με τη μεταβλητή C. Οι μεταβλητές μπορούν να παίρνουν τιμές διακριτές ή συνεχόμενες.



Σχήμα 9.10 *Μπαϋεσιανό Δίκτυο*

Ο γράφος των Μπαϋεσιανών Δικτύων καταγράφει τις σχέσεις μεταξύ των μεταβλητών. Οι σχέσεις αυτές ποσοτικοποιούνται με τον Πίνακα Υπό Συνθήκη Πιθανοτήτων (Conditional Probability Table (CPT)). Στον πίνακα CPT καταγράφεται για κάθε μεταβλητή X η κατανομή πιθανοτήτων $P(X|Par(X))$, όπου $Par(X)$ οι γονείς της μεταβλητής X. Αν τα δεδομένα περιέχουν n μεταβλητές $X_1,X_2,\ldots,X_n$, τότε η πιθανότητα εμφάνισης μιας παρατήρησης με τιμές $x_1,x_2,\ldots,x_n$ για τις αντίστοιχες μεταβλητές δίνεται από την Εξίσωση 9.16

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P\big(x_i | Par(X_i)\big) \tag{9.16}$$

Ένα Μπαϋεσιανό Δίκτυο μπορεί να χρησιμοποιηθεί ως κατηγοριοποιητής. Ένας από τους κόμβους αντιπροσωπεύει τη μεταβλητή της κλάσης. Για μια παρατήρηση υπολογίζονται οι πιθανότητες για κάθε δυνατή τιμή της κλάσης, και η παρατήρηση εκχωρείται στην πιο πιθανή κλάση. Έχουν προταθεί διάφοροι κατηγοριοποιητές Μπαϋεσιανών Δικτύων. Ένας από τους πιο γνωστούς είναι ο Tree Augmented Naïve Bayes (Friedman, Geiger & Goldszmidt, 1997), ο οποίος προβλέπει ότι μια μεταβλητή έχει οπωσδήποτε γονέα της τη μεταβλητή της κλάσης και πιθανώς να έχει γονέα της μια επιπλέον μεταβλητή.

Η δημιουργία ενός μοντέλου Μπαϋεσιανού Δικτύου περιλαμβάνει δύο εργασίες:

- την κατασκευή του γράφου,
- τον υπολογισμό του πίνακα πιθανοτήτων CPT.

Ο υπολογισμός του CPT είναι ευκολότερο καθήκον, ειδικά εάν δεν υπάρχουν χαμένες τιμές. Εάν δεν υπάρχουν κρυφά δεδομένα, ο υπολογισμός του CPT είναι απλός και γίνεται με τρόπο αντίστοιχο με τον υπολογισμό των πιθανοτήτων στον Αφελή Μπαϋεσιανό κατηγοριοποιητή. Η ύπαρξη χαμένων τιμών περιπλέκει τον υπολογισμό του CPT. Για την κατασκευή του γράφου υπάρχουν δύο εκδοχές. Κατά την πρώτη εκδοχή ο γράφος σχεδιάζεται από ανθρώπους, οι οποίοι είναι ειδικοί στο πρόβλημα το οποίο εξετάζεται. Η δεύτερη εκδοχή είναι να εξαχθεί ο γράφος από τα δεδομένα με αυτοματοποιημένο τρόπο. Η αυτόματη δημιουργία του γράφου είναι ένα δύσκολο καθήκον. Σε επιστημονικές εργασίες έχουν προταθεί διάφορες μέθοδοι για την εξαγωγή του γράφου. Ενδεικτικά αναφέρουμε τις Heckerman, Geiger and Chickering (1995), Cooper and Herskovits (1992) και Cheng, Greiner, Kelly, Bell and Liu (2002).

9.9.3 Πλεονεκτήματα και μειονεκτήματα των Μπαϋεσιανών Δικτύων

Τα Μπαϋεσιανά Δίκτυα Πίστης συγκεντρώνουν πολλά **πλεονεκτήματα**:

- Δημιουργούν ένα μοντέλο για την κατανομή πιθανοτήτων για ένα πρόβλημα. Υπό αυτήν την έννοια είναι ιδιαίτερα κατάλληλα για περιπτώσεις όπου υπάρχουν σύνθετες εξαρτήσεις μεταξύ της μεταβλητής της κλάσης και των μεταβλητών εισόδου ή και ακόμα μεταξύ των μεταβλητών εισόδου.
- Ο γράφος που δημιουργείται οπτικοποιεί τις σχέσεις μεταξύ της κλάσης και των μεταβλητών εισόδου. Για τον λόγο αυτό, τα Μπαϋεσιανά Δίκτυα Πίστης είναι εύκολα κατανοητά από τους ανθρώπους.
- Για τον σχεδιασμό του γράφου μπορεί να χρησιμοποιηθεί προηγούμενη γνώση ειδικών. Ακόμα και εάν ο γράφος εξαχθεί από τα δεδομένα με αυτοματοποιημένο τρόπο, υπάρχει δυνατότητα μεταβολής του από τους ειδικούς. Στην περίπτωση αυτή επιτρέπεται ο συγκερασμός της γνώσης ειδικών με το αποτέλεσμα του αλγορίθμου εξαγωγής του γράφου.
- Μπορούν να χειριστούν και αριθμητικές και ονομαστικές μεταβλητές.
- Μπορούν να επιτύχουν υψηλούς ρυθμούς ακρίβειας.
- Διαθέτουν στιβαρή θεωρητική θεμελίωση βασισμένη στη Στατιστική.

Μειονεκτήματα των Μπαϋεσιανών Δικτύων είναι τα ακόλουθα:

- Το σημαντικότερο μειονέκτημα τους είναι το γεγονός ότι δεν υπάρχει ένας καθιερωμένος και γενικά αποδεκτός τρόπος εξαγωγής του γράφου από τα δεδομένα.
- Για τον υπολογισμό των πιθανοτήτων ενός κλάδου του δικτύου απαιτείται ο υπολογισμός όλων των άλλων κλάδων επιφέροντας σημαντικό υπολογιστικό κόστος.
- Στους Αφελείς Μπαϋεσιανούς Κατηγοριοποιητές η υπόθεση ανεξαρτησίας των μεταβλητών εισόδου ισχύει σπάνια.

9.10 Μελέτη περίπτωσης – Εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων με χρήση μεθόδων κατηγοριοποίησης.

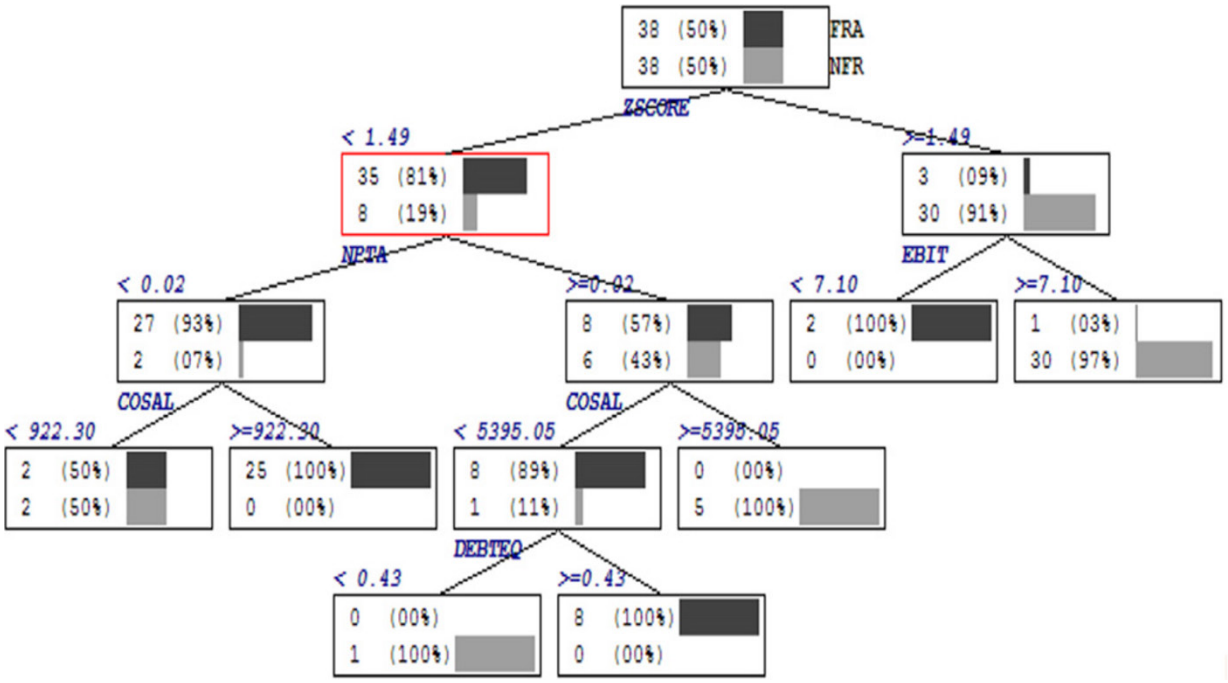
Η παραποίηση των χρηματοοικονομικών καταστάσεων από τη διοίκηση των επιχειρήσεων είναι ένα σημαντικό πρόβλημα της παγκόσμιας οικονομίας. Ο Wells (1997) εκτιμά ότι η απάτη κοστίζει στην αμερικανική οικονομία 400 δισεκατομμύρια δολάρια ετησίως, ενώ ο Koskivaara (2004) αποκαλεί το έτος 2002 «φριχτή χρονιά» ως προς την τήρηση των βιβλίων και ισχυρίζεται ότι η χειραγώγηση συνεχίζεται. Οι Spathis, Doumpos and Zopounidis (2002) επισημαίνουν ότι οι παραποιημένες χρηματοοικονομικές καταστάσεις αυξάνονται τα τελευταία χρόνια. Ο εντοπισμός των περιπτώσεων διοικητικής απάτης είναι ιδιαίτερα δύσκολος, καθώς τα έμπειρα διοικητικά στελέχη γνωρίζουν τα όρια των τυπικών ελεγκτικών διαδικασιών και ηθελημένα προσπαθούν να παραπλανήσουν τους ελεγκτές. Αυτοί οι περιορισμοί συνιστούν την εφαρμογή νέων, περίτεχνων αναλυτικών διαδικασιών. Οι Kirkos, Spathis and Manolopoulos (2007) διερευνούν τη δυνατότητα των μεθόδων εξόρυξης δεδομένων, και ειδικότερα των μεθόδων κατηγοριοποίησης, να εντοπίσουν περιπτώσεις παραποιη-

μένων χρηματοοικονομικών καταστάσεων.

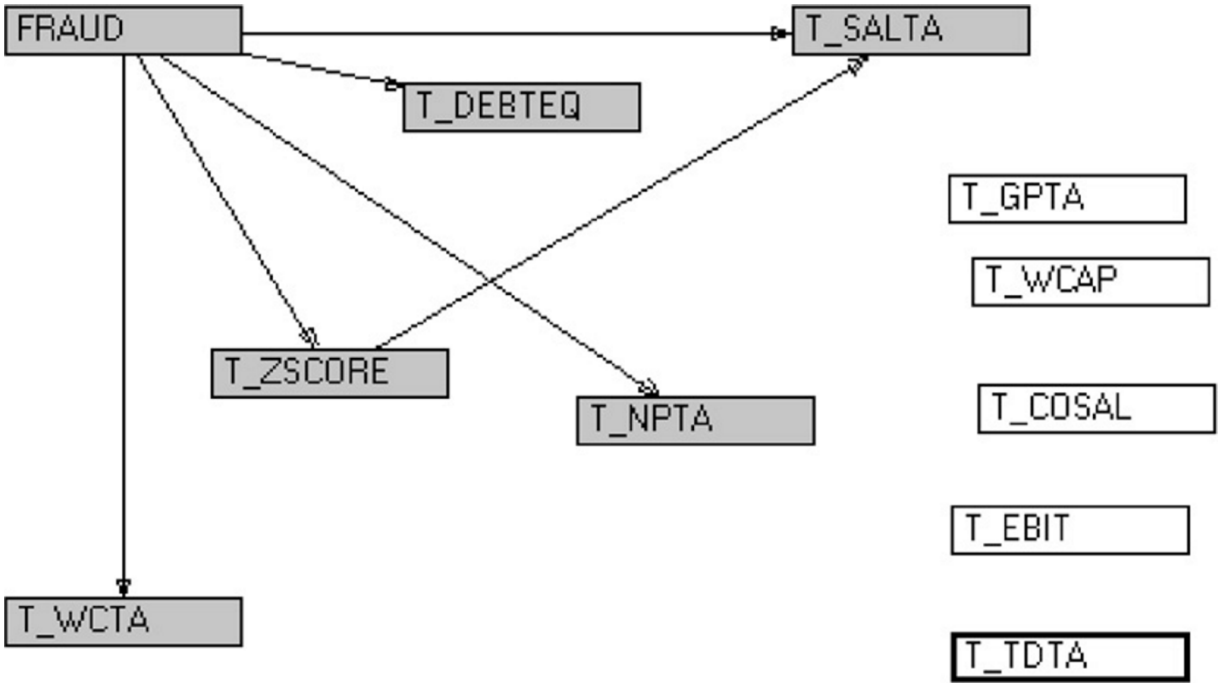
Τα δεδομένα αφορούν 76 περιπτώσεις ελληνικών, μη χρηματοπιστωτικών επιχειρήσεων. Οι μισές από αυτές τις επιχειρήσεις είχαν αποδεδειγμένα παραποιήσει τις χρηματοοικονομικές τους καταστάσεις. Το γεγονός αυτό προκύπτει από παρατηρήσεις των εξωτερικών ελεγκτών, από αποφάσεις δικαστηρίων, από αποφάσεις της Επιτροπής Κεφαλαιαγοράς του ΧΑΑ, καθώς και από αποφάσεις φορολογικών αρχών. Για τις υπόλοιπες 38 επιχειρήσεις δεν υπήρχε καμία απόδειξη ή ένδειξη παραποίησης, και για τον λόγο αυτό οι χρηματοοικονομικές τους καταστάσεις θεωρήθηκαν ειλικρινείς και σύννομες. Ως μεταβλητή κλάσης χρησιμοποιήθηκε μια δυαδική μεταβλητή που διαφοροποιούσε τις δύο κατηγορίες επιχειρήσεων. Σημειωτέον ότι τα δεδομένα αφορούν δημοσίως διαθέσιμα στοιχεία, τα οποία δημοσιεύτηκαν σε ισολογισμούς και σε αποτελέσματα χρήσης.

Η αρχική επιλογή μεταβλητών βασίστηκε στην προηγούμενη ερευνητική βιβλιογραφία. Μεγάλος αριθμός ερευνητικών εργασιών ασχολήθηκε με το θέμα της παραποίησης των χρηματοοικονομικών καταστάσεων. Στις εργασίες αυτές προτείνονται διάφοροι αριθμοδείκτες, οι οποίοι παρέχουν ενδείξεις παραποίησης και μπορούν να χρησιμοποιηθούν ως μεταβλητές εισόδου σε ένα μοντέλο πρόβλεψης. Ενδεικτικά αναφέρουμε τις εργασίες των Fanning and Cogger (1998), Loebbecke, Eining and Willingham (1989) και Persons (1995). Συνολικά επιλέχτηκαν 27 αρχικοί αριθμοδείκτες. Σε αυτές περιλαμβάνονταν και ο αριθμοδείκτης Z-Score του Altman. Ακολούθησε ανάλυση επιλογής σημαντικών χαρακτηριστικών με εφαρμογή της μεθόδου ANOVA. Σύμφωνα με τα αποτελέσματα, 10 αριθμοδείκτες παρουσίασαν χαμηλή τιμή *p (p<=0,05)*. Οι αριθμοδείκτες αυτοί επιλέχτηκαν, ώστε να αποτελέσουν τις μεταβλητές εισόδου στα μοντέλα που αναπτύχθηκαν.

Τρεις διαφορετικές μέθοδοι κατηγοριοποίησης εφαρμόστηκαν για την ανάπτυξη αντίστοιχων μοντέλων. Το πρώτο μοντέλο ήταν ένα Δένδρο Αποφάσεων τύπου ID3. Το δεύτερο μοντέλο ήταν ένα Νευρωνικό Δίκτυο τύπου Multilayer Perceptron με ένα κρυφό στρώμα, το οποίο περιείχε πέντε νευρώνες. Το τρίτο μοντέλο ήταν ένα Μπαϋεσιανό Δίκτυο. Το λογισμικό που χρησιμοποιήθηκε για το Μπαϋεσιανό Δίκτυο ήταν ικανό να εξάγει τον γράφο από τα δεδομένα. Το λογισμικό επέτρεπε στον χρήστη να τροποποιήσει τη δομή του γράφου ο οποίος κατασκευάστηκε αυτόματα, ωστόσο επιλέξαμε να χρησιμοποιήσουμε τον αυτοματοποιημένο γράφο χωρίς μεταβολές. Τα μοντέλα του Δένδρου Αποφάσεων και του Μπαϋεσιανού Δικτύου παρουσιάζονται στα σχήματα 9.11 και 9.12 αντίστοιχα.



Σχήμα 9.11 *Εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων με Δένδρο Αποφάσεων*



Σχήμα 9.12 *Εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων με Μπαϋεσιανό Δίκτυο.*

Παρατηρούμε ότι το Δένδρο Αποφάσεων χρησιμοποιεί ως μεταβλητή διαχωρισμού πρώτου επιπέδου τη μεταβλητή Z-Score. Σύμφωνα με το κριτήριο μείωσης της εντροπίας, ο δείκτης Z-Score είναι αυτός που διαχωρίζει με τον καλύτερο τρόπο τις δύο κλάσεις. Σαράντα τρεις επιχειρήσεις είχαν τιμή Z-Score μικρότερη από 1,49. Οι τριάντα πέντε από αυτές τις επιχειρήσεις είχαν παραποιήσει τις χρηματοοικονομικές τους καταστάσεις. Αντιθέτως, από τις τριάντα τρεις επιχειρήσεις με δείκτη Z-Score μεγαλύτερο από ή ίσο με 1,49 μόνο οι τρεις είχαν παραποιήσει τις χρηματοοικονομικές τους καταστάσεις. Υπενθυμίζουμε ότι ο Altman είχε ορίσει την τιμή Z-Score=1,81 ως τιμή διαχωρισμού μεταξύ των υγειών και προβληματικών επιχειρήσεων για την αμερικανική βιομηχανία. Με βάση αυτό το γεγονός, μπορούμε να συμπεράνουμε ότι επιχειρήσεις σε οικονομική δυσπραγία τείνουν να χειραγωγήσουν τις χρηματοοικονομικές τους καταστάσεις. Μια άλλη ενδιαφέρουσα παρατήρηση είναι ότι και οι δύο αριθμοδείκτες που χρησιμοποιούνται ως μεταβλητές διαχωρισμού δεύτερου επιπέδου σχετίζονται με την κερδοφορία. Η μεταβλητή NPTA είναι τα καθαρά κέρδη προς σύνολο ενεργητικού (Net Profit to Total Assets) και η μεταβλητή EBIT είναι τα κέρδη προ φόρων και τόκων (Earnings Before Interest and Tax). Από τους δύο κλάδους του κόμβου Z-Score και τις μεταβλητές διαχωρισμού δευτέρου επιπέδου προκύπτει ότι επιχειρήσεις εμπλεκόμενες σε πρακτικές χειραγώγησης και με χαμηλό Z-Score παρουσιάζουν χαμηλή κερδοφορία. Αντιθέτως, η μη χειραγώγηση σχετίζεται κυρίως με επιχειρήσεις που έχουν υψηλό Z-Score και ικανοποιητική κερδοφορία.

Σύμφωνα με το μοντέλο του Μπαϋεσιανού Δικτύου υπάρχει σχέση εξάρτησης μεταξύ της μεταβλητής της κλάσης και πέντε αριθμοδεικτών. Είναι ενδιαφέρον ότι κάθε μια από αυτές τις μεταβλητές αναφέρεται σε μια διαφορετική διάσταση των οικονομικών στοιχείων της επιχείρησης. Ειδικότερα, η μεταβλητή Z-Score αναφέρεται στην οικονομική ευρωστία, η μεταβλητή DEBTEQ (Debt to Equity) στη μόχλευση, η μεταβλητή NPTA (Net Profit to Total Assets) στην κερδοφορία, η μεταβλητή SALTA (Sales to Total Assets) στο ύψος των πωλήσεων και η μεταβλητή WCTA (Working Capital to Total Assets) στη ρευστότητα. Φαίνεται ότι το Μπαϋεσιανό Δίκτυο οικοδομεί μια πιο γενικευμένη άποψη και συσχετίζει τη χειραγώγηση των χρηματοοικονομικών καταστάσεων με διάφορες εκφάνσεις της οικονομικής κατάστασης της επιχείρησης.

Για την εκτίμηση της ρεαλιστικής επίδοσης των τριών μεθόδων, δηλαδή της ικανότητας τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων, εφαρμόστηκε η τεχνική της επικύρωσης 10 τμημάτων (10 fold cross validation), η οποία παρουσιάζεται αναλυτικά στο Κεφάλαιο 10. Η συγκεκριμένη τεχνική θεωρείται ιδιαίτερα αξιόπιστη και κατάλληλη για σχετικά μικρά σύνολα παρατηρήσεων. Σύμφωνα με τα αποτελέσματα, το Μπαϋεσιανό Δίκτυο είχε γενική επίδοση 90,3% και επέτυχε να κατηγοριοποιήσει σωστά το 91,7% των περιπτώσεων απάτης και το 88,9% των περιπτώσεων μη απάτης. Οι αντίστοιχες επιδόσεις για το Νευρωνικό Δίκτυο ήταν 80%, 82,5% και 77,5%, ενώ για το Δένδρο Αποφάσεων ήταν 73,6%, 75% και 72,5%. Παρατηρούμε ότι το Μπαϋεσιανό Δίκτυο επέτυχε εξαιρετικά υψηλές επιδόσεις, ακολουθούμενο από το Νευρωνικό Δίκτυο και το Δένδρο Αποφάσεων. Παρατηρούμε επίσης ότι και τα τρία μοντέλα προβλέπουν σε μεγαλύτερο

ποσοστό τις περιπτώσεις απάτης από τις περιπτώσεις μη απάτης. Το γεγονός αυτό είναι σημαντικό. Η σημασία των διαφορετικών τύπων σφάλματος παρουσιάζεται αναλυτικά στο Κεφάλαιο 10.

Βιβλιογραφία / Αναφορές

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC Press.

Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian Networks from Data: An Information-Theory Based Approach. *Artificial Intelligence, 137*(1-2), 43-90. doi: 10.1016/s0004-3702(02)00191-1

Cooper, G., & Herskovits, E. (1992). A Bayesian Method for the Induction Probabilistic Networks from Data. *Machine Learning, 9*(4), 309-347. doi: 10.1007/bf00994110

Fanning, K., & Cogger, K. (1998). Neural Network Detection of Management Fraud Using Published Financial Data. *International Journal of Intelligent Systems in Accounting, Finance and Management, 7*(1), 21-41. doi: 10.1002/(sici)1099-1174(199803)7:1<21::aid-isaf138>3.0.co;2-k

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In: U. Fayyad, G. Piatetsky-Shapiro & P. Smyth (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1-34). Menlo Park, CA: AAAI/MIT Press.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning, 29*(2-3), 131-163. doi: 10.1023/A:1007465528199

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Boston: Academic Press.

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

Heckerman, D. (1997). Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery, 1*(1), 79-119. doi: 10.1023/A:1009730122752

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning, 20*(3), 197-243. doi: 10.1007/bf00994016

Hwang, J., Lay, S., & Lippman, A. (1994). Nonparametric Multivariate Density Estimation: A Comparative Study. *IEEE Transactions on Signal Processing, 42*(10), 2795-2810. doi: 10.1109/78.324744

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics, 29*(2), 119-127. doi: 10.2307/2986296

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications, 32*(4), 995-1003. doi: 10.1016/j.eswa.2006.02.016

Koskivaara, E. (2004). Artificial Neural Networks in Analytical Review Procedures. *Managerial Auditing Journal, 19*(2), 191–223. doi: 10.1108/02686900410517821

Loebbecke, J., Eining, M., & Willingham, J. (1989). Auditor's Experience with Material Irregularities: Frequency, Nature and Detectability. *Auditing: A Journal of Practice and Theory, 9*, 1-28.

Loh, W. Y., & Shih, X. (1997). Split Selection Methods for Classification Trees. *Statistica Sinika, 7*, 815-840.

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

Olaru, C., & Wehenkel, L. (2003). A Complete Fuzzy Decision Tree Technique. *Fuzzy Sets and Systems, 138*(2), 221-254. doi: 10.1016/S0165-0114(03)00089-7

Parker, D. B. (1985). Learning-logic: Casting the Cortex of the Human Brain in Silicon. *Technical Report TR-47*. Boston, MA: Center for Computational Research in Economics and Management Science, MIT.

Persons, O. (1995). Using Financial Statements Data to Identify Factors Associated with Fraudulent Financial Reporting. *Journal of Applied Business Research, 11*(3), 38-46.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81-106. doi: 10.1007/bf00116251

Quinlan, J. R. (1987). Simplifying Decision Trees. *International Journal of Man-Machine Studies, 27*(3), 221-234. doi: 10.1016/s0020-7373(87)80053-6

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufman.

Quinlan, J. R., & Rivest, R. L. (1989). Inferring Decision Trees Using the Minimum Description Length Principle. *Information and Computation, 80*(3), 227-248. doi: 10.1016/0890-5401(89)90010-2

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-propagating Errors. *Letters to Nature, 323*(6088), 533-536. doi: 10.1038/323533a0

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York, NY: Springer-Verlag.
Sonquist, J. A., Baker, E. L., & Morgan, J. N. (1971). *Searching for Structure*. An Arbor, MI: Institute for Social Research, University of Michigan.
Spathis, C., Doumpos, M., & Zopounidis, C. (2002). Detecting Falsified Financial Statements: A Comparative Study Using Multicriteria Analysis and Multivariate Statistical Techniques. *The European Accounting Review, 11*(3), 509-535. doi: 10.1080/0963818022000000966
Wells, J. T. (1997). *Occupational Fraud and Abuse*. Austin, TX: Obsidian Publishing.
Wong, B. K., Bodnovich, T. A., & Selvi, Y. (1997). Neural Networks Applications in Business: A Review and Analysis of the Literature (1988-1995). *Decision Support Systems, 19(4), 301-320.* doi: 10.1016/s0167-9236(96)00070-x

Κριτήρια Αξιολόγησης

Άσκηση Υπολογισμών 9.1

Χρησιμοποιήστε τα δεδομένα του Πίνακα 9.1 και υπολογίστε το Κέρδος Πληροφορίας σε περίπτωση που επιλεγεί το πεδίο «Εισόδημα» ως μεταβλητή διαχωρισμού για τη δημιουργία Δένδρου Αποφάσεων ID3.

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

Πίνακας 9.1 *Δεδομένα Άσκησης 1*

Λύση

Αρχικά πρέπει να υπολογιστεί η Εντροπία του συνόλου E(S), σύμφωνα με την Εξίσωση 9.1. Θεωρούμε ως θετική κλάση την έγκριση του δανείου. Στο σύνολο δεδομένων υπάρχουν εννέα θετικές και πέντε αρνητικές παρατηρήσεις. Η Εντροπία υπολογίζεται ως εξής:

$E(S)= - (9/14)*\log_2(9/14) – (5/14)*\log_2(5/14) = 0,94.$

Εάν επιλεγεί το Εισόδημα ως μεταβλητή διαχωρισμού, τότε το σύνολο δεδομένων θα διαχωριστεί σε τρία υποσύνολα, όπου στο πρώτο υποσύνολο S_1 θα περιλαμβάνονται οι υποψήφιοι με χαμηλό εισόδημα, στο δεύτερο υποσύνολο S_2 οι υποψήφιοι με υψηλό εισόδημα και στο τρίτο υποσύνολο S_3 οι υποψήφιοι με μεσαίο εισόδημα. Η συνολική Εντροπία διαχωρισμού θα υπολογιστεί σύμφωνα με την Εξίσωση 9.3. Αρχικά πρέπει να υπολογιστούν οι Εντροπίες των τριών υποσυνόλων.

Το υποσύνολο S_1 περιέχει τρεις θετικές και μια αρνητική παρατήρηση. Η Εντροπία του υπολογίζεται ως εξής:

$E(S_1) = -(3/4)*\log_2(3/4) – (1/4)*\log_2(1/4) = 0,811.$

Το υποσύνολο S_2 περιέχει δύο θετικές και δύο αρνητικές παρατηρήσεις. Η Εντροπία του υπολογίζεται ως εξής:

$E(S_2) = -(2/4)*\log_2(2/4) – (2/4)*\log_2(2/4) = 1.$

Το υποσύνολο S_3 περιέχει τέσσερεις θετικές και δύο αρνητικές παρατηρήσεις. Η Εντροπία του υπολογίζεται ως εξής:

$E(S_3) = -(4/6)*\log_2(4/6) – (4/6)*\log_2(4/6) = 0,918.$

Το υποσύνολο S_1 περιέχει τέσσερεις παρατηρήσεις, το υποσύνολο S_2 περιέχει τέσσερεις παρατηρήσεις, το υποσύνολο S_3 περιέχει έξι παρατηρήσεις, και το αρχικό σύνολο περιέχει δέκα τέσσερεις παρατηρήσεις. Σύμφωνα με την Εξίσωση 9.3, η Εντροπία διαχωρισμού θα υπολογιστεί ως εξής:

$E(S, Εισόδημα) = (4/14)*E(S_1) + (4/14)*E(S_2) + (6/14)*E(S_3) = 0,911.$

Το Κέρδος Πληροφορίας υπολογίζεται σύμφωνα με την Εξίσωση 9.4
IG(S, Εισόδημα) = E(S) - E(S, Εισόδημα) = 0,94 – 0,911 = 0,029.

Άσκηση Υπολογισμών 9.2

Χρησιμοποιήστε τα δεδομένα του παρακάτω πίνακα. Εφαρμόστε το θεώρημα του Bayes για να προβλέψετε τις πιθανότητες έγκρισης και απόρριψης της αίτησης ενός υποψηφίου με μέσο εισόδημα και μικρή ηλικία.

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

Πίνακας 9.2 *Δεδομένα Άσκησης 2*

Λύση

Οι πιθανότητες έγκρισης και απόρριψης του δανείου θα υπολογιστούν σύμφωνα με την Εξίσωση 9.13. Ο πίνακας περιέχει 14 περιπτώσεις και στις τρεις από αυτές οι υποψήφιοι έχουν μέσο εισόδημα και μικρή ηλικία. Η πιθανότητα P(X) υπολογίζεται ως εξής:

P(X) = 3/14 = 0,21

Για τις οκτώ περιπτώσεις του συνόλου το δάνειο εγκρίνεται, ενώ για τις τέσσερεις απορρίπτεται. Οι αντίστοιχες πιθανότητες υπολογίζονται ως εξής:

P(Yes) = 8/14 = 0,57

P(No) = 6/14 = 0,429

Από τις οκτώ περιπτώσεις όπου το δάνειο εγκρίνεται, στις δύο το εισόδημα είναι μέσο και η ηλικία μικρή. Από τις έξι περιπτώσεις όπου το δάνειο δεν εγκρίνεται, στη μια περίπτωση το εισόδημα είναι μέσο και η ηλικία μικρή. Οι αντίστοιχες πιθανότητες υπολογίζονται ως εξής:

P(X|Yes) = 2/8 = 0,25

P(X|No) = 1/6 = 0,167

Σύμφωνα με την Εξίσωση 9.13, οι πιθανότητες έγκρισης και απόρριψης του δανείου για μέσο εισόδημα και μικρή ηλικία είναι:

P(Yes|X)= (0,25*0,57)/0,21 = 0,667

P(No|X)= (0,167*0,429)/0,21 = 0,333

Η πιθανότητα έγκρισης του δανείου είναι διπλάσια από την πιθανότητα απόρριψης του.

Άσκηση Εφαρμογής 9.3

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 μοντέλα πρόβλεψης χρεοκοπίας με χρήση των μεθόδων α)Δένδρου Αποφάσεων C4.5, β)Νευρωνικού Δικτύου Multilayer Perceptron, γ)Μπαϋεσιανού Δικτύου. Στην επιλογή «Test Option» επιλέξτε «Cross-validation». Με την επιλογή αυτή τα μοντέλα δοκιμάζονται χρησιμοποιώντας άγνωστες παρατηρήσεις. Μελετήστε τα αποτελέσματα των μοντέλων και επιλέξτε το μοντέλο που επιτυγχάνει τις καλύτερες επιδόσεις κατηγοριοποίησης. Μελετήστε το Δένδρο Αποφάσεων. Παρατηρήστε τα βάρη των συνδέσεων του Νευρωνικού Δικτύου. Προβάλετε το Δένδρο Αποφάσεων με γραφικό τρόπο. Επαναλάβετε το πείραμα αφού προηγουμένως έχετε επιλέξει την επιλογή «Use training set» στο πεδίο «Test Options». Με την επιλογή αυτή τα μοντέλα δοκιμάζονται χρησιμοποιώντας τις παρατηρήσεις με τις οποίες εκπαιδεύτηκαν. Συγκρίνετε τις επιδόσεις των μοντέλων με τις προηγούμενες επιδόσεις που είχαν υπολογιστεί με την τεχνική «Cross-validation». Τι παρατηρείτε;

Λύση

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

Στο tab «Preprocess» εμφανίζονται διάφορες πληροφορίες για τα δεδομένα. Παρατηρήστε τα πεδία (Attributes). Ως πεδίο κλάσης ορίζεται αυτόματα το τελευταίο πεδίο, δηλαδή το πεδίο «Bankrupt». Κάνοντας κλικ σε ένα αριθμητικό πεδίο εμφανίζονται η ελάχιστη και μέγιστη τιμή, η μέση τιμή και η τυπική απόκλιση. Επίσης, εμφανίζεται η κατανομή των τιμών. Μπορείτε να κάνετε αρχική διερευνητική ανάλυση παρατηρώντας την κατανομή των τιμών και των κλάσεων.

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

Βήμα 2. Μεταβείτε στο tab «Classify».

Επιλέξτε μέθοδο κατηγοριοποίησης πιέζοντας το κουμπί «Choose» στο πεδίο «Classifier». Επιλέξτε πρώτα τη μέθοδο weka/classifiers/trees/J48 για το Δένδρο Αποφάσεων C4.5 και πατήστε το κουμπί «Start». Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Μπορείτε να δείτε το Δένδρο Αποφάσεων και τις επιδόσεις του μοντέλου. Το μοντέλο κατηγοριοποιεί σωστά το 78% του συνόλου των παρατηρήσεων, το 68% της κλάσης «0» και το 88% της κλάσης «1». Κάνοντας δεξί κλικ στο μοντέλο στο πεδίο «Results list» και επιλέγοντας «Visualize tree» παρουσιάζεται το δένδρο με γραφικό τρόπο.

Βήμα 3. Επιλέξτε τη μέθοδο weka/classifiers/functions/MultilayerPerceptron και πατήστε το κουμπί «Start». Εμφανίζονται τα βάρη των συνδέσεων. Το μοντέλο κατηγοριοποιεί σωστά το 90% των συνολικών περιπτώσεων, το 88% της κλάσης «0» και το 92% της κλάσης «1».

Βήμα 4. Επιλέξτε τη μέθοδο weka/classifiers/bayes/BayesNet και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά το 88% των συνολικών περιπτώσεων, και το 88% των κλάσεων «0» και «1».

Καλύτερες επιδόσεις επιτυγχάνει το Νευρωνικό Δίκτυο.

Βήμα 5. Επαναλάβετε τα τρία προηγούμενα βήματα έχοντας ενεργοποιήσει την επιλογή «Use training set». Τα αποτελέσματα που λαμβάνετε έχουν υπολογιστεί χρησιμοποιώντας τις παρατηρήσεις με τις οποίες εκπαιδεύτηκαν τα μοντέλα. Παρατηρήστε τις σημαντικές αυξήσεις των επιδόσεων. Για παράδειγμα, το Δένδρο Αποφάσεων αυξάνει το ποσοστό ακρίβειας στο 96% (από 78%). Ωστόσο, οι επιδόσεις αυτές είναι πλασματικές. Η πραγματική αξία των μοντέλων βρίσκεται στην ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων.

Άσκηση Εφαρμογής 9.4

Χρησιμοποιήστε το αρχείο «credit-a» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή UCI repository). Το σύνολο δεδομένων προσφέρθηκε από τον καθηγητή Ross Quinlan και χρησιμοποιήθηκε στο Quinlan (1987). Πρόκειται για δεδομένα αιτήσεων για πιστωτικές κάρτες. Υπάρχουν συνολικά 16 πεδία, όπου το τελευταίο είναι το πεδίο της

κλάσης. Στο πεδίο κλάσης χρησιμοποιούνται τα σύμβολα «+» και «-» για τις «καλές» και τις «κακές» αιτήσεις αντίστοιχα. Επίσης το αρχείο περιέχει 690 παρατηρήσεις, εκ των οποίων οι 307 ανήκουν στην κλάση «+» και 383 στην κλάση «-». Τα ονόματα των πεδίων έχουν αλλαχθεί και οι τιμές των δεδομένων έχουν κωδικοποιηθεί για λόγους απορρήτου.

Εφαρμόστε επιλογή χαρακτηριστικών με τη μέθοδο CFS Subset Evaluator. Στη συνέχεια αναπτύξτε μοντέλο νευρωνικού δικτύου τύπου Multilayer Perceprton και επικυρώστε το με τη μέθοδο «Cross-validation». Πειραματιστείτε με τις παραμέτρους της μεθόδου και προσπαθήστε να αυξήσετε τις επιδόσεις.

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «credit-a.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. Εάν επιθυμείτε, μπορείτε να επιλέξτε μια άλλη μέθοδο κάνοντας κλικ στα περιεχόμενα του πεδίου «Filter». Αφού ορίσετε τη μέθοδο που επιθυμείτε (τα αποτελέσματα που ακολουθούν ισχύουν για τη CFS), κάνετε κλικ στο κουμπί «Apply». Θα διαπιστώσετε ότι στο πεδίο «Attributes» μειώνεται το πλήθος των στηλών. Ειδικότερα, παραμένουν οι στήλες A4, A6, A8, A9, A11, A14, A15 και 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/functions/MultilayerPerceptron.

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».

Θα πρέπει να περιμένετε μερικά δευτερόλεπτα. Η εκπαίδευση των Νευρωνικών δικτύων είναι αργή. Παρατηρήστε το μικρό πουλί στο κάτω δεξιά μέρος της οθόνης. Όσο το πουλί κινείται, το λογισμικό εκτελεί υπολογισμούς.

Όταν ολοκληρωθεί η εκπαίδευση και η επικύρωση θα εμφανιστούν τα αποτελέσματα στο πεδίο «Classifier Output». Το μοντέλο κατηγοριοποιεί σωστά το 83,4783% των συνολικών παρατηρήσεων, το 82,1% της κλάσης «+» και το 84,6% της κλάσης «-».

Βήμα 3. Κάντε κλικ στα περιεχόμενα του πεδίου «Classifier» και στο όνομα «MultilayerPerceptron». Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων του νευρωνικού δικτύου. Μεταβάλετε τις τιμές ορισμένων παραμέτρων και επαναλάβετε την εκπαίδευση και επικύρωση του μοντέλου. Οι σημαντικότερες παράμετροι είναι οι «hiddenLayers», «learningRate», «momentum», και «trainingTime». Οδηγίες για τις παραμέτρους μπορείτε να λάβετε κάνοντας κλικ στο κουμπί «More».

Η παράμετρος «hiddenLayers» ορίζει το πλήθος των κρυφών στρωμάτων και των κρυφών νευρώνων. Η προκαθορισμένη τιμή είναι η «a». Η κωδική αυτή τιμή σημαίνει ότι το μοντέλο έχει ένα κρυφό στρώμα με πλήθος νευρώνων ίσο με το άθροισμα των μεταβλητών εισόδου και του πλήθους των τιμών κλάσης δια δύο ((attributes + classes) /2). Για περισσότερες πληροφορίες συμβουλευτείτε τον οδηγό του WEKA στο τελευταίο κεφάλαιο, ή κάντε κλικ στο κουμπί «More».

Πειραματιστείτε αρκετές φορές με διάφορες τιμές παραμέτρων προσπαθώντας να αυξήσετε τις επιδόσεις του Νευρωνικού Δικτύου.

Ορίζοντας Learning rate = 0.2, momentum=0.1, Training Time=700 και Hidden Layers=7 Το Νευρωνικό Δίκτυο επιτυγχάνει ακρίβεια 84.6377% και κατηγοριοποιεί σωστά το 83.7% των περιπτώσεων της κλάσης «+» και το 85.4% των περιπτώσεων της κλάσης «-».