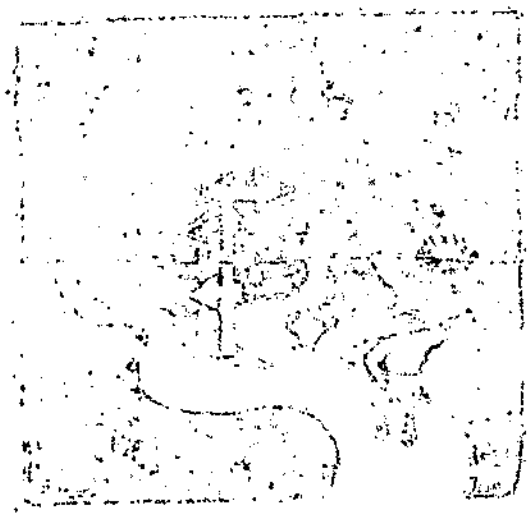
Ανώτατο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Πάτρας
Σχολή Διοίκησης και Οικονομίας
Τμήμα Λογιστικής

Συστήματα Υποστήριξης Αποφάσεων Έγκρισης Τραπεζικού Δανείου

ΟΜΑΔΑ ΕΡΓΑΣΙΑΣ
ΖΑΦΕΙΡΟΠΟΥΛΟΥ ΟΥΡΑΝΙΑ
ΖΩΓΟΓΙΑΝΝΗ ΔΗΜΗΤΡΑ
ΣΟΦΟΓΙΑΝΝΗ ΒΑΣΙΛΙΚΗ
ΕΙΣΗΓΗΤΗΣ: ΚΩΤΣΙΑΝΤΗΣ ΣΩΤΗΡΗΣ

ΠΑΤΡΑ 2004



Περιεχόμενα

Πρόλογος

Τι είναι Συστήματα Υποστήριξης αποφάσεων ;

Τα Συστήματα Υποστήριξης αποφάσεων παρέχουν τη δυνατότητα στις εταιρείες να μετατρέψουν μεγάλες ποσότητες επιχειρηματικών πληροφοριών σε κερδοφόρα αποτελέσματα. Γι' αυτό το λόγο αποτελούν το σημαντικότερο μέρος ενός επιχειρηματικού πληροφοριακού συστήματος.

Πιο συγκεκριμένα, οι επιχειρήσεις συγκεντρώνουν καθημερινά πληροφορίες που αφορούν τις συναλλαγές τους με τους πελάτες. Οι μάνατζερ χρησιμοποιούν αυτές τις πληροφορίες για να παίρνουν οποιεσδήποτε αποφάσεις. Επομένως, τα Συστήματα Υποστήριξης αποφάσεων στηρίζονται σ' αυτή την καθημερινή συλλογή πληροφοριών για να μπορούν να παρέχουν υπηρεσίες υποστήριξης αποφάσεων.

Η καρδιά των Συστημάτων Υποστήριξης αποφάσεων είναι η Επιχειρησιακή Νοημοσύνη. Πρόκειται για μια λειτουργία που έχει σκοπό την τροφοδότηση της διοίκησης ή άλλων τομέων σχετικών με την λήψη αποφάσεων, με πληροφορίες που αφορούν συναλλαγές με πελάτες, συμπεριφορά ανταγωνιστών κτλ.

Η εξόρυξη γνώσης είναι μία διαδικασία που σαν ρόλο έχει να εφαρμόζει μεθόδους ανάλυσης με μεγάλο όγκο δεδομένων. Πρόκειται για μία πολύ πρόσφατη τεχνολογία που βοηθάει τους μάνατζερ να εστιάζουν μόνο στα πιο σημαντικά δεδομένα από τις αποθήκες δεδομένων τους. Σκοπός αυτού του εργαλείου είναι η ανακάλυψη που είναι οι πιο χρήσιμες για τις επιχειρήσεις.

Λόγω της νέας τεχνολογίας που χρησιμοποιεί μπορεί και προβλέπει τάσεις και συμπεριφορές ώστε να παίρνονται κάθε φορά οι σωστές αποφάσεις. Αναγνωρίζει

επίσης τις μορφές των δεδομένων και μ' αυτό τον τρόπο αποκαλύπτει την ύπαρξη ενός γεγονότος. Ταξινομεί τα δεδομένα και βελτιστοποιεί όλους τους πόρους που έχει στα χέρια της μία εταιρεία.

Ο σκοπός της εργασίας μας είναι να χρησιμοποιήσει αλγόριθμους εξόρυξης γνώσης για την έγκριση πιστωτικών καρτών βασιζόμενη σε στοιχεία όπως ηλικία, εισόδημα, πιστωτική ιστορία και ιδιοκτησία κατοικίας κλπ. Αξιολογήσαμε αρκετούς αλγόριθμους εξόρυξης γνώσης και επιλέξαμε τον καλύτερο, χρησιμοποιώντας το εργαλείο εξόρυξης δεδομένων WEKA. Εν συνεχεία, παρουσιάζουμε το εμπορικό πακέτο εξόρυξης γνώσης Knowledge Seeker που χρησιμοποιεί τον καλύτερο αλγόριθμο βάση των πειραμάτων μας.

Αυτή η εργασία δείχνει ότι μπορούμε να χρησιμοποιήσουμε εργαλεία εξόρυξης γνώσης για να ανακαλύψουμε χρήσιμη γνώση όπως οι εγγυητικοί κανόνες πιστωτικού ορίου για αιτούντες πιστωτικών καρτών.

1. Εισαγωγή

Τα Συστήματα Υποστήριξης Αποφάσεων (ΣΥΑ) αναπτύχθηκαν από δυο κύριες ερευνητικές δραστηριότητες – τις θεωρητικές μελέτες πάνω στην οργανωτική λήψη αποφάσεων που έγιναν στο Carnegie Institute of Technology (Simon, Cyert, March και άλλοι) στα τέλη του 1950 και στις αρχές του 1960 και στην τεχνική εργασία του MIT (Gerrity, Ness και άλλοι) το 1960. Τα συστήματα υποστήριξης αποφάσεων σχεδιάζονται για να βοηθούν αυτούς που λαμβάνουν αποφάσεις να τις πάρουν με βραχυπρόθεσμο ή μακροπρόθεσμο ορίζοντα. Θα πρέπει εξαρχής να τονιστεί ότι δεν αναλαμβάνουν να πάρουν αποφάσεις αφού αξιολογήσουν κάποια στοιχεία. Ο ρόλος τους είναι καθαρά συμβουλευτικός και υποστηρικτικός.

Κάποιος θα μπορούσε να αναλογιστεί τι γινόταν τα περασμένα χρόνια. Πως δηλαδή λαμβάνονταν τότε οι αποφάσεις χωρίς τη χρήση τέτοιων συστημάτων. Η απάντηση είναι ότι ασφαλώς λαμβάνονταν αποφάσεις απλά αυτό απαιτούσε πολύ χρόνο και πάρα πολύ κόπο γιατί απασχολούσε πολύ προσωπικό. Την εποχή που τα Πληροφοριακά Συστήματα δεν είχαν την σημερινή διάδοση η λήψη αποφάσεων απαιτούσε τη χειροκίνητη συλλογή πληροφοριών από τα τμήματα της εταιρίας και τη δημιουργία των λεγόμενων εκθέσεων (reports). Αυτά αξιολογούσαν οι μάνατζερ για να μπορέσουν να αποφασίσουν ποια πορεία θα ακολουθήσουν.

Γίνεται επομένως σαφές ότι τέτοιες διαδικασίες δεν θα μπορούσαν να επιζήσουν στο σημερινό επιχειρηματικό περιβάλλον που απαιτούνται γρήγορες ταχύτητες για να αντεπεξέλθει μια εταιρία στον ανταγωνισμό. Σήμερα οι πληροφορίες που έχουν στη διάθεση τους οι επιχειρήσεις κρατούνται σε βάσεις δεδομένων και είναι διαθέσιμες σε όλα τα τμήματα ανά πάσα στιγμή. Οι εφαρμογές υποστήριξης αποφάσεων έχουν πρόσβαση σε όλο το εύρος αυτών και χρησιμοποιούν κατάλληλα εργαλεία και σύγχρονες τεχνολογίες ώστε να παρουσιάζουν τα δεδομένα έγκαιρα, περιληπτικά και με γραφικό τρόπο στους αποφασίζοντες ώστε να μπορούν αυτοί να πάρουν κατευθυνόμενες από τη γνώση αποφάσεις.

Αυτή η παρουσίαση των δεδομένων δεν γίνεται αυτόματα. Προκύπτει μετά από μια έντονη διαλεκτική σχέση μεταξύ χρήστη και συστήματος. Ο χρήστης έχει στη διάθεση του διάφορες δυνατότητες τις οποίες εφαρμόζει πάνω στα δεδομένα και παίρνει τα ανάλογα αποτελέσματα. Έτσι, ανάλογα με τις δυνατότητες που προσφέρουν, έχουν προκύψει τρεις τύποι συστημάτων υποστήριξης αποφάσεων.

- Κατευθυνόμενα από μοντέλο (model-driven). Πρόκειται για ολοκληρωμένα συστήματα που έχουν τη δυνατότητα εκτέλεσης what-if σεναρίων καθώς και άλλων τύπων ανάλυσης..
- Κατευθυνόμενα από τα δεδομένα (data-driven). Επιτρέπουν στο χρήστη να εξάγει και να αναλύει χρήσιμη πληροφορία από μεγάλες βάσεις δεδομένων.
- Εξόρυξης Γνώσης (data mining). Εύρεση κρυμμένων τυποποιημένων μορφών (patterns) και κάποιων σχέσεων(relationships) σε μεγάλες βάσεις δεδομένων. Το αποτέλεσμα είναι η εξαγωγή κάποιων κανόνων ώστε να είναι δυνατή η πρόβλεψη μελλοντικών συμπεριφορών.

Στις δυο πρώτες κατηγορίες ανήκουν κυρίως τα σημερινά ΣΥΑ. Η εργασία θα εξερευνήσει την τρίτη κατηγορία και θα προσπαθήσει να δείξει τη δυναμική που μπορούν να προσδώσουν οι τεχνολογίες που σχετίζονται με την εξόρυξη γνώσης στην υποστήριξη αποφάσεων.

Η υποστήριξη αποφάσεων οπωσδήποτε ταυτίζεται με την πληροφορία. Αυτή είναι που βοηθά κάποιον να καταλήξει σε μια πολιτική, σε μια απόφαση, σε μια πορεία. Η πληροφορία προέρχεται από το πρόσφατο παρελθόν. Η μελέτη αυτού κρύβει μυστικά που αν αποκαλυφθούν θα αποτελέσουν σημαντικό οδηγό για το πώς πρέπει να κινηθεί μελλοντικά μια εταιρία ή ένας οργανισμός.

2. Τα συστήματα υποστήριξης αποφάσεων στις επιχειρήσεις – Η φιλοσοφία της επιχειρηματικής νοημοσύνης

Τα συστήματα υποστήριξης αποφάσεων αποτελούν το σημαντικότερο κομμάτι στην υποδομή ενός επιχειρησιακού πληροφοριακού συστήματος (σχήμα 2.1) γιατί παρέχουν τη δυνατότητα στις εταιρίες να μετατρέψουν μεγάλες ποσότητες επιχειρηματικών πληροφοριών σε χειροπιαστά και επικερδή αποτελέσματα. Παρόλα αυτά η συλλογή, συντήρηση και ανάλυση τεράστιων ποσοτήτων δεδομένων αποτελούν δύσκολες εργασίες και απαιτούν τεχνικές δεξιότητες, κόστος και οργανωτικές δεσμεύσεις.

Σχήμα 2.1. Η εικόνα μιας οποιασδήποτε επιχείρησης-Ο ρόλος των OLTP,ΣΥΑ

Ας αναλύσουμε λίγο το σχήμα για να εξηγήσουμε την κυκλική πορεία που παρουσιάζεται. Τα OLTP (On Line Transaction Processing Systems) συστήματα επιτρέπουν στους οργανισμούς να συγκεντρώνουν πληροφορίες για τις καθημερινές συναλλαγές τους με τους πελάτες (για παράδειγμα πληροφορίες από τις πωλήσεις). Οι OLTP εφαρμογές τυπικά αυτοματοποιούν δομημένες και επαναληπτικές διαδικασίες επεξεργασίας δεδομένων όπως είναι η εισαγωγή στοιχείων και οι τραπεζικές συναλλαγές.

Λεπτομερέστερα, «φρέσκα» (up-to-date) δεδομένα από διάφορα, ανεξάρτητα μεταξύ τους σημεία θα πρέπει να συγκεντρωθούν σε μια τοποθεσία πριν μπορέσουν οι αναλυτές να εξάγουν χρήσιμα συμπεράσματα. Οι μάνατζερ χρησιμοποιούν καθημερινά αυτό το σύνολο δεδομένων για να παίρνουν αποφάσεις για οτιδήποτε.

Για παράδειγμα ένας αγοραστής βιομηχανικού εξοπλισμού επιθυμεί να έχει μια λίστα με τους -μέχρι αυτή τη στιγμή- παγκόσμιους προμηθευτές καθώς και τις κάθε λεπτό ανανεώσιμες τιμές που δίνει κάθε πωλητής. Η άμεση εμφάνιση αυτών των δεδομένων στην επιφάνεια εργασίας του μαζί με όλες τις σχετικές πληροφορίες θα βοηθήσουν στην έγκαιρη και σωστή απόφαση αγοράς.

Φαίνεται λοιπόν και στο σχήμα 2.1. ότι οι απλές καθημερινές συναλλαγές δημιουργούν δεδομένα στα οποία στηρίζονται τα ΣΥΑ για να μπορέσουν να παρέχουν υπηρεσίες υποστήριξης αποφάσεων. Οι αποφάσεις αυτές θα οδηγήσουν ίσως την εταιρία σε μια νέα πορεία. Ο κύκλος επαναλαμβάνεται αφού και μετά τις όποιες αλλαγές θα συνεχιστούν οι συναλλαγές με τους πελάτες κ.ο.κ.

Το κλειδί της επιτυχίας αυτών των συστημάτων είναι το κατά πόσο η παρεχόμενη πληροφορία καλύπτει τις ανάγκες ενός λήπτη αποφάσεων. Είναι επομένως κατανοητό ότι ένα ΣΥΑ για έναν αντιπρόεδρο μιας τράπεζας ο οποίος ενδιαφέρεται για παράδειγμα για τους ρυθμούς χορήγησης δανείων είναι διαφορετικό από ένα σύστημα που θα είναι χρήσιμο σε ένα ταξιδιωτικό γραφείο το οποίο αναζητά τις χαμηλότερες τιμές σε ξενοδοχεία μιας περιοχής.

Στις επόμενες παραγράφους του κεφαλαίου αυτού περιγράφεται η νοητή διαδικασία η οποία θα ωθούσε πιθανόν μια εταιρία να εκμεταλλευτεί τα δεδομένα της για να λάβει αποφάσεις για το μέλλον.

2.1 Τα δεδομένα μιας εταιρίας είναι χρυσός

Το πλήθος των δεδομένων οδήγησε μερικές επιχειρήσεις στην δημιουργία αυτόνομων επιχειρησιακών μονάδων που ως σκοπό έχουν την τροφοδότηση της διοίκησης ή άλλων τομέων σχετικούς με την λήψη αποφάσεων, με πληροφορίες σχετικές με πελάτες, ανταγωνιστές κλπ. Η λειτουργία αυτή ονομάζεται Επιχειρηματική Νοημοσύνη (Business Intelligence). Η Επιχειρηματική Νοημοσύνη χρησιμοποιεί λογισμικό και τα δεδομένα της εταιρίας ώστε να υποστηρίξει την διαδικασία λήψης αποφάσεων. Σε ότι αφορά το λογισμικό οι διάφορες λύσεις προσφέρουν τη δυνατότητα εξερεύνησης και ανάλυσης των δεδομένων με σκοπό την αποκάλυψη διαφόρων ροπών και την απάντηση ζωτικών επιχειρηματικών ερωτήσεων.

2.2 Τι μπορεί να κάνει η Επιχειρηματική Νοημοσύνη για μια επιχείρηση;

Οι υπέρ-ανταγωνιστικές αγορές, οι αυξανόμενες απαιτήσεις των πελατών, οι ραγδαίες τεχνολογικές αλλαγές και η αλματώδης ανάπτυξη των επιχειρήσεων προκαλούν το ενδιαφέρον πολλών βιομηχανικών τομέων. Για να είναι δυνατή η βελτίωση της απόδοσης της εταιρίας και η ικανοποίηση των απαιτήσεων των πελατών θα πρέπει να αξιοποιηθούν καλύτερα οι κρίσιμες επιχειρηματικές πληροφορίες και μάλιστα ταχύτερα από ότι οι ανταγωνιστές. Αυτό εξηγεί γιατί η Επιχειρηματική Νοημοσύνη αποτελεί το στρατηγικό πλεονέκτημα για κάθε οργανισμό. Αυτό σημαίνει ότι εξοπλίζει το προσωπικό της εταιρίας με τα πλέον ενήμερα και ακριβή στοιχεία.

2.3 Σε ποιους απευθύνεται η Επιχειρηματική Νοημοσύνη;

Η ΕΝ είναι κατάλληλη για τον οικονομικό διευθυντή μιας νομικής εταιρίας όπως επίσης για τον διευθυντή μιας κλινικής όπως και για τον διευθυντή του τμήματος μάρκετινγκ προϊόντων σε μια εταιρία υψηλής τεχνολογίας. Απευθύνεται, με άλλα λόγια, σε ανθρώπους που χρειάζονται στοιχεία και πληροφορίες για να λάβουν αποφάσεις.

2.4 Πώς λειτουργεί;

Η Επιχειρηματική Νοημοσύνη ενοποιεί τις παραδοσιακά διακριτές λειτουργίες της πρόσβασης σε δεδομένα, της εξερεύνησης και της ανάλυσης. Αυτή η ενοποίηση παρέχει τα μέσα για την μετατροπή απομονωμένων νησίδων πληροφορίας σε περιεχτική γνώση που αποτελεί τη βάση των δυναμικών επιχειρηματικών αποφάσεων. Αυτή η γνώση επιτρέπει γρήγορη, μη χρονοβόρα δράση επειδή τα διαθέσιμα δεδομένα είναι εύκολο να εντοπιστούν.

2.5 Τι χρειάζεται μια εταιρία;

Οι περισσότερες εταιρίες έχουν ήδη στήσει τα θεμέλια της Επιχειρηματικής Νοημοσύνης – υπάρχουν βάσεις δεδομένων που περιέχουν στοιχεία σχετικά με την εταιρία, τους πελάτες της, τα προϊόντα της, το προσωπικό και τα τμήματα της. Το πρόβλημα είναι ότι οι πληροφορίες αυτές είναι συνήθως «θαμμένες» σε διαφορετικές πηγές και η διαδικασία ανάκτησης αυτών είναι χρονοβόρα και επίπονη.

Οι εφαρμογές ΕΝ προσφέρουν τον μηχανισμό της απλής και γρήγορης ενοποίησης επιχειρηματικών στοιχείων από ανόμοιες πηγές ώστε να είναι δυνατός ο διαμοιρασμός πληροφοριών στο εσωτερικό της εταιρίας μεταξύ υπαλλήλων αλλά και στο εξωτερικό με πελάτες, προμηθευτές. Με αυτό τον τρόπο μπορούμε να χρησιμοποιήσουμε αποτελεσματικά τα δεδομένα για ανάλυση, αναφορές ή για υποστήριξη αποφάσεων ώστε να βελτιώσουμε την λειτουργικότητα και για να χτίσουμε πιο επικερδής πελατειακές σχέσεις.

2.6 Εργαλεία Επιχειρηματικής Νοημοσύνης

Υπάρχει κατάλληλο λογισμικό που επιτρέπει την παρακολούθηση και χρήση μεγάλων ποσοτήτων δεδομένων. Οι τρεις παρακάτω τύποι εργαλείων είναι γνωστά ως Εργαλεία Επιχειρηματικής Νοημοσύνης:

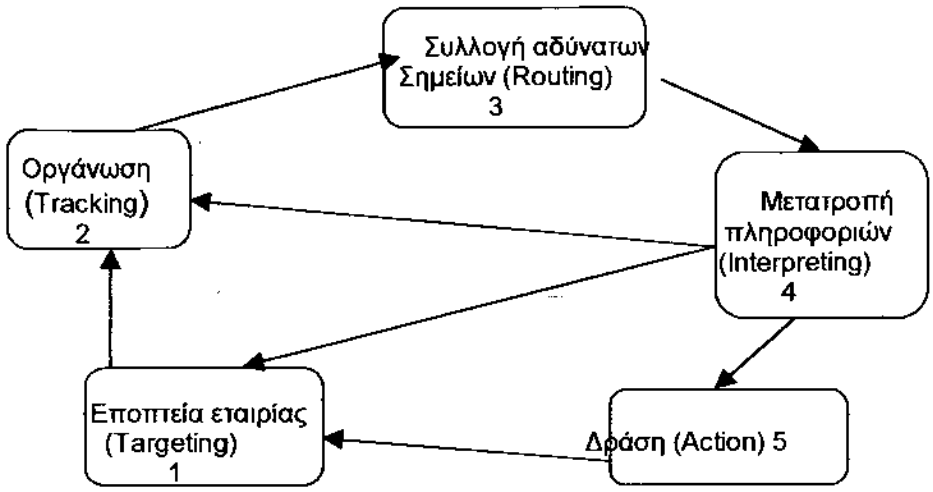
- **Λογισμικό Πολυδιάστατης Ανάλυσης** – γνωστό και ως OLAP (On- Line Analytical Processing) – .δίνει την δυνατότητα στον χρήστη να κοιτάξει τα δεδομένα από πολλές διαφορετικές διαστάσεις.
- **Εργαλεία ερωτήσεων (queries)** – Λογισμικό που επιτρέπει στον χρήστη να κάνει ερωτήσεις σχετικά με τυποποιημένες μορφές των δεδομένων ή ακόμα και για λεπτομέρειες.
- **Εργαλεία Εξόρυξης Γνώσης** – Λογισμικό που αυτομάτως αναζητά σημαντικές τυποποιημένες μορφές ή συσχετίσεις στα δεδομένα.

Ο Lesca[1] το 1994 περιέγραψε την Επιχειρηματική Νοημοσύνη ως μια κυκλική διαδικασία αποτελούμενη από πέντε φάσεις.

Στην πρώτη φάση "targeting", γίνεται η εποπτεία της εταιρίας με σκοπό να γίνει κάποια κατανομή των προτεραιοτήτων της. Η δεύτερη φάση "tracking", είναι περισσότερο οργανωτική αλλά περιλαμβάνει και τον εντοπισμό των πιο κρίσιμων αδύνατων σημείων. Στην τρίτη φάση "routing", γίνεται με επαναλαμβανόμενο τρόπο η συλλογή των αδύνατων σημείων της εταιρίας από έξω προς τα μέσα. Η τέταρτη φάση "interpreting" αποτελείται από την μετατροπή των πληροφοριών που συλλέχθηκαν σε χρήσιμη πληροφορία. Αν αυτή η φάση αποδώσει η διαδικασία προχωράει στη φάση 5. Διαφορετικά θα πρέπει η πληροφορία να επεξεργαστεί ξανά με πιο ιδιαίτερο τρόπο (επιστροφή στη φάση 2) αν αυτή είναι ανακριβής ή θα πρέπει να επανεξεταστεί ξανά ο στόχος (επιστροφή στη φάση 1) αν η πληροφορία είναι πολύ μεγάλη.

1 Lesca, H., 1994. "Veille strategique pour le manager strategique etat de la question et axes de recherché", Economies et societes, serie Science de Gestion 5 (20), 31 – 50

Από αυτές τις φάσεις, η πιο σημαντική και πιο δύσκολη είναι η τέταρτη (interpreting). Ανεπαρκείς μετατροπές θα οδηγήσουν σε λανθασμένη χρήση της Επιχειρηματικής Νοημοσύνης.



Σχήμα 2.2. Οι φάσεις της επιχειρηματικής νοημοσύνης

Η εργασία στο εξής θα περιγράψει το πώς μπορεί να επιτευχθεί η φιλοσοφία της Επιχειρηματικής Νοημοσύνης η οποία είναι και η καρδιά των συστημάτων υποστήριξης αποφάσεων. Πιο συγκεκριμένα θα αναλυθούν σε αρκετό βάθος οι τεχνολογίες εκείνες οι οποίες αποτελούν τα συστατικά στοιχεία των συστημάτων υποστήριξης αποφάσεων που ανήκουν στην τρίτη από τις κατηγορίες που παρουσιάστηκαν στην εισαγωγή.

Θα θέλαμε να τονίσουμε στον αναγνώστη την μέχρι τώρα ιεραρχία ώστε να κατανοήσει τον τρόπο με τον οποίο προσεγγίζουμε το θέμα. Η εργασία ασχολείται με τον χώρο των **Συστημάτων Υποστήριξης Αποφάσεων** τα οποία βασίζονται στην φιλοσοφία της **Επιχειρηματικής Νοημοσύνης** η οποία μπορεί να υλοποιηθεί και να επιτευχθεί και με χρήση τεχνολογιών **Αποθήκευσης Δεδομένων, Πολυδιάστατης Ανάλυσης** και **Εξόρυξης Γνώσης**.

3. Οι νέες τεχνολογίες στον χώρο των Βάσεων δεδομένων στην υπηρεσία των συστημάτων υποστήριξης αποφάσεων

Στις αρχές της δεκαετίας του 1990 τρία πανίσχυρα εργαλεία εμφανίστηκαν στην περιοχή της ανάπτυξης συστημάτων υποστήριξης αποφάσεων. Το πρώτο νέο εργαλείο ήταν οι αποθήκες δεδομένων. Τα δυο επόμενα που ακολούθησαν ήταν η επεξεργασία δεδομένων σε πραγματικό χρόνο (OLAP) και η εξόρυξη γνώσης. Ο παρακάτω πίνακας αναπαριστά τα χαρακτηριστικά τεχνολογικά βήματα της κάθε εποχής.

Εξελικτικό βήμα	Επιχειρηματική Ερώτηση	Βοηθητικές Τεχνολογίες	Κατασκευαστές προϊόντων	Χαρακτηριστικά
Συλλογή Δεδομένων (1960)	«Ποια ήταν τα συνολικά μου έσοδα τα τελευταία 5 χρόνια;»	Υπολογιστές, ταινίες, δισκέτες	IBM, CDC	Αναδρομική, στατική ανάκτηση δεδομένων
Πρόσβαση σε δεδομένα (1980)	«Ποιες ήταν οι πωλήσεις μου στην Πάτρα τον τελευταίο Μάρτιο;»	Σχεδιαστικές βάσεις δεδομένων (RDBMS), γλώσσα SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Αναδρομική, δυναμική ανάκτηση δεδομένων σε επίπεδο εγγραφής
Αποθήκες Δεδομένων & Υποστήριξη Αποφάσεων (1990)	«Ποιες ήταν οι πωλήσεις μου στην Πάτρα τον τελευταίο Μάρτιο; Βάση αυτών παίρνω αποφάσεις για την Αθήνα»	Επεξεργασία σε πραγματικό χρόνο (OLAP), πολυδιάστατες βάσεις δεδομένων, αποθήκες δεδομένων	Pilot, Comshare, Arbor, Cognos, Microstrategy	Αναδρομική, δυναμική ανάκτηση δεδομένων σε πολλαπλά επίπεδα
Εξόρυξη γνώσης (Ανερχόμενος τομέας σήμερα)	«Ποιες είναι οι πιθανές πωλήσεις του επόμενου μήνα στην Αθήνα;»	Ανεπτυγμένοι αλγόριθμοι, πολυεπεξεργαστικά υπολογιστικά συστήματα, μεγάλες βάσεις δεδομένων	Pilot, Lockheed, IBM, SGI	Προφητική ανάκτηση πληροφορίας

Θα εστιάσουμε την προσοχή μας στις δύο τελευταίες κατηγορίες και θα δείξουμε πως μπορούν να βασιστούν τα συστήματα υποστήριξης αποφάσεων πάνω στις τεχνολογίες αυτές.

Στον παραπάνω πίνακα φαίνεται ότι ο τομέας της εξόρυξης γνώσης είναι ανερχόμενος. Σε αυτό το αποτέλεσμα όμως δε φτάσαμε τυχαία. Οι επιχειρήσεις θεώρησαν πολύ σημαντικό να διαθέτουν απαντήσεις σε ερωτήσεις όπως αυτή που φαίνεται στον πίνακα. Αυτή ακριβώς η ανάγκη επέβαλε την ανάπτυξη συστημάτων υποστήριξης αποφάσεων βασισμένα σε εξόρυξη γνώσης.

Ας θεωρήσουμε μια εταιρία που κατασκευάζει υποδήματα και αναλύσουμε τις δικές της ανάγκες ώστε να δούμε αν αυτή χρειάζεται αυτές τις τεχνολογίες για να βρει απαντήσεις στα ερωτήματα της. Η εταιρία αυτή πουλάει τα προϊόντα της με δυο τρόπους. Είτε κατευθείαν στους πελάτες, είτε μέσω μεταπωλητών. Οι ειδικοί του τμήματος μάρκετινγκ της εταιρίας χρειάζεται να εξάγουν τις παρακάτω πληροφορίες από το «βουνό» πληροφοριών της εταιρίας:

- τις πέντε μεγαλύτερες αυξήσεις σε πωλήσεις στην κατηγορία νέων προϊόντων για τα περασμένα χρόνια,
- τις συνολικές πωλήσεις σε υποδήματα στη Νέα Υόρκη τον τελευταίο μήνα ανά προϊόν παραγωγής,
- τις πενήντα πόλεις με τον μεγαλύτερο αριθμό «καλών» πελατών, ένα εκατομμύριο πελάτες που αποτελούν τους πιο πιθανούς αγοραστές του νέου τύπου Walk-On-Air.

Για να βρεθούν οι απαντήσεις σε αυτά είναι σαφές ότι δεν αρκεί μια απλή ανάγνωση των δεδομένων που διαθέτει η εταιρία. Χρειάζεται μια διαφορετική προσέγγιση διαχείρισης των δεδομένων ώστε να προκύψουν πληροφορίες που ουσιαστικά είναι κρυμμένες.

Έτσι λοιπόν η ανάγκη θα την ωθούσε στην επέκταση ή ακόμα και αλλαγή του πιθανού υπάρχοντος συστήματος υποστήριξης αποφάσεων. Ο στόχος είναι να βρεθούν

απαντήσεις σε σύνθετα ερωτήματα. Την λύση μπορεί να την παρέχει η παρακάτω διαδικασία.

- δημιουργία αποθήκης δεδομένων
- Εφαρμογή OLAP πράξεων
- Εφαρμογή αλγορίθμων εξόρυξης γνώσης πάνω στην αποθήκη δεδομένων

Αυτή τη διαδικασία ακολουθεί μια εταιρία που επιθυμεί να προσδώσει στο σύστημα υποστήριξης αποφάσεων της δυνατότητες εξόρυξης γνώσης. Φαίνεται λοιπόν ότι η πολυπόθητη εξόρυξη γνώσης δεν μπορεί να εφαρμοστεί αμέσως στα δεδομένα της εταιρίας.

Αποθήκη δεδομένων(data warehouse) : Περιλαμβάνει δεδομένα που συσσωρεύονται εκεί από τις βάσεις δεδομένων της επιχείρησης και συχνά το μέγεθος τους φτάνει τα gigabytes ή ακόμα και terabytes. Τυπικά η αποθήκη δεδομένων συντηρείται ξεχωριστά από τις βάσεις δεδομένων του οργανισμού γιατί οι απαιτήσεις των εφαρμογών ανάλυσης δεν συμπίπτουν με τις δυνατότητες των βάσεων δεδομένων. Οι αποθήκες δεδομένων εξυπηρετούν τα συστήματα υποστήριξης αποφάσεων γιατί παρέχουν ιστορικά, ομαδοποιημένα και συγκεντρωτικά δεδομένα αντί για λεπτομερείς εγγραφές.

Υπάρχουν εμφανείς διαφορές μεταξύ των κλασικών βάσεων δεδομένων και των αποθηκών δεδομένων. Στην αποθήκη δεδομένων καταλήγουν κατάλληλα επεξεργασμένα δεδομένα των επιμέρους βάσεων δεδομένων, διαφοροποιώντας την έτσι ως προς το περιεχόμενο των πληροφοριών, πολλές φορές μάλιστα αυτά τα δεδομένα αποτελούν δομημένες πληροφορίες και όχι απλά μια καταγραφή απλών στοιχείων –πράγματα που εμφανίζονται στις απλές βάσεις δεδομένων.

Επειδή όμως η κατασκευή μιας αποθήκης δεδομένων μπορεί να διαρκέσει πολλά χρόνια, μερικοί οργανισμοί αντί αυτών κτίζουν τα λεγόμενα data marts που περιλαμβάνουν πληροφορίες για κάποια συγκεκριμένα τμήματα. Έτσι μπορεί ένα data mart μπορεί να ανήκει στο τμήμα Μάρκετινγκ, ένα άλλο στο Λογιστήριο. Όλα αυτά μαζί

αποτελούν την κεντρική αποθήκη δεδομένων. Μια ακόμα σημαντική παράμετρος είναι και ο τρόπος υλοποίησης της αποθήκης δεδομένων. Αν δηλαδή θα βασίζεται στο σχεσιακό ή το πολυδιάστατο μοντέλο(ROLAP εναντίον MOLAP). Η επιλογή παίζει ρόλο στην απόδοση της αποθήκης δεδομένων όχι όμως και στις δυνατότητες που αυτή μπορεί να προσφέρει.

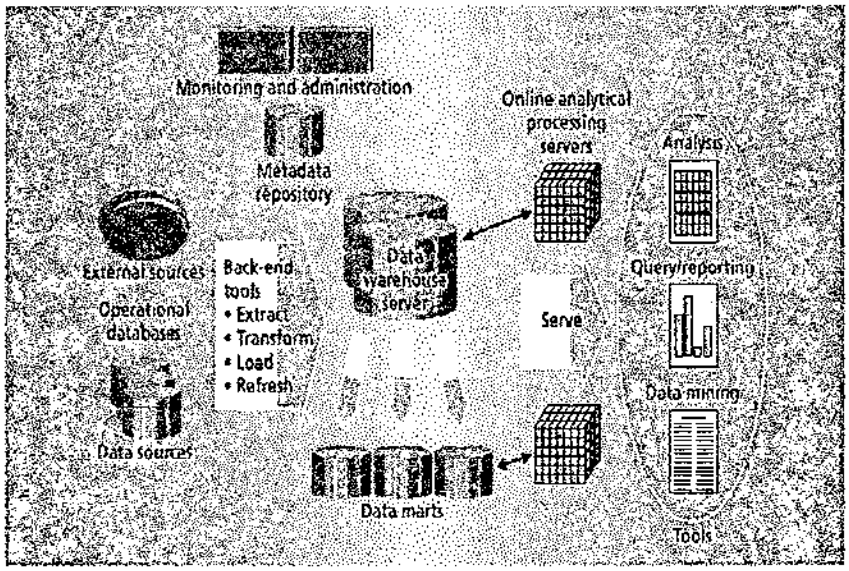
OLAP : Οι αποθήκες δεδομένων παρέχουν τη δυνατότητα για Συνεχή Αναλυτική Επεξεργασία (On-Line Analytical Processing – OLAP) των δεδομένων περιέχοντας ιστορικά και συγκεντρωτικά δεδομένα χρήσιμα για υποστήριξη αποφάσεων.

Η ανάπτυξη και η εξέλιξη της συνεχούς αναλυτικής διαδικασίας (OLAP) οφείλεται κυρίως σε δύο λόγους: Στη ραγδαία αύξηση των ποσοτήτων των δεδομένων και την ταυτόχρονη ανάγκη για ταχεία ανάλυση τους. Με τα εργαλεία OLAP παρέχονται περισσότερες δυνατότητες από αυτές που οι απλές ερωτήσεις και οι αναφορές(reports) μπορούν να δώσουν. Βοηθούν τους αναλυτές, τους μάνατζερ και τα υψηλόβαθμα στελέχη των επιχειρήσεων στη ταχεία πρόσβαση και πολυδιάστατη επεξεργασία των δεδομένων τους με σκοπό τη παρουσίαση και τη λύση των προβλημάτων της επιχείρησης στις πραγματικές τους διαστάσεις.

Βοηθούν τον χρήστη να δημιουργεί αναλύσεις μέσα από πολλαπλές ερωτήσεις του τύπου "what-if" και έτσι να μοντελοποιεί το σενάριο του. Οι εφαρμογές OLAP έχουν γίνει συνώνυμα με τη πολυδιάστατη παρουσίαση των δεδομένων. Αυτή η πολυδιάστατη παρουσίαση ενισχύεται και υποστηρίζεται από τις πολυδιάστατες βάσεις δεδομένων παρέχοντας έτσι στις OLAP εφαρμογές τη βάση για τον υπολογισμό και την ανάλυση των δεδομένων. Η ανάγκη για πολυδιάστατη ανάλυση αναδεικνύει τις αποθήκες δεδομένων ως την κύρια πηγή άντλησης πληροφοριών.

Εξόρυξη Γνώσης: Αφού λοιπόν έχουμε δημιουργήσει την αποθήκη δεδομένων και έχουμε εκμεταλλευτεί τις δυνατότητες που προσφέρει η τεχνολογία OLAP, μπορούμε να προχωρήσουμε ακόμα ένα βήμα και να ψάξουμε με εξελιγμένους αλγόριθμους για

κρυμμένη πληροφορία που βρίσκεται στην αποθήκη δεδομένων. Η εξόρυξη γνώσης από αποθήκη δεδομένων είναι ότι πιο σύγχρονο χρησιμοποιούν οι αναλυτές σήμερα.



Σχήμα 2.3. Η αρχιτεκτονική ενός συστήματος υποστήριξης Αποφάσεων που αποτελείται από τρία μέρη: έναν data warehouse server,εργαλεία ανάλυσης και εξόρυξης γνώσης καθώς και back – end εργαλεία για την αποθήκη δεδομένων

Τώρα λοπόν που υπάρχει μια σχετική εξοικείωση με τους μέχρι τώρα άγνωστους όρους μπορούμε να δούμε καλύτερα το πώς μπορεί ένα ΣΥΑ να στηρίζεται σε μια αποθήκη δεδομένων που μπορεί να αποτελείται από πολλά data marts και η οποία γεμίζει με στοιχεία που προέρχονται μετά από επεξεργασία των βάσεων δεδομένων της εταιρίας ή από άλλες εξωτερικές πηγές(από το internet για παράδειγμα). Θέματα που έχουν να κάνουν με φυσική διαχείριση της αποθήκης και των μεταδεδομένων της δεν θα μας απασχολήσουν. Παρατηρούμε όμως ότι η αποθήκη δεδομένων μπορεί να εξυπηρετήσει τόσο την εφαρμογή OLAP πράξεων καθώς επίσης και την εξόρυξη γνώσης.

Η παραπάνω συνοπτική παρουσίαση ασφαλώς δεν έχει αγγίξει βαθύτερα θέματα των τεχνολογιών αυτών. Έγινε μια πρώτη εισαγωγή για να μπορεί ο αναγνώστης να

παρακολουθήσει την ακόλουθη ιεραρχική(από χαμηλό σε υψηλότερο επίπεδο) υλοποίηση ενός ΣΥΑ.

Για να γίνουμε πιο συγκεκριμένοι, θα προχωρήσουμε με **θέματα σχεδίασης** και **εισαγωγής στοιχείων** σε μια **αποθήκη δεδομένων**. Θα αναφερθούμε σε θέματα εφαρμογής **OLAP πράξεων** πάνω στην αποθήκη αφού αυτές εξυπηρετούν την υποστήριξη αποφάσεων και θα φθάσουμε να δούμε **αλγόριθμους εξόρυξης γνώσης** όπου έχουν την ικανότητα να φανερώνουν σχέσεις και εξαρτήσεις που δεν είναι ορατές.

4. Το μέσο: οι Αποθήκες δεδομένων

Από τα μέσα της δεκαετίας του '70, η αλματώδης παραγωγή πολύ ισχυρών συστημάτων διαχείρισης βάσεων δεδομένων βοήθησε στην ανάπτυξη πληροφοριακών συστημάτων που καλύπτουν τις λειτουργικές ανάγκες οργανισμών και επιχειρήσεων. Τα μεγαλύτερα και ισχυρότερα συστήματα αναπτύχθηκαν με στόχο τον αυτοματισμό βασικών αναγκών των οργανισμών όπως η διεκπεραίωση των τραπεζικών εργασιών και τα λογιστικά συστήματα. Η λειτουργία αυτών των πληροφοριακών συστημάτων είναι πλέον κρίσιμη και πολύτιμη για τη ζωή των οργανισμών στους οποίους έχουν εγκατασταθεί, η δε βάση δεδομένων ενός τέτοιου συστήματος αποτελεί τον πυρήνα τους. Η ορθή σχεδίαση, ανάπτυξη και λειτουργία της βάσης είναι ο σημαντικότερος παράγοντας για την επιτυχία ενός πληροφοριακού συστήματος. Τα συστήματα αυτά παρέχουν τη δυνατότητα επεξεργασίας μεγάλου αριθμού δοσοληψιών που διαχειρίζονται τα δεδομένα του οργανισμού (OLTP). Ένα άλλο είδος πληροφοριακών συστημάτων που αναπτύσσονται στους οργανισμούς είναι τα συστήματα υποστήριξης αποφάσεων που σκοπό έχουν να βοηθήσουν τα στελέχη των οργανισμών να σχεδιάσουν τις δραστηριότητές του. Η επιτυχία των συστημάτων αυτών είναι επίσης βασικός παράγοντας επιτυχίας του οργανισμού. Μία βασική απαίτηση των συστημάτων υποστήριξης αποφάσεων είναι η αποδοτική πρόσβαση στα δεδομένα των συστημάτων αυτοματισμού. Το πρόβλημα που προκύπτει, όμως, είναι ότι τα συστήματα αυτοματισμού έχουν ήδη πολύ σοβαρό υπολογιστικό φορτίο από μόνα τους και επιπλέον, είναι σχεδιασμένα για την εκτέλεση διαφορετικών λειτουργιών.

Ένας τηλεπικοινωνιακός οργανισμός, για παράδειγμα, συνήθως διαθέτει ένα μεγάλο πληροφοριακό σύστημα ελέγχου του τηλεφωνικού δικτύου του. Αυτό το σύστημα ελέγχει την ομαλή λειτουργία του δικτύου και παράλληλα των παροχή υπηρεσιών και την χρέωση των συνδρομητών του. Η βάση δεδομένων του συστήματος περιέχει όλα τα δεδομένα των παραπάνω εργασιών. Είναι σαφές ότι αυτό το σύστημα λειτουργεί συνεχώς (24 ώρες ημερησίως) με μεγάλο όγκο δοσοληψιών (transactions) να εξυπηρετούνται στη βάση δεδομένων. Από αυτή τη βάση θα πρέπει να αντλήσει και ένα

σύστημα υποστήριξης αποφάσεων τα απαραίτητα δεδομένα, για να μπορέσει να βοηθήσει στο σχεδιασμό της λειτουργίας του οργανισμού. Μελετώντας λίγο πιο προσεκτικά την περίπτωση αυτή, θα δούμε ότι είναι πρακτικά αδύνατο το σύστημα ελέγχου του δικτύου και το σύστημα υποστήριξης αποφάσεων να λειτουργούν, χρησιμοποιώντας την ίδια βάση δεδομένων. Διάφορα προβλήματα κάνουν αδύνατη την εφαρμογή αυτού του σεναρίου. Τα κυριότερα από αυτά τα προβλήματα είναι τα παρακάτω:

1. Τα δύο συστήματα αναπτύχθηκαν πιθανότατα από διαφορετικούς ανθρώπους και κυρίως με τη χρήση διαφορετικών τεχνολογιών. Είναι πιθανό η τεχνολογία του συστήματος αποφάσεων να αδυνατεί να επιτρέψει άμεση πρόσβαση (on-line) στη βάση δεδομένων του συστήματος ελέγχου του δικτύου. Πολύ συχνά, σε μεγάλα συστήματα, όπως στην προκειμένη περίπτωση, το σύστημα ελέγχου του δικτύου έχει αναπτυχθεί με τη χρήση παρωχημένης τεχνολογίας, όπως, για παράδειγμα, αρχεία COBOL. Εφαρμογές που χρησιμοποιούν μοντέρνα τεχνολογία αντιμετωπίζουν προβλήματα στο να διαχειριστούν πληροφορία που προέρχεται από μια βάση δεδομένων παλαιάς τεχνολογίας.

2. Η βάση δεδομένων του συστήματος ελέγχου του δικτύου σχεδιάστηκε με βάση αποκλειστικά τις απαιτήσεις αυτής της εφαρμογής. Βασικό χαρακτηριστικό σε εφαρμογές αυτού του είδους είναι η όσο το δυνατό αποδοτικότερη ικανοποίηση μικρών δοσοληψιών που εισάγουν ή τροποποιούν πολύ μικρό αριθμό εγγραφών της βάσης. Μία τυπική δοσοληψία που θα αφορούσε τη χρέωση μίας υπεραστικής συνδιάλεξης θα εισήγαγε μία εγγραφή με τον κωδικό του συνδρομητή και τη διάρκεια της συνδιάλεξης. Στη σχεδίαση μιας τέτοιας βάσης δεδομένων, με την εφαρμογή των κανόνων κανονικοποίησης, καταλήγουμε σε μεγάλο αριθμό από πίνακες που ο κάθε ένας έχει περιορισμένο αριθμό πεδίων. Σε αντίθεση με τα παραπάνω, μία εφαρμογή που αντλεί στοιχεία λειτουργίας του δικτύου για λόγους ανάλυσης και λήψης αποφάσεων, δεν κάνει καμία αλλαγή στη βάση του δικτύου αλλά απαιτεί αποδοτική απόκριση από το σύστημα στις

ερωτήσεις που θέτει. Αυτές οι ερωτήσεις συνήθως απαιτούν πρόσβαση σε μεγάλο αριθμό δεδομένων, θέτοντας διαφορετικούς κανόνες σχεδίασης της βάσης δεδομένων του συστήματος. Για μία ερώτηση σχετική με τη στρατηγική του οργανισμού, που θα είχε πρόσβαση σε μεγάλο αριθμό δεδομένων το κόστος σε μία βάση με πολλούς πίνακες θα ήταν σημαντικό, καθώς θα έπρεπε να εκτελεστεί μεγάλο αριθμός από πράξεις join μεταξύ των πινάκων αυτών.

3. Κάθε σύστημα υποστήριξης αποφάσεων εκτελεί μεγάλο αριθμό ερωτήσεων, θα δεσμεύσει μεγάλο αριθμό πόρων του συστήματος διαχείρισης της βάσης δεδομένων με αποτέλεσμα να μειώσει την απόδοση του συστήματος ελέγχου του δικτύου. Για παράδειγμα, μία ερώτηση σχετική με τις χρεώσεις των πελατών του δικτύου, που απαιτεί κάποιο χρονικό διάστημα για να εκτελεστεί, θα κλείδωνε τον πίνακα με τις χρεώσεις των πελατών εμποδίζοντας οποιαδήποτε μεταβολή από το σύστημα ελέγχου (π.χ. μια νέα χρέωση). Από τα παραπάνω γίνεται σαφές ότι είναι εξαιρετικά δυσχερής η χρήση των βάσεων δεδομένων των πληροφοριακών συστημάτων των οργανισμών από τα συστήματα υποστήριξης αποφάσεων. Όμως, η αποδοτική χρήση των συστημάτων υποστήριξης αποφάσεων απαιτεί όπως προαναφέρθηκε, πρόσβαση σε αυτά τα δεδομένα. Η εισαγωγή των αποθηκών δεδομένων είναι η λύση στο κρίσιμο αυτό πρόβλημα.

4.1 Σχεδίαση Αποθηκών Δεδομένων

Το μεγαλύτερο ποσοστό των επιχειρήσεων επιλέγουν τον σχεδιασμό και την εφαρμογή κεντρικών και ανεξάρτητων αποθηκών δεδομένων. Οι σχεδιαστές των συστημάτων αποσκοπούν σε:

- Μία απλή και ανεξάρτητη εφαρμογή που να διαχειρίζεται εύκολα μεγάλο όγκο δεδομένων
- Συγκεντρωτική παρουσίαση των δεδομένων της επιχείρησης στους αποφασίζοντες, συλλέγοντας τα διάσπαρτα δεδομένα από όλες τις πτυχές της.

- Διαχωρισμό ανά κατηγορία και θέμα των δεδομένων (πολλές φορές σε διαφορετικές αποθήκες δεδομένων), αφού πολλές φορές η συγκέντρωση σε μία και μόνη αποθήκη δυσχεραίνει την πρόσβαση.

Είναι αυτονόητο ότι μία εφαρμογή αποθήκης δεδομένων -η ακόμα περισσότερο μία έτοιμη εμπορική εφαρμογή- δεν ταιριάζει σε μία επιχείρηση. Για αυτό τον λόγο και τις περισσότερες φορές ο σχεδιασμός από την αρχή είναι επιβεβλημένος. Πολλές εταιρίες παρέχουν την δυνατότητα προσαρμογής ήδη υπαρχόντων εφαρμογών, ανάλογα με τις απαιτήσεις της κάθε επιχείρησης.

Σε κάθε περίπτωση όμως, οι παράγοντες που θα πρέπει να λαμβάνονται υπόψη είναι:

1. Πολλές επιχειρήσεις γνωρίζουν την αναγκαιότητα μιας αποθήκης δεδομένων, άλλοι δεν είναι σε θέση να θέσουν τις προτεραιότητες και τις προδιαγραφές, πράγματα που μόνο οι εταιρίες εφαρμογών μπορούν να ξέρουν και να θέσουν. Σαν τέτοιες χαρακτηρίζονται το μέγεθος, η εγκατάσταση, η συχνότητα χρήσης και η συντήρηση. Για πολλές εταιρίες, όμως, μία ολοκληρωμένη μελέτη και εφαρμογή κρίνεται αντιοικονομική.

2. Η κατανόηση του ήδη υπάρχοντος διαχειριστικού συστήματος είναι ένας σημαντικός παράγοντας. Πάνω σε αυτό το σύστημα θα στηριχτεί η όποια εφαρμογή αποθήκης δεδομένων. Έτσι, πρώτος στόχος είναι να διασαφηνιστεί η ακριβής προέλευση των δεδομένων. Έτσι, πρώτος στόχος είναι να διασαφηνιστεί η ακριβής προέλευση των δεδομένων και της δομής τους, ώστε να επιλεχθούν τα κατάλληλα εργαλεία για την εφαρμογή.

3. Η ανάγκη για μετακίνηση των δεδομένων είναι ένας άλλος παράγοντας που θα πρέπει να εξεταστεί. Πολλές φορές κρίνεται συμφέρουσα και οικονομική η παραμονή των δεδομένων στις αρχικές βάσεις δεδομένων. Τα κριτήρια για την μετακίνηση ή όχι των δεδομένων είναι:

- Η ποιότητα των δεδομένων
- Το μέγεθος και η χρηστικότητά τους
- Η δομή τους
- Η δυνατότητα να εφαρμοστούν στη νέα εφαρμογή
- Η ευκολία πρόσβασης
- Η αναγκαιότητα τους στο ήδη υπάρχον σύστημα

4. Ένα πολύ σημαντικό σημείο στη μελέτη και τον σχεδιασμό είναι και ο τελικός αποδέκτης των μετακινούμενων δεδομένων. Πριν καν μελετηθεί η μετακίνηση των δεδομένων, *πρέπει* να γίνει ουσιαστικά ο σχεδιασμός της δομής των δεδομένων που θα συμπεριληφθούν στην αποθήκη δεδομένων.

5. Για να γίνει περισσότερο κατανοητή η χρήση της αποθήκης δεδομένων πρέπει να εντάσσεται σε ένα πληροφοριακό σύστημα διοίκησης ή σε ένα σύστημα υποστήριξης αποφάσεων. Τα ήδη υπάρχοντα συστήματα είναι ένας παράγοντας που πρέπει να λαμβάνεται υπόψη κατά τον σχεδιασμό μιας αποθήκης δεδομένων.

6. Ο βαθμός μετατροπής των δεδομένων, καθώς και τα εργαλεία που απαιτούνται για αυτό.

7. Η επιλογή των κατάλληλων εργαλείων (στην περίπτωση όπου δεν υπάρχει εφαρμογή ΣΥΑ) για την επεξεργασία των δεδομένων της αποθήκης δεδομένων. Τις περισσότερες των περιπτώσεων απαιτούνται διαφορετικά εργαλεία, ανάλογα σε τι είδους χρήστη απευθύνεται (για παράδειγμα σε έμπειρους χρήστες αποθηκών δεδομένων, βελτιωτές εφαρμογών, υψηλόβαθμα στελέχη της επιχείρησης κ.α.).

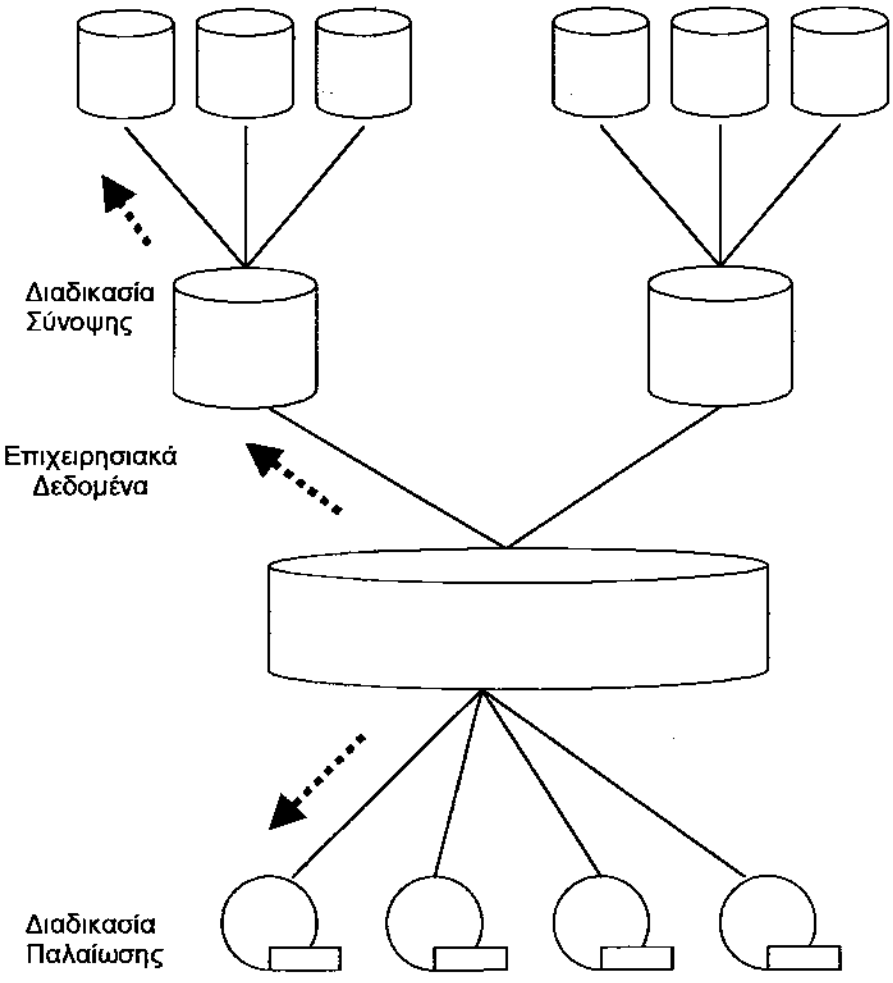
8. Καθορισμός των κανόνων χρήσης και λειτουργίας της αποθήκης δεδομένων. Καθορισμός των χρηστών με πρόσβαση μερική ή ολική σε αυτή. Οι παραπάνω

παράγοντες δεν είναι οι μοναδικοί για τον σχεδιασμό μιας αποθήκης δεδομένων. Η ευελιξία που δίνεται στον σχεδιαστή μέσω των διαφόρων τεχνολογικών εφαρμογών που ποικίλουν πλέον στην αγορά, καθώς και η πληθώρα των λογισμικών που είναι διαθέσιμα για τον σχεδιασμό βάσεων δεδομένων, έχουν ως αποτέλεσμα την εμφάνιση πολλών και διαφορετικών αποθηκών δεδομένων.

4.2 Λειτουργία της αποθήκης δεδομένων

Στο σχήμα 4 παρουσιάζεται η ροή των δεδομένων από την αρχική πηγή τους μέχρι τον χρηστή, συμπεριλαμβανομένων και των εφαρμογών που παρεμβάλλονται. Λόγω των ετερογενών πηγών των δεδομένων (διαφορετικές διαχειριστικές βάσεις δεδομένων), και λοιπόν προβλημάτων που προκύπτουν σε τέτοια ανομοιογενή συστήματα, παρεμβάλλονται μηχανισμοί που τα αποκαθιστούν, μετατρέποντας τα και παραδίδοντας τα στις αποθηκευτικές βάσεις δεδομένων. Αυτός ο μηχανισμός βασίζεται σε προεπιλεγμένο μοντέλο (σε αυτό βασίζεται ουσιαστικά και ο σχεδιασμός των αποθηκών δεδομένων). Κύριο συστατικό – περιγράφει το μοντέλο ενώ παράλληλα διευκρινίζει τα στοιχεία των δεδομένων-είναι το metadata.

Σχήμα 4.1. Ροή δεδομένων μέσα σε αποθήκη δεδομένων



Τα περισσότερα δεδομένα εισέρχονται στην αποθήκη δεδομένων απ' ευθείας από τις διαχειριστικές βάσεις δεδομένων. Έτσι, τα δεδομένα διακρίνονται σε:

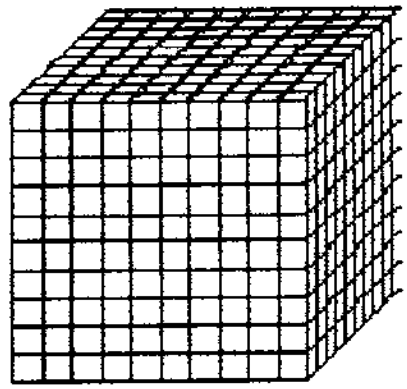
- Διαγραφέντα (τα δεδομένα προς διαγραφή, μικρή χρηστικότητα)
- Συνοπτικά (είναι δεδομένα που χρησιμοποιούνται σε μεγαλύτερη συχνότητα)
- Δεδομένα-αρχεία.

Η διαδικασία "παλαίωσης" (aging process) σε μία βάση δεδομένων μετατρέπει τα δεδομένα ανάλογα με την ηλικία τους από σύγχρονα δεδομένα (current data) σε παλαιά (older data) και για αυτό τον λόγο, τις περισσότερες φορές τα αποθέτει σε αποθηκευτικούς δίσκους. Η διαδικασία σύνοψης (summarized process) χρησιμοποιεί προσεκτικά επιλεγμένα δεδομένα με σκοπό να τα μετατρέψει σε συνοπτικά δεδομένα (lightly summarized data), είτε πολύ συνοπτικά δεδομένα (hightly summarized data).

Όσο πιο συγκεντρωτικά και συνοπτικά είναι τα δεδομένα σε μία αποθήκη δεδομένων, τόσο ταχύτερη και αποτελεσματικότερη είναι η πρόσβαση για τη χρήση τους.

4.3 Η αποθήκη δεδομένων όπως την γνωρίζουν οι αναλυτές

Όταν κάποιος ασχολείται με την αποθήκη δεδομένων μιας εταιρίας δεν βλέπει μπροστά του πίνακες συνδεδεμένους μεταξύ τους, όπως θα έκανε κάποιος ειδικός επί των βάσεων δεδομένων. Ανεξάρτητα με το πώς έχει γίνει η φυσική αποθήκευση των δεδομένων οι αναλυτές γνωρίζουν ότι έχουν μπροστά τους ένα κύβο με πολλές ίσως διαστάσεις, δηλαδή όπως λένε ένα υπερκύβο.



Σχήμα 4.2. Ένας κύβος. Τα κελιά του αντιστοιχούν σε μία τιμή από κάθε διάσταση του.

Κάθε διάσταση ουσιαστικά αντιπροσωπεύει και μια παράμετρο που λαμβάνεται υπόψη κατά την ανάλυση. Έτσι για παράδειγμα ο παραπάνω κύβος θα μπορούσε να έχει τις εξής τρεις διαστάσεις: προϊόν, χρόνος και τοποθεσία αν ασχολούμασταν με την αποθήκη δεδομένων ενός super-market. Στις διαστάσεις αυτές κυριαρχεί η ιεραρχία. .δηλαδή στη διάσταση χρόνος μπορούμε να συναντήσουμε έτος – τρίμηνο – μήνας – ημέρομηνία.

Τα κελιά του υπερκύβου περιέχουν συγκεντρωτικά δεδομένα όπως έχει ήδη αναφερθεί. Αυτό σημαίνει πως εκεί υπάρχουν μέσοι όροι, σύνολα, μέγιστα και ελάχιστα κτλ. Το κάθε κελί αντιστοιχεί σε ένα συγκεκριμένο τόπο, για ένα συγκεκριμένο χρονικό διάστημα και για κάποιο από τα προϊόντα του super-market. Τα κελιά μπορούν να γεμίσουν με την εκτέλεση κατάλληλων SQL ερωτήσεων, συγκεκριμένα SQL aggragated ερωτήσεων(SELECT AVG, COUNT, MIN, MAX), προς τη βάση δεδομένων. Οι δυνατότητες υποστήριξης αποφάσεων που προσφέρουν οι αποθήκες δεδομένων είναι μεγάλες. Πολλά στελέχη επιχειρήσεων όταν βρίσκονται μακρυά από την έδρα της εταιρίας τους, άρα και μακρυά από τις βάσεις δεδομένων, φροντίζουν να έχουν στο φορητό υπολογιστή τους συγκεντρωτικά, ιστορικά στοιχεία για την τελευταία εβδομάδα για παράδειγμα ώστε να έχουν εικόνα της επιχείρησης. Με άλλα λόγια μια αποθήκη δεδομένων είναι εύκολα μεταφέρσιμη αφού το μέγεθος της μπορεί να καθοριστεί μιας και ο χρήστης επιλέγει σε τι βάθος των ενδιαφέρουν τα στοιχεία.

Θα μπορούσε κάποιος να αναρωτηθεί; Τώρα που έχουμε την αποθήκη δεδομένων πως θα μπορέσουμε να βγάλουμε συμπεράσματα. Υπάρχουν απλές πράξεις που μπορούν να γίνουν όπως να αναζητηθούν ο συνολικός αριθμός προϊόντων που πωλήθηκαν το μήνα Μάρτιο, στην κατηγορία ειδών διατροφής στην περιοχή της Πάτρας. Μπορούμε όμως να εφαρμόσουμε ποιο σύνθετες πράξεις όπως αυτές που περιγράφονται στην επόμενη ενότητα με βαθύτερο σκοπό πάντα να πάρουμε την σωστή αποφάση.

5. Το εργαλείο: η Εξόρυξη Γνώσης

Είδαμε λοιπόν με πιο τρόπο μπορούν οι σχετικές με τις αποθήκες δεδομένων τεχνολογίες να παίξουν πρωταγωνιστικό ρόλο στο χώρο της υποστήριξης αποφάσεων. Υπάρχουν όμως κατηγορίες αποφάσεων που για να ληφθούν σωστά απαιτούν να έχει ο αποφασίζων στη διάθεση του απαντήσεις και πληροφορίες που δεν είναι εύκολο να προκύψουν με τις τεχνολογίες που μέχρι τώρα αναλύθηκαν. Αυτές τις κατηγορίες αποφάσεων καλείται να υποστηρίξει η εξόρυξη γνώσης.

Η εξόρυξη γνώσης, δηλαδή η διαδικασία εφαρμογής μεθόδων ανάλυσης σε μεγάλο όγκο δεδομένων, είναι μια πολύ ισχυρή νέα τεχνολογία που μπορεί να βοηθήσει τις εταιρίες να εστιάσουν μόνο στα πιο σημαντικά δεδομένα των αποθηκών δεδομένων τους. Τα εργαλεία εξόρυξης γνώσης δίνουν τη δυνατότητα στο χρήστη να προφητεύει μελλοντικές συμπεριφορές και ροπές επιτρέποντας έτσι στις επιχειρήσεις να παίρνουν κατευθυνόμενες από τη γνώση αποφάσεις. Το αποτέλεσμα είναι να μπορούν τα εργαλεία αυτά να απαντήσουν σε επιχειρηματικές ερωτήσεις που παραδοσιακά απαιτούσαν πολύ χρόνο ανάλυσης. Οι περισσότερες εταιρίες ήδη συλλέγουν και επεξεργάζονται τεράστιες ποσότητες δεδομένων. Οι τεχνικές εξόρυξης γνώσης μπορούν να αναπτυχθούν γρήγορα χωρίς να χρειάζονται αλλαγές στην υλικοτεχνική υποδομή και ως σκοπό έχουν την αξιοποίηση των πηγών πληροφοριών.

5.1 Τα αποτελέσματα της εξόρυξης γνώσης

Η εφαρμογή των μεθόδων εξόρυξης γνώσης αποσκοπεί στην ανακάλυψη πληροφοριών που είναι πολύ χρήσιμες για τις επιχειρήσεις. Πληροφορίες για **συσχετίσεις** όπως «όταν ένας πελάτης αγοράζει βίντεο τότε αγοράζει επίσης κάποια άλλη ηλεκτρονική συσκευή» ή για **τυποποιημένες μορφές** όπως «ο πελάτης που θα ψωνίσει περισσότερο από δύο φορές σε περίοδο εκπτώσεων είναι πιθανό να αγοράσει τουλάχιστο μία φορά κατά τη διάρκεια των Χριστουγέννων» αποτελούν πραγματικό θησαυρό για τους διοικούντες που μπορούν έτσι να αποφασίσουν για διάφορα θέματα λειτουργίας της επιχείρησης τους, όπως είναι το ωράριο, το ύψος και η διάρκεια των

εκπτώσεων και η τοποθέτηση των πραγμάτων μέσα στα καταστήματα αν βέβαια μιλάμε για εμπορικού τύπου επιχειρήσεις. Τέτοιες πληροφορίες μπορούν επίσης να χρησιμοποιηθούν για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων, για τον συνδυασμό διαφόρων πραγμάτων(βίντεο-ηλεκτρική σκούπα για παράδειγμα) στις διαφημίσεις ή για τη σχεδίαση ανάλογα την εποχή διαφορετικών στρατηγικών μάρκετινγκ.

5.2 Οι στόχοι της εξόρυξης γνώσης

Παρακάτω αναλύονται αυτά που μπορεί να προσφέρει η εξόρυξη. Τις δυνατότητες αυτές καλείται να εκμεταλλευτεί το μάνατζμεντ της εταιρίας ή ενός οργανισμού και να προχωρήσει σε αποφάσεις που θα μετατρέψουν τη γνώση σε χειροπιαστά αποτελέσματα. Αν το πετύχει τότε οι αρχές της επιχειρηματικής νοημοσύνης, που αποτελούν και την κεντρική ιδέα των συστημάτων υποστήριξης αποφάσεων, εφαρμόζονται και είναι σίγουρο ότι τα οφέλη θα είναι μεγάλα.

- **Πρόβλεψη τάσεων και συμπεριφορών**. Δηλαδή η προσπάθεια ανακάλυψης κάποιων μελλοντικών συμπεριφορών ώστε να παρθούν οι κατάλληλες αποφάσεις με σκοπό τη μεγιστοποίηση του κέρδους ή την πρόληψη δυσμενών καταστάσεων. Τα αποτελέσματα αυτού του είδους εξόρυξης μπορεί να είναι η πρόβλεψη για το που θα φτάσουν οι πωλήσεις ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο και το αν το κλείσιμο μιας γραμμής παραγωγής προϊόντων θα ενεργούσε θετικά σε ότι αφορά αυτές (τις πωλήσεις). Σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών ακολουθιών μπορεί ίσως να οδηγήσει στην πρόβλεψη, με υψηλά ποσοστά επιτυχίας, σεισμικής δραστηριότητας.
- **Αναγνώριση**. Οι τυποποιημένες μορφές ανάμεσα στα δεδομένα μπορούν να χρησιμοποιηθούν για να αποκαλύψουν την ύπαρξη ενός γεγονότος, μια δραστηριότητας. Για παράδειγμα οι εισβολείς στη προσπάθεια να σπάσουν ένα σύστημα ασφαλείας μπορούν να αναγνωριστούν από τα προγράμματα που

εκτέλεσαν, τα αρχεία που προσπέλασαν και τον χρόνο που απασχόλησαν την CPU.

- **Ταξινόμηση**. Η εξόρυξη γνώσης μπορεί να διαχωρίσει έτσι τα δεδομένα ώστε να προκύψουν διαφορετικές κλάσεις ή κατηγορίες βάση κάποιων παραμέτρων. Για παράδειγμα οι πελάτες ενός super-market μπορούν να χωριστούν σε κατηγορίες, όπως φίλοι-των-εκπτώσεων, παρορμητικοί, πιστοί-κανονικοί, και σπάνιοι πελάτες. Αυτή η κατηγοριοποίηση μπορεί να χρησιμοποιηθεί στην ανάλυση των πωλήσεων ώστε να μπορεί για παράδειγμα ο μάνατζερ να λάβει αποφάσεις για να προσελκύσει σε μεγαλύτερο βαθμό κάποια από τις παραπάνω κατηγορίες.
- **Βελτιστοποίηση**. Ένας τελικός στόχος της εξόρυξης γνώσης μπορεί να είναι και η βέλτιστη χρήση περιορισμένων πόρων όπως είναι ο χρόνος, ο χώρος, το χρήμα ή τα υλικά και η μεγιστοποίηση, κάτω από ορισμένους περιορισμούς, κάποιων «ποσοτήτων» όπως είναι οι πωλήσεις ή τα κέρδη. Έτσι σε ότι αφορά τουλάχιστον αυτό το στόχο, η εξόρυξη γνώσης έχει κοινά στοιχεία με την επιχειρησιακή έρευνα που επίσης ασχολείται με θέματα βελτιστοποίησης κάτω από περιορισμούς.

5.3 Οι κυριότερες τεχνικές

Τα εργαλεία εξόρυξης γνώσης συνήθως δίνουν τη δυνατότητα στον χρήστη να επιλέξει ποια τεχνική-αλγόριθμο θέλουν να εφαρμόσουν. Παρακάτω γίνεται μια σύντομη γνωριμία με αυτούς τους αλγόριθμους.

- **Τα νευρωνικά δίκτυα:** Μη γραμμικά, προφητικά και μπορούν να εκπαιδευτούν. Σε ότι αφορά τη δομή μοιάζουν στα βιολογικά νευρωνικά δίκτυα.
- **Δένδρα απόφασης:** Δενδρικές δομές που αναπαριστούν σύνολα απόφασης. Αυτές οι αποφάσεις γεννούν κανόνες για τη ταξινόμηση ενός συνόλου δεδομένων.
- **Γενετικοί αλγόριθμοι:** Τεχνικές βελτιστοποίησης που χρησιμοποιούν διαδικασίες όπως γενετικοί συνδυασμοί, μετάλλαξη.

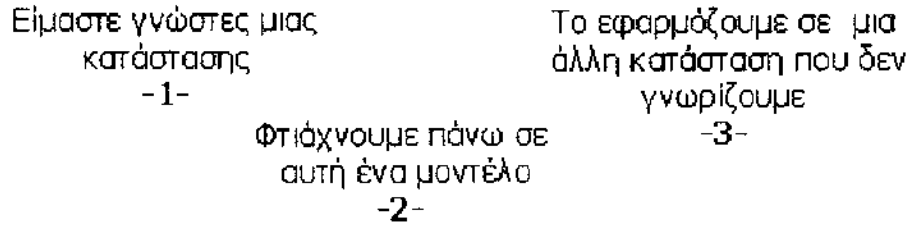
- **Επαγωγή κανόνα:** Η εξαγωγή χρήσιμων, και με στατιστική σημασία , if-then κανόνων από τα δεδομένα.

5.4 Πως λειτουργεί η εξόρυξη γνώσης

Πως ακριβώς μπορεί να μας πει η εξόρυξη γνώσης πράγματα που δεν ξέρουμε ή που θα συμβούν στο μέλλον; Η τεχνική που χρησιμοποιείται για να επιτευχθούν αυτά λέγεται μοντελοποίηση. Με άλλα λόγια η σκέψη του κτισίματος ενός μοντέλου για μια κατάσταση όπου γνωρίζουμε την απάντηση και στη συνέχεια η εφαρμογή του σε μια άλλη που δεν τη ξέρουμε. Για παράδειγμα, αν αναζητούσαμε μια βυθισμένη ισπανική γαλέρα στην ανοικτή θάλασσα το πρώτο πράγμα που ίσως σκεφτόμασταν θα ήταν να ερευνήσουμε όλες τις περασμένες περιπτώσεις εύρεσης ισπανικών θησαυρών από άλλους. Ίσως λοιπόν να παρατηρούσαμε ότι αυτά τα πλοία στην πλειονότητα τους βρέθηκαν στις ακτές Βερμούδα και ότι υπήρχαν κάποιες βέβαιες πορείες που ακολουθούσαν οι καπετάνιοι των πλοίων αυτών εκείνη την εποχή. Αυτές οι ομοιότητες σημειώνονται και κτίζεται ένα μοντέλο που περιλαμβάνει τα χαρακτηριστικά που είναι κοινά στις τοποθεσίες αυτών των βυθισμένων θησαυρών. Με αυτό το μοντέλο αρχίζει το ψάξιμο σε περιοχές που δείχνει αυτό ότι είναι πιθανό να υπήρξε μια παρόμοια κατάσταση στο παρελθόν. Αν το μοντέλο είναι καλό ο θησαυρός θα βρεθεί.

Επομένως τη σκέψη του κτισίματος μοντέλων την είχαν οι άνθρωποι εδώ και πολύ καιρό και σίγουρα πριν την έλευση των υπολογιστών και της τεχνολογίας της εξόρυξης γνώσης. Πάντως αυτό που συμβαίνει στους υπολογιστές δεν διαφέρει πολύ από τον τρόπο με τον οποίο οι άνθρωποι κτίζουν μοντέλα. Οι υπολογιστές φορτώνονται με πληροφορίες για μια ποικιλία καταστάσεων ενώ μια απάντηση είναι γνωστή. Τότε το λογισμικό εξόρυξης γνώσης τρέχει πάνω σε αυτά τα δεδομένα και ξεχωρίζει εκείνα τα χαρακτηριστικά που πρέπει να συμπεριληφθούν στο μοντέλο. Όταν τελειώσει η διαδικασία κτισίματος μπορεί το μοντέλο να χρησιμοποιηθεί σε παρόμοιες καταστάσεις που όμως η απάντηση δεν είναι γνωστή.

Εξόρυξη Γνώσης = Μοντελοποίηση



Σχήμα 5.1. Η φιλοσοφία της εξόρυξης γνώσης.

Για παράδειγμα ας υποθέσουμε ότι βρισκόμαστε στη θέση του διευθυντή μάρκετινγκ μιας εταιρίας τηλεπικοινωνιών και θέλουμε να αποκτήσουμε μερικούς πελάτες που κάνουν τηλεφωνήματα μεγάλων αποστάσεων. Βρισκόμαστε δηλαδή αντιμέτωποι με ένα πρόβλημα απόφασης, σε ποιους να απευθυνθούμε. Θα μπορούσαμε να ταχυδρομήσουμε με τυχαίο τρόπο κουπόνια στο γενικό πληθυσμό όπως θα μπορούσαμε να ταξιδεύουμε στις θάλασσες ψάχνοντας για βυθισμένους θησαυρούς. Πάντως σε καμιά από τις δυο περιπτώσεις δεν θα είχαμε τα επιθυμητά αποτελέσματα. Αντί αυτού θα μπορούσαμε να χρησιμοποιήσουμε την εμπειρία της εταιρίας που βρίσκεται αποθηκευμένη στις βάσεις δεδομένων και να κτίσουμε ένα μοντέλο.

Ο διευθυντής μάρκετινγκ έχει πρόσβαση σε πολλές πληροφορίες σχετικές με τους πελάτες μας: την ηλικία τους, το φύλο τους, το αν είναι καλοί πληρωτές, το πόσα τηλεφωνήματα μεγάλων αποστάσεων κάνουν. Το καλό είναι ότι υπάρχουν πληροφορίες και για τους πιθανούς πελάτες της εταιρίας: την ηλικία τους, το φύλο τους, το πόσο γρήγορα θα πληρώνουν κτλ. Το πρόβλημα είναι ότι δεν γνωρίζουμε πόσο πολύ θα κάνουν χρήση τηλεφωνημάτων σε απομακρυσμένες περιοχές. Επειδή θέλουμε αυτούς που κάνουν πολλά τέτοια τηλεφωνήματα μπορούμε να το πετύχουμε αυτό κτίζοντας ένα μοντέλο.

Ένα απλό μοντέλο που θα ταίριαζε σε μια τηλεπικοινωνιακή εταιρία είναι το παρακάτω:

98% των πελατών που έχουν λογαριασμό μεγαλύτερο
από 60.000$ το χρόνο δαπανούν περισσότερα
από 80$ το μήνα για τηλεφωνήματα σε μακρινές περιοχές

Αυτό το μοντέλο θα μπορούσε να εφαρμοστεί στα δεδομένα των πιθανών πελατών και να δοθεί απάντηση στο πρόβλημα απόφασης. Αφού γίνει αυτό θα ξέρει σε ποιους να απευθυνθεί η εταιρία.

5.5 Λογισμικό Εξόρυξης Γνώσης

Η εξόρυξη γνώσης είναι κατά κάποιο τρόπο μια επέκταση της στατιστικής με κάποια στοιχεία τεχνητής νοημοσύνης και μηχανική μάθηση(machine learning). Όπως και η στατιστική, η εξόρυξη γνώσης δεν αποτελεί επιχειρηματική λύση. Είναι απλά μια τεχνολογία. Για παράδειγμα φανταστείτε ότι πρέπει από ένα κατάλογο εμπόρων λιανικής να αποφασιστεί σε ποιους θα σταλούν πληροφορίες για κάποιο νέο προϊόν. Η πληροφορία που αναζητείται από την διαδικασία εξόρυξης γνώσης περιλαμβάνεται σε βάσεις ιστορικών δεδομένων προηγούμενων συναλλαγών με τους πελάτες και στα χαρακτηριστικά των πελατών όπως η ηλικία, ο ταχυδρομικός τους κώδικας, το αν αποκρίθηκαν στο παρελθόν. Το λογισμικό εξόρυξης γνώσης θα χρησιμοποιήσει αυτές τις πληροφορίες από το παρελθόν για να χτίσει ένα μοντέλο συμπεριφοράς πελάτη που θα μπορεί να χρησιμοποιηθεί για να προβλέψουμε ποιοι πελάτες θα ήταν πιθανό να ανταποκριθούν στο νέο προϊόν. Ένας διευθυντής μάρκετινγκ κάνοντας χρήση αυτής της πληροφορίας μπορεί να επιλέξει μόνο τους πελάτες που είναι πιο πιθανό να ανταποκριθούν. Το λογισμικό της επιχείρησης μπορεί τότε να τροφοδοτήσει με τα αποτελέσματα της απόφασης τα κατάλληλα «σημεία επαφής» (τηλεφωνικά κέντρα, web servers, e-mails κτλ) ώστε οι κατάλληλοι πελάτες να λαμβάνουν τις κατάλληλες πληροφορίες.

5.6 Η διαδικασία εξόρυξης γνώσης από μια αποθήκη δεδομένων

Σε αυτή την παράγραφο θα ασχοληθούμε με πιο τεχνικά θέματα των συστημάτων υποστήριξης αποφάσεων ώστε ο αναγνώστης να έχει συνολική εικόνα του θεμάτων με τα οποία ασχολείται η εργασία. Πιο συγκεκριμένα θα δούμε αναλυτικά τα στάδια που μεσολαβούν μέχρι να είναι η δυνατή η ανάλυση και η ερμηνεία των αποτελεσμάτων. Η ανακάλυψη γνώσης – η διαδικασία καθορισμού και επίτευξης ενός σκοπού μέσω επαναληπτικής εξόρυξης γνώσης – τυπικά αποτελείται από τρεις φάσεις:

- Προετοιμασία των δεδομένων,
- Υλοποίηση και αποτίμηση του μοντέλου και
- Ανάπτυξη του μοντέλου

5.6.1 Η φάση της προετοιμασίας των δεδομένων

Στη φάση της προετοιμασία των δεδομένων, ο αναλυτής προετοιμάζει ένα σύνολο δεδομένων που περιλαμβάνει αρκετές πληροφορίες για να κτιστεί ένα σωστό μοντέλο σε ακόλουθες φάσεις. Προσδιορίζοντας αυτές τις απαραίτητες πληροφορίες για μια εταιρία, ένα αποτελεσματικό μοντέλο θα μπορούσε να προβλέψει αν υπάρχει πιθανότητα να αγοράσει κάποιος πελάτης προϊόντα που διαφημίζονται σε ένα νέο κατάλογο. Επειδή οι προβλέψεις βασίζονται σε παράγοντες που πιθανότατα επηρεάζουν τις αγορές των πελατών, ένα μοντέλο συνόλου δεδομένων θα μπορούσε να περιλάμβανε όλους τους πελάτες που ανταποκρίθηκαν σε καταλόγους μέσω e-mails, ταχυδρομείων κτλ τα τελευταία τρία χρόνια, τις δημογραφικές πληροφορίες τους, τα δέκα πιο ακριβά προϊόντα που αγόρασε κάθε πελάτης και πληροφορίες για τους καταλόγους από τους οποίους έγιναν οι αγορές.

Η προετοιμασία των δεδομένων μπορεί να περιλαμβάνει πολύπλοκες ερωτήσεις με τεράστια αποτελέσματα-απαντήσεις. Για παράδειγμα στην υποθετική εταιρία που αναφέρθηκε και στα προηγούμενα παραδείγματα, η προετοιμασία του μοντέλου περιλαμβάνει joins μεταξύ του πίνακα των πελατών και του πίνακα των πωλήσεων

καθώς επίσης και τον προσδιορισμό των δέκα κορυφαίων προϊόντων για κάθε πελάτη. Όλα τα θέματα που έχουν να κάνουν με την αποτελεσματική επεξεργασία ερωτήσεων υποστήριξης αποφάσεων σχετίζονται με το περιβάλλον της εξόρυξης γνώσης.

Η εξόρυξη γνώσης τυπικά περιλαμβάνει επαναληπτικό κτίσιμο μοντέλων πάνω σε ένα ήδη προετοιμασμένο σύνολο δεδομένων και στη συνέχεια την ανάπτυξη ενός ή περισσότερων μοντέλων. Επειδή το κτίσιμο των μοντέλων σε μεγάλα σύνολα δεδομένων μπορεί να είναι δαπανηρό, οι αναλυτές συχνά εργάζονται επαναληπτικά με δείγματα συνόλων δεδομένων.

5.6.2 Η φάση της υλοποίησης και υπολογισμός του μοντέλου

Μόνο όταν έχει αποφασιστεί ποιο μοντέλο θα αναπτυχθεί, κτίζει ο αναλυτής το μοντέλο πάνω στο συνολικά προετοιμασμένο σύνολο δεδομένων. Ο σκοπός της φάσης της υλοποίησης είναι ο εντοπισμός των τυποποιημένων μορφών που καθορίζουν ένα χαρακτηριστικό-στόχο(target attribute). Ένα παράδειγμα τέτοιου χαρακτηριστικού-στόχου σε ένα σύνολο δεδομένων θα μπορούσε να ήταν το αν αγόρασε ένας πελάτης τουλάχιστον ένα προϊόν από ένα περασμένο κατάλογο.

Μερικές κλάσεις μοντέλων εξόρυξης γνώσης βοηθούν την πρόβλεψη τόσο ρητά καθορισμένων όσο και κρυφών χαρακτηριστικών. .υο σημαντικά θέματα που επηρεάζουν την επιλογή του μοντέλου είναι η ακρίβεια του και η αποτελεσματικότητα του αλγορίθμου κατασκευής του μοντέλου πάνω σε μεγάλα σύνολα δεδομένων. Από στατιστικής πλευράς η ακρίβεια των περισσότερων μοντέλων βελτιώνεται με το πλήθος των δεδομένων που χρησιμοποιούνται, οπότε οι αλγόριθμοι που επηρεάζουν τα μοντέλα εξόρυξης πρέπει να κάνουν αποτελεσματική και κλιμακωτή επεξεργασία μεγάλων συνόλων δεδομένων σε ένα λογικό χρονικό διάστημα.

6. Τύποι Μοντέλων

Τα μοντέλα ταξινόμησης είναι προφητικά. Μπορούν να προβλέψουν αν μία νέα πλειάδα ανήκει σε ένα από τα σύνολα των κλάσεων-στόχων. Κάνοντας εφαρμογή πάνω στον κατάλογο μιας εταιρίας για παράδειγμα, ένα μοντέλο ταξινόμησης θα μπορούσε να προσδιορίσει, βασιζόμενο σε παλιότερες συμπεριφορές, αν υπάρχει πιθανότητα να αγοράσει κάποιος πελάτης από τον κατάλογο αυτό. Τα δέντρα αποφάσεων και τα μοντέλα του Bayes αποτελούν δυο δημοφιλείς τύπους μοντέλων ταξινόμησης.

Οι αναλυτές χρησιμοποιούν τα λεγόμενα βασισμένα στους κανόνες(rulebased) μοντέλα για να εξερευνήσουν αν για παράδειγμα η αγορά ενός καθορισμένου συνόλου προϊόντων υποδημάτων είναι ενδεικτική, με κάποιο βαθμό εμπιστοσύνης, της αγοράς κάποιου άλλου προϊόντος.

Κατά τον προγραμματισμό των βασισμένων στους κανόνες μοντέλων χρησιμοποιούνται οι αλγόριθμοι. Αλγόριθμος είναι ένας μαθηματικός όρος που σημαίνει κάθε συστηματική μέθοδο με την οποία βρίσκει κανείς το ζητούμενο, τη λύση δηλαδή ενός προβλήματος, όταν ακολουθήσει κανονικά, με τη σειρά, ορισμένες πράξεις και συγκεκριμένους κανόνες.

6.1 Δέντρα Αποφάσεων

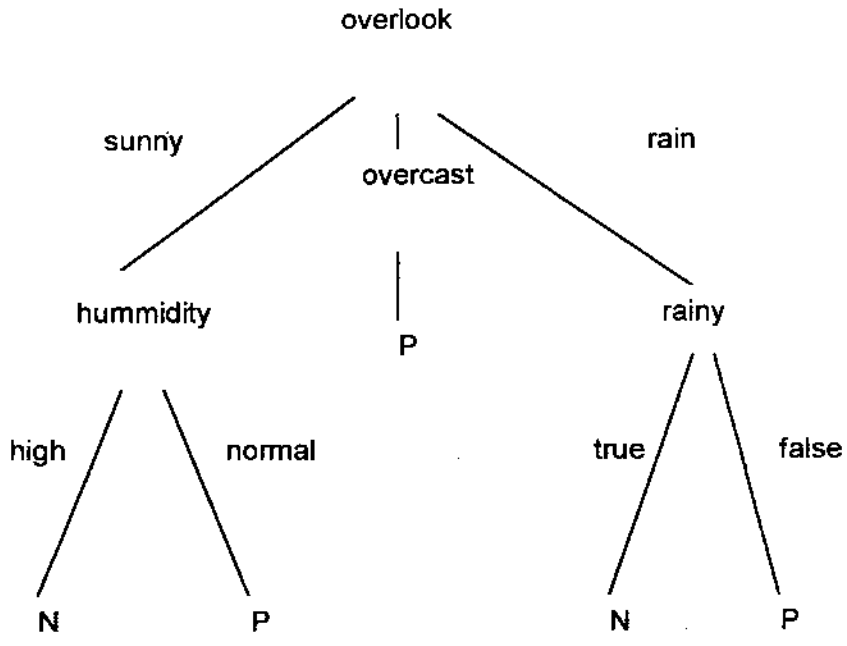
Τα *Δένδρα Αποφάσεων* (Decision Trees) είναι πολύ ισχυρά και δημοφιλή εργαλεία για classification και prediction. Τα Δένδρα Αποφάσεων αντιπροσωπεύουν κανόνες, οι οποίοι μπορούν εύκολα να διατυπωθούν σε φυσική γλώσσα ώστε να είναι εύκολα κατανοητοί από τους ανθρώπους ή να διατυπωθούν σε μία γλώσσα προσπέλασης βάσεων δεδομένων π.χ. σε SQL. Υπάρχει μια πληθώρα αλγορίθμων που αναλαμβάνουν να φτιάξουν Δένδρα Αποφάσεων, όπως : *CART* (Classification and Regression Trees), *CHAID* (CHi-squared Automation Interaction Detection), ένας πιο πρόσφατος πολλά υποσχόμενος αλγόριθμος είναι ο *C4.5*.

Γενικά ένα *Δένδρο Απόφασης* αντιπροσωπεύει μια σειρά από **IF THEN** κανόνες που συνδυάζονται μεταξύ τους από τη ρίζα του δένδρου προς τα φύλλα. Οι κόμβοι του δέντρου χαρακτηρίζονται με τα ονόματα των χαρακτηριστικών, οι ακμές ονομάζονται με τις δυνατές τιμές που μπορεί να πάρει ένα χαρακτηριστικό και τα φύλλα με τις διάφορες κλάσεις. Τα αντικείμενα ταξινομούνται ακολουθώντας ένα μονοπάτι που οδηγεί προς τα κάτω στο δέντρο, λαμβάνοντας τις ακμές που αντιστοιχούν στις τιμές των χαρακτηριστικών ενός αντικειμένου.

Μία εγγραφή εισέρχεται στο δέντρο από τον κόμβο της κορυφής. Στην ρίζα, εφαρμόζεται έλεγχος για να καθορισθεί ποιο κόμβο παιδί θα ακολουθήσει στην συνέχεια η εγγραφή. Υπάρχουν διάφοροι αλγόριθμοι για την επιλογή του αρχικού ελέγχου, αλλά ο στόχος είναι πάντα ο ίδιος, δηλαδή, να επιλέξουμε τον έλεγχο ο οποίος διαχωρίζει καλύτερα τις τελικές κλάσεις. Η επεξεργασία αυτή επαναλαμβάνεται μέχρι η εγγραφή να φτάσει στο κόμβο φύλλο. Όλες οι εγγραφές οι οποίες καταλήγουν σε ένα συγκεκριμένο φύλλο ταξινομούνται με τον ίδιο τρόπο. Υπάρχει ένα μοναδικό μονοπάτι που οδηγεί από την ρίζα σε κάθε φύλλο. Το μονοπάτι αυτό είναι μία έκφραση του κανόνα που χρησιμοποιείται για να ταξινομήσουμε τις εγγραφές.

Πολλά διαφορετικά φύλλα μπορούν να οδηγούν στην ίδια ταξινόμηση, αλλά κάθε φύλλο κάνει την ταξινόμηση αυτή για διαφορετικό λόγο. Για παράδειγμα, σε ένα δέντρο το οποίο ταξινομεί φρούτα και λαχανικά με βάση το χρώμα, οι τελικοί κόμβοι του δέντρου απόφασης για τα μήλα, ντομάτες και κεράσια θα πρέπει όλα να προβλέπουν "κόκκινο", παρά τον διαφορετικό βαθμό πίστης καθώς υπάρχουν πράσινα μήλα και μαύρα κεράσια.

Στο σχήμα 6.1. παρουσιάζεται ένα παράδειγμα αντικειμένων το οποίο περιγράφει τον καιρό σε μία δεδομένη στιγμή. Κάποια αντικείμενα τα οποία είναι θετικά παραδείγματα δηλώνονται ως P και άλλα τα οποία είναι αρνητικά δηλώνονται ως N. Το classification στην περίπτωση αυτή είναι η κατασκευή ενός δέντρου το οποίο μπορεί να χρησιμοποιηθεί για να ταξινομήσει τα αντικείμενα με σωστό τρόπο.



Σχήμα 6.1. Δέντρα Αποφάσεων

Στα θετικά σημεία της μεθόδου αυτής συγκαταλέγονται:

- Η ευρωστία που επιδεικνύει αναφορικά με το θόρυβο που ενδέχεται να παρουσιαστεί στα δεδομένα που απαρτίζουν το χώρο του προβλήματος.
- Η ανοχή στην απουσία τιμών (missing values), σε κάποια χαρακτηριστικά του σώματος εκπαίδευσης.
- Η χρήση ακόμα και συνεχών (μη διακριτών) χαρακτηριστικών και η προσέγγιση μη διακριτών συναρτήσεων στόχου, μέσω εξειδικευμένων τεχνικών που αναλαμβάνουν τη διακριτοποίησή τους (discretization), τη διαδικασία δηλαδή της μετατροπής συνεχών αριθμητικών χαρακτηριστικών σε ονομαστικά.
- Η δυνατότητα μεταφοράς του παραγόμενου μοντέλου από δένδρο απόφασης σε ένα σύνολο κανόνων συμπερασμού (if – then rules), προς διευκόλυνση της κατανόησής του.

6.1.1 Αλγόριθμος δημιουργίας του δέντρου

Το δέντρο γεννάται με την επαναλαμβανόμενη διάσπαση του δοσμένου συνόλου δεδομένων σύμφωνα με τις διάφορες ανεξάρτητες μεταβλητές. Η σειρά με την οποία χρησιμοποιούνται οι ανεξάρτητες μεταβλητές στη δόμηση του δέντρου εξαρτάται από το μέτρο ταξινόμησης της κάθε ανεξάρτητης μεταβλητής. Ο αλγόριθμος σταματά όταν φτάσει σε κόμβο από τον οποίο δεν είναι δυνατό να ξεκινήσει μία νέα διάσπαση. Τότε ο κόμβος αυτός δεν έχει παιδιά και αποτελεί φύλλο του δέντρου. Ο αλγόριθμος δημιουργίας του δέντρου είναι δυαδικός (binary), δηλαδή σε κάθε διάσπαση δημιουργεί δύο μόνο κλαδιά.

6.1.2 Η περίπτωση διακριτής εξαρτημένης μεταβλητής

Στη περίπτωση δημιουργίας δέντρου ταξινόμησης – *όταν η εξαρτημένη μεταβλητή της ανάλυσής μας είναι διακριτή* – η επιλογή της ανεξάρτητης μεταβλητής σε κάθε επίπεδο – στάδιο δόμησης του δέντρου, γίνεται σύμφωνα με την πληροφορία που περιέχεται σε κάθε ανεξάρτητη μεταβλητή σε σχέση πάντα με την διακριτή εξαρτημένη μεταβλητή. Αυτή η περιεχόμενη πληροφορία μετράται με βάση την τιμή της εντροπίας (entropy) της κάθε ανεξάρτητης μεταβλητής, με τη μέγιστη πληροφορία να αντιστοιχεί στην ελάχιστη τιμή εντροπίας. Σε κάθε νέα διάσπαση επιλέγεται η ανεξάρτητη τυχαία μεταβλητή με τη μικρότερη τιμή εντροπίας. Η ανεξάρτητη αυτή μεταβλητή με τη μικρότερη εντροπία παρουσιάζει την αμέσως μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή. Όσον αφορά την εσωτερική διάσπαση μίας ανεξάρτητης μεταβλητής (διακριτής ή συνεχής), αυτή γίνεται με τρόπο τέτοιο ώστε να ελαχιστοποιείται η τιμή της εντροπίας της. Η εντροπία υπολογίζεται με βάση τον ακόλουθο τύπο:

$$Entropy = \sum_{i=1}^{2} \frac{\sum_{j=1}^{c} f_{ij}}{R} \sum_{j=1}^{c} P(c_{ij}) Ln(P(c_{ij}))$$

Όπου c είναι ο αριθμός των τιμών της εξαρτημένης μεταβλητής που έχουν προκύψει, f είναι η συχνότητα της τιμής j της εξαρτημένης μεταβλητής στο κλαδί i, R είναι ο συνολικός αριθμός εγγραφών (παρατηρήσεων) και στα δύο κλαδιά και το $P(c_{ij})$ δίνεται από την ακόλουθη σχέση:

$$P(c_{ij}) = \frac{f_{ij}}{\sum_{k=1}^{c} f_{kj}}$$

Ο αλγόριθμος δημιουργίας του δέντρου σταματά όταν δεν είναι δυνατή η παραγωγή δύο νέων κλαδιών που το καθένα να περιέχει αριθμό εγγραφών ίσο ή μεγαλύτερο από έναν ελάχιστο αριθμό παρατηρήσεων σε κάθε κλαδί, τον οποίο έχει καθορίσει ο χρήστης στην αρχή της διαδικασίας.

6.2 Μαθαίνοντας ένα σύνολο κανόνων

Τα δέντρα απόφασης μπορούν να μεταφραστούν σε ένα σύνολο κανόνων με τη δημιουργία ενός ξεχωριστού κανόνα για κάθε πορεία από τη ρίζα σ' ένα φύλλο στο δέντρο (Quinlan, 1993). Εντούτοις οι κανόνες μπορούν επίσης να προκληθούν άμεσα από τα στοιχεία κατάρτισης χρησιμοποιώντας ποικίλους αλγόριθμους βασισμένους στους κανόνες. Ο Furnkranz (1999) παρέχει μια άριστη επισκόπηση της υπάρχουσας εργασίας σε μεθόδους βασισμένες στον κανόνα.

Οι κανόνες ταξινόμησης αντιπροσωπεύουν κάθε κατηγορία από την διαζευκτική κανονική μορφή (DNF). Ο στόχος είναι να κατασκευαστεί το μικρότερο σύνολο κανόνων που είναι σύμφωνο με τα στοιχεία κατάρτισης. Ένας μεγάλος αριθμός μαθημένων κανόνων είναι συνήθως ένα σημάδι που ο αλγόριθμος εκμάθησης προσπαθεί να θυμηθεί το σύνολο κατάρτισης, αντί να ανακαλύπτει τις υποθέσεις που το κυβερνούν.

Η διαφορά μεταξύ πρακτικών για την εκμάθηση κανόνα και πρακτικών για τα δέντρα απόφασης είναι ότι τα τελευταία αξιολογούν τη μέση ποιότητα ενός αριθμού από ασυνάρτητα σύνολα (ένα για κάθε αξία του χαρακτηριστικού γνωρίσματος που εξετάζεται), ενώ οι μαθητές κανόνων αξιολογούν μόνο την ποιότητα του συνόλου περιπτώσεων που καλύπτεται από τον υποψήφιο κανόνα.

Ο RIPPER είναι ένας πολύ γνωστός αλγόριθμος βασισμένος σε κανόνες (Cohen,1995). Διαμορφώνει τους κανόνες μέσω μιας διαδικασίας της επαναλαμβανόμενης ανάπτυξης και περικοπής. Κατά τη διάρκεια της αυξανόμενης φάσης οι κανόνες γίνονται πιο περιοριστικοί προκειμένου να ταιριάζουν με τα στοιχεία κατάρτισης όσο το δυνατόν περισσότερο. Κατά τη διάρκεια της φάσης περικοπής, οι κανόνες γίνονται λιγότερο περιοριστικοί για να αποφύγουν την υπερβολική, το οποίο μπορεί να προκαλέσει την κακή απόδοση στις απαρατήρητες περιπτώσεις. Η πρακτική που χρησιμοποιείται στο RIPPER είναι h λειτουργία κέρδους πληροφοριών. Το RIPPER χειρίζεται πολλαπλές κατηγορίες ταξινομώντας αυτές από τις λιγότερο στις περισσότερο επικρατούσες και έπειτα με τη μεταχείριση κάθε μιας με τη σειρά ως ευδιάκριτο πρόβλημα δύο-κατηγοριών.

Υπάρχουν πολυάριθμοι άλλοι αλγόριθμοι εκμάθησης βασισμένοι στους κανόνες. Ο Furnkranz (1999) αναφέρεται στους περισσότερους από αυτούς. Ο αλγόριθμος PART διαμορφώνει κανόνες από τμήματα των δέντρων απόφασης σε μία προσπάθεια να αποφευχθεί η υπερβολική περικοπή. Μόλις χτιστεί ένα επί μέρους δέντρο, ένας ενιαίος κανόνας εξάγεται από αυτό (Frank και Witten, 1998)

Για τη μάθηση του συνόλου κανόνων έχουν επίσης χρησιμοποιηθεί γενετικοί αλγόριθμοι (GAs). Ο GABIL (Dejong et al.1993) χρησιμοποίησε τον γενετικό αλγόριθμο για να μάθει τις δυαδικές έννοιες που αναπαρίσταντο από ένα διαζευκτικό σύνολο όλων των κανόνων και βρέθηκε για να είναι συγκρίσιμο σε γενικευμένη ακρίβεια με τον αλγόριθμο εκμάθησης C4.5. δέντρων απόφασης. Υποθέτοντας δύο δυαδικά

χαρακτηριστικά γνωρίσματα X1, X2 και η δυαδική αξία στόχων C, η αντιπροσώπευση κανόνων στα χρωμοσώματα είναι:

Εάν X1=σωστό	X2=λάθος	ΤΟΤΕ c=σωστό	ΑΝ X1=λάθος	X2=σωστό	ΤΟΤΕ c = λάθος
10	.01.	1.	01.	10.	0

Σημειώστε ότι υπάρχει μια σταθερή αντιπροσώπευση σειράς μήκους κομματιών για κάθε κανόνα. Ο στόχος του γενετικού αλγορίθμου είναι να βρεθούν καλά χρωμοσώματα. Η καλή ποιότητα ενός χρωμοσώματος αντιπροσωπεύεται στο GA από μια λειτουργία, η οποία ονομάζεται λειτουργία ικανότητας (Banzhaf et Al το 1998). Για το στόχο ταξινόμησης, η λειτουργία ικανότητας σημειώνει χαρακτηριστικά την ακρίβεια ταξινόμησης του κανόνα πάνω σ' ένα σύνολο παρεχόμενων περιπτώσεων κατάρτισης. Στην καρδιά του αλγόριθμου υπάρχουν διαδικασίες, που παίρνουν τον πληθυσμό στη σημερινή γενιά και παράγουν τον πληθυσμό στο επόμενο βήμα κατά τέτοιο τρόπο ώστε η γενική ικανότητα του πληθυσμού αυξάνεται. Αυτές οι λειτουργίες επαναλαμβάνονται έως ότου ικανοποιείται κάποιο κριτήριο διακοπής , όπως ένας ορισμένος αριθμός χρωμοσωμάτων επεξεργάστηκε ή ένα χρωμόσωμα ορισμένης ποιότητας έχει παραχθεί. Τρεις διαδικασίες παίρνουν τον πληθυσμό στην γενιά t και παράγουν το νέο πληθυσμό στην γενιά t + 1: επιλογή, διασταύρωση (Banzhafet Al 1998)

6.3 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (NEURAL NETWORKS)

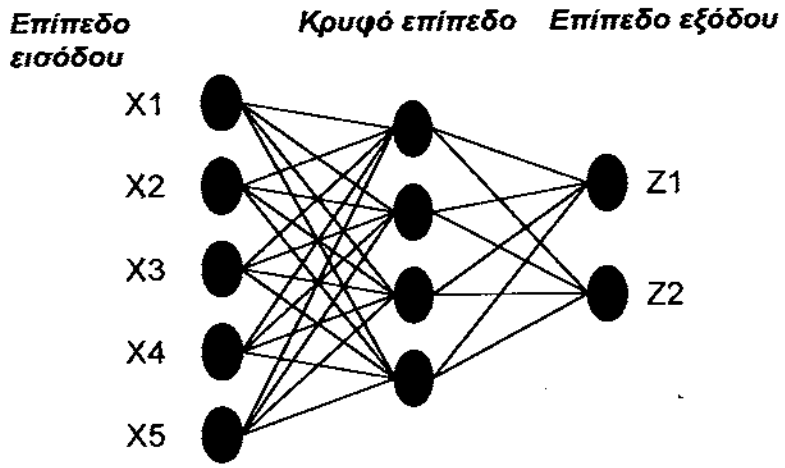
Τα νευρωνικά δίκτυα αποτελούν μία πολύ δυνατή, γενικού σκοπού τεχνική η οποία μπορεί να εφαρμοστεί για *πρόβλεψη(prediction)*, *classification* και *clustering*. Η εμφάνιση των νευρωνικών δικτύων έχει σαν στόχο να γεφυρώσει το κενό μεταξύ των υπολογιστών και του ανθρώπινου μυαλού. Οι άνθρωποι μπορούν να εξάγουν συμπεράσματα με βάση την εμπειρία τους ενώ οι υπολογιστές βασίζονται σε συγκεκριμένες οδηγίες. Τα νευρωνικά δίκτυα στοχεύουν στο να μειώσουν αυτό το κενό. Όταν χρησιμοποιούνται σε καλά ορισμένο περιβάλλον, η ικανότητα τους να παράγουν και να μαθαίνουν από τα δεδομένα, μιμείται την ικανότητα των ανθρώπων να μαθαίνουν

από τις εμπειρίες τους. Αυτή η ικανότητα είναι χρήσιμη για το data mining κάνοντας συγχρόνως τα νευρωνικά δίκτυα μία σημαντική περιοχή για έρευνα, υποσχόμενα νέα και καλύτερα αποτελέσματα στο μέλλον.

Τα νευρωνικά δίκτυα είναι μία προσέγγιση ανάπτυξης και εκτίμησης μαθηματικών δομών με την δυνατότητα να μαθαίνουν. Οι μέθοδοι αυτοί είναι αποτελέσματα ακαδημαϊκών ερευνών με στόχο την μοντελοποίηση συστημάτων μάθησης. Τα νευρωνικά δίκτυα έχουν την ικανότητα να εξάγουν κάποιο συμπέρασμα από πολύπλοκα ή μη ακριβή δεδομένα και μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να προσδιορίζουν τάσεις οι οποίες είναι πολύ πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να αντιμετωπιστεί ως ένας "ειδικός" για την κατηγορία της πληροφορίας που του δόθηκε να αναλύσει. Έτσι μπορεί να χρησιμοποιηθεί για να κάνει κάποιες προβλέψεις, όταν προκύψουν κάποιες νέες περιπτώσεις.

Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι. Αυτό διακρίνει τα νευρωνικά δίκτυα από τα παραδοσιακά προγράμματα υπολογιστών, τα οποία απλά ακολουθούν οδηγίες σύμφωνα με μία καλά ορισμένη σειρά.

Η δομή των νευρωνικών δικτύων είναι ανάλογη με αυτή του σχήματος 6.2.



Σχήμα 6.2. *Δομή ενός νευρωνικού δικτύου.*

Το αριστερό επίπεδο αναπαριστά το επίπεδο εισόδου, στην περίπτωση του σχήματος έχουμε πέντε εισόδους με ετικέτες X1, X2,..., X5. Το μεσαίο επίπεδο είναι αυτό που καλείται *κρυφό επίπεδο (hidden level)*, το οποίο έχει μεταβλητό αριθμό κόμβων. Το μεσαίο επίπεδο είναι και αυτό που εκτελεί το μεγαλύτερο μέρος της εργασίας του δικτύου. Το επίπεδο εξόδου (επίπεδο στα δεξιά) έχει δύο κόμβους στο παράδειγμά μας Z1 και Z2, οι οποίες αναπαριστούν τις τιμές εξόδου που προσπαθούμε να προσδιορίσουμε από τις εισόδους. Για παράδειγμα, μπορεί με την βοήθεια ενός κατάλληλα εκπαιδευμένου δικτύου να προβλέψουμε τις πωλήσεις (*έξοδος*) βασιζόμενοι στις παλιές πωλήσεις, την τιμή και την εποχή (*είσοδοι*).

Τα τεχνητά νευρωνικά δίκτυα διακρίνονται για την ικανότητά τους να προσεγγίζουν τόσο διακριτές όσο και συνεχείς, πραγματικές, ακόμα και διανυσματικές συναρτήσεις στόχου, για την ευρωστία τους όσον αφορά την παρείσφρηση θορύβου στα δεδομένα εκπαίδευσης, καθώς και για την ταχύτητά τους κατά την ταξινόμηση άγνωστων στιγμιοτύπων. Απαιτούν ωστόσο μεγάλους χρόνους εκπαίδευσης, ενώ τις περισσότερες φορές το εξαγόμενο μοντέλο δεν παρέχεται σε καταληπτή μορφή.

6.3.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Μονόδρομα δίκτυα (feed-forward)

Η μονόδρομη αρχιτεκτονική δικτύων επιτρέπει στα σήματα να κινούνται προς την μια κατεύθυνση, από τη είσοδο στην έξοδο. Δεν υπάρχουν αμφίδρομες καταστάσεις (loops), παραδείγματος χάρη το αποτέλεσμα ενός επιπέδου δεν επηρεάζει το ίδιο το επίπεδο. Η μορφή αυτή των νευρωνικών δικτύων χρησιμοποιείται αρκετά στην αναγνώριση μοντέλων και συχνά τη συναντάμε και ως από πάνω προς τα κάτω ή από κάτω προς τα πάνω διαδικασία.

Αμφίδρομα δίκτυα

Στα αμφίδρομα δίκτυα τα σήματα ταξιδεύουν και προς τις δυο κατευθύνσεις. Τα αμφίδρομα δίκτυα είναι αρκετά ισχυρά αλλά και περίπλοκα. Χαρακτηρίζονται από τη διαρκή τους εξέλιξη μέχρι να φτάσουν σε κάποια κατάσταση ισορροπίας. Αυτή η κατάσταση διατηρείται ως ότου νέα δεδομένα προστεθούν όποτε το δίκτυο αναζητεί μια νέα κατάσταση ισορροπίας.

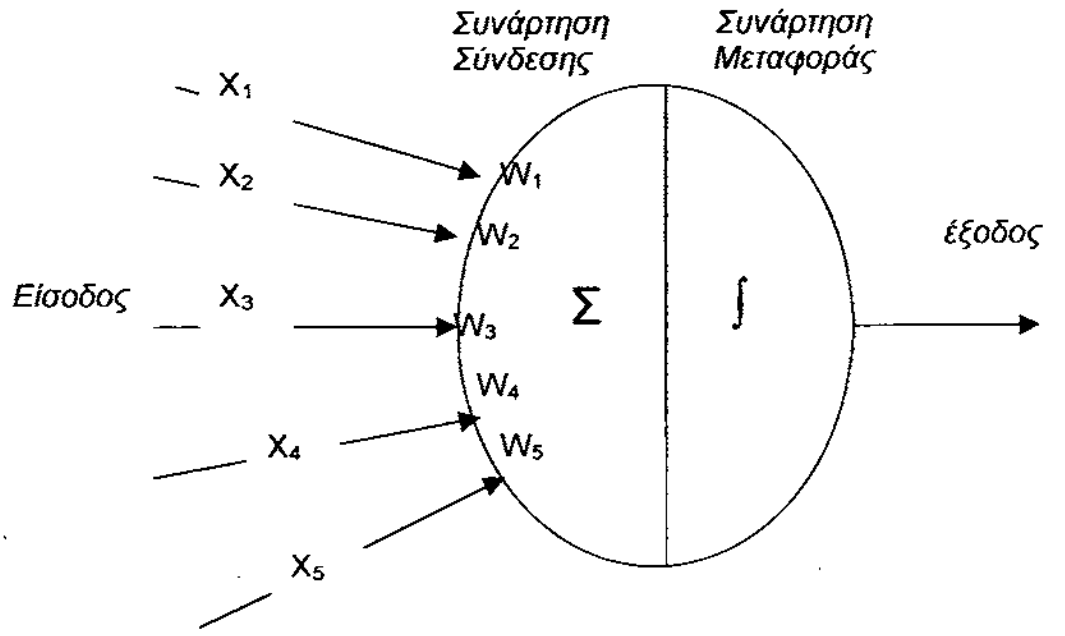
Κόμβοι Νευρωνικού Δικτύου

Τα νευρωνικά δίκτυα, όπως προαναφέραμε, αποτελούνται από βασικές μονάδες (κόμβους) που σχεδιάζονται για να μοντελοποιήσουν την συμπεριφορά των βιολογικών νευρώνων (σχήμα 1.5). Κάθε κόμβος στο μεσαίο επίπεδο είναι πλήρως συνδεδεμένος με τις εισόδους, γεγονός που σημαίνει ότι το κρυφό πεδίο βασίζεται σε όλες τις εισόδους τις οποίες και συνδυάζει στις τιμές εξόδου. Ο συνδυασμός αυτός καλείται *συνάρτηση ενεργοποίησης* του κόμβου.

Η συνάρτηση ενεργοποίησης έχει δύο μέρη. Το πρώτο μέρος είναι *η συνάρτηση σύνδεσης (combination function)* η οποία συνδυάζει όλες τις εισόδους σε μία απλή τιμή. Κάθε είσοδος έχει το δικό της βάρος. Η πιο κοινή συνάρτηση σύνδεσης είναι το άθροισμα όλων των εισόδων πολλαπλασιασμένων με το αντίστοιχο βάρος τους (X_1*W_1 + X_2*W_2 + ... + X_N*W_N). Σε ορισμένες περιπτώσεις είναι χρήσιμες άλλες συναρτήσεις και

περιλαμβάνουν το μέγιστο των εισόδων πολλαπλασιασμένων με το βάρος τους, το ελάχιστο, ή το λογικό AND ή OR των τιμών. Ωστόσο, η συνάρτηση που βασίζεται στο άθροισμα των εισόδων πολλαπλασιασμένων με τα βάρη τους δουλεύει καλύτερα στην πράξη.

Το δεύτερο μέρος της συνάρτησης ενεργοποίησης είναι η *συνάρτηση μεταφοράς(transfer function)*, η οποία μεταφέρει την τιμή της συνάρτησης σύνδεσης στην έξοδο. Υπάρχουν τρία είδη συναρτήσεων μεταφοράς: η *σιγμοειδής*, *γραμμική* και η συνάρτηση *υπερβολικής εφαπτομένης (hyperbolic tangent)*. Η γραμμική συνάρτηση έχει περιορισμένη πρακτική σημασία αντίθετα με τις άλλες δύο (μη γραμμικές συναρτήσεις) οι οποίες παρουσιάζουν μη γραμμική συμπεριφορά.



Σχήμα 6.3. *Η μονάδα επεξεργασίας(κόμβος) του νευρωνικού δικτύου*

Σε κάθε περιοχή του χάρτη συσχέτισης το δίκτυο αποθηκεύει τις σχέσεις μεταξύ των σχεδίων με ακρίβεια ώστε κάθε μονάδα να έχει ξεχωριστή σημασία. Αυτός ο τύπος της διαδικασίας εκπαίδευσης είναι κατάλληλος για ανακάλυψη χαρακτηριστικών και αναπαράσταση γνώσης. Κάθε νευρωνικό δίκτυο έχει την γνώση του αποθηκευμένη στα βάρη σύνδεσης. Μεταβάλλοντας την αποθηκευμένη γνώση του δικτύου μεταβάλλονται και οι τιμές των βαρών βάσει μιας εμπειρικής συνάρτησης.

Οι πληροφορίες για τα βάρη του νευρωνικού δικτύου αποθηκεύονται σε ένα πίνακα W. Η εκπαίδευση είναι ο προσδιορισμός των βαρών. Ακολουθούν οι πιο αποδοτικοί τρόποι εκπαίδευσης για τις δυο μεγάλες κατηγόριες νευρωνικών δικτύων που είναι:

- Τα συγκεκριμένα δίκτυα στα οποία τα βάρη δεν μπορούν να αλλάξουν, δηλαδή dw/dt=0. Στα περισσότερα δίκτυα τα βάρη είναι συγκεκριμένα και αφορούν ένα συγκεκριμένο πρόβλημα.
- Προσαρμοσμένα δίκτυα που είναι ικανά να αλλάζουν τα βάρη τους (dw/dt≠0)

  Όλες οι μέθοδοι μάθησης που χρησιμοποιούνται στα νευρωνικά δίκτυα μπορούν να διακριθούν σε δυο βασικές κατηγόριες.
- Επιτηρούμενη εκπαίδευση όπου ένας εξωτερικός δάσκαλος βοηθάει ώστε κάθε έξοδος να είναι η επιθυμητοί απάντηση βάση των σημάτων εισόδου. Κατά τη διάρκεια της εκπαίδευσης γενικές πληροφορίες μπορεί να ζητηθούν.

  Παραδείγματα εκπαίδευσης υπό επιτήρησης περιλαμβάνουν εκμάθηση εντοπισμού λαθών και διόρθωσης και στοχαστική εκμάθηση.

  Ένα σημαντικό θέμα που αφορά την εκπαίδευάη υπό επιτήρηση είναι το πρόβλημα σύγκλισης του λάθους ανάμεσα στο επιθυμητό και υπολογιζόμενο. Ο σκοπός είναι να καθορίσουμε Ένα σύνολο βαρών το οποίο ελαχιστοποιεί το λάθος. Μια αρκετά γνωστή μέθοδος η οποία είναι κοινή σε πολλά Παραδείγματα εκπαίδευσης είναι η σύγκλιση του ελάχιστου μέσου τετράγωνου.

Λέμε ότι ένα νευρωνικό δίκτυο εκπαιδεύεται off line όταν η φάση εκπαίδευσης και η φάση λειτουργίας είναι ξεχωριστές. ένα νευρωνικό δίκτυο εκπαιδεύεται on line αν μαθαίνει και λειτουργεί ταυτόχρονα. Συνήθως, η υπό επιτήρηση εκπαίδευση γίνεται off line ενώ η εκπαίδευση χωρίς επιτήρηση γίνεται on line.

Συνάρτηση μεταφοράς

Η συμπεριφορά ενός νευρωνικού δικτύου βασίζεται και στα βάρη και στη συνάρτηση εισόδου-εξόδου (συνάρτηση μεταφοράς) η οποία συγκεκριμενοποιείται για τις μονάδες. Αυτή η συνάρτηση χωρίζεται στις παρακάτω πληροφορίες.

- Γραμμική
- Οριακή
- Σιγμοειδείς

Για τις γραμμικές μονάδες η δραστηριότητα εξόδου είναι ανάλογη της συνολικής εξόδου των βαρών. Για τις οριακές μονάδες η έξοδος είναι τοποθετημένοι σε ένα από τα δυο επίπεδα και εξετάζει αν η γενική έξοδος είναι μεγαλύτερη ή μικρότερη από μια Οριακή τιμή. Για σιγμοειδείς μονάδες η έξοδος ποικίλει συνεχώς αλλά όχι Γραμμική καθώς η είσοδος αλλάζει. Οι σιγμοειδείς μονάδες μοιάζουν περισσότερο με τους ανθρώπινους νευρώνες από ότι οι γραμμικές ή οι οριακές μονάδες αλλά και οι τρεις πρέπει να θεωρηθούν ως διαδικασίες προσέγγισης.

Για να φτιάξουμε ένα νευρωνικό δίκτυο το οποίο να κάνει κάποιες συγκεκριμένες εργασίες πρέπει να επιλέξουμε με πιο τρόπο οι μονάδες θα είναι συνδεδεμένες μεταξύ τους και να υπολογίσουμε τα βάρη της σύνδεσης ορθά. Οι συνδέσεις καθορίζουν το κατά πόσο είναι δυνατό η μια μονάδα να επηρεάζει την άλλη. Τα βάρη προσδιορίζουν το μέγεθος της επιρροής.

Για να μάθουμε σε ένα δίκτυο τριών επιπέδων να κάνει μια συγκεκριμένη εργασία μπορούμε να χρησιμοποιήσουμε την εξής διαδικασία:

1. παρουσιάζουμε στο δίκτυο προγράμματα εκπαίδευσης τα οποία περιέχουν ένα σχέδιο από δραστηριότητες για τις μονάδες εισόδου και ένα επιθυμητό σχέδιο για τις μονάδες εξόδου.
2. εξετάζουμε πόσο στενά η πραγματική έξοδος του δικτύου ταιριάζει με την επιθυμητοί έξοδο.
3. αλλάζουμε το βάρος κάθε σύνδεσης έτσι ώστε το δίκτυο να παράγει μια καλύτερη προσέγγιση της επιθυμητής εξόδου.

6.4 Δίκτυα Bayes

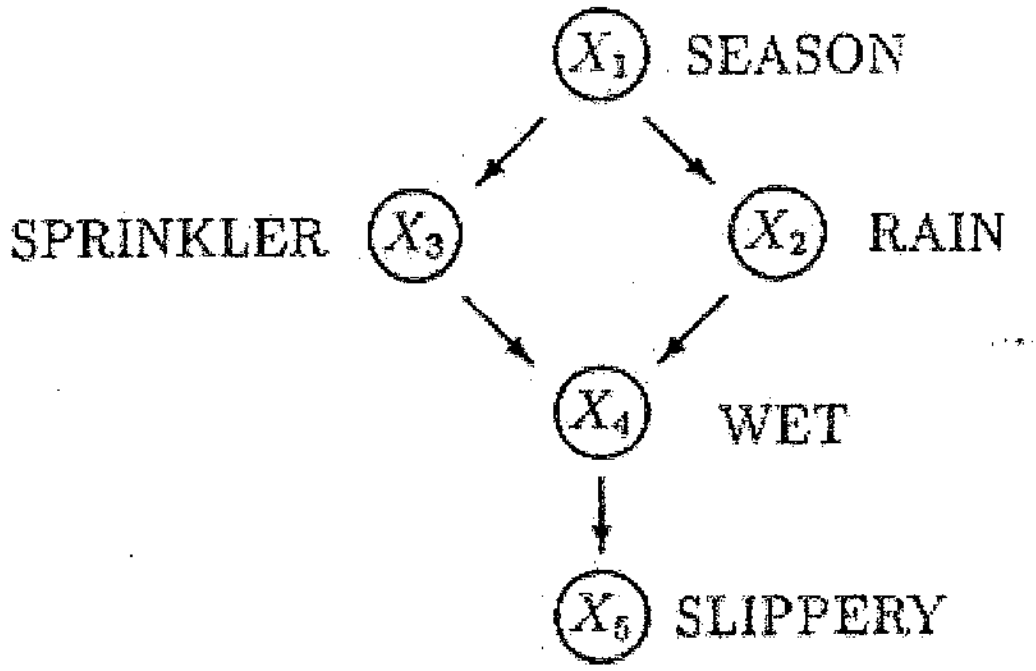
Η Μάθηση κατά Bayes αποτελεί μια ιδιαίτερα δημοφιλή προσέγγιση για την επαγωγική κατασκευή ταξινομητών, αφενός διότι εκπορεύεται από τον οικείο χώρο του Πιθανοτικού Λογισμού, αφετέρου διότι έχει επιδείξει σημαντικά αποτελέσματα σε ένα ευρύτατο φάσμα εφαρμογών. Η λειτουργία αυτής της κατηγορίας αλγορίθμων στηρίζεται στην υπόθεση ότι η υπό εκμάθηση έννοια σχετίζεται άμεσα με την κατανομή των πιθανοτήτων που παρουσιάζουν τα στιγμιότυπα του προβλήματος αναφορικά με την κλάση στην οποία ανήκουν. Ως βασικότερα πλεονεκτήματα της προσέγγισης αυτής μπορούμε να αναφέρουμε:

- Τη δυνατότητα αξιολόγησης των υποθέσεων στις οποίες καταλήγει ο αλγόριθμος μάθησης, μέσω της συσχέτισης ενός βαθμού εμπιστοσύνης της ορθότητάς τους, που αντιστοιχεί στην υπολογισθείσα πιθανότητα να είναι συνεπείς με την πλειοψηφία των παρατηρούμενων δεδομένων. Το χαρακτηριστικό αυτό συνεισφέρει στην παραγωγή εύρωστων μοντέλων, που εξασφαλίζουν ότι η αλήθεια μιας υπόθεσης δεν αμφισβητείται από μεμονωμένες περιπτώσεις στιγμιοτύπων για τις οποίες η υπόθεση κρίνεται ασυνεπής.
- Τη συμβολή της στη βαθύτερη κατανόηση και ανάλυση αλγορίθμων μάθησης οι οποίοι δε χειρίζονται απ' ευθείας πιθανότητες. Ένα χαρακτηριστικό παράδειγμα της ιδιότητας αυτής αποτελεί η μελέτη της επαγωγικής προδιάθεσης (inductive bias) ενός αλγορίθμου, του συνόλου των υποθέσεων δηλαδή στις οποίες στηρίζεται ο αλγόριθμος, ώστε να παράγει ένα μοντέλο ικανό να γενικεύει τις υποθέσεις στις οποίες κατέληξε κατά το χειρισμό άγνωστων στιγμιοτύπων.
- Την παροχή ενός μέτρου σύγκρισης έναντι άλλων μεθόδων Μ.Μ., καθώς οι αλγόριθμοι της κατηγορίας αυτής εγγυώνται τη βέλτιστη επίλυση ενός προβλήματος, δεδομένου ενός συνόλου υποθέσεων που απλοποιούν την κατασκευή του μοντέλου.

Το δίκτυο Bayes μπορεί να οριστεί ως ένα ζεύγος (G,p) όπου $G=(V, E)$ είναι ένας κατευθυνόμενος άκυκλος γράφος (directed acyclic graph-DAG) του οποίου οι κόμβοι

αναπαριστούν τυχαίες μεταβλητές και οι σύνδεσμοι αντιπροσωπεύουν τις αιτιολογικές επιδράσεις μεταξύ των μεταβλητών και p μια διακριτή συνάρτηση πιθανότητας με . Η ισχύς της επίδρασης αναπαρίσταται από δεσμευμένες πιθανότητες. Στον γράφο αυτό οι μεταβλητές είναι υπό συνθήκη ανεξάρτητες από αυτές με τις οποίες δεν ενώνονται, δεδομένων των γονιών τους.

Η παρακάτω εικόνα παρουσιάζει ένα απλό αλλά τυπικό δίκτυο Bayes. Περιγράφει τις αιτιολογικές σχέσεις ανάμεσα στην εποχή του έτους (X_1), εάν βρέχει(X_2) κατά την διάρκεια της εποχής, αν υπάρχει αυτόματο πότισμα (X_3), αν το πεζοδρόμιο είναι βρεγμένο(X_4) και τέλος αν το πεζοδρόμιο γλιστράει(X_5). Όλες οι μεταβλητές στο δίκτυο αυτό είναι δύτιμες 0/1(true/false), εκτός από την πρώτη που παίρνει 4 τιμές (άνοιξη - καλοκαίρι - φθινόπωρο - χειμώνας). Εδώ η απουσία μια απευθείας σύνδεσης ανάμεσα στις X_1 και X_5μας δείχνει ότι η επιρροή της μεταβλητότητας των εποχών στο εάν γλιστράει το πεζοδρόμιο, επιτυγχάνεται δια μεσολαβήσεως άλλων παραγόντων (π.χ. από το εάν είναι υγρό το πεζοδρόμιο).



Σχήμα 6.4. Τυπικό δίκτυο Bayes

Όπως δείχνει αυτό το παράδειγμα, ένα δίκτυο Bayes αποτελεί ένα μοντέλο του περιβάλλοντος. Στην πραγματικότητα προσομοιώνει τον αιτιολογικό μηχανισμό που λειτουργεί στο περιβάλλον και έτσι επιτρέπει στον ερευνητή να δώσει απάντηση σε μια σειρά ερωτημάτων όπως: 'Έχοντας παρατηρήσει το Α, τι μπορούμε να περιμένουμε για το Β ;' ή 'Τι θα συμβεί εάν επέμβουμε στο περιβάλλον;'. Απαντήσεις σε ερωτήσεις του πρώτου είδους βασίζονται μόνο στις πιθανότητες ενώ προκειμένου να απαντήσουμε σε ερωτήσεις του δευτέρου είδους πρέπει να βασιστούμε σε αιτιολογικές γνώσεις που περιέχει το δίκτυο. Και για τα δυο τις πληροφορίες τις παίρνουμε από το δίκτυο Bayes.

Το πιο σημαντικό χαρακτηριστικό που έχουν τα δίκτυα Bayes, είναι η ικανότητα τους να αναπαριστούν και να ανταποκρίνονται στις διάφορες αλλαγές της διαμόρφωσης. Κάθε τοπική αλλαγή στη διαμόρφωση του υπό μελέτη περιβάλλοντος , μπορεί να μεταφραστεί σε μια ισομορφική αναδιαμόρφωση της τοπολογίας του δικτύου. Για

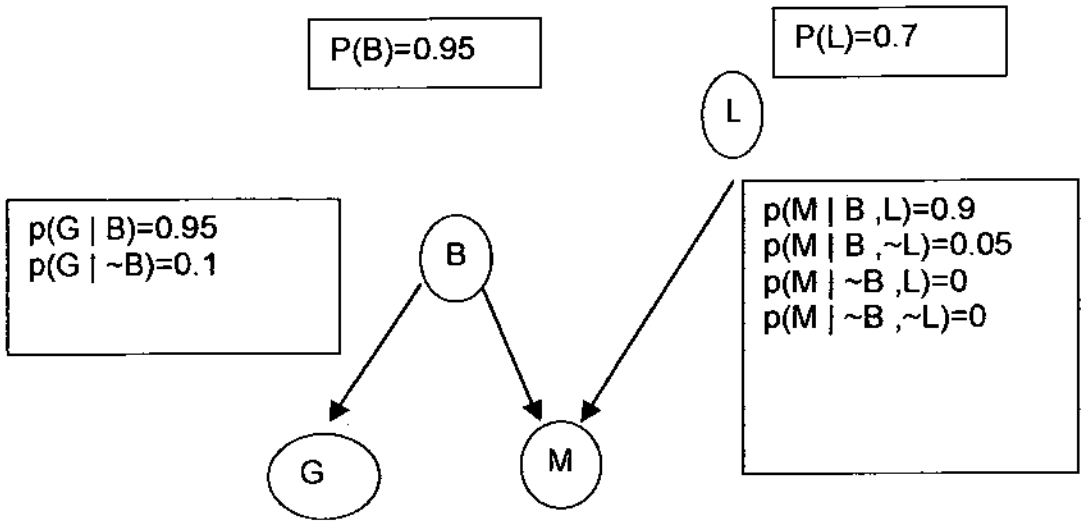
παράδειγμα, για να αναπαραστήσουμε ένα πεζοδρόμιο καλυμμένο με τέντα , απλώς διαγράφουμε την γραμμή που συνδέει την βροχή(X_2) με το υγρό πεζοδρόμιο(X_3).

Δεδομένου ενός DAG G και μιας από κοινού κατανομής P των διακριτών μεταβλητών $X_1, X_2 X_3, X_4, X_5$ λέμε ότι το G αντιπροσωπεύει την P μόνο εάν υπάρχει μια '1-1' σχέση μεταξύ των μεταβλητών X και των κόμβων του G, έτσι ώστε το P να παίρνει την εξής μορφή: $P(X_1, X_2 X_3, X_4, X_5) = \prod_{i=1}^{n} (P(X_i \mid pa_i))$

Όπου pa_i (γονείς) είναι οι κομβόι που βρίσκονται στο αμέσως προηγούμενο επίπεδο από την μεταβλητή X και συνδέονται με αυτήν , στο δίκτυο G. Για το συγκεκριμένο παράδειγμα που μελετάμε :

$$P(X_1, X_2 X_3, X_4, X_5) = P(X_1) * P(X_2 | X_1) * P(X_3 | X_1) * P(X_4 | X_2, X_3) * P(X_5 | X_4)$$

Δηλαδή με δεδομένους τους γονείς της, κάθε μεταβλητή X_i είναι ανεξάρτητη από όλες τις άλλες μεταβλητές που προηγούνται από αυτήν $\{X_1, X_2, ..., X_{i-1}) \setminus pa_i$. Αυτού του είδους οι ανεξαρτησίες καλούνται Μαρκοβιανές γιατί εκφράζουν τις Μαρκοβιανές συνθήκες για τις μεταβάσεις μεταξύ των καταστάσεων: Κάθε κατάσταση είναι ανεξάρτητη από το παρελθόν, δεδομένης της αμέσως προηγούμενης κατάστασης.
Έστω το δίκτυο:.



Σχήμα 6.5

Όπως συμπεραίνουμε από το σχήμα:

$$P(G,B,M,L)=P(G|B)*P(M|B,L)*P(B)*P(L).$$

Ο κανόνας της αλυσίδας θα μας έδινε:

$$P(G,B,M,L)=P(G|B,M,L)*P(M|B,L)*P(B|L)*P(L).$$

Παρατηρούμε λοιπόν ότι ενώ με την απλή εφαρμογή του νόμου της αλυσίδας θα απαιτούνταν 16 υπολογισμοί διαφόρων πιθανοτήτων, με το δίκτυο Bayes απαιτούνται μόνο 8. Αυτό μας επιτρέπει όπως καταλαβαίνουμε γρηγορότερο υπολογισμό της από κοινού πυκνότητας πιθανότητας και προβλήματα που μέχρι πρότινος ήταν NP τώρα μπορούν να λυθούν σε λιγότερο χρόνο.

6.4.1 Αφελής ταξινομητής Bayes

Ο αφελής ταξινομητής Bayes είναι η απλούστερη μορφή του Bayesian δικτύου. Σ' αυτό το άκυκλο γράφημα δεν υπάρχει άκρο μεταξύ των αριθμών και τα άκρα υπάρχουν μόνο μεταξύ της τάξης των αριθμών και των αριθμών. Αυτό το δίκτυο υποθέτει ότι κάθε χαρακτηριστικό (κάθε διακλάδωση του δικτύου)είναι ανεξάρτητο από τα υπόλοιπα χαρακτηριστικά, με δεδομένη την κατάσταση της τάξης των χαρακτηριστικών (την αρχή). Έτσι το ανεξάρτητο μοντέλο (αφελής Bayes) είναι βασισμένο στον υπολογισμό:

$$R= \frac{P(i/X)}{P(j/X)} = \frac{P(i)P(x/i)}{P(j)P(x/j)} = \frac{P(i)\prod P(X/i)}{P(j)\prod P(x/j)}$$

Συγκρίνοντας αυτές τις δύο πιθανότητες, η μεγαλύτερη πιθανότητα δηλώνει την τιμή της αξίας της τάξης που είναι η πιο πιθανή να είναι η πραγματική τιμή (αν R>1: προβλέπεται i, αν όχι j). Από τη στιγμή που ο αλγόριθμος ταξινόμησης Bayes χρησιμοποιεί το αποτέλεσμα μιας συνάρτησης για να υπολογίσει τις πιθανότητες P(x, i), έχει ιδιαίτερα την τάση να δέχεται υπερβολική επίδραση από την πιθανότητα του 0. Αυτό μπορεί να αποφευχθεί με τον τύπο του Laplace , προσθέτοντας τη μονάδα σε όλους τους αριθμητές και προσθέτοντας το αποτέλεσμα στον παρονομαστή.

Η υπόθεση της ανεξαρτησίας είναι καθαρά σχεδόν πάντα λάθος και γι' αυτό το λόγο ο αφελής ταξινομητής Bayes είναι συνήθως λιγότερο ακριβής απ' ότι άλλοι περισσότερο περίπλοκοι αλγόριθμοι μάθησης (όπως ANNs). Ωστόσο οι Domingo's & Pazzani (1997) παρουσίασαν μια μακροσκελής σύγκριση του αφελή ταξινομητή Bayes με σύγχρονους αλγόριθμους και βρήκαν ότι μερικές φορές είναι ανώτερος από τους άλλους αλγόριθμους εκμάθησης ακόμα και σε σειρά δεδομένων με ουσιώδης εξαρτώμενες μεταβλητές.

Το μεγάλο πλεονέκτημα του αφελή ταξινομητή Bayes είναι ο μικρός υπολογιστικός του χρόνος για εξάσκηση. Ωστόσο ο αφελής ταξινομητής Bayes είναι λιγότερο αποτελεσματικός στη χρήση της μνήμης της.

6.5 Μάθηση βασισμένη στα στιγμιότυπα

Οι βασισμένες στα στιγμιότυπα (instance-based, για συντομία IB) μέθοδοι μάθησης έχουν μια θεμελιώδη διαφορά από τις άλλες μεθόδους μάθησης που έχουν αναπτυχθεί: δεν κατασκευάζουν ένα γενικό ρητά διατυπωμένο μοντέλο που προσεγγίζει τη συνάρτηση-στόχο καθολικά. Το μόνο που κάνουν στη φάση της μάθησης είναι να αποθηκεύουν τα δεδομένα εκπαίδευσης, γι' αυτό είναι γνωστές και ως μέθοδοι βασισμένες στη μνήμη (memory-based). Η γενίκευση πέρα από τα παρατηρηθέντα δεδομένα γίνεται κάθε φορά που εμφανίζεται ένα νέο στιγμιότυπο προς κατάταξη. Τότε, ένα σύνολο από σχετιζόμενα με αυτό γνωστά στιγμιότυπα ανακαλείται από τη μνήμη και χρησιμοποιείται για την κατάταξη του νέου στιγμιοτύπου. Έτσι, αυτό που συμβαίνει ουσιαστικά είναι να παρέχεται μια τοπική προσέγγιση στη συνάρτηση-στόχο αντί μίας καθολικής [Aha et al. 1991]. Το κύριο πλεονέκτημα των IB μεθόδων είναι πως μπορούν να προσεγγίσουν πολύ καλύτερα από άλλες μεθόδους τη συνάρτηση-στόχο αν αυτή είναι πολύπλοκη καθολικά, αλλά μπορεί να περιγραφεί ως μια συλλογή λιγότερο σύνθετων τοπικών προσεγγίσεων. Το κύριο μειονέκτημα τους είναι πως το υπολογιστικό κόστος κατά την ταξινόμηση νέων στιγμιότυπων μπορεί να είναι πολύ υψηλό. Ο λόγος είναι πως σχεδόν όλοι οι υπολογισμοί λαμβάνουν χώρα τότε και όχι κατά τη φάση εκπαίδευσης.

Η IB μάθηση αναφέρεται και ως οκνηρή μάθηση (lazy learning), ακριβώς για το λόγο ότι αναβάλει τους υπολογισμούς μέχρι την αίτηση για κατάταξη ενός νέου στιγμιοτύπου (query). Έτσι, ένα σημαντικό πρακτικό ζήτημα είναι η ανάπτυξη τεχνικών αποδοτικής ευρετηριοποίησης των στιγμιότυπων εκπαίδευσης, για να μειωθεί ο χρόνος ανάκτησης τους κατά τη φάση κατάταξης.

Από τα παραπάνω, γίνεται αντιληπτό ότι οι αλγόριθμοι της κατηγορίας αυτής δεν κατασκευάζουν ένα καθολικό μοντέλο που να αναπαριστά τη γνώση που απέκτησαν από τα δεδομένα της εκπαίδευσης, αλλά ο προσδιορισμός της συνάρτησης στόχου γίνεται τοπικά, με κάθε ταξινόμηση ενός άγνωστου στιγμιοτύπου, αντλώντας πληροφορίες από τα χαρακτηριστικά της ομάδας στιγμιοτύπων με τα οποία συγγενεύει.

Αυτή ακριβώς η διαφοροποίηση της συγκεκριμένης κατηγορίας αλγορίθμων αποτελεί ένα από τα σημαντικότερα πλεονεκτήματα και συνάμα μειονεκτήματά τους. Ο τοπικός προσδιορισμός της συνάρτησης στόχου κατά την ταξινόμηση κάθε στιγμιοτύπου κρίνεται επιθυμητός όταν μια συνάρτηση στόχου, καθολικά συνεπής με το σώμα εκπαίδευσης, είναι ιδιαίτερα περίπλοκη. .ωστόσο, η μεταφορά του προσδιορισμού της συνάρτησης στόχου στο στάδιο της λήψης της απόφασης έχει ως αποτέλεσμα την αύξηση του κόστους ταξινόμησης νέων στιγμιοτύπων, τόσο ως προς τον χρόνο που απαιτείται όσο και ως προς την υπολογιστική πολυπλοκότητα. Ο παράγοντας αυτός μπορεί σε κάποιο βαθμό να αντισταθμισθεί χρησιμοποιώντας τεχνικές ευρετηριοποίησης των στιγμιοτύπων εκπαίδευσης. Σημαντικό χαρακτηριστικό επίσης για την αποτελεσματικότητα των αλγορίθμων αυτών αποτελεί η επιλογή της συνάρτησης απόστασης, αλλά και των χαρακτηριστικών εκείνων που θα χρησιμοποιηθούν κατά την εύρεση της ομάδας συγγενών στιγμιοτύπων, καθώς ενδέχεται ένα μικρό υποσύνολο των χαρακτηριστικών να είναι αρκετό, ενώ η χρήση περισσοτέρων να κριθεί επιζήμια για την ικανότητα γενίκευσης της μεθόδου. Τέλος, οι εν λόγω αλγόριθμοι χαρακτηρίζονται εν γένει για την αστάθειά τους στην ύπαρξη θορύβου στα δεδομένα εκπαίδευσης.

6.5.1 Αλγόριθμος των k κοντινότερων γειτόνων (k-Nearest Neighbor)

Ο αλγόριθμος ταξινόμησης με βάση τους k κοντινότερους γείτονες (k-Nearest Neighbor Algorithm – k-NN) είναι η πιο βασική IB μέθοδος μάθησης. Η κεντρική ιδέα είναι πως η τιμή της συνάρτησης-στόχου για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των k πιο "κοντινών" του στιγμιότυπων εκπαίδευσης, τα οποία αποτελούν τους "γείτονές" του. Τρία ζητήματα πρέπει να αποφασιστούν προκειμένου να καθοριστεί πλήρως ο αλγόριθμος:

- Ο ορισμός της απόστασης μεταξύ δύο στιγμιότυπων, δηλαδή μιας μετρικής πάνω στο χώρο των στιγμιότυπων (instance space), που θα εκφράζει την εγγύτητα, ή αλλιώς την "ομοιότητα" μεταξύ των στιγμιότυπων.
- Ο τρόπος συνδυασμού των τιμών των k κοντινότερων γειτόνων.
- Η τιμή του k.

Για το πρώτο ζήτημα, υπάρχουν πολλές εναλλακτικές επιλογές. Η απόφαση εξαρτάται από τα ειδικά χαρακτηριστικά του χώρου στιγμιότυπων του προβλήματος. Ιδιαίτερη σημασία έχει το αν στην αναπαράσταση των στιγμιότυπων περιλαμβάνονται αριθμητικά ή συμβολικά χαρακτηριστικά. Στον "παραδοσιακό" k-NN αλγόριθμο, στον οποίο τα στιγμιότυπα θεωρούνται πως ανήκουν στον n-διάστατο χώρο Rn, μια μετρική που υιοθετείται συχνά είναι η γνωστή Ευκλείδεια απόσταση. Πολυάριθμες άλλες μετρικές έχουν παρουσιαστεί. Οι περισσότερες απ΄ αυτές παρουσιάζονται στον πίνακα 1

Minkowsky: $D(x,y)=\left(\sum_{i=1}^{m}\lvert x_i-y_i\rvert^{r}\right)^{1/r}$	Euclidean: $D(x,y)=\left(\sum_{i=1}^{m}\lvert x_i-y_i\rvert^{2}\right)^{1/2}$
Manhattan: $D(x,y)=\sum_{i=1}^{m}\lvert x_i-y_i\rvert$	Camberra: $D(x,y)=\sum_{i=1}^{m}\frac{\lvert x_i-y_i\rvert}{\lvert x_i+y_i\rvert}$
Chebychev: $D(x,y)=\max_{i=1}^{m}\lvert x_i-y_i\rvert$	Kendall' s Rank Correlation: $D(x,y)=1-\frac{2}{m(m-1)}\sum_{i=j}^{m}\sum_{j=1}^{i-1}sign(x_i-x_j)sign(y_i-y_j)$

Πίνακας 6.1: Προσεγγίσεις για να καθορίσεις την απόσταση μεταξύ των περιπτώσεων (x και y)

Επίσης γενική είναι η προσέγγιση της βελτιστοποίησης παραμέτρων μέσω της χρησιμοποίησης του ίδιου του αλγορίθμου μάθησης ως μέσο για την αξιολόγηση της απόδοσης, η οποία αναφέρεται ως προσέγγιση περιτυλίγματος (wrapper approach).

Ο k-NN είναι ένας πολύ αποτελεσματικός αλγόριθμος μάθησης, τόσο για αριθμητικά όσο και για συμβολικά δεδομένα, ιδιαίτερα όταν γίνεται με αποτίμηση χαρακτηριστικών και γειτόνων. Είναι ανθεκτικός σε θορυβώδη στιγμιότυπα εκπαίδευσης, ειδικά για μεγαλύτερες τιμές του k, καθώς τα απομονωμένα λανθασμένα δεδομένα "απορροφώνται" κατά τον υπολογισμό του μέσου όρου. Η επαγωγική κλίση του k-NN είναι η υπόθεση πως η τιμή της συνάρτησης-στόχου ενός στιγμιοτύπου είναι παρόμοια με αυτή των γειτονικών του. Ένα πρακτικό θέμα κατά την εφαρμογή του k-NN, όπως αναφέρθηκε και παραπάνω για τις IB μεθόδους γενικότερα, είναι η αποδοτική ευρετηριοποίηση των στιγμιότυπων στη μνήμη. Σε μια απλή υλοποίηση, η υπολογιστική πολυπλοκότητα για την κατάταξη ενός νέου στιγμιοτύπου είναι ανάλογη του αριθμού των στιγμιότυπων εκπαίδευσης, αφού χρειάζεται να υπολογιστεί η απόσταση του νέου με κάθε στιγμιότυπο εκπαίδευσης, για να επιλεχθούν στη συνέχεια τα k κοντινότερα. Κάτι τέτοιο έχει υψηλότατο κόστος για μεγάλα σύνολα δεδομένων. Για το λόγο αυτό έχουν αναπτυχθεί διάφορες μέθοδοι ευρετηριοποίησης, όπως τα k-d δέντρα (k-d trees) [Friedman et al. 1977], που σκοπό έχουν τον πιο γρήγορο εντοπισμό των κοντινότερων γειτόνων με κάποιο επιπλέον κόστος στη μνήμη.

6.6. ΓΡΑΜΜΙΚΑ SUPPORT VECTOR MACHINES

Ας προσδιορίσουμε τα δεδομένα μας με $\{\mathbf{x}_i, y_i\}$, $i=1,\ldots,k$, $y_i \in \{-1, 1\}$ και $x_i \in R^d$. Έστω ότι έχουμε κάποια υπερεπίπεδα που διαχωρίζουν τα θετικά από τα αρνητικά παραδείγματα Τα σημεία $\mathbf{x}$ τα οποία βρίσκονται στο υπερεπίπεδο ικανοποιούν την εξίσωση $\mathbf{wx} + b = 0$ όπου w το κανονικό υπερεπίπεδο , $|b| / \|w\|$ η κάθετη απόσταση του υπερεπιπέδου από την αρχή και $\|w\|$ η ευκλείδεια νόρμα .
Έστω d_+ και d_- η ελάχιστη απόσταση του υπερεπιπέδου διαχώρισης και του πλησιέστερου θετικού ή αρνητικού παραδείγματος αντίστοιχα .

Για τη γραμμική διαχωρίσιμη περίπτωση ο SV αλγόριθμος αναζητά τα υπερεπίπεδα με τη μεγαλύτερη απόσταση (**margin**). Το παραπάνω παίρνει την εξής μορφή προβλήματος:

Έστω ότι όλα τα δεδομένα εκπαίδευσης ικανοποιούν τους παρακάτω περιορισμούς

$\mathbf{x}_i \mathbf{w}+b \geq +1$ για $y_i=+1$ (10)

$\mathbf{x}_i \mathbf{w}+b \leq -1$ για $y_i=-1$ (11)

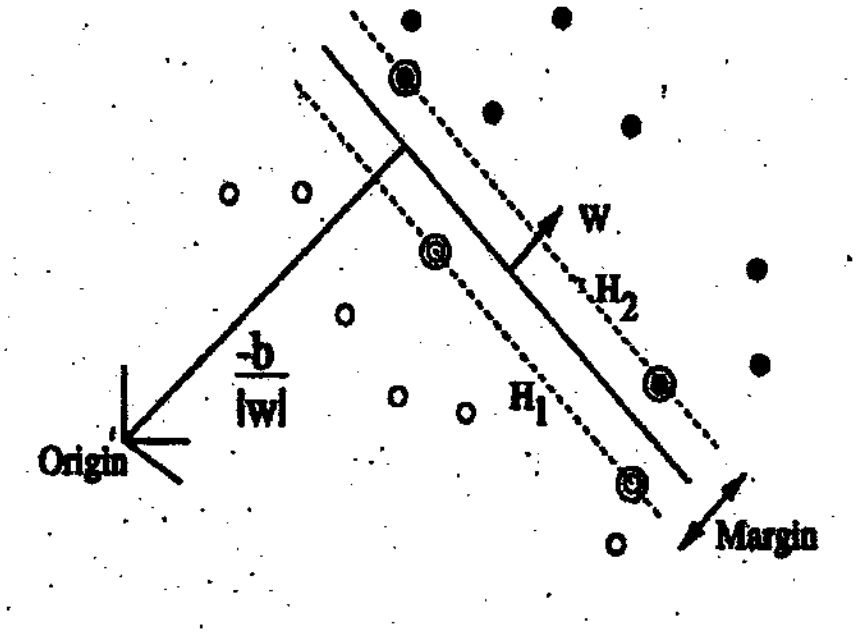
τα οποία μπορούν να συνδυαστούν σε ένα σύνολο ανισώσεων

$y_i (x_i \mathbf{w}+b)-1 \geq 0 \; \forall i$ (12).

Τώρα ας θεωρήσουμε τα σημεία που ορίζονται από την (10) . Αυτά τα σημεία βρίσκονται στο υπερεπίπεδο H_1 $\mathbf{x}_i \mathbf{w}+b = +1$ με κανονικό **w** και κάθετη απόσταση από την αρχή $|b-1|/\|\mathbf{w}\|$.Ανάλογα τα σημεία που ικανοποιούν την **(11)** βρίσκονται στο υπερεπίπεδο H_2 $\mathbf{x}_i \mathbf{w}+b = -1$, με κανονικό **w** και κάθετη απόσταση από την αρχή ίση με $|-b-1|/\|\mathbf{w}\|$.Έτσι d_+ ,$d_-=1/\|\mathbf{w}\|$ και το margin είναι απλά $2/\|\mathbf{w}\|$. Σημειώνουμε δε ότι τα H_1 , H_2 είναι προφανώς παράλληλα και ότι πλέον είναι πολύ εύκολο να βρούμε ένα ζεύγος από υπερεπίπεδα που δίνει τη maximum (margin) απόσταση με την ελαχιστοποίηση το $\|\mathbf{w}\|^2$, υπό τους περιορισμούς **(12).** Όπως θα περίμενε κανείς η λύση μιας τυπικής περίπτωσης σε δύο διαστάσεις έχει τη μορφή που παρουσιάζεται στο (**σχήμα 6.7**) .Τα σημεία μάθησης για τα οποία η ανισότητα **(12)** ισχύει και η αφαίρεση των οποίων θα αλλάξει τη λύση που βρίσκουμε καλούνται Support Vectors (διανύσματα υποστήριξης) και παρουσιάζονται στο (**σχήμα 6.7**) μέσα σε κύκλους.

Τώρα θα αντιμετωπίσουμε το πρόβλημα με τη βοήθεια των πολλαπλασιαστών **Lagrange .**

Εισάγουμε τους θετικούς πολλαπλασιαστές Lagrange α_i με i= 1,...,k έναν δηλαδή για κάθε ένα από τους περιορισμούς του (12).Εδώ θα ήταν πολύ χρήσιμο να θυμηθούμε τον κανόνα , πως οι περιορισμοί της μορφής $c_i \geq 0$ οι εξισώσεις των περιορισμών πολλαπλασιάζονται με θετικούς πολλαπλασιαστές Lagrange και αφαιρούμενοι από την αντικειμενική συνάρτηση σχηματίζουμε τη Lagrangian (λαγκρανζιανή) .



Σχήμα 6.6

Αντίστοιχα για ισοτικούς περιορισμούς οι πολλαπλασιαστές Lagrange δε περιορίζονται Έτσι παίρνουμε τη Lagrangian

$$L_p = 1/2\ \|w\|^2 - \sum_{i=1}^{k} \alpha_i y_i\ (x_i \cdot w + b) + \sum_{i=1}^{k} \alpha_i \qquad (13).$$

Τώρα πρέπει να ελαχιστοποιήσουμε την L_p ως προς τα w, b και αυτομάτως να απαιτήσουμε οι παράγωγοι ως προς όλα τα α_i να εξαφανίζονται φυσικά πάντα από τους περιορισμούς $\alpha_i \geq 0\ \forall\ i$.

Ουσιαστικά καταλήξαμε σε ένα πρόβλημα κυρτού τετραγωνικού προγραμματισμού αφού η αντικειμενική μας συνάρτηση είναι κυρτή και τα σημεία τα οποία ικανοποιούν τους περιορισμούς αποτελούν και αυτά κυρτό σύνολο.

Αυτό μας οδηγεί στο συμπέρασμα ότι μπορούμε ισοδυνάμως να λύσουμε το εξής δυϊκό πρόβλημα : Μεγιστοποίησε το L_p υπό τον περιορισμό το gradient του L_p να χάνεται και

$\alpha_i \geq 0$.Αυτή η συγκεκριμένη αναδιατύπωση του προβλήματος καλείται και δυϊκό WOLVE .

Απαιτώντας το ∇L_p να εξαφανίζεται έχω $\mathbf{w}=\sum_{i=1}^{k} \alpha_i y_i \cdot \mathbf{x}_i$ (14)

$$\sum_{i=1}^{k} \alpha_i y_i=0 \quad (15).$$

Αφού αυτοί είναι ισοτικοί περιορισμοί στο δυϊκό , μπορούμε να αντικαταστήσουμε στην εξίσωση (13) και να πάρουμε $L_D= \sum_{i} \alpha_i - 1/2 \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$ (16)

Ο SMO είναι ένας πρόσφατος αλγόριθμος για τη SVM , που προτάθηκε από τον Platt (1999). Ο Platt χωρίζει το μεγάλο πρόβλημα σε μικρά υπό – προβλήματα κάτω από ορισμένες συνθήκες και λύνει κάθε υπό – πρόβλημα χωριστά. Οι Keerthi & Gilbert (2002) πρότειναν δύο τροποποιημένες εκδόσεις του SMO που είναι αξιοσημείωτα γρηγορότερες από το αυθεντικό SMO στις περισσότερες περιπτώσεις.

7. Χρήση WEKA

Ο σκοπός της εργασίας μας είναι να χρησιμοποιήσει αλγόριθμους εξόρυξης γνώσης για την έγκριση πιστωτικών καρτών βασιζόμενη σε στοιχεία όπως ηλικία, εισόδημα, πιστωτική ιστορία και ιδιοκτησία κατοικίας κλπ. Αξιολογήσαμε αρκετούς αλγόριθμους εξόρυξης γνώσης και επιλέξαμε τον καλύτερο, χρησιμοποιώντας το εργαλείο εξόρυξης δεδομένων WEKA.

7.1 Περίπτωση: Credit-A

Η κάθε περίπτωση των δεδομένων μας αντιπροσωπεύει μια αίτηση για εγκαταστάσεις πιστωτικών καρτών που περιγράφονται από οκτώ «φρόνημα» και έξι συνεχή χαρακτηριστικά, με δύο περιπτώσεις απόφασης (Αποδέχομαι/ Απορρίπτω). Στο UCI Repository τα αρχικά ονόματα των χαρακτηριστικών έχουν αλλαχθεί σε σύμβολα χωρίς νόημα (A1 – A14) με την αιτιολογία ότι προστατεύουν το απόρρητο των δεδομένων. Ωστόσο τα πραγματικά ονόματα των χαρακτηριστικών είναι διαθέσιμα στη σελίδα της Rulequest Research (http:/www.rulequest.com/see5-examples.html).

Τα χαρακτηριστικά των βάσεων δεδομένων φαίνονται στον πίνακα 7.1. παρακάτω. Τα αρχικά ονόματα των χαρακτηριστικών (που μας παρέχονται από τη σελίδα της Rulequest Research) δίνονται στις παρενθέσεις. Η στήλη πεδίο δείχνει την σειρά ή την κατηγορία των πιθανών αξιών για κάθε χαρακτηριστικό. Στη στήλη τύπος κάνουμε μια διάκριση μεταξύ κατηγορικών και αριθμητικών μεταβλητών.

Χαρακτηριστικό	Πεδίο	Τύπος
A1 (φύλο)	0, 1	Αριθμητικός
A2(ηλικία)	13,75-80,25	Συνεχής
A3(Παρούσα διεύθυνση)	0-28	Συνεχής

A4(Οικογενειακή κατάσταση)	1,2,3	Αριθμητικός
A5(Παρούσα απασχόληση)	1-14	Αριθμητικός
A6(Παρούσα εργασιακή κατάσταση)	1-9	Αριθμητικός
A7(Φορέας απασχόλησης)	0-28,5	Συνεχής
A8(Άλλες επενδύσεις)	0, 1	Αριθμητικός
A9(Τραπεζικός λογ/μός)	0, 1	Αριθμητικός
A 10(Παρούσα τράπεζα)	0-67	Συνεχής
A11(Αναφορά υπευθυνότητας)	0, 1	Αριθμητικός
A12(Οικονομική ενσωμάτωση)	1,2,3	Αριθμητικός
A13(Μηνιαία εξοδα κατοικίας)	0-2000	Συνεχής
A14(Οικονομική ισορροπία αποταμιεύσεων)	1-100001	Συνεχής
Τάξη (Απορρίπτω/ Αποδέχομαι)	0, 1	Αριθμητικός

Πίνακας 7.1 Χαρακτηριστικά βάσεων δεδομένων

Εν συνεχεία παρουσιάζουμε τους παραγόμενους ταξινομητές και την ακρίβεια πρόγνωσης (τις σωστές προβλέψεις/το σύνολο των προβλέψεων) για κάθε εξεταζόμενο αλγόριθμο εξόρυξης γνώσης.

7.1.1. Αφελής ταξινομητής Bayes

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class +: Prior probability = 0.45

A1: Discrete Estimator. Counts = 207 99 (Total = 306)

A2: Normal Distribution. Mean = 33.723 StandardDev = 12.7816 WeightSum = 305 Precision = 0.1910919540229885

A3: Normal Distribution. Mean = 5.9075 StandardDev = 5.4649 WeightSum = 307 Precision = 0.1308411214953271

A4: Discrete Estimator. Counts = 257 46 3 1 (Total = 307)

A5: Discrete Estimator. Counts = 257 46 3 (Total = 306)

A6: Discrete Estimator. Counts = 63 8 30 15 4 15 17 3 52 34 33 15 20 8 (Total = 317)

A7: Discrete Estimator. Counts = 170 88 26 4 3 7 3 9 2 (Total = 312)

A8: Normal Distribution. Mean = 3.4242 StandardDev = 4.1179 WeightSum = 307 Precision = 0.21755725190839695

A9: Discrete Estimator. Counts = 285 24 (Total = 309)

A10: Discrete Estimator. Counts = 210 99 (Total = 309)

A11: Normal Distribution. Mean = 4.6823 StandardDev = 6.5274 WeightSum = 307 Precision = 3.0454545454545454

A12: Discrete Estimator. Counts = 147 162 (Total = 309)

A13: Discrete Estimator. Counts = 288 6 16 (Total = 310)

A14: Normal Distribution. Mean = 164.1864 StandardDev = 161.3686 WeightSum = 301 Precision = 11.834319526627219

A15: Normal Distribution. Mean = 2027.9939 StandardDev = 7655.808 WeightSum = 307 Precision = 418.41004184100416

Class -: Prior probability = 0.55

A1: Discrete Estimator. Counts = 263 113 (Total = 376)

A2: Normal Distribution. Mean = 29.8068 StandardDev = 10.9057 WeightSum = 373 Precision = 0.1910919540229885

A3: Normal Distribution. Mean = 3.8409 StandardDev = 4.3316 WeightSum = 383 Precision = 0.1308411214953271

A4: Discrete Estimator. Counts = 264 119 1 1 (Total = 385)

A5: Discrete Estimator. Counts = 264 119 1 (Total = 384)

A6: Discrete Estimator. Counts = 76 24 13 46 8 38 23 2 28 32 7 12 36 47 (Total = 392)

A7: Discrete Estimator. Counts = 231 52 35 6 3 3 5 50 2 (Total = 387)

A8: Normal Distribution. Mean = 1.2525 StandardDev = 2.1128 WeightSum = 383 Precision = 0.21755725190839695

A9: Discrete Estimator. Counts = 78 307 (Total = 385)

A10: Discrete Estimator. Counts = 87 298 (Total = 385)

A11: Normal Distribution. Mean = 0.6043 StandardDev = 1.9863 WeightSum = 383 Precision = 3.0454545454545454

A12: Discrete Estimator. Counts = 171 214 (Total = 385)

A13: Discrete Estimator. Counts = 339 4 43 (Total = 386)

A14: Normal Distribution. Mean = 199.7986 StandardDev = 181.3128 WeightSum = 376

A15: Normal Distribution. Mean = 186.8097 StandardDev = 675.6316 WeightSum = 383 Precision = 418.41004184100416

Time taken to build model: 0.72 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	536	77.6812 %
Incorrectly Classified Instances	154	22.3188 %
Kappa statistic	0.534	
Mean absolute error	0.2228	
Root mean squared error	0.4356	
Relative absolute error	45.0979 %	
Root relative squared error	87.6429 %	
Total Number of Instances	690	

Άρα η ακρίβεια του Naïve Bayes στα δεδομένα ήταν 77.68%.

7.1.2. Δέντρο απόφασης C4.5

=== Classifier model (full training set) ===

C4.5

A9 = t
| A10 = t: + (228.0/21.0)
| A10 = f
| | A15 <= 444
| | | A7 = v
| | | | A4 = u
| | | | | A14 <= 112: + (16.57/1.57)
| | | | | A14 > 112
| | | | | | A15 <= 70: - (30.0/10.0)
| | | | | | A15 > 70: + (2.0)
| | | | A4 = y
| | | | | A13 = g: - (12.0/2.0)
| | | | | A13 = p: - (0.0)
| | | | | A13 = s: + (3.0/1.0)
| | | | A4 = l: - (0.0)
| | | | A4 = t: - (0.0)
| | | A7 = h: + (27.24/8.24)
| | | A7 = bb
| | | | A3 <= 1.375: + (5.0/1.0)
| | | | A3 > 1.375: - (9.13/1.0)
| | | A7 = j: - (1.01)
| | | A7 = n: + (0.0)
| | | A7 = z: + (0.0)

| | | A7 = dd: + (1.01/0.01)
| | | A7 = ff: - (5.05/1.0)
| | | A7 = o: + (0.0)
| | A15 > 444: + (21.0/1.0)
A9 = f
| A3 <= 0.165
| | A7 = v
| | | A2 <= 35.58: - (18.72/3.44)
| | | A2 > 35.58: + (3.6/0.16)
| | A7 = h: - (0.0)
| | A7 = bb: + (1.24/0.08)
| | A7 = j: + (1.24/0.08)
| | A7 = n: + (1.24/0.08)
| | A7 = z: - (0.0)
| | A7 = dd: - (0.0)
| | A7 = ff: - (4.96/0.64)
| | A7 = o: - (0.0)
| A3 > 0.165: - (298.0/12.0)

Number of Leaves: 30

Size of the tree: 42

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	594	86.087 %
Incorrectly Classified Instances	96	13.913 %
Kappa statistic	0.718	

Mean absolute error	0.1924
Root mean squared error	0.3313
Relative absolute error	38.9417 %
Root relative squared error	66.6637 %
Total Number of Instances	690

Άρα η ακρίβεια του C4.5 στα δεδομένα ήταν 86.09%.

7.1.3. Κανόνες αποφάσεων από τον αλγόριθμο RIPPER

RIPPER rules:

==========

(A9 = t) and (A15 >= 234) => class=+ (157.0/7.0)

(A9 = t) and (A10 = t) => class=+ (99.0/18.0)

(A9 = t) and (A14 <= 110) and (A15 <= 0) => class=+ (31.0/5.0)

=> Class=- (403.0/50.0)

Number of Rules: 4

Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	592	85.7971 %
Incorrectly Classified Instances	98	14.2029 %
Kappa statistic	0.7136	
Mean absolute error	0.2145	
Root mean squared error	0.345	

Relative absolute error	43.4166 %
Root relative squared error	69.4163 %
Total Number of Instances	690

Άρα η ακρίβεια του RIPPER στα δεδομένα ήταν 85.79%. Πρέπει επίσης ότι αυτή η ακρίβεια επιτεύχθηκε χρησιμοποιώντας τέσσερις μόνο κανόνες.

7.1.4. Αλγόριθμος SMO

BinarySMO

Machine linear: showing attribute weights, not support vectors.

```
      0.0002 * (normalized) A1
+     -0.0042 * (normalized) A2
+      0.001  * (normalized) A3
+      0.3316 * (normalized) A4=u
+      0.3325 * (normalized) A4=y
+     -0.664  * (normalized) A4=l
+      0.3316 * (normalized) A5=g
+      0.3325 * (normalized) A5=p
+     -0.664  * (normalized) A5=gg
+     -0.0033 * (normalized) A6=c
+      0      * (normalized) A6=d
+     -0.0053 * (normalized) A6=cc
+     -0.0028 * (normalized) A6=i
+      0.0135 * (normalized) A6=j
+      0.001  * (normalized) A6=k
+     -0.0016 * (normalized) A6=m
```

```
+     0      * (normalized) A6=r
+    -0.0038 * (normalized) A6=q
+    -0.0038 * (normalized) A6=w
+    -0.0068 * (normalized) A6=x
+    -0.0051 * (normalized) A6=e
+    -0.0011 * (normalized) A6=aa
+     0.0191 * (normalized) A6=ff
+     0.0015 * (normalized) A7=v
+     0.0004 * (normalized) A7=h
+     0.0032 * (normalized) A7=bb
+    -0.0143 * (normalized) A7=j
+     0.012  * (normalized) A7=z
+     0.0019 * (normalized) A7=dd
+    -0.0063 * (normalized) A7=ff
+     0.0015 * (normalized) A7=o
+    -0.0087 * (normalized) A8
+     2.0008 * (normalized) A9
+     0.0013 * (normalized) A10
+    -0.0255 * (normalized) A11
+     0.0003 * (normalized) A12
+     0.5003 * (normalized) A13=g
+    -1      * (normalized) A13=p
+     0.4997 * (normalized) A13=s
+     0.0195 * (normalized) A14
+    -0.0919 * (normalized) A15
-    2.162

Number of kernel evaluations: 146567

Time taken to build model: 3.09 seconds
```

=== Stratified cross-validation ===		
=== Summary ===		
Correctly Classified Instances	586	84.9275 %
Incorrectly Classified Instances	104	15.0725 %
Kappa statistic	0.7003	
Mean absolute error	0.1507	
Root mean squared error	0.3882	
Relative absolute error	30.5133 %	
Root relative squared error	78.1202 %	
Total Number of Instances	690	

Άρα η ακρίβεια του SMO στα δεδομένα ήταν 84.93%.

7.1.5. Αλγόριθμος BP για νευρωνικά δίκτυα

=== Classifier model (full training set) ===

Sigmoid Node 0
  Inputs Weights
  Threshold 1.6104258980164028
  Node 2 -4.443436741637387
Sigmoid Node 1
  Inputs Weights
  Threshold -1.6104258980164028
  Node 2 4.4434367416373854
Sigmoid Node 2
  Inputs Weights
  Threshold 0.770654936922311
  Attrib A1 2.4686158578559754

Attrib A2 0.8708952741587124
Attrib A3 -0.7549003156668521
Attrib A4=u -1.252814575011715
Attrib A4=y 1.2731626928266144
Attrib A4=l -0.8256421775758617
Attrib A4=t -0.018649069530998742
Attrib A5=g -1.2232136702861873
Attrib A5=p 1.2793061063753015
Attrib A5=gg -0.8028974465710356
Attrib A6=c -2.1191703452723525
Attrib A6=d 1.1120027067707585
Attrib A6=cc -2.6843318698573273
Attrib A6=i -1.4713722840422685
Attrib A6=j -1.6563898371098598
Attrib A6=k 5.520794265236609
Attrib A6=m 0.1863790932329909
Attrib A6=r -0.8054759846117279
Attrib A6=q 2.164039086823643
Attrib A6=w 0.5658601329308847
Attrib A6=x -2.514722554169417
Attrib A6=e -10.62883677481209
Attrib A6=aa 3.134327320776049
Attrib A6=ff 2.0765942952825274
Attrib A7=v -4.252076488931497
Attrib A7=h -3.9745622887954943
Attrib A7=bb 4.422678836776461
Attrib A7=j -1.6640772168090898
Attrib A7=n -0.5309832459227579
Attrib A7=z -1.1535365212607271
Attrib A7=dd 1.528578732692125

Attrib A7=ff 2.030109631120246
Attrib A7=o -0.5576360965635871
Attrib A8 -2.940881884083295
Attrib A9 19.84931142649782
Attrib A10 7.595041544872093
Attrib A11 -2.2558404973487423
Attrib A12 2.0871426254550425
Attrib A13=g -1.5003977682739393
Attrib A13=p -0.7431097370472662
Attrib A13=s 1.5371222088707703
Attrib A14 5.939779629008369
Attrib A15 -1.5960341690201199
Class +
Input
Node 0
Class -
Input
Node 1

Time taken to build model: 4.39 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	593	85.942 %
Incorrectly Classified Instances	97	14.058 %
Kappa statistic	0.7157	
Mean absolute error	0.205	
Root mean squared error	0.3436	
Relative absolute error	41.5017 %	

Root relative squared error	69.147 %
Total Number of Instances	690

Άρα η ακρίβεια του BP στα δεδομένα ήταν 85.94%.

7.2 Περίπτωση: German Credit

Η German Credit σειρά δεδομένων (διαθέσιμο στο ftp.ics.uci.edu/pub/mashine-learning-databesees/statlog/) περιέχει παρατηρήσεις σε 30-μεταβλητές-για 1000 αιτώντες στο παρελθόν για πίστωση. Κάθε αιτών έχει χαρακτηριστεί ως «καλός πιστωτής» (700 Περιπτώσεις) ή «κακός πιστωτής» (300 περιπτώσεις)

Νέοι αιτούντες για πίστωση μπορούν επίσης να εξεταστούν σ' αυτές τις 30 μεταβλητές που προβλέπουν. Θέλουμε να αναπτύξουμε ένα επιτυχημένο κανόνα πίστωσης που μπορεί να χρησιμοποιηθεί για να καθορίσει αν ένας νέος αιτώντας είναι ένας καλός πιστωτικός κίνδυνος ή κακός πιστωτικός κίνδυνος, βασιζόμενοι σε αξίες για μία ή περισσότερες από τις μεταβλητές που προβλέπουν. Όλες οι μεταβλητές εξηγούνται στον πίνακα 7.2.

Α/Α	Όνομ. Μεταβλητής	Περιγραφή	Τύπος Μεταβλητής	Κωδ. Περιγραφής
1	OBS#	Αριθμός παρατήρησης	Κατηγορηματικός	Αλληλουχία αριθμών στη σειρά δεδομένων
2	CHK_ACCT	Κατάσταση ελέγχου λογ/μού	Κατηγορηματικός	0:<0 DM 1:0<=...<DM 2:=>200DM 3: μη ελενχόμενος λογ/μος
3	DURATION	Διάρκεια	Αριθμητικός	

		πίστωσης σε μήνες		
4	HISTORY	Ιστορικό πίστωσης	Κατηγορηματικός	0: δεν έχουν παρθεί πιστώσεις 2:όλες οι πιστώσεις σ' αυτή την τράπεζα έχουν πληρωθεί 3:καθυστερήσει στην εξόφληση στο παρελθόν 4:κρίσιμος λογ/μος
5	NEW_CAR	Αιτία πίστωσης	Δυαδικός	Αυτοκίνητο(νέο):0:όχι, 1:ναι
6	USED_CAR	Αιτία πίστωσης	Δυαδικός	Αυτοκίνητο(παλιό): 0:όχι, 1:ναι
7	FURNITURE	Αιτία πίστωσης	Δυαδικός	Επίπλωση/ εξοπλισμός: 0:όχι, 1:ναι
8	RADIO/TV	Αιτία πίστωσης	Δυαδικός	Ράδιο/τηλεόραση: 0:όχι, 1:ναι
9	EDUCATION	Αιτία πίστωσης	Δυαδικός	Μόρφωση: 0:όχι, 1:ναι
10	RETRAINING	Αιτία πίστωσης	Δυαδικός	Μετεκπαίδευση: 0:όχι, 1:ναι
11	AMOUNT	Ποσό πίστωσης	Αριθμητικός	
12	SAV_ACCT	Μέση ισορροπία σε αποταμιευτικούς λογ/μους	Κατηγορηματικός	0:<100DM 1:100<=...<500 DM 2:500<=...<1000DM 3:=> 1000 DM 4: άγνωστος/μη αποθεματικός λογ/μος

13	EMPLOYMENT	Παρούσα εργασία	Κατηγορηματικός	0: άνεργος 1:< 1 χρόνο 2:1<=...<4 χρόνο 3:4<=...<7 χρόνια 4: >= 7 χρόνια
14	INSTALL_RATE	Ποσοστό δόσης επί %στο διατιθέμενο εισόδημα	Αριθμητικός	0:όχι, 1:ναι
15	MALE_DIV	Ο αιτών είναι άνδρας και διαζευγμένος	Δυαδικός	0:όχι, 1:ναι
16	MALE_SINGLE	Ο αιτών είναι άνδρας και ανύπαντρος	Δυαδικός	0:όχι, 1:ναι
17	MALE_MAR_WID	Ο αιτών είναι άνδρας και παντρεμένος ή χήρος	Δυαδικός	0:όχι, 1:ναι
18	CO-APPLICANT	Η αίτηση έχει και δεύτερο αιτούντα	Δυαδικός	0:όχι, 1:ναι
19	GUARANTOR	Ο αιτών έχει εγγυητή	Δυαδικός	0:όχι, 1:ναι
20	PRESENT_RESIDENT	Παρούσα κατοικία από πότε	Κατηγορηματικός	0: <=1 χρόνο 1: <...<= 2 χρόνια 2: <...<= 3 χρόνια 3: > 4 χρόνια
21	REAL_ESTATE	Ο αιτών κατέχει	Δυαδικός	0:όχι, 1:ναι

		ακίνητο		
22	PROP_UNKN_ NONE	Ο αιτών δεν έχει ιδιοκτησία(ή είναι άγνωστη)	Δυαδικός	0:όχι, 1:ναι
23	AGE	Ηλικία σε χρόνια	Αριθμητικός	0:όχι, 1:ναι
24	OTHER_ INSTALL	Ο αιτών έχει και άλλα πιστωτικά σχέδια	Δυαδικός	0:όχι, 1:ναι
25	RENT	Ο αιτών νοικιάζει	Δυαδικός	0:όχι, 1:ναι
26	OWN_RES	Ο αιτών ιδιοκατοικεί	Δυαδικός	0:όχι, 1:ναι
27	NUM_CREDITS	Αριθμός από υπάρχουσες πιστώσεις στην τράπεζα	Αριθμητικός	
28	JOB	Φύση της δουλείας	Κατηγορηματικός	0:άνεργος/ ανειδίκευτος/ όχι μόνιμος 1: ανειδίκευτος/ μόνιμος 2:ειδικευμένος/ μόνιμος 3:στέλεχος/ ελευθ. Επαγγελματίας/ υψηλών προσόντων εργαζόμενος/ αξιωματούχος
29	NUM_ DEPENDENTS	Αριθμός ατόμων από τα οποία εξαρτάται η	Αριθμητικός	

		μονιμοποίηση		
30	TELEPHONE	Ο αιτών έχει αριθμό τηλεφώνου στο όνομά του/ της	Δυαδικός	0:όχι, 1:ναι
31	FOREIGN	Αλλοδαπός εργοδότης	Δυαδικός	0:όχι, 1:ναι
32	RESPONSE	Το ποσοστό πίστωσης είναι καλό	Δυαδικός	0:όχι, 1:ναι

Πίνακας 7.2 Μεταβλητές που καθορίζουν αν ένας νέος αιτώντας είναι ένας καλός πιστωτικός κίνδυνος ή κακός πιστωτικός κίνδυνος.

Εν συνεχεία παρουσιάζουμε τους παραγόμενους ταξινομητές και την ακρίβεια πρόγνωσης για κάθε εξεταζόμενο αλγόριθμο εξόρυξης γνώσης.

7.2.1. Αφελής ταξινομητής Bayes

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class good: Prior probability = 0.7

checking_status: Discrete Estimator. Counts = 140 165.50 349 (Total = 704)
Duration: Normal Distribution. Mean = 19.1766 StandardDev = 10.9817 WeightSum = 700 Precision = 2.125

Credit_history: Discrete Estimator. Counts = 16 22 362 61 244 (Total = 705)
Purpose: Discrete Estimator. Counts = 146 87 124 219 9 15 29 1 9 64 8 (Total = 711)
Credit_amount: Normal Distribution. Mean = 2985.6721 StandardDev = 2399.7801 WeightSum = 700 Precision = 19.754347826086956
Savings_status: Discrete Estimator. Counts = 387 70 53 43 152 (Total = 705)
Employment: Discrete Estimator. Counts = 40 103 236 136 190 (Total = 705)
Installment_commitment: Normal Distribution. Mean = 2.92 StandardDev = 1.1273 WeightSum = 700 Precision = 1.0
Personal_status: Discrete Estimator. Counts = 31 202 403 68 1 (Total = 705)
Other_parties: Discrete Estimator. Counts = 636 24 43 (Total = 703)
Residence_since: Normal Distribution. Mean = 2.8429 StandardDev = 1.1076 WeightSum = 700 Precision = 1.0
Property_magnitude: Discrete Estimator. Counts = 223 162 231 88 (Total = 704)
Age: Normal Distribution. Mean = 36.1723 StandardDev = 11.4005 WeightSum = 700 Precision = 1.0769230769230769
Other_payment_plans: Discrete Estimator. Counts = 83 29 591 (Total = 703)
Housing: Discrete Estimator. Counts = 110 528 65 (Total = 703)
Existing_credits: Normal Distribution. Mean = 1.4243 StandardDev = 0.5843 WeightSum = 700 Precision = 1.0
Job: Discrete Estimator. Counts = 16 145 445 98 (Total = 704)
Num_dependents: Normal Distribution. Mean = 1.1557 StandardDev = 0.3626 WeightSum = 700 Precision = 1.0
Own_telephone: Discrete Estimator. Counts = 410 292 (Total = 702)
Foreign_worker: Discrete Estimator. Counts = 668 34 (Total = 702)

Class bad: Prior probability = 0.3

Checking_status: Discrete Estimator. Counts = 136 106 15 47 (Total = 304)
Duration: Normal Distribution. Mean = 24.8129 StandardDev = 13.3608 WeightSum = 300 Precision = 2.125

Credit_history: Discrete Estimator. Counts = 26 29 170 29 51 (Total = 305)

Purpose: Discrete Estimator. Counts = 90 18 59 63 5 9 23 1 2 35 6 (Total = 311)

Credit_amount: Normal Distribution. Mean = 3938.1609 StandardDev = 3529.4788 WeightSum = 300 Precision = 19.754347826086956

Savings_status: Discrete Estimator. Counts = 218 35 12 7 33 (Total = 305)

Employment: Discrete Estimator. Counts = 24 71 105 40 65 (Total = 305)

Installment_commitment: Normal Distribution. Mean = 3.0967 StandardDev = 1.0866 WeightSum = 300 Precision = 1.0

Personal_status: Discrete Estimator. Counts = 21 110 147 26 1 (Total = 305)

Other_parties: Discrete Estimator. Counts = 273 19 11 (Total = 303)

Residence_since: Normal Distribution. Mean = 2.85 StandardDev = 1.0928 WeightSum = 300 Precision = 1.0

Property_magnitude: Discrete Estimator. Counts = 61 72 103 68 (Total = 304)

Age: Normal Distribution. Mean = 33.9267 StandardDev = 11.259 WeightSum = 300 Precision = 1.0769230769230769

Other_payment_plans: Discrete Estimator. Counts = 58 20 225 (Total = 303)

Housing: Discrete Estimator. Counts = 71 187 45 (Total = 303)

Existing_credits: Normal Distribution. Mean = 1.3667 StandardDev = 0.5588 WeightSum = 300 Precision = 1.0

Job: Discrete Estimator. Counts = 8 57 187 52 (Total = 304)

Num_dependents: Normal Distribution. Mean = 1.1533 StandardDev = 0.3603 WeightSum = 300 Precision = 1.0

Own_telephone: Discrete Estimator. Counts = 188 114 (Total = 302)

Foreign_worker: Discrete Estimator. Counts = 297 5 (Total = 302)

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	754	75.4 %
Incorrectly Classified Instances	246	24.6 %
Kappa statistic	0.3813	
Mean absolute error	0.2936	
Root mean squared error	0.4201	
Relative absolute error	69.8801 %	
Root relative squared error	91.6718 %	
Total Number of Instances	1000	

Άρα η ακρίβεια του NB στα δεδομένα ήταν 75.40%.

7.2.2. Δέντρο απόφασης C4.5

=== Classifier model (full training set) ===

C4.5 pruned tree

Checking_status = <0
| Foreign_worker = yes
| | Credit_history = no credits/all paid: bad (11.0/3.0)
| | Credit_history = all paid: bad (9.0/1.0)
| | Credit_history = existing paid
| | | Other_parties = none
| | | | Savings_status = <100
| | | | | job = unemp/unskilled non res: bad (1.0)
| | | | | job = unskilled resident: bad (18.0/8.0)
| | | | | job = skilled

```
|  |  |  |  |  |  own_telephone = none
|  |  |  |  |  |  |  purpose = new car: bad (6.0/1.0)
|  |  |  |  |  |  |  purpose = used car: good (1.0)
|  |  |  |  |  |  |  purpose = furniture/equipment: good (16.0/8.0)
|  |  |  |  |  |  |  purpose = radio/tv: bad (11.0/5.0)
|  |  |  |  |  |  |  purpose = domestic appliance: bad (1.0)
|  |  |  |  |  |  |  purpose = repairs: bad (0.0)
|  |  |  |  |  |  |  purpose = education: bad (2.0)
|  |  |  |  |  |  |  purpose = vacation: bad (0.0)
|  |  |  |  |  |  |  purpose = retraining: bad (0.0)
|  |  |  |  |  |  |  purpose = business: bad (1.0)
|  |  |  |  |  |  |  purpose = other: bad (0.0)
|  |  |  |  |  |  own_telephone = yes: bad (9.0)
|  |  |  |  |  job = high qualif/self emp/mgmt: good (10.0/3.0)
|  |  |  |  savings_status = 100<=X<500: bad (8.0/3.0)
|  |  |  |  savings_status = 500<=X<1000: good (1.0)
|  |  |  |  savings_status = >=1000: good (2.0)
|  |  |  |  savings_status = no known savings: bad (12.0/5.0)
|  |  |  other_parties = co applicant: good (4.0/2.0)
|  |  |  other_parties = guarantor: good (8.0/1.0)
|  |  credit_history = delayed previously: bad (7.0/2.0)
|  |  credit_history = critical/other existing credit: good (38.0/10.0)
|  foreign_worker = no: good (12.0/2.0)
checking_status = 0<=X<200
|  other_parties = none
|  |  credit_history = no credits/all paid: bad (9.0/1.0)
|  |  credit_history = all paid: bad (10.0/4.0)
|  |  credit_history = existing paid
|  |  |  credit_amount <= 8858: good (70.0/21.0)
|  |  |  credit_amount > 8858: bad (8.0)
```

| | credit_history = delayed previously: good (25.0/6.0)

| | credit_history = critical/other existing credit: good (26.0/7.0)

| other_parties = co applicant: bad (7.0/1.0)

| other_parties = guarantor: good (18.0/4.0)

checking_status = >=200: good (44.0/9.0)

checking_status = no checking: good (262.0/31.0)

Number of Leaves : 36

Size of the tree : 47

Time taken to build model: 0.03 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	775	77.5 %
Incorrectly Classified Instances	225	22.5 %
Kappa statistic	0.4022	
Mean absolute error	0.3248	
Root mean squared error	0.4032	
Relative absolute error	77.3008 %	
Root relative squared error	87.976 %	
Total Number of Instances	1000	

Άρα η ακρίβεια του C4.5 στα δεδομένα ήταν 77.50%.

7.2.3. Κανόνες αποφάσεων από τον αλγόριθμο RIPPER

RIPPER rules:

==========

(checking_status = <0) and (job = skilled) => class=bad (172.0/76.0)

(checking_status = 0<=X<200) and (duration >= 24) and (savings_status = <100) => class=bad (61.0/19.0)

=> class=good (767.0/162.0)

Number of Rules : 3

Time taken to build model: 0.5 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	717	71.7 %
Incorrectly Classified Instances	283	28.3 %
Kappa statistic	0.2513	
Mean absolute error	0.3781	
Root mean squared error	0.4472	
Relative absolute error	89.9974 %	
Root relative squared error	97.5906 %	
Total Number of Instances	1000	

Άρα η ακρίβεια του RIPPER στα δεδομένα ήταν 71.70%. Πρέπει επίσης να σημειωθεί ότι αυτή η ακρίβεια επιτεύχθηκε με την χρήση μόνο 3 κανόνων.

7.2.4. Αλγόριθμος SMO

SMO

Classifier for classes: good, bad

BinarySMO

Machine linear: showing attribute weights, not support vectors.

```
      0.6805 * (normalized) checking_status=<0
+     0.3347 * (normalized) checking_status=0<=X<200
+    -0.4616 * (normalized) checking_status=>=200
+    -0.5537 * (normalized) checking_status=no checking
+     1.6987 * (normalized) duration
+     0.5398 * (normalized) credit_history=no credits/all paid
+     0.6015 * (normalized) credit_history=all paid
+    -0.109  * (normalized) credit_history=existing paid
+    -0.3182 * (normalized) credit_history=delayed previously
+    -0.7141 * (normalized) credit_history=critical/other existing credit
+     0.5673 * (normalized) purpose=new car
+    -0.5615 * (normalized) purpose=used car
+    -0.1464 * (normalized) purpose=furniture/equipment
+    -0.0798 * (normalized) purpose=radio/tv
+     0.5456 * (normalized) purpose=domestic appliance
+     0      * (normalized) purpose=repairs
+     0.4441 * (normalized) purpose=education
+    -0.3951 * (normalized) purpose=retraining
+    -0.0823 * (normalized) purpose=business
```

+ -0.2919 * (normalized) purpose=other
+ 1.1473 * (normalized) credit_amount
+ 0.4056 * (normalized) savings_status=<100
+ 0.115 * (normalized) savings_status=100<=X<500
+ 0.1378 * (normalized) savings_status=500<=X<1000
+ -0.3775 * (normalized) savings_status=>=1000
+ -0.2809 * (normalized) savings_status=no known savings
+ 0.2887 * (normalized) employment=unemployed
+ 0.1663 * (normalized) employment=<1
+ 0.0021 * (normalized) employment=1<=X<4
+ -0.3348 * (normalized) employment=4<=X<7
+ -0.1222 * (normalized) employment=>=7
+ 0.6503 * (normalized) installment_commitment
+ 0.3335 * (normalized) personal_status=male div/sep
+ 0.1177 * (normalized) personal_status=female div/dep/mar
+ -0.3697 * (normalized) personal_status=male single
+ -0.0815 * (normalized) personal_status=male mar/wid
+ 0.0514 * (normalized) other_parties=none
+ 0.5697 * (normalized) other_parties=co applicant
+ -0.6211 * (normalized) other_parties=guarantor
+ -0.0001 * (normalized) residence_since
+ -0.2247 * (normalized) property_magnitude=real estate
+ -0.0544 * (normalized) property_magnitude=life insurance
+ -0.0795 * (normalized) property_magnitude=car
+ 0.3586 * (normalized) property_magnitude=no known property
+ -0.4191 * (normalized) age
+ 0.0697 * (normalized) other_payment_plans=bank
+ 0.159 * (normalized) other_payment_plans=stores
+ -0.2287 * (normalized) other_payment_plans=none
+ 0.3271 * (normalized) housing=rent

```
+     -0.0702 * (normalized) housing=own
+     -0.257  * (normalized) housing=for free
+      0.4503 * (normalized) existing_credits
+     -0.2026 * (normalized) job=unemp/unskilled non res
+      0.1501 * (normalized) job=unskilled resident
+      0.1027 * (normalized) job=skilled
+     -0.0502 * (normalized) job=high qualif/self emp/mgmt
+      0.0198 * (normalized) num_dependents
+     -0.1394 * (normalized) own_telephone
+     -0.9888 * (normalized) foreign_worker
-     1.5398

Number of kernel evaluations: 436644
```

```
Time taken to build model: 5.67 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         751               75.1    %
Incorrectly Classified Instances       249               24.9    %
Kappa statistic                          0.3654
Mean absolute error                      0.249
Root mean squared error                  0.499
Relative absolute error                 59.2607 %
Root relative squared error            108.8905 %
Total Number of Instances              1000
```

Άρα η ακρίβεια του SMO στα δεδομένα ήταν 75.10%.

7.2.5. Αλγόριθμος BP για νευρωνικά δίκτυα

Sigmoid Node 0

Inputs Weights

Threshold -0.07324978292862763

Node 2 3.0044957011533477

Sigmoid Node 1

Inputs Weights

Threshold 0.07324978292862741

Node 2 -3.0044957011533455

Sigmoid Node 2

Inputs Weights

Threshold -0.1524383241253404

Attrib checking_status=<0 -7.127308909968609

Attrib checking_status=0<=X<200 -5.568840800067419

Attrib checking_status=>=200 5.536442678431966

Attrib checking_status=no checking 7.448517624737233

Attrib duration -15.869684583430612

Attrib credit_history=no credits/all paid -6.817339661283083

Attrib credit_history=all paid -3.6648681519473008

Attrib credit_history=existing paid -0.42855909699610395

Attrib credit_history=delayed previously 6.067056265086362

Attrib credit_history=critical/other existing credit 5.202479053504914

Attrib purpose=new car -4.358809242887763

Attrib purpose=used car -0.16391374319658863

Attrib purpose=furniture/equipment 0.7622694371506793

Attrib purpose=radio/tv 1.7737583840894737

Attrib purpose=domestic appliance 1.8504336534889916

Attrib purpose=repairs 2.1256495964623667

Attrib purpose=education 0.8104421363684363

Attrib purpose=vacation 0.01806673914461429

Attrib purpose=retraining 0.6635877207353198

Attrib purpose=business -2.9018486603268867

Attrib purpose=other 0.5412025537531738

Attrib credit_amount -14.471604796632404

Attrib savings_status=<100 -8.491660261965142

Attrib savings_status=100<=X<500 -3.0368882447270185

Attrib savings_status=500<=X<1000 2.1008069748721

Attrib savings_status=>=1000 6.738265810863285

Attrib savings_status=no known savings 3.058206882235268

Attrib employment=unemployed -6.519641320196984

Attrib employment=<1 -1.799287155988702

Attrib employment=1<=X<4 0.6247469663446439

Attrib employment=4<=X<7 4.375385310436861

Attrib employment=>=7 3.809204026168733

Attrib installment_commitment -5.807958324468605

Attrib personal_status=male div/sep -2.713296402011672

Attrib personal_status=female div/dep/mar 0.6876305771791159

Attrib personal_status=male single 0.6271535791111095

Attrib personal_status=male mar/wid 1.6984781745224942

Attrib personal_status=female single 0.04660287398503703

Attrib other_parties=none 1.2286702730465122

Attrib other_parties=co applicant -7.090495452923499

Attrib other_parties=guarantor 5.9829579169630716

Attrib residence_since 3.3348346927199923

Attrib property_magnitude=real estate 3.0872430162523985

Attrib property_magnitude=life insurance 1.7090352341067732

Attrib property_magnitude=car 0.584304141243327

Attrib property_magnitude=no known property -5.05061522200528

Attrib age 0.8517640853787177

Attrib other_payment_plans=bank -2.5652543834404966

Attrib other_payment_plans=stores -1.0643671748128702

Attrib other_payment_plans=none 3.7371110599404647

Attrib housing=rent -3.292778487618977

Attrib housing=own -0.5720561151656457

Attrib housing=for free 4.048407500585945

Attrib existing_credits -0.301979734122562

Attrib job=unemp/unskilled non res 2.0646391791137657

Attrib job=unskilled resident 2.424464680108377

Attrib job=skilled -0.9748219219521777

Attrib job=high qualif/self emp/mgmt -3.2410608868256805

Attrib num_dependents 2.8533458905397975

Attrib own_telephone 4.980271719159508

Attrib foreign_worker 9.605799763342299

Class good

Input

Node 0

Class bad

Input

Node 1

Time taken to build model: 9.72 seconds			
=== Stratified cross-validation ===			
=== Summary ===			
Correctly Classified Instances	729	72.9	%
Incorrectly Classified Instances	271	27.1	%
Kappa statistic	0.3767		
Mean absolute error	0.3226		
Root mean squared error	0.4313		

Relative absolute error	76.7887 %
Root relative squared error	94.1153 %
Total Number of Instances	1000

Άρα η ακρίβεια του BP στα δεδομένα ήταν 72. 90%.

Όπως έχει ήδη αναφερθεί ο σκοπός της εργασίας μας ήταν να χρησιμοποιήσει αλγόριθμους εξόρυξης γνώσης για την έγκριση πιστωτικών καρτών βασιζόμενη σε στοιχεία όπως ηλικία, εισόδημα, πιστωτική ιστορία και ιδιοκτησία κατοικίας κλπ. Τα δέντρα αποφάσεων αποδείχθηκε ότι έχουν την καλύτερη ακρίβεια στο πρόβλημα αυτό. Εν συνεχεία, παρουσιάζουμε το εμπορικό πακέτο εξόρυξης γνώσης Knowledge Seeker που επιτρέπει μεγάλη παραμετροποίηση στα δέντρα αποφάσεων.

8.Ανάλυση δεδομένων με το Knowledge SEEKER

8.1. Εισαγωγή

Το Knowledge Seeker είναι ένα υπολογιστικό πακέτο που επεξεργάζεται και αναλύει δεδομένα (data mining tool) και επιτρέπει στους χρήστες του να επεξεργαστούν και να καταλάβουν τις σχέσεις που υπάρχουν μεταξύ των μεταβλητών σε ένα σύνολο δεδομένων. Παρουσιάζει μία πλήρη ανάλυση γραμμικών και μη γραμμικών σχέσεων και εκθέτει τα αποτελέσματα γραφικά με στατιστικά δέντρα αποφάσεων, χρησιμοποιώντας ποσοστά ή μέσους όρους δίνοντας σημαντικές πληροφορίες. Υποστηρίζεται ότι η ανάλυση που παρουσιάζει θα διαρκούσε πολλές μέρες από ένα στατιστικό για να ολοκληρωθεί. Είναι φιλικό με το χρήστη και εύχρηστο ακόμη και σε μη στατιστικούς. Ψάχνει για τις σχέσεις που θα καθορίσει ο χρήστης, έχει τη δυνατότητα γρήγορων ελέγχων υποθέσεων και μπορεί επίσης να εκφράσει τις σχέσεις που έχει βρει με τη μορφή κανόνων. Αυτοί οι κανόνες βοηθούν για τις προβλέψεις, τον προγραμματισμό ή τις διαγνώσεις που μπορεί να γίνουν. Παρέχει τέλος, μία μέθοδο για τον προκαθορισμό της διάταξης και της τυποποίησης των δεδομένων.

Το υπολογιστικό αυτό πακέτο εξόρυξης δεδομένων μπορεί να δώσει απαντήσεις σε σημαντικές ερωτήσεις έχοντας στη διάθεσή του μεγάλο πλήθος από δεδομένα. Αρχικά επεξεργάζεται αυτόματα όλα τα δεδομένα, συνοψίζει τα στατιστικά σημαντικά πεδία και τις σχέσεις μεταξύ τους και παρουσιάζει τα αποτελέσματα με δέντρα αποφάσεων παρέχοντας ταυτόχρονα έλεγχο αξιοπιστίας και εγκυρότητας αυτών. Έτσι όχι μόνο ανακαλύπτει και υπογραμμίζει τις σχέσεις των δεδομένων, αλλά εξασφαλίζει και επιβεβαιώνει την εγκυρότητά τους.

Μπορεί να χρησιμοποιηθεί για την ανάλυση δεδομένων από βάσεις δεδομένων, λογιστικά ή στατιστικά φύλλα ή από κειμενογράφους, σε υπολογιστικά συστήματα μεγάλης ισχύος ή μικροϋπολογιστές. Χρησιμοποιείται λοιπόν σε ανάλυση αγοράς, όπου

καθορίζει τους παράγοντες που επηρεάζουν τις πωλήσεις των προϊόντων (γεωγραφικούς, τιμές, χαρακτηριστικά πελάτη), σε έλεγχο ποιότητας, όπου αναγνωρίζει τους σημαντικούς παράγοντες για ελαττωματικά προϊόντα, σε θέματα υγείας, όπου εξετάζει τα δεδομένα για να ανακαλύψει συνδυασμένα αποτελέσματα που συμβάλλουν στην υγεία και την αρρώστια, σε θέματα διοικητικής ανάλυσης, όπου καθορίζει τους παράγοντες που επηρεάζουν το μισθό σε μεγάλο δείγμα εργαζομένων και καθορίζει πως σχετίζονται, σε θέματα επιστημονικής έρευνας, όπου αναλύει αποτελέσματα πειραμάτων και καθορίζει τους παράγοντες που επηρεάζουν την έρευνα και σε θέματα εξυπηρέτησης πελατών, όπου ανακαλύπτει προβλήματα στο χώρο παραγωγής πριν γίνουν επιδημικά.

8.2. Εισαγωγή δεδομένων (import)

Ο χρήστης έχει δυνατότητα εισαγωγής δεδομένων στο Knowledge Seeker IV από βάσεις δεδομένων, λογιστικά φύλλα και άλλα στατιστικά πακέτα, όπως από τα: dBase II (*.dbf), Paradox (*.db), Sawtooth (*.alp), SmartWare (*.db), SAS (*.tpt, *.ssd, *.sd2), SPSS (*.sav, *.por), Gauss (*.dat), Excel (*.xls), Lotus (*.w??), QuattroPro (*.wql), Splus (*.*), Stata (*.dta), Systat (*.sys) και παλαιότερες εκδόσεις του Knowledge Seeker (*.fmt). Όμως, και στην περίπτωση που το Knowledge Seeker δεν υποστηρίζει απευθείας κάποιο άλλο πρόγραμμα διαχείρισης δεδομένων, μπορούμε να εισάγουμε δεδομένα στο πρώτο υπό τη μορφή αρχείου κειμένου χαρακτήρων ASCII, αρκεί το δεύτερο να μπορεί να εξάγει δεδομένα σε αυτή τη μορφή. Ακόμα, παρέχεται η δυνατότητα επεξεργασίας ODBC και SQL βάσεων δεδομένων, καθώς και επεξεργασία δεδομένων από το clipboard. Όλη η διαδικασία εισαγωγής δεδομένων γίνεται μέσω ενός αρκετά φιλικού οδηγού που ζητά από το χρήστη τον καθορισμό της πηγής δεδομένων, των πεδίων, το μέγεθος, τον τύπο και το χαρακτήρα διαχωρισμού τους, της εξαρτημένης μεταβλητής (Dependent Variable, DV), κ.α. Έχουμε επίσης τη δυνατότητα, από τον οδηγό εισαγωγής δεδομένων, να επιλέξουμε τόσο δείγμα δεδομένων από το συνολικό πληθυσμό, όσο και τον τύπο κωδικοποίησης του προς εισαγωγή αρχείου. Εάν το μέγεθος της βάσης δεδομένων είναι πολύ μεγάλο μπορούμε να μειώσουμε τον αριθμό

των παρατηρήσεων επιλέγοντας δείγμα. Σε αυτή την περίπτωση, η δειγματοληψία δεν αλλοιώνει το αποτέλεσμα της ανάλυσης κυρίως στα "πάνω" επίπεδα του δέντρου αποφάσεων. Παίρνοντας λοιπόν ως δείγμα το 25% των παρατηρήσεων της βάσεως, το πρόγραμμα εκτιμά αυτό το ποσοστό επιλογής με τέτοιο τρόπο ώστε ο πραγματικός αριθμός των εγγραφών να ποικίλει.

Σε περίπτωση που θέλουμε να εισάγουμε δεδομένα από λογιστικά φύλλα, αυτά πρέπει να είναι στοιχισμένα πάνω αριστερά σε κάθε κελί του φύλλου. Οι διαφορετικές εγγραφές (παρατηρήσεις) πρέπει να είναι διατεταγμένες σε γραμμές και τα πεδία (μεταβλητές) σε στήλες. Η πρώτη εγγραφή (γραμμή) θα πρέπει να περιέχει τα ονόματα των πεδίων. Εάν δεν τα περιέχει το Knowledge Seeker δίνει αριθμητικό όνομα στα πεδία. Επίσης, μέσα στη βάση δεδομένων δεν πρέπει να υπάρχουν κενές γραμμές ή στήλες. Σε περίπτωση τέτοιων κενών το Knowledge Seeker δεν μπορεί να αναγνωρίσει τα όρια (μέγεθος) των δεδομένων. Κατά την εισαγωγή δεδομένων από εφαρμογές που υποστηρίζουν πολλαπλή επεξεργασία λογιστικών φύλλων εισάγεται μόνο το πρώτο φύλλο.

Στο Knowledge Seeker μπορούμε να πραγματοποιήσουμε εισαγωγή δεδομένων από το στατιστικό πακέτο SAS. Το πρόγραμμα μπορεί να διαβάσει αρχεία του SAS με επέκταση (*.tpt), που είναι συμβατά σχεδόν με όλα τα λειτουργικά συστήματα (portable file). Επίσης διαβάζει αρχεία του SAS έκδοση 6.0x με επέκταση (*.ssd) που είναι συμβατά με πλατφόρμες Windows και Sun, καθώς και αρχεία SAS έκδοση 6.11 που είναι συμβατά μόνο με πλατφόρμες Windows.

Στο Knowledge Seeker μπορούμε να πραγματοποιήσουμε εισαγωγή δεδομένων από το στατιστικό πακέτο SPSS. Το SPSS αποθηκεύει αρχεία σε δυαδική μορφή (binary format). Για να εισαχθεί αρχείο δεδομένων του SPSS στο Knowledge Seeker πρέπει πρώτα αυτό να έχει εξαχθεί από το SPSS σε μορφή αρχείου με επέκταση (*.por, portable file). Τέτοιου είδους αρχεία είναι γραμμένα υπό τη μορφή κειμένου χαρακτήρων ASCII.

8.3. Εξαγωγή δεδομένων (Export)

Το Knowledge Seeker IV μπορεί να εξάγει δεδομένα σε μορφή αρχείου κειμένου (titled – no titled, delimited), καθώς και σε μορφή αρχείου έτοιμο προς χρήση από τις εφαρμογές dBase III, SAS (ver. 6.12) και Excel. Μπορούν να εξαχθούν όλα τα δεδομένα ή μέρος αυτών (δεδομένα που αντιστοιχούν σε συγκεκριμένο κόμβο).

8.4. Δέντρα αποφάσεων

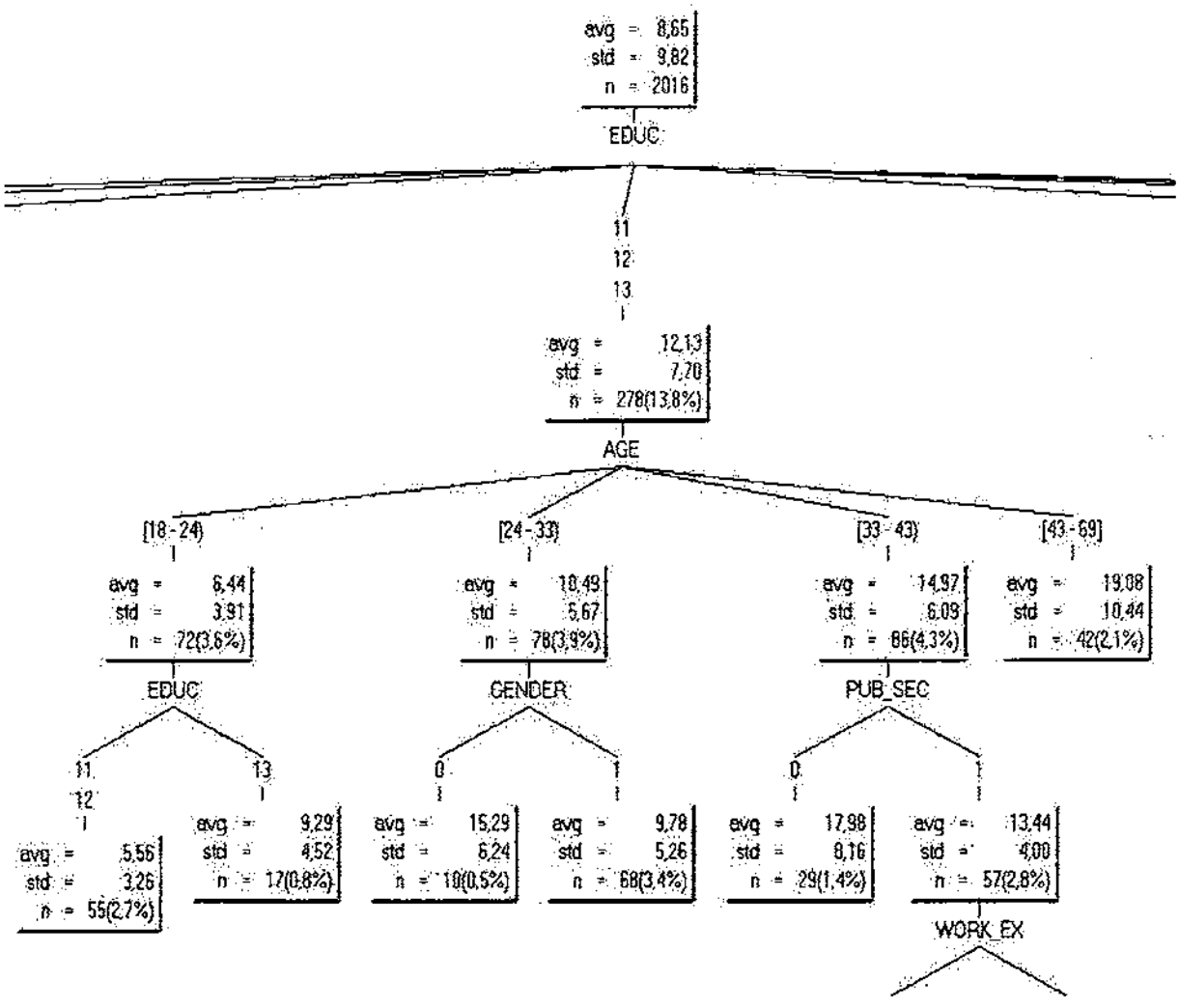
Όπως ήδη αναφέραμε το Knowledge Seeker χρησιμοποιεί τη μέθοδο της ανάλυσης με δέντρα αποφάσεων. Τα δέντρα αποφάσεων αναπτύχθηκαν κυρίως για την εξουδετέρωση κάποιων προβλημάτων που είχαν προκύψει με τη χρήση της πολυμεταβλητής παλινδρόμησης στην Εξόρυξη Δεδομένων. Με αφορμή τέτοια προβλήματα οι Morgan και Sonquist ανέπτυξαν μία στατιστική τεχνική ανάλυσης δεδομένων, τον «αυτόματο αλληλεπιδρών ανιχνευτή», την AID (Automatic Interactive Detector) μέθοδο, με την οποία ανακαλύπτονται ακόμη και «κρυμμένες» σχέσεις των μεταβλητών. Τα δέντρα, έτσι των αποφάσεων, παρέχουν αυτές τις σχέσεις των μεταβλητών και δημιουργούν ένα μοντέλο αυτών που είναι έγκυρο αλλά και εύκολο να το ερμηνεύσει κανείς.

Μετά την εισαγωγή των δεδομένων μπορούμε να δημιουργήσουμε το δέντρο, το οποίο θα δώσει πληροφορίες για την ανάλυσή μας. Ο πρώτος κόμβος ή ρίζα του δέντρου, που εμφανίζεται αρχικά περιέχει τις τιμές της εξαρτημένης μεταβλητής που έχει καθοριστεί από το χρήστη. Προχωρώντας το Knowledge Seeker ψάχνει για όλες τις πιθανές σχέσεις των άλλων μεταβλητών με τη ρίζα και τις παρουσιάζει, αρχίζοντας από τις πιο σημαντικές. Εξετάζει όλα τα πεδία που μπορούν να χρησιμοποιηθούν για να περιγράψει την εξαρτημένη μεταβλητή και επιλέγει αυτά που μπορούν καλύτερα να εξηγήσουν ή να προβλέψουν τις μεταβλητότητες στη μεταβλητή. Κάθε μία από αυτές τις μεταβλητές είναι ανεξάρτητη με τις υπόλοιπες και με τον ίδιο τρόπο μπορεί να χωριστεί

σε κατηγορίες, στις οποίες συνεχίζεται η ίδια διαδικασία ώστε να έχουμε περισσότερες πληροφορίες.

Στο παράδειγμα που θα χρησιμοποιήσουμε για την παρακάτω ανάλυση, δουλεύουμε με δεδομένα από τον χώρο της εργασίας, που περιέχει πληροφορίες για το εισόδημα των εργαζομένων, το παρελθόν τους όσον αφορά την εργασία τους καθώς και πληροφορίες για την εκπαιδευτική και οικογενειακή τους κατάσταση. Θα προσδιορίσουμε ένα προφίλ χαμηλόμισθων και υψηλόμισθων ανθρώπων όσο το δυνατό καθαρό και κατανοητό. Τα συμπεράσματα θα προκύψουν από τις τιμές των μεταβλητών που θα καθορίσουμε και τις σχέσεις μεταξύ τους και θα πρέπει να είναι ουσιώδη και έγκυρα.

Έχουμε ορίσει δηλαδή, ως *Wage* το ωριαίο εισόδημα του ερωτηθέντα, *Age* την ηλικία (σε χρόνια) κάθε ερωτηθέντα, *Work_ex* την προϋπηρεσία (σε χρόνια), *Gender* την εικονική (dummy) μεταβλητή που παίρνει τιμές 0 και 1, αν ο ερωτούμενος είναι θηλυκού ή αρσενικού γένους αντίστοιχα, *Pub_sec* την εικονική μεταβλητή με τιμές 0 και 1, αν ο ερωτούμενος εργάζεται στον ιδιωτικό ή δημόσιο τομέα αντίστοιχα, *Educ* το επίπεδο εκπαίδευσης (σε χρόνια), *Fath_ed* το επίπεδο εκπαίδευσης (σε χρόνια) του πατέρα και *Moth_ed* το επίπεδο εκπαίδευσης (σε χρόνια) της μητέρας του ερωτηθέντα.



Σχήμα 8.1

Η βασική-εξαρτημένη μεταβλητή που χρησιμοποιείται γι'αυτή την ανάλυση είναι το ωριαίο εισόδημα των εργαζομένων, *Wage*. Σκοπός μας είναι να προσδιορίσουμε τους παράγοντες που προβλέπουν υψηλό ή χαμηλό εισόδημα αντίστοιχα. Στην αρχή, στη ρίζα του δέντρου εμφανίζονται βασικές πληροφορίες, από τις οποίες βρίσκουμε τις σχέσεις που μπορούν να εξηγήσουν υψηλό ή χαμηλό εισόδημα στη βάση δεδομένων. Όπως βλέπουμε από το παραπάνω σχήμα από τις 2016 εγγραφές, το μέσο εισόδημα ενός εργαζομένου είναι $8,65 ανά ώρα, με τυπική απόκλιση $9,82. Για να μελετήσουμε πως σχετίζεται το εισόδημα με τα άλλα πεδία προχωράμε στον επόμενο κόμβο. Στον επόμενο κόμβο εμφανίζεται η ανεξάρτητη μεταβλητή, το επίπεδο εκπαίδευσης.

ότι απευθύνονται και σε ρυθμίσεις που εξισορροπούν τον αριθμό των περιττών πεδίων των δεδομένων. Δίνεται έτσι η δυνατότητα, η ανάλυση που βασίζεται σε κατηγοριοποιημένες ομάδες πεδίων να γίνεται με υψηλό επίπεδο εμπιστοσύνης. Με αυτές τις ρυθμίσεις μπορεί να διατηρηθεί η ποσοστιαία αναλογία λάθους (α) πρώτου βαθμού, δηλαδή η πιθανότητα να βρεθεί αξιόπιστη διάσπαση σε κάποιο κόμβο του δέντρου, ενώ στην πραγματικότητα δεν υπάρχει τέτοια σχέση μεταξύ των μεταβλητών, σε επιτρεπτά όρια, όπως 0.05%.

8.7. Δημιουργία ταμπλό (generate crosstable)

Με αυτή την εντολή ο χρήστης μπορεί να απεικονίσει τις πληροφορίες όλου ή μέρους του στατιστικού δέντρου αποφάσεων που έχει παράγει υπό τη μορφή πίνακα κειμένου. Στον πίνακα φαίνονται οι ιδιότητες των πεδίων (μεταβλητών) που διασπούν (split) τον αμέσως προηγούμενο κόμβο (για παράδειγμα φαίνεται εάν η μεταβλητή είναι συνεχής ή κατηγορική, εάν η ομαδοποίηση έχει γίνει με βάση τη διάταξη ή όχι, εάν η μεταβλητή έχει παραλειπόμενες τιμές, κ.α.). Επίσης παρουσιάζονται οι συχνότητες του αριθμού των παρατηρήσεων που εμφανίζονται σε κάθε κόμβο του δέντρου, παρουσιάζονται οι συστάδες , στις οποίες χωρίζεται η μεταβλητή, που στη συνέχεια θα παράγουν την επόμενη διάσπαση και τα επί τις εκατό ποσοστά των κόμβων που επιλέχτηκαν για τη δημιουργία του ταμπλό, τοποθετημένα σε γραμμές και στήλες. Ανά γραμμές παρουσιάζονται ποσοστά που αφορούν τον τρέχον κόμβο, ενώ ανά στήλες ποσοστά που αφορούν απογόνους του. Τέλος, καταγράφονται στατιστικά μεγέθη που υπολογίστηκαν για να προσδιορισθεί η διάσπαση, όπως το επίπεδο εμπιστοσύνης, η τιμή της στατιστικής συνάρτησης ελέγχου, οι βαθμοί ελευθερίας, κ.α.

8.8. Δημιουργία κανόνων (generate rules)

Το Knowledge Seeker μπορεί να μετατρέψει το παραγόμενο στατιστικό δέντρο αποφάσεων, ή μέρος αυτού, σε μια ακολουθία κανόνων (rules) ή δηλώσεων της SQL. Ένα τέτοιο αποτέλεσμα – έξοδος (output) μπορεί να χρησιμοποιηθεί στη δημιουργία

συστημάτων διεξαγωγής πληροφοριών από ένα σύνολο δεδομένων με βάση μια σειρά κανόνων.

Το Knowledge Seeker μπορεί να δημιουργήσει κανόνες γενικής ή αναλυτικής μορφής, για το SPSS, για το SAS, τη Visual Basic και την Prolog. Σε αυτή την περίπτωση το δέντρο μετατρέπεται σε μια σειρά IF...THEN δηλώσεων ή άλλων εντολών γλώσσας προγραμματισμού, που εκφράζουν τις διασπάσεις και γενικότερα τις πληροφορίες του δέντρου. Η έξοδος στέλνεται σε ένα κειμενογράφο, όπου εκεί ο χρήστης έχει τη δυνατότητα να τροποποιήσει και να αποθηκεύσει το αρχείο.

Το Knowledge Seeker μπορεί να δημιουργήσει κανόνες υπό τη μορφή SQL δηλώσεων. Οι κανόνες αυτοί επιλέγουν εγγραφές (παρατηρήσεις) που αφορούν ένα συγκεκριμένο – ιδιαίτερο γκρουπ (κόμβο) από ένα μεγαλύτερο σύνολο δεδομένων. Για παράδειγμα, φανταστείτε ότι αναλύεται η επιτυχημένη αποστολή συγκεκριμένου μηνύματος μέσο ηλεκτρονικού ταχυδρομείου σε 1000 ανθρώπους. Τότε ανακαλύπτεται ένα γκρουπ ανθρώπων που πιθανώς το έλαβαν. Οι SQL κανόνες μπορούν να δημιουργήσουν μια ερώτηση (query) που θα επιλέξει ανθρώπους από μια διαφορετική βάση δεδομένων που θα είχαν την ίδια πιθανότητα λήψης αυτού του μηνύματος.

8.9. Εύρεση συνεισφοράς (leverage) και πίνακας αποτελεσμάτων (gains chart)

Η εντολή "leverage" υπολογίζει τη σχετική συνεισφορά συγκεκριμένων τιμών σε μια κατηγορική εξαρτημένη μεταβλητή (η διεργασία αυτή δεν εφαρμόζεται σε συνεχή εξαρτημένη μεταβλητή) σε σχέση με το συνολικό αριθμό εμφάνισης αυτών των τιμών στο σύνολο δεδομένων μας. Η λειτουργία αυτή βασίζεται στην Αρχή Βελτιστοποίησης του Pareto. Με βάση αυτή την αρχή, το 80% της επίδρασης μιας συγκεκριμένης τιμής παράγεται από το 20% της εμφάνισής της στο σύνολο των δεδομένων. Αναλογικά, μπορούμε να βρούμε έναν ή περισσότερους κόμβους (κατά τη διάσπαση μίας μεταβλητής σε κατηγορίες δημιουργούνται περισσότεροι από ένα κόμβο) στο δέντρο

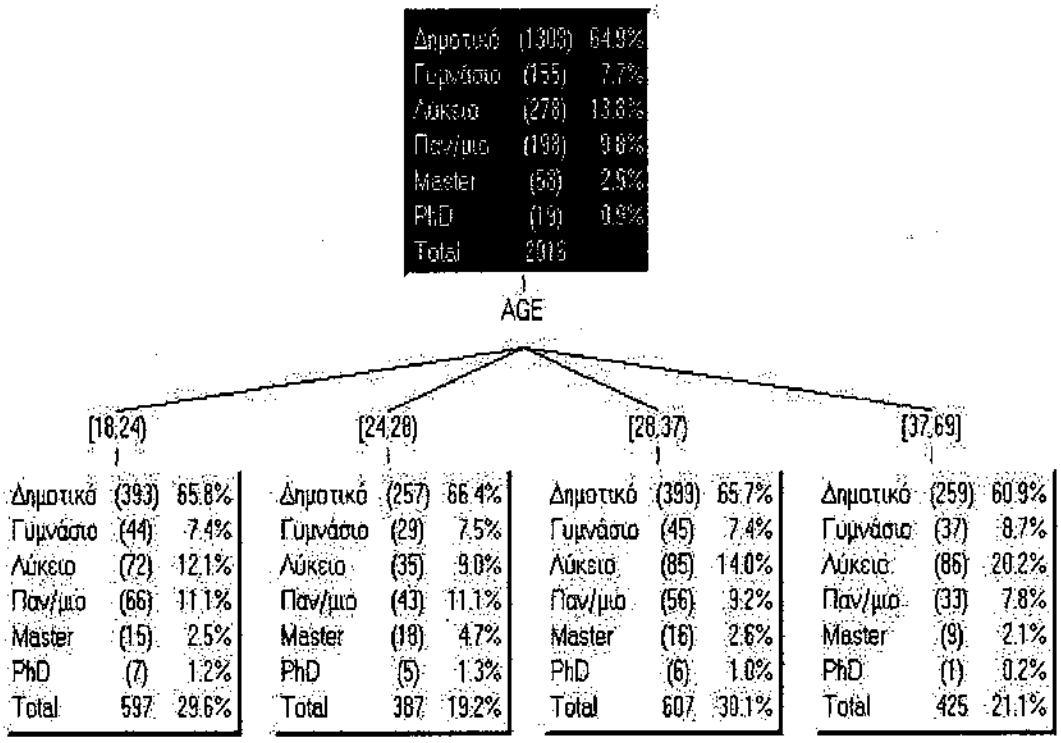
αποφάσεων που να περιέχουν το 20% των συνολικών παρατηρήσεων των δεδομένων, το οποίο συνεισφέρει κατά 80% στην επίδραση των τιμών αυτών. Τέτοιου είδους διαδικασίες είναι πολύ χρήσιμες σε μια μεγάλη γκάμα εφαρμογών, όπως από την αναγνώριση των χαρακτηριστικών του καλύτερου πελάτη μιας επιχείρησης έως την αναγνώριση των χαρακτηριστικών ανθρώπων που ανήκουν σε ομάδα ύψιστου κινδύνου υγείας.

Για την καλύτερη κατανόηση, θα χρησιμοποιήσουμε το σύνολο δεδομένων μας, με τρόπο τέτοιο που θα μας επιτρέψει να παρουσιάσουμε τη χρησιμότητα της διεργασίας αυτής. Ας υποθέσουμε λοιπόν ότι έχουμε ορίσει ως εξαρτημένη μεταβλητή την κατηγορική μεταβλητή *Educ*. Αυτή έχει κατηγοριοποιηθεί, για τις ανάγκες του παραδείγματος, σε έξι κατηγορίες με τους εξής χαρακτηρισμούς:

Κατηγορία	*Χρόνια Εκπαίδευσης*	*Επίπεδο Εκπαίδευσης (Χαρακτηρισμός)*
1	0 – 6	Δημοτικό
2	8 – 10	Γυμνάσιο
3	11 – 13	Λύκειο
4	14 – 17	Πανεπιστήμιο
5	18 – 19	Master
6	20 – 23	PhD

Πίνακας 8.1

Στη συνέχεια, καθοδηγούμε το πρόγραμμα να βρει σχέσεις μεταξύ της *Educ* και της συνεχούς μεταβλητής *Age*, διασπώντας τη δεύτερη σε τέσσερις κατηγορίες. Έτσι έχουμε το ακόλουθο δέντρο αποφάσεων:



Σχήμα 8.5. Δέντρο αποφάσεων

Ας επιλέξουμε την κατηγορία "Master" της εξαρτημένης μεταβλητής *Educ* για την περαιτέρω ανάλυσή μας. Χρησιμοποιώντας την εντολή εύρεσης συνεισφοράς "leverage" θα προσπαθήσουμε να αναγνωρίσουμε το 20% των παρατηρήσεων που συνεισφέρουν 80% στην απόκτηση μεταπτυχιακού τίτλου σπουδών Master. Με αυτό τον τρόπο επιδιώκουμε την εύρεση των χαρακτηριστικών των ανθρώπων που βρίσκονται σε επίπεδο επιστημονικής κατάρτισης Master.

Εφαρμόζοντας λοιπόν την εντολή "leverage" για την κατηγορία "Master" στο πιο πάνω δέντρο παίρνουμε τον ακόλουθο πίνακα αποτελεσμάτων:

Group	*Size of Group*	*Number of Responses*	*Cumulative Responses*	*Response Rate*	*Cumulative Rate*	*Cumulative Lift*
AGE [24,28)	387	18	18	4.7%	4.7%	162
AGE [28,37)	607	16	34	2.6%	3.4%	119
AGE [18,24)	597	15	49	2.5%	3.1%	107
AGE [37,69]	425	9	58	2.1%	2.9%	100
TOTAL	2016	58		2.9%		

Πίνακας 8.2

Η στήλη "size of group" δείχνει τον αριθμό των συνολικών παρατηρήσεων κάθε τελικού κόμβου. Η στήλη "number of responses" δείχνει τον αριθμό των παρατηρήσεων, που αντιστοιχούν στην επιλεγείσα κατηγορία κατά την ανάλυση, κάθε κόμβου. Η στήλη "cumulative responses" δείχνει τον αθροιστικό αριθμό των παρατηρήσεων της επιλεγείσας κατηγορίας από κόμβο σε κόμβο. Η στήλη "response rate" δείχνει το επί τοις εκατό ποσοστό των παρατηρήσεων της επιλεγείσας κατηγορίας σε σχέση με το συνολικό αριθμό των παρατηρήσεων του κόμβου. Αυτά τα ποσοστά επί του συνόλου θα ισούνται με το ποσοστό της επιλεγείσας κατηγορίας σε όλες τις παρατηρήσεις της ανάλυσης. Η στήλη "cumulative rate" δείχνει το ποσοστό κατά μέσο όρο της επιλεγείσας κατηγορίας διαμέσου όλων των κόμβων. Ανταποκρίνεται στο ποσοστό του αθροιστικού αριθμού των παρατηρήσεων της επιλεγείσας κατηγορίας στο συνολικό αριθμό παρατηρήσεων της ανάλυσης. Η στήλη "cumulative lift" δείχνει την αναλογία των κατηγοριών καθώς κινούμαστε στο δέντρο από πάνω προς τα κάτω, από κόμβο σε κόμβο. Υπολογίζεται σε σχέση με το συνολικό επί τοις εκατό ποσοστό που αντιπροσωπεύει η επιλεγείσα κατηγορία σε ολόκληρο το σύνολο δεδομένων.

Αναλύοντας τον πίνακα αποτελεσμάτων που προέκυψε, παρατηρούμε ότι η ομάδα δεδομένων που σχετίζεται περισσότερο με την κατηγορία "Master" της εξαρτημένης μεταβλητής *Educ* είναι άτομα ηλικίας 24-28 ετών (πληροφορίες για την οποία έχουμε στην πρώτη γραμμή του πίνακα). Στην ομάδα αυτή ανήκουν 18 παρατηρήσεις εκ των 58 συνολικών παρατηρήσεων που ανήκουν στην κατηγορία (Master) και στις τέσσερις ομάδες ηλικίας. Από τις 387 συνολικές παρατηρήσεις της ομάδας ανθρώπων ηλικίας 24-28, 18 (ή το 4.7%) ανήκουν στην κατηγορία "Master". Η σχετική συνεισφορά αυτού του κόμβου στην κατηγορία "Master" σε σύγκριση με τον συνολικό αριθμό υποθέσεων που λαμβάνονται υπόψη σε όλες τις παρατηρήσεις του κόμβου, παρουσιάζεται στην τελευταία στήλη του πίνακα, που εδώ έχει τιμή 162%. Αυτή είναι η αναλογία του επί τοις εκατό ποσοστού της κατηγορίας "Master" (18 από τις 58 παρατηρήσεις ή περίπου το 4.7%) προς το εκατοστιαίο ποσοστό του αριθμού των παρατηρήσεων της ομάδας ανθρώπων ηλικίας 24-28 που αντιπροσωπεύει το συνολικό αριθμό παρατηρήσεων σε όλα τα δεδομένα (387 από τις 2016 παρατηρήσεις ή περίπου το 2.9%). Αυτό γεννά την αναλογία $\frac{18/58}{387/2016} \cong 1.62$ ή ένα "lift" 162%.

Εύκολα διαπιστώνουμε ότι το συγκεκριμένο παράδειγμα ξεφεύγει του 80/20 κανόνα του Pareto. Και αυτό γιατί το 4.7% (ποσοστό ανθρώπων ηλικίας 24-28 που κατέχουν μεταπτυχιακό τίτλο σπουδών Master) προέρχεται από το 2.9% των παρατηρήσεων. Μοιάζει λοιπόν περισσότερο με κανόνα 50/25. Η αναφορά του όμως σε αυτό το σημείο της παρούσας εργασίας είναι σκόπιμη, μια και βοηθάει στην παρουσίαση της διεργασίας εύρεσης συνεισφοράς.

8.10. Συνθήκες βάρους

Το Knowledge Seeker χρησιμοποιεί συνθήκες βάρους*, ως ένα βήμα στην προεργασία των δεδομένων για να «ζυγίσει» υποθέσεις ή στατιστικά αποτελέσματα με δύο τρόπους:

Με δειγματοληψία βάρους (sampling weights). Η δειγματοληψία αυτή χρησιμοποιείται για να εξασφαλίσει ότι οι παρατηρήσεις του δείγματος συμπεριλαμβάνονται στην ανάλυση σε αναλογία τέτοια, που είναι σχετική με την κατανομή των παρατηρήσεων στον πληθυσμό από τον οποίο έχει επιλεγεί το δείγμα. Για παράδειγμα αν ο αριθμός των αντρών υπερκαλύπτει το δείγμα σε σχέση με την πραγματική κατανομή αντρών και γυναικών στον πληθυσμό, οι συνθήκες βάρους μπορούν να χρησιμοποιηθούν για να προσαρμόσουν τη συμβολή κάθε γένους στην ανάλυση, ώστε τα αποτελέσματα να αντιπροσωπεύουν την αληθινή αναλογία. Τέτοια δείγματα είναι τα στρωματοποιημένα.

Με συχνότητες βάρους (frequency weights). Αυτές χρησιμοποιούνται ως μία συνθήκη βάρους για να ελεγχθεί πόσες φορές εμφανίζεται η δοσμένη υπόθεση – παρατήρηση στην ανάλυση. Μπορεί να χρησιμοποιηθεί όταν εμφανίζονται πολλές ξεχωριστές παρατηρήσεις, οι οποίες βρίσκονται στη βάση δεδομένων ως μία παρατήρηση. Τότε οι παρατηρήσεις που έχουν τις ίδιες τιμές, αντικαθίστανται από μία σταθμική παρατήρηση, η οποία αντιπροσωπεύει τις συχνότητες των παρατηρήσεων στα δεδομένα. Αυτές οι συχνότητες βάρους εμφανίζονται για παράδειγμα σε πειραματικές έρευνες, όπου ο αριθμός ίδιων συνθηκών έχουν συλλεχθεί ως επαναλαμβανόμενες παρατηρήσεις.

Οι συνθήκες, λοιπόν, βάρους που χρησιμοποιεί το Knowledge Seeker έχουν σαν αποτέλεσμα την αλλαγή των συχνοτήτων κάποιων παρατηρήσεων. Αυτό επηρεάζει και τον υπολογισμό των X^2 και F στατιστικών μεγεθών. Όταν υπολογίζεται X^2 τιμή ο αριθμός

* Το Knowledge Seeker δεν υπολογίζει το βάρος, αλλά ο χρήστης δηλώνει μία μεταβλητή να ληφθεί υπόψη ως σταθμική.

των παρατηρήσεων αντικαθίστανται από το άθροισμα των σταθμικών παρατηρήσεων. Η X^2 είναι κατανομή με (d-1)(k-1) βαθμούς ελευθερίας – οι βαθμοί ελευθερίας είναι ανεξάρτητοι από τις συνθήκες βάρους της εξαρτημένης μεταβλητής. Οι έλεγχοι που βασίζονται σε αυτή την τιμή όταν οι σταθμικές τιμές είναι αποτελέσματα στρωματοποιημένης δειγματοληψίας είναι αρκετά συντηρητικοί.

8.11. Ορισμός κόστους – μεγιστοποίηση κέρδους

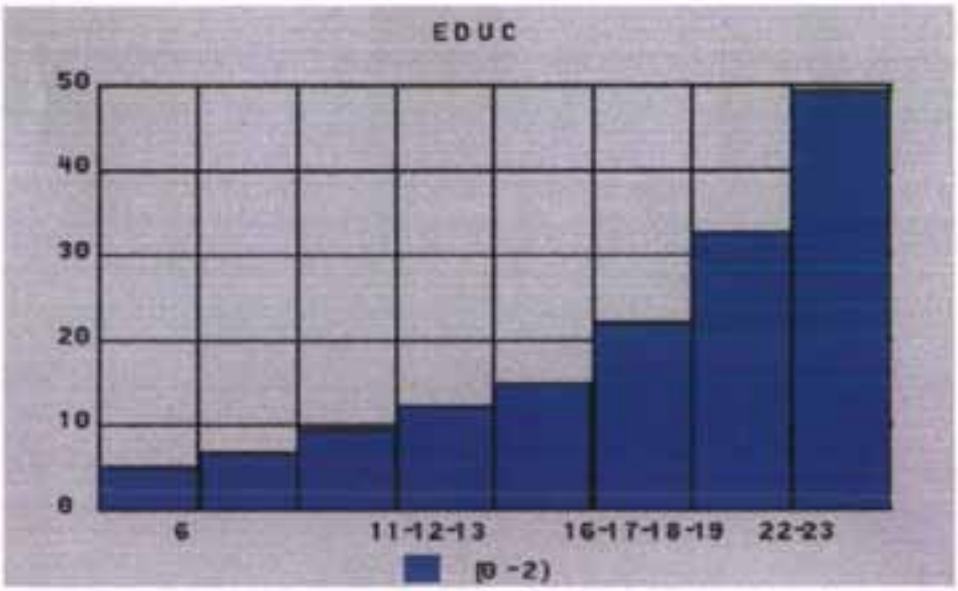
Το Knowledge Seeker μπορεί να αναθέσει κόστος ή κέρδος σε απαντήσεις που αντιστοιχούν σε κατηγορίες της εξαρτημένης μεταβλητής. Για παράδειγμα αν η εξαρτημένη μεταβλητή έχει τις απαντήσεις "αγοράζω" / "δεν αγοράζω", κωδικοποιημένες με τις τιμές 1 και 0 αντίστοιχα, μπορούμε να αναθέσουμε κέρδος στην τιμή 1 και κόστος στην τιμή 0. Το συνολικό κέρδος στο συγκεκριμένο κόμβο υπολογίζεται πολλαπλασιάζοντας το κέρδος με τον αριθμό των τιμών 1 και αφαιρώντας το κόστος πολλαπλασιασμένο με τον αριθμό των τιμών 0. Μπορεί έτσι να παράγεται στη ρίζα του δέντρου το άθροισμα των κατανομών συχνοτήτων των τιμών σε κάθε απάντηση – πεδίο. Μπορούμε τελικά να μετατρέψουμε το χάσιμο σε κέρδος, αρκεί να μπορούμε να αναγνωρίσουμε το τμήμα αγοράς που έχει μπει σε στόχο και το οποίο έχει ένα σχετικά υψηλό ποσοστό σε σχέση με το κέρδος των απαντήσεων.

Ας υποθέσουμε για παράδειγμα ότι έχουμε τις κωδικοποιημένες απαντήσεις τριών τιμών, απάντηση: πληρωμή, απάντηση: όχι πληρωμή, καμία απάντηση. Η πρώτη τιμή δίνει έσοδα $15, η δεύτερη κοστίζει $15 και η τρίτη κοστίζει $5. Τα συνολικά έσοδα / έξοδα για κάποια ανάλυση μπορούν να προσδιοριστούν από τα αποτελέσματα των πεδίων των απαντήσεων. Επιλέγοντας το πεδίο που μας ενδιαφέρει ως εξαρτημένη μεταβλητή υπολογίζεται αυτόματα το κόστος / κέρδος που μας ενδιαφέρει και παρουσιάζεται σε σχέση με την κατανομή των παρατηρήσεων.

8.12. Γραφήματα

Οι πληροφορίες που υπάρχουν σε κάθε κόμβο του δέντρου μπορούν να αναπαρασταθούν γραφικά. Τα γραφικά αποτελέσματα μιας μεταβλητής – κόμβου δείχνουν την κατανομή συχνοτήτων των αμέσως επόμενων κόμβων (απογόνων) ή αν δεν υπάρχουν απόγονοι την κατανομή των τιμών της εξαρτημένης μεταβλητής σε σχέση με την εκάστοτε υπό μελέτη μεταβλητή – κόμβο. Αυτά απεικονίζονται με δισδιάστατα και τρισδιάστατα ραβδογράμματα, τομεογραφήματα, γραφήματα στρωμάτων, καθώς και απλά γραφήματα γραμμών.

Το παρακάτω σχήμα, για παράδειγμα απεικονίζει το πως κατανέμεται η μεταβλητή *Educ* σε σχέση με την εξαρτημένη μεταβλητή *Wage*. Στον οριζόντιο άξονα απεικονίζονται οι τιμές της μεταβλητής της εκπαίδευσης σε χρόνια και στον κάθετο το μέσο ωριαίο εισόδημα των ανθρώπων που ανήκουν στην αντίστοιχη κατηγορία ή επίπεδο εκπαίδευσης. Το ίδιο γράφημα μπορεί να παρουσιαστεί και στον χώρο καθώς και με άλλη μορφή όπως γραμμογράφημα ή τομεογράφημα.



Σχήμα 8.6

8.13. Έλεγχοι ακρίβειας του δέντρου αποφάσεων

Το Knowledge Seeker μπορεί να εκτελέσει τρεις τύπους ελέγχων ακρίβειας του δέντρου: επαναντικατάστασης (resubstitution), επικύρωσης (validation) και σύγκρισης (comparison). Ο έλεγχος τύπου επαναντικατάστασης είναι αρκετά γρήγορος , αλλά υποτιμά το βαθμό σφάλματος ακρίβειας (error rate) μια και τα ίδια δεδομένα που χρησιμοποιήθηκαν για τη δημιουργία του δέντρου χρησιμοποιούνται και από τον έλεγχο. Με παρόμοιο τρόπο λειτουργίας, ο έλεγχος τύπου επικύρωσης υπολογίζει το βαθμό σφάλματος με μεγαλύτερη ακρίβεια μια και χρησιμοποιεί ένα εναλλακτικό σύνολο δεδομένων για την εφαρμογή του ελέγχου. Τέλος, ο έλεγχος τύπου σύγκρισης, που είναι έλεγχος επικύρωσης, εξάγει αποτελέσματα με περισσότερες λεπτομέρειες. Είναι σχεδιασμένος για να δείχνει τις διαφορές μεταξύ του παρόντος δέντρου και ενός εναλλακτικού συνόλου δεδομένων και όχι μόνο το συνολικό βαθμό σφάλματος, που παρουσιάζουν οι άλλοι δύο τύποι.

8.13.1. Έλεγχος με τη χρήση μεθόδου επαναντικατάστασης (re-substitution)

Αν η εξαρτημένη μεταβλητή είναι κατηγορική: Ο αριθμός ακρίβειας (accuracy number) δηλώνει πόσο συχνά είναι σωστός ο προσδιορισμός των εγγραφών – παρατηρήσεων στο δείγμα της επικρατούσας κατηγορίας της εξαρτημένης μεταβλητής. Για παράδειγμα αν είναι 60% ,τότε θα έχει προσδιοριστεί σωστά, περίπου 60% των φορών, μια μη ταξινομημένη εγγραφή της επικρατούσας κατηγορίας του κόμβου της μεταβλητής, η οποία συμφωνεί με τα χαρακτηριστικά των μη ταξινομημένων εγγραφών. Μπορούμε να επαληθεύσουμε την ακρίβεια αυτού του υπολογισμού κοιτάζοντας τα ποσοστά των παρατηρήσεων των "κατώτερων" κόμβων του δέντρου αποφάσεων που δε βρίσκονται στην επικρατούσα κατηγορία. Το σταθμικό (weighted) άθροισμα αυτών των ποσοστών όλων των "κατώτερων" κόμβων του δέντρου είναι ο συνολικός βαθμός σφάλματος ακρίβειας. Ο βαθμός σφάλματος δεδομένου κόμβου αντισταθμίζεται από τον

αριθμό των παρατηρήσεων στον κόμβο, σε σχέση με το συνολικό μέγεθος του δέντρου. Γενικότερα, μια ταξινόμηση λαμβάνεται σωστή όταν η τιμή του προτύπου (αρχέτυπου) συνόλου δεδομένων συμφωνεί με την τιμή της επικρατούσας κατηγορίας (modal value) της εξαρτημένης μεταβλητής για ένα δοσμένο κόμβο. Ο υπολογισμός αυτός γίνεται για κάθε παρατήρηση του συνόλου δεδομένων και αθροίζεται διαμέσου όλων των παρατηρήσεων του πρωτότυπου συνόλου δεδομένων, για να προσδιορισθεί ποια τιμή της εξαρτημένης μεταβλητής χρησιμοποιείται για να προσδιορίσει αν η ταξινόμηση είναι σωστή ή όχι.

8.13.2 Έλεγχος με τη χρήση της μεθόδου επικύρωσης (validation)

Το Knowledge Seeker παρέχει μια διεργασία επικύρωσης για να διαβεβαιώσει ότι χρησιμοποιείται μία ακριβής επιστημονική προσέγγιση για την ανάλυση δεδομένων στην αναγνώριση (εύρεση) σχέσεων που περιγράφονται από το στατιστικό δέντρο αποφάσεων. Μια ακριβής επιστημονική προσέγγιση ανάλυσης δεδομένων απαιτεί τα αποτελέσματα να είναι τόσο ακριβή όσο και έγκυρα. Το Knowledge Seeker χρησιμοποιεί στατιστικό έλεγχο υποθέσεων ως θεμελιώδη (πρωταρχική) μέθοδο αναγνώρισης ισχυρών σχέσεων που εμφανίζονται στο δέντρο αποφάσεων. Οι έλεγχοι υποθέσεων βασίζονται σε στατιστική θεωρία και σε εμπειρικούς κανόνες, που είναι απόρροια στατιστικών διαδικασιών, για να υπολογίσουν την πιθανότητα λάθους που σχετίζεται άμεσα με κάθε σχέση που παρουσιάζεται στο δέντρο. Ουσιαστικά, μια τέτοια προσέγγιση σημαίνει ότι εάν ένα αποτέλεσμα αναφέρεται ότι έχει επίπεδο εμπιστοσύνης 0.01 τότε βάσει στατιστικής θεωρίας κατά 99% είναι σωστό.

Όταν μία σχέση έχει αναγνωρισθεί, συνήθως επιχειρείται η διεξαγωγή μιας ένδειξης της ισχύος της. Για παράδειγμα, μπορεί κάποιος να πει ότι υπάρχει στατιστική διαφορά σημαντικότητας (διαφορά επιπέδου εμπιστοσύνης) στο ύψος των ανδρών και των γυναικών. Αλλά γνωρίζοντας αυτό, πόσο καλή μπορεί να είναι η πρόβλεψη του ύψους ενός συγκεκριμένου άνδρα σε σχέση με μιας γυναίκας; Φυσικά, γνωρίζοντας για παράδειγμα ότι η διάκριση μεταξύ ανδρών και γυναικών είναι σημαντική και γνωρίζοντας

τη δομή της σχέσης, π.χ. εάν το γένος είναι αρσενικό τότε το ύψος είναι 20 εκατοστά πάνω από το μέσο όρο.

Το θέμα της ακρίβειας μιας προβλεπόμενης σχέσης είναι κρίσιμο. Για παράδειγμα ας πάρουμε μία ιατρική ή επενδυτική περίπτωση. Μπορεί να γνωρίζουμε ότι η υψηλή χοληστερίνη επηρεάζει την πιθανότητα περιστατικού καρδιακής προσβολής. Αφού λοιπόν έχει επικυρωθεί η ισχύς της σχέσης αυτής, η πραγματική ερώτηση που μας απασχολεί είναι πόσο η υψηλή χοληστερίνη επηρεάζει και με τι πιθανότητα. Τυχούσα διαφορά βαθμών σημαντικότητας ακρίβειας είναι πολύ σημαντική, μιας και συγκεκριμένη εκτίμηση θα καθορίσει το είδος της θεραπείας που απαιτείται, καθώς και το αν η θεραπεία θα είναι επιτυχημένη ή όχι. Παρεμφερώς, εάν αποδείξουμε ότι υπάρχει μια ισχυρή σχέση του επιτοκίου και της μορφής κατανάλωσης δανείου, τότε η επόμενη πρόκληση είναι να προσδιορίσουμε πόσο πρέπει να αυξηθεί ο τόκος πριν από την κατανάλωση κατά μέσο όρο.

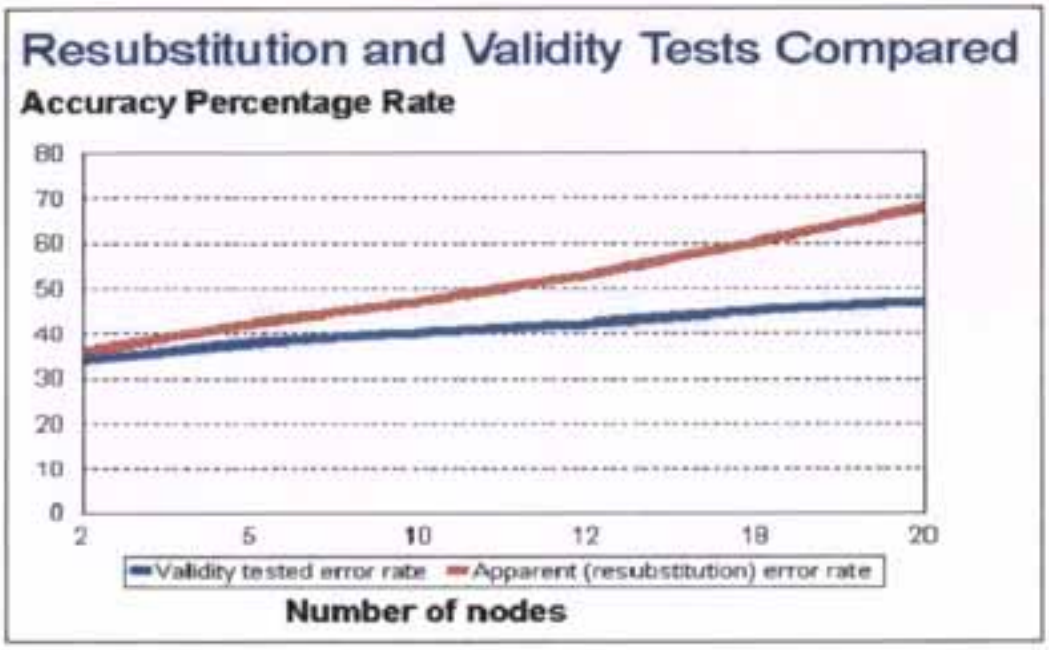
Τα πιο πάνω παραδείγματα απαιτούν βαθμό ακρίβειας που δεν μπορεί υπό κανονικές συνθήκες να παρέχει η συμβατική στατιστική θεωρία και τα εμπειρικά στατιστικά αποτελέσματα. Στην πραγματικότητα τα εμπειρικά στατιστικά αποτελέσματα μπορεί να είναι παραπλανητικά σε τέτοιες περιπτώσεις, υποστηρίζουν οι κατασκευαστές αυτού του πακέτου επεξεργασίας δεδομένων. Αυτό μπορεί να συμβεί γιατί τα ίδια δεδομένα που χρησιμοποιήθηκαν για να αναγνωρισθούν ισχυρές στατιστικές σχέσεις, αργότερα επαναχρησιμοποιούνται για να καθοριστεί η ισχύς των σχέσεων που αναγνωρίσθηκαν πριν. Η εμπειρία έχει δείξει ότι αυτού του είδους η εκτίμηση ακρίβειας, που εξήχθηκε με τη μέθοδο της επαναντικατάστασης (re-substitution), δίνει αρκετά χρήσιμα αποτελέσματα, όταν όμως εφαρμόζεται σε προηγούμενα μη ταξινομημένα δεδομένα. Το να χρησιμοποιηθούν τα ίδια δεδομένα δύο φορές, μία για την διεξαγωγή του στατιστικού μοντέλου και μετά για τον έλεγχο ακρίβειας του μοντέλου, έχει ως αποτέλεσμα να λαμβάνονται υπόψη ιδιοσυγκρασίες των δεδομένων, κάτι που δε θα συνέβαινε εάν ένα καινούριο σύνολο παρατηρήσεων χρησιμοποιούταν κατά τον υπολογισμό της ακρίβειας.

Ο πραγματικός έλεγχος επανακατασκευαστικότητας ενός δεδομένου στατιστικού μοντέλου επιτυγχάνεται όταν η ακρίβειά του εκτιμάται από εντελώς καινούριες παρατηρήσεις. Υπολογίζεται ένα επικυρωμένο ποσοστό ακρίβειας διαβεβαιώνοντας ότι διαφορετικό σύνολο δεδομένων χρησιμοποιείται στην δημιουργία και στον έλεγχο του στατιστικού μοντέλου, που αποτελεί τη βάση δημιουργίας του στατιστικού δέντρου αποφάσεων. Αυτό γίνεται δημιουργώντας σύνολο δεδομένων για έλεγχο (test data = hold back sample) πριν τη δημιουργία του δέντρου. Οπότε ένα σύνολο δεδομένων χρησιμοποιείται για την εκμάθηση της δομής των στατιστικών σχέσεων των δεδομένων και ένα άλλο νέο σύνολο για τον έλεγχο της ακρίβειας του μοντέλου.

Το Knowledge Seeker ακολουθεί τέτοια διαδικασία με την εντολή "partition data". Ως αποτέλεσμα της διαδικασίας έχουμε τη δημιουργία δέντρου αποφάσεων με τη χρήση μικρότερου μεγέθους δεδομένα εκμάθησης δομής, στο οποίο αν και οι κατανομές των ποσοστών σε κάθε κόμβο διαφέρουν, τα γενικά αποτελέσματα παραμένουν τα ίδια. Με αυτό τον τρόπο, το πρόγραμμα έχει στη διάθεσή του ένα δεύτερο σύνολο δεδομένων για ελέγχους. Αυτό είναι το σύνολο δεδομένων χρησιμοποιείται για να υπολογιστούν οι στατιστικοί έλεγχοι και δίνει ποσοστό ακρίβειας περίπου 60% και βαθμό σφάλματος ακρίβειας 40%. Όπως λοιπόν γίνεται φανερό, το σφάλμα ακρίβειας στην περίπτωση της επικύρωσης είναι μικρότερο από αυτό στην περίπτωση της επαναντικατάστασης. Έτσι χρησιμοποιώντας ανεξάρτητα επιλεγμένα σύνολα δεδομένων αρχικά για τη δημιουργία του δέντρου και στη συνέχεια για τον έλεγχο αυτού, παρέχεται μια πολύ καλύτερη εκτίμηση της ακρίβειας και του ποσοστού της λανθασμένης ταξινόμησης των παρατηρήσεων του δέντρου αποφάσεων.

Αυτή η διαφορά μεταξύ του "φαινομενικού" (που προέκυψε από την εκτίμηση με τη χρήση της μεθόδου επαναντικατάστασης) και του επικυρωμένου βαθμού σφάλματος ακρίβειας (που προέκυψε από τη χρήση της μεθόδου επικύρωσης) αυξάνεται όλο και περισσότερο, όσο αυξάνονται οι κόμβοι του δέντρου, όπως φαίνεται και από το ακόλουθο διάγραμμα. Έτσι, όσο μεγαλώνει το δέντρο, τόσο περισσότερο σημαντική

είναι η επικύρωση για την καλύτερη εξαγωγή ακριβών εκτιμήσεων της πραγματικής ακρίβειας ταξινόμησης του δέντρου αποφάσεων.



Σχήμα 8.7

9. ΕΠΙΛΟΓΟΣ

Ανάλογα με τις δυνατότητες που προσφέρουν, έχουν προκύψει τρεις τύποι συστημάτων υποστήριξης αποφάσεων.

- Κατευθυνόμενα από μοντέλο (model-driven). Πρόκειται για ολοκληρωμένα συστήματα που έχουν τη δυνατότητα εκτέλεσης what-if σεναρίων καθώς και άλλων τύπων ανάλυσης..
- Κατευθυνόμενα από τα δεδομένα (data-driven). Επιτρέπουν στο χρήστη να εξάγει και να αναλύει χρήσιμη πληροφορία από μεγάλες βάσεις δεδομένων.
- Εξόρυξης Γνώσης (data mining). Εύρεση κρυμμένων τυποποιημένων μορφών (patterns) και κάποιων σχέσεων(relationships) σε μεγάλες βάσεις δεδομένων. Το αποτέλεσμα είναι η εξαγωγή κάποιων κανόνων ώστε να είναι δυνατή η πρόβλεψη μελλοντικών συμπεριφορών.

Στις δυο πρώτες κατηγορίες ανήκουν κυρίως τα σημερινά ΣΥΑ. Η εργασία μελετά την τρίτη κατηγορία και προσπαθεί να δείξει τη δυναμική που μπορούν να προσδώσουν οι τεχνολογίες που σχετίζονται με την εξόρυξη γνώσης στην υποστήριξη αποφάσεων. Αναλυτικότερα, αυτή η εργασία δείχνει ότι μπορούμε να χρησιμοποιήσουμε εργαλεία εξόρυξης γνώσης για να ανακαλύψουμε χρήσιμη γνώση όπως οι εγγυητικοί κανόνες πιστωτικού ορίου για αιτούντες πιστωτικών καρτών.

Είναι γνωστό ότι στα σύστημα υποστήριξης αποφάσεων ένα πρόβλημα μπορεί να λυθεί συγκριτικά καλά χρησιμοποιώντας διάφορες μεθόδους, π.χ. διαφορετικούς τύπους τεχνητών νευρονικών δικτύων, δέντρων απόφασης, Μπεϋζιανών ταξινομητών, κ.λπ. Επιπλέον, μπορεί να συμβεί μια μέθοδος να προβλέπει καλύτερα ορισμένα μέρη του διαστήματος των περιπτώσεων από τις άλλες. Κατά συνέπεια, η επιλογή της καταλληλότερης μεθόδου για την τελική λύση είναι ένα περίπλοκο πρόβλημα. Διαφορετικές αναπαραστάσεις (ταξινομητές) είναι κατάλληλες για διαφορετικά προβλήματα.

Γι' αυτό το λόγο αξιολογήσαμε τους γνωστότερους αλγόριθμους εξόρυξη γνώσης στο πρόβλημα μας χρησιμοποιώντας το WEKA και παρατηρήσαμε ότι τα δένδρα αποφάσεων συμπεριφέρονται καλύτερα. Έτσι αποφασίσαμε να δοκιμάσουμε το εμπορικό πακέτο –Knowledge SEEKER – που επιτρέπει μεγάλη παραμετροποίηση στα δέντρα αποφάσεων και εξοικειωθήκαμε με αυτό.

10. ΑΝΑΦΟΡΕΣ

Acid, S. and de Campos. L. M. (2003). Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs. *Journal of Artificial Intelligence Research* 18: 445-490.

Aha, D. (1997). *Lazy Learning.* Dordrecht: Kluwer Academic Publishers.

Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. 2(2): 1-47.

Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* 137: 43-90.

Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge.

Dejong, K. A., Spears, W. M., & Gordon, D. F. (1993). Using genetic algorithms for concept learning. *Machine Learning* 13: 161-188.

De Mantaras & Armengol E. (1998). *Machine learning* from examples: Inductive and Lazy methods. Data & Knowledge Engineering 25: 99-123.

Domingo's, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29: 103-130.

Dutton, D. & Conroy, G. (1996). A review of machine learning, *Knowledge Engineering Review* 12: 341-367.

Famili, A., Shen, W., Weber, R., Simoudis, E. (1997), Data Preprocessing and Intelligent Data Analysis, *Intelligent Data Analysis* 1: 3-23.

Furnkranz, J. (1999). Separate-and-Conquer Rule Learning. *Artificial Intelligence Review* 13: 3-54.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood: London.

Mitchell, T. (1997*). Machine Learning.* McGraw Hill.

Murthy, (1998), Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining and Knowledge Discovery* 2: 345-389.

Neocleous, C. & Schizas, C., (2002), Artificial Neural Network Learning: A Comparative Review, In Vlahavas, I. P. & Spyropoulos C. D. (ed.), *SETN 2002*, 300-313, LNAI 2308, Springer-Verlag Berlin Heidelberg.

Platt, J. (1999). Using sparseness and analytic QP to speed training of support vector machines. In Kearns, M. S., Solla, S. A. & Cohn, D. A. (ed.), *Advances in neural information processing systems.* MA: MIT Press.

Pyle, D. (1999). *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, Los Altos, CA.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco.

Witten, I. & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Mateo, 2000.

Πηγές στο διαδίκτυο

Άρθρο του Dan Greening για τις τεχνικές εξόρυξης γνώσης από δεδομένα που προέρχονται από το Internet με σκοπό την ανάλυση κα τη λήψη αποφάσεων: "Data Mining on the Web"
www.webtechniques.com/archives/2000/01/greening

Άρθρο του Kurt Thearling: "An Introduction to Data Mining"
http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm

Πληροφορίες για την Επιχειρηματική Νοημοσύνη:
http://www.infocube.co.uk/business_intelligence.htm