ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Μηχανικών Παράγωγης και Διοίκησης

ΜΕΘΟΔΟΛΟΓΙΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΑΠΑΤΗΣ: ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΣΤΟ ΧΩΡΟ ΤΩΝ ΑΣΦΑΛΕΙΩΝ

Διατριβή που υπεβλήθη για τη μερική ικανοποίηση των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης

Υπό

ΚΟΝΣΟΛΑΚΗ ΕΥΑΓΓΕΛΙΑ

Επίβλεψη

ΔΟΥΜΠΟΣ ΜΙΧΑΛΗΣ

ΧΑΝΙΑ 2012

Στους γονείς μου..

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Δούμπο Μιχάλη, για την καθοδήγηση του και τις πολύτιμες συμβουλές που μου παρείχε και καλλιέργησε άρτιο κλίμα συνεργασίας καθ' όλη τη διάρκεια της παρούσας μεταπτυχιακής εργασίας.

Ένα πολύ μεγάλο ευχαριστώ στην οικογένεια μου, είναι το λιγότερο που θα μπορούσα να πω για τη διαρκή υποστήριξη τους

.

ΠΕΡΙΛΗΨΗ

Η φύση της απάτης αλλάζει με το χρόνο, η έκτασή της είναι άγνωστη εκ των προτέρων και μπορεί να απλώνεται σε περισσότερες της μιας υπηρεσίες. Η απάτη αυξάνεται δραματικά με την εξάπλωση της τεχνολογίας και των παγκόσμιων λεωφόρων τηλεπικοινωνιών, και επιφέρει απώλεια δισεκατομμυρίων δολαρίων παγκοσμίως κάθε χρόνο. Αν και οι τεχνολογίες πρόληψης είναι ο καλύτερος τρόπος μείωσης της απάτης, ωστόσο οι «απατεώνες» τείνουν να είναι προσαρμοστικοί και σχεδόν πάντα βρίσκουν τρόπους να διαφεύγουν. Οι μεθοδολογίες ανίχνευσης της απάτης είναι σημαντικές για την σύλληψη των «απατεώνων» στην περίπτωση που η πρόληψη έχει αποτύχει. Η στατιστική και η εκμάθηση των μηχανών (machine learning), παρέχουν ικανοποιητική τεχνολογία για την ανίχνευση της απάτης και έχουν εφαρμοσθεί επιτυχώς στην ανίχνευση «ύποπτων» δραστηριοτήτων όπως σε πιστωτικές κάρτες ηλεκτρονικού εμπορίου, στις τηλεπικοινωνίες, στην παραβίαση ασφάλειας ηλεκτρονικών υπολογιστών κ.ά.

Για την επίλυση του προβλήματος της ανίχνευσης απάτης προτείνουμε την εφαρμογή αλγορίθμων εξόρυξης δεδομένων (ΕΔ). Παρουσιάζεται η λειτουργία τους και τέλος γίνεται εφαρμογή με δεδομένα από το χώρο της ασφάλισης αυτοκίνητων.

Λέξεις κλειδιά : ανίχνευση απάτης, εξόρυξη δεδομένων, ασφάλιση αυτοκίνητων, κατηγοριοποίηση, weka

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

«ΕΙΣΑΓΩΓΗ»

1.1 Γενικά

Η φύση της απάτης αλλάζει με το χρόνο, η έκτασή της είναι άγνωστη εκ των προτέρων και απλώνεται σε όλα τα επίπεδα της επιχειρηματικής δραστηριότητας. Ο εντοπισμός της απάτης είναι πλέον ιδιαίτερα περίπλοκος λόγω της πολυπλοκότητας των σύγχρονων τεχνολογικών και διοικητικών συστημάτων αλλά και του μεγάλου όγκου πληροφοριών και δεδομένων που αφορούν επιχειρηματικές συναλλαγές.

Η απάτη αυξάνεται δραματικά με την εξάπλωση της τεχνολογίας και των παγκόσμιων δικτύων τηλεπικοινωνιών, και επιφέρει απώλεια δισεκατομμυρίων δολαρίων παγκοσμίως κάθε χρόνο. Αν και οι τεχνολογίες πρόληψης είναι ο καλύτερος τρόπος αντιμετώπισης του προβλήματος, ωστόσο οι «απατεώνες» τείνουν να είναι προσαρμοστικοί και σχεδόν πάντα βρίσκουν τρόπους να διαφεύγουν. Οι μεθοδολογίες ανίχνευσης της απάτης είναι σημαντικές στην περίπτωση που η πρόληψη έχει αποτύχει. Η στατιστική και η εκμάθηση των μηχανών (machine learning), παρέχουν ικανοποιητική τεχνολογία για την ανίχνευση της απάτης και έχουν εφαρμοσθεί επιτυχώς στην ανίχνευση «ύποπτων» δραστηριοτήτων όπως σε πιστωτικές κάρτες, στο ηλεκτρονικό εμπόριο, στις τηλεπικοινωνίες, στην παραβίαση ασφάλειας ηλεκτρονικών υπολογιστών κ.ά.

Με τον όρο «**απάτη**» (fraud**)** ορίζεται κάθε ενέργεια που μέσω της χρήσης ψευδών, λανθασμένων ή ημιτελών αναπαραστάσεων, εγγράφων και πληροφοριών αποσκοπεί στην απόκτηση παράνομου πλεονεκτήματος, συνήθως οικονομικού.

Η απάτη απαντάται σε πάρα πολλές μορφές και σε πολλούς τομείς της κοινωνικής ζωής. Ωστόσο, η ραγδαία εξέλιξη της τεχνολογίας σε συνδυασμό με την αύξηση της καταναλωτικής ισχύος και τις προηγμένες μεθόδους επικοινωνίας, έχει οδηγήσει σε πολλαπλά είδη απάτης. Αρχικά, λόγω της λεπτής γραμμής διαχωρισμού ανάμεσα στην έρευνα ύποπτων αιτημάτων και στην παρενόχληση νόμιμων ατόμων, κάθε προσπάθεια για ανίχνευση της απάτης θεωρείτο αντικαταναλωτική κίνηση. Μόλις στη δεκαετία του 1990 άρχισε συντονισμένη δράση κατά της απάτης σε διάφορους τομείς.

Υπάρχει σαφής διαχωρισμός ανάμεσα στην πρόληψη και στην ανίχνευση της απάτης. Η πρόληψη συνίσταται σε λήψη μέτρων αποτρεπτικού χαρακτήρα (π.χ. πολύπλοκα σχέδια, φθορίζουσες ίνες, υδατογραφήματα, ολογράμματα σε χαρτονομίσματα, κάρτες SIM για τα κινητά τηλέφωνα, PIN και κωδικοί ασφαλείας για τραπεζικούς λογαριασμούς ή πρόσβαση σε Η/Υ, βιομετρικές μέθοδοι αναγνώρισης προσώπου/δακτυλικού αποτυπώματος/αμφιβληστροειδούς), τα οποία όμως δεν είναι πάντα αποτελεσματικαά. Η ανίχνευση είναι μια στρατηγική post-hoc και έρχεται συνήθως αφότου η πρόληψη έχει αποτύχει, όμως είναι κάτι που πρέπει να ακολουθεί υποχρεωτικά την πρόληψη, έτσι κι αλλιώς, διότι σχεδόν ποτέ κανείς δεν αντιλαμβάνεται άμεσα τη διάπραξη απάτης. Είναι μια συνεχώς εξελισσόμενη διαδικασία, διότι από τη στιγμή που μια μέθοδος ανίχνευσης χρησιμοποιηθεί, σχεδόν άμεσα θα μπορεί σε μεγάλο βαθμό να αντιμετωπισθεί από τους επαγγελματίες «απατεώνες». Παρόλα αυτά όμως συχνά παραγκωνίζονται παλαιότερες και αποτελεσματικότερες μέθοδοι ανίχνευσης προς όφελος νεώτερων, επομένως τα πρώτα εργαλεία ανίχνευσης απάτης θα πρέπει να χρησιμοποιούνται σε συνδυασμό με τα νέα πιο εξελιγμένα εργαλεία.

Στην εργασία αυτή, λοιπόν, θα πραγματοποιηθεί μια γενική επισκόπηση του φαινόμενου της απάτης στους διαφόρους τομείς όπου μπορεί να εμφανιστεί. Επιπλέον, θα γίνει μελέτη για την ανίχνευση και την πρόληψη της απατής με δεδομένα που προέρχονται από το χώρο των ασφαλιστήριων αυτοκίνητων.

1.2 Διάρθρωση εργασίας

Στο κεφάλαιο 2 που ακολουθεί, γίνεται μια σύντομη αναφορά στην έννοια της απάτης και σε κάποιους όρους που σχετίζονται με αυτή. Επιπλέον, περιγράφονται οι μέθοδοι ανίχνευσης της απάτης και παρατίθενται μια βιβλιογραφική ανασκόπηση εφαρμογής των μεθόδων αυτών. Τέλος παρουσιάζονται οι πιο δημοφιλείς μορφές απάτης που έχουν εμφανιστεί τα τελευταία, ειδικά, χρόνια.

Στο τρίτο κεφαλαίο γίνεται αναφορά στον ορισμό της εξόρυξης δεδομένων, παρουσιάζεται μια ιστορική ανάδρομη της πορείας της εξόρυξης δεδομένων και στα προβλήματα που οδήγησαν στη χρήση της και κυρίως στη στατιστική. Στο τελευταίο μέρος του κεφαλαίου παρουσιάζονται οι σημαντικότεροι αλγόριθμοι εξόρυξης δεδομένων και οι περιοχές εφαρμογής των αλγορίθμων αυτών.

Στο τέταρτο κεφάλαιο περιγράφεται η εφαρμογή της εξόρυξης δεδομένων στο πρόγραμμα WEKA. Επιπλέον, παρουσιάζονται τα δεδομένα του προβλήματος, τα οποία προέρχονται από τον κλάδο της ασφάλισης αυτοκίνητων, και οι διάφοροι παράμετροι που επιλεχτήκαν κατά την υλοποίηση των αλγορίθμων. Τέλος, πραγματοποιείται ανάλυση και σχολιασμός των αποτελεσμάτων που πρόεκυψαν από την εφαρμογή της εξόρυξης δεδομένων.

Στο πέμπτο και τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα της παρούσας εργασίας και προτείνονται θέματα μελέτης για το μέλλον.

ΚΕΦΑΛΑΙΟ 2

«ΑΝΙΧΝΕΥΣΗ ΤΗΣ ΑΠΑΤΗΣ»

2.1 Εισαγωγή

Η απάτη, παλιά όσο και η ανθρωπότητα, απαντάται σε πάρα πολλές μορφές και σε πολλούς τομείς της κοινωνικής ζωής. Ωστόσο, τα φαινόμενα απάτης τείνουν να κυριαρχήσουν τις τελευταίες δεκαετίες σε κάθε τομέα. Η ραγδαία εξέλιξη της τεχνολογίας σε συνδυασμό με την αύξηση της καταναλωτικής ισχύος και τις προηγμένες μεθόδους επικοινωνίας, έχει οδηγήσει σε πολλαπλά είδη απάτης.

Αρχικά, λόγω της λεπτής γραμμής διαχωρισμού ανάμεσα στην έρευνα ύποπτων περιπτώσεων, και στην παρενόχληση νόμιμων ατόμων, κάθε προσπάθεια για ανίχνευση της απάτης θεωρείτο αντικαταναλωτική κίνηση. Μόλις στη δεκαετία του 1990 άρχισε συντονισμένη δράση κατά της απάτης σε διάφορους τομείς. Αυτή συνίσταται στην πρόληψη και στην ανίχνευση της απάτης.

Τόσο για τον δημόσιο, όσο και για τον ιδιωτικό τομέα, η διενέργεια απάτης είναι ένα ζήτημα με επιπτώσεις:

- Οικονομικές: μειωμένη λειτουργικότητα αποτελεσματικότητας λόγω έλλειψης πόρων
- Νομικές: στέρηση πόρων από τους νόμιμους κατόχους
- Ψυχολογικές: ηθική κατάπτωση και έλλειψη εμπιστοσύνης

Τα μεγαλύτερο ίσως εμπόδιο στην καταπολέμηση της απάτης είναι το γεγονός ότι αυτή μπορεί να περιοριστεί αλλά όχι να εξαλειφτεί, μπορεί να διαχειριστεί αλλά όχι να απομακρυνθεί.

2.2 Ορισμός της απάτης

2.2.1 Εισαγωγή

Στη βιβλιογραφία συναντά κανείς διάφορους ορισμούς για το τι είναι απάτη. Οι Thomas et al.(2004) αναφέρουν ότι ανάλογα με το πεδίο στο οποίο χρησιμοποιείται η

έννοια του όρου αυτού μπορεί να αλλάζει ελαφρώς περιεχόμενο και να αποδίδεται με την ορολογία του συγκεκριμένου τομέα,. Με τον όρο «απάτη» (fraud) ορίζεται κάθε ενέργεια που μέσω της χρήσης ψευδών, λανθασμένων ή ημιτελών αναπαραστάσεων, εγγράφων και πληροφοριών αποσκοπεί στην απόκτηση παράνομου πλεονεκτήματος, συνήθως οικονομικού. Ένας άλλος ορισμός, σύμφωνα με τον Hollmen (2000) εκφράζει το πνεύμα της απάτης ως "την κακή χρήση, τις ανέντιμες προθέσεις και την ανάρμοστη συμπεριφορά χωρίς την προϋπόθεση της ύπαρξης νομικών συνεπειών". Επιπλέον, υπάρχει μια σειρά στοιχείων που πρέπει να ληφθούν υπόψη όταν μελετάται μια περίπτωση απάτης όπως (Davia et al., 2000) :

- Το θύμα της απάτης.
- Λεπτομέρειες για τις ενέργειες που θεωρούνται ύποπτες.
- Τη ζημία που υπέστη το θύμα της απάτης.
- Αποδείξεις ότι ο δράστης έδρασε με δόλο.
- Αποδείξεις ότι ο δράστης είχε όφελος από τη διάπραξη της απάτης.

Η ανάπτυξη των σύγχρονων τεχνολογιών έχει εξελίξει τις παλιές απάτες και έχει βοηθήσει στη δημιουργία καινούριων. Σε αυτό το πλαίσιο, η νομοθεσία αναγκαστικά εκσυγχρονίζεται διαρκώς προσπαθώντας να συμπεριλάβει όλες τις νέες μορφές απάτης που εμφανίζονται, όπως π.χ. αυτές του ηλεκτρονικού εμπορίου. Παράλληλα όμως εκσυγχρονίζονται και οι απατεώνες κάνοντας δυσκολότερο το έργο των νομοθετών.

Υπάρχουν δύο τρόποι για το χειρισμό της απάτης: ο ένας είναι η *πρόληψη* (*prevention*) και ο άλλος η *ανίχνευση* (*detection*). Η *πρόληψη* είναι ο καλύτερος και πιο αποτελεσματικός τρόπος για την αντιμετώπιση της απάτης και περιλαμβάνει το σύνολο των μέτρων που λαμβάνονται για την αποσόβηση της απάτης. Η πρόοδος στην τεχνολογία έχει επιτρέψει την ανάπτυξη μέσων, μηχανημάτων αλλά και ολόκληρων συστημάτων που βοηθούν στην πρόληψη της απάτης. Για παράδειγμα, σε συστήματα ηλεκτρονικού υπολογιστή, η χρήση firewall εμποδίζει τη μη εξουσιοδοτημένη εκτέλεση εφαρμογών, ενώ στην περίπτωση των χαρτονομισμάτων ο περίπλοκος σχεδιασμός, τα ολογράμματα και οι ταινίες ασφαλείας δυσκολεύουν την παραχάραξή τους. Το μειονέκτημα αυτού του τρόπου αντιμετώπισης της απάτης είναι ότι μπορεί να εμποδίσει τις ήδη γνωστές απάτες και μόνο αυτές. Όμως οι απατεώνες, εκμεταλλευόμενοι και την εξέλιξη της τεχνολογίας, βρίσκουν συχνά νέους τρόπους απάτης που παρακάμπτουν τους μηχανισμούς της πρόληψης. Κατά συνέπεια καμία μέθοδος προφύλαξης δεν είναι απόλυτα ασφαλής.

Η *ανίχνευση* της απάτης εμπεριέχει όλες εκείνες τις ενέργειες και τα εργαλεία που βοηθούν στην αποκάλυψη της διαπραχθείσας απάτης. Αξίζει να τονιστεί ότι η ανίχνευση δεν αποτελεί ένα δεύτερο (βοηθητικό) εργαλείο το οποίο θέτουμε σε λειτουργία με την αποτυχία της πρόληψης. Αυτό διότι δεν είναι δυνατόν να γνωρίζουμε πότε τα συστήματα προφύλαξης έχουν υποστεί διάρρηξη ώστε να περάσουμε στην ανίχνευση. Καθώς οι απάτες εξελίσσονται με γοργούς ρυθμούς κάνοντας τα συστήματα πρόληψης να φαίνονται ανεπαρκή και καθιστώντας σαφές

ότι δεν αποτελούν οριστική λύση στο πρόβλημα, τα εργαλεία ανίχνευσης έχουν ως στόχο την ανακάλυψη της απάτης καλύπτοντας το κενό ασφαλείας που δημιουργείται από τις νέες περιπτώσεις απάτης. Επιπλέον, η ανίχνευση μιας απάτης, αν και δεν μας προφυλάσσει, είναι σημαντική και για την πρόληψη. Πέρα από το σταμάτημά λοιπόν της συγκεκριμένης απάτης, η ανίχνευση βοηθά στη λήψη όλων εκείνων των μέτρων που εξασφαλίζουν την προστασίας από την ίδια απάτη στο μέλλον. Συμβάλλει έτσι στη βελτίωση και τον εκσυγχρονισμό της πρόληψης. Προφανώς, η χρήση των εργαλείων ανίχνευσης οφείλει να είναι συνεχής και να γίνεται ταυτόχρονα με αυτών της πρόληψης.

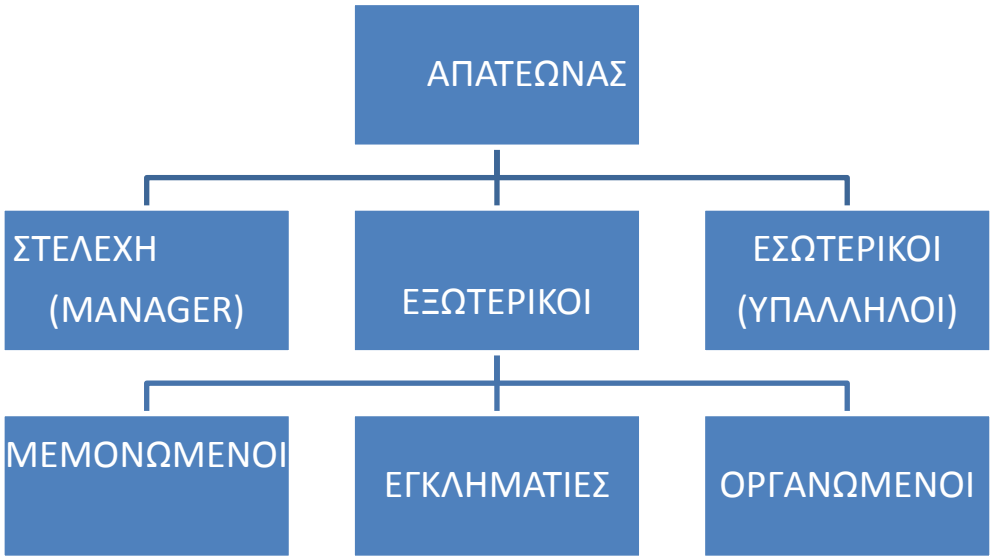
Η ανάπτυξη νέων μεθόδων για την ανίχνευση της απάτης δυσχεραίνεται από το γεγονός ότι η ανταλλαγή ιδεών στον τομέα αυτό είναι εξαιρετικά περιορισμένη. Δεν έχει νόημα να δοθούν στη δημοσιότητα εκτενής λεπτομέρειες για τις τεχνικές ανίχνευσης της απάτης, καθώς αυτό δίνει τη δυνατότητα στους εγκληματίες να συλλέξουν τις πληροφορίες που χρειάζονται, προκειμένου να ξεφύγουν από τη διαδικασία της ανίχνευσης. Για το λόγο αυτό, το σύνολο των δεδομένων, σε αρκετές περιπτώσεις, δεν είναι προσβάσιμα και τα αποτελέσματα συχνά λογοκρίνονται, γεγονός που καθιστά δύσκολη την αξιολόγηση (π.χ., Leonard, 1993).

2.2.2 Η έννοια του απατεώνα

Ως απατεώνας ορίζεται αυτός που διαπράττει την απάτη. Το οικονομικό κίνητρο θεωρείται ο πιο δημοφιλής λόγος για την διάπραξη της άπατης με ελάχιστες εξαιρέσεις (π.χ. τρομοκράτες). Σε αυτή την περίπτωση ο απατεώνας, συνήθως, έχει άμεση σχέση με τον περιβάλλον της πληγείσας επιχείρησης.

Γενικά μπορούμε να διαχωρίσουμε τους απατεώνες ανάμεσα σε εσωτερικούς και εξωτερικούς, ανάλογα με το εάν εργάζονται ή όχι στο σύστημα στο οποίο διαπράττεται η απάτη. Παραδοσιακά, κάθε επιχείρηση είναι ευαίσθητη σε θέματα εσωτερικής απάτης η οποία προέρχεται είτε από τα υψηλά επίπεδα (στελέχη) είτε από τα χαμηλότερα επίπεδα (υπάλληλοι) της εκάστοτε επιχείρησης.

Οι εξωτερικοί απατεώνες όπως φαίνεται και από το Σχήμα 2.1 διαιρούνται σε δύο κατηγορίες: τους απλούς και τους εγκληματίες που εμφανίζουν περιστασιακή παραβατική συμπεριφορά είτε εκμεταλλευόμενοι μια ευκαιρία είτε από ξαφνικό πειρασμό, είτε εξαιτίας οικονομικών δυσκολιών. Η υποκατηγορία των εγκληματιών είναι η πιο επικίνδυνη. Θα μπορούσαν να χαρακτηριστούν και ως επαγγελματίες αφού η δράση τους είναι συστηματική και συνεχίζουν μέχρι να ξεπεράσουν το εμπόδιο της πρόληψης αποφεύγοντας παράλληλα και την ανίχνευση. Οι εγκληματίες λειτουργούν μεμονωμένα ή οργανωμένοι σε ομάδες και αποτελούν μόνιμη απειλή. Οι Phua et al. (2005) επισημαίνουν ότι εσωτερικές ή ασφαλιστικές απάτες διαπράττουν συχνότερα οι απλοί απατεώνες, ενώ απάτες με πιστωτικές κάρτες ή απάτες τηλεπικοινωνιών οι εγκληματίες.



Σχήμα 2.1 : Διάγραμμα ταξινόμησης των απατεώνων

2.3 Διαθέσιμα δεδομένα

Τα δεδομένα που υπάρχουν για την ανίχνευση της απάτης, τις περισσότερες φορές είναι οργανωμένα σε βάσεις δεδομένων με ποικίλα στοιχεία για κάθε εγγραφή. Αποτελούνται από δυαδικά, ονομαστικά ή αριθμητικά αλλά και περισσότερο σύνθετα δεδομένα όπως εικόνες (Phua et al., 2005). Υπάρχουν δεδομένα με το πολύ 500 εγγραφές που αφορούν σε στελέχη επιχειρήσεων αλλά συνήθως είναι αρκετά μεγαλύτερα όπως οι 40.000 εγγραφές που προκύπτουν από τα ασφαλιστικά δεδομένα του Williams (1999). Στον κλάδο των τηλεπικοινωνιών και των συναλλαγών συναντώνται τα μεγαλύτερα σύνολα δεδομένων, όπως για παράδειγμα, αυτά της εταιρίας AT&T που διεκπεραιώνει 275 εκατομμύρια τηλεφωνήματα κάθε εβδομάδα (Cortes και Pregibon, 1998). Από την προηγούμενη αναφορά φαίνεται ότι τα δεδομένα εκτός από ογκώδη μπορεί να είναι και μη στατικά, καθώς ενδέχεται να αυξάνονται (και μάλιστα σημαντικά) στην πάροδο του χρόνου. Πρέπει επίσης να αναφερθεί ότι τα δεδομένα εμφανίζουν ανισορροπία σε σχέση με τις νόμιμες ή μη εγγραφές καθώς οι περιπτώσεις απάτης αποτελούν συνήθως λιγότερο από το 10% του συνόλου. Μάλιστα, δεν είναι λίγες οι φορές που τα ποσοστά απάτης στα δεδομένα πέφτουν μέσα στα όρια του στατιστικού σφάλματος, με αποτέλεσμα η κλασσική στατιστική να μην αρκεί.

Οι Bolton και Hand (2002) επικαλούνται μια έκθεση του Αμερικάνικου γραφείου αποτίμησης της τεχνολογίας (OTA, Office of Technology Assessment) στην οποία εκτιμάται ότι περίπου το 0,05% με 0,1% των συναλλαγών για μεταφορά χρημάτων αφορούν στο ξέπλυμα χρημάτων, ενώ ο Hassibi (2000) αναφέρει ότι στα δικά του δεδομένα 1 στις 1200 συναλλαγές μέσω πιστωτικών καρτών ήταν αποτέλεσμα απάτης. Οι συχνότητες αυτές είναι σημαντικά μικρότερες από το 5% ή ακόμη και το 1% που θεωρούνται αποδεκτά όρια λάθους σύμφωνα με τη στατιστή θεωρία.

Τα ποσοστά της απάτης που αναφέρθηκαν παραπάνω δεν πρέπει να λειτουργούν παραπλανητικά ως προς την αξία της ανίχνευσης απάτης. Το πρόβλημα της απάτης είναι σημαντικό και τα οφέλη πίσω από την ανίχνευση μιας απάτης μεγάλα. Μπορεί τα μικρά ποσοστά να σημαίνουν ότι οι απάτες γίνονται σπάνια αλλά αυτό είναι ένα σχετικό μέτρο. Εάν συνδυαστεί με το μεγάλο αριθμό συναλλαγών που πραγματοποιούνται, ώστε να υπολογίσουμε την απάτη σε απόλυτα νούμερα τότε εύκολα τα ποσά γίνονται πολύ μεγάλα. Μόνο στο παράδειγμα του ξεπλύματος χρημάτων, το OTA για το 1995 αναφέρει ότι οι καθημερινές συναλλαγές είναι περίπου 750.000 και αντιστοιχούν σε περίπου 2 τρισεκατομμύρια δολάρια. Ετησίως λοιπόν έχουμε περίπου 300 δισεκατομμύρια δολάρια να αφορούν την παγκόσμια επιχείρηση ξεπλύματος χρημάτων.

2.4 Μετρικές Απόδοσης

Γενικά, ο αντικειμενικός στόχος ενός συστήματος ανίχνευσης απάτης είναι η μεγιστοποίηση του αριθμού των σωστών προβλέψεων αλλά και η διατήρηση των εσφαλμένων προβλέψεων σε αποδεκτά επίπεδα. Τα περισσότερα συστήματα ανίχνευσης απάτης προσπαθούν να ποσοτικοποιήσουν την απάτη και να της αποδώσουν κάποια χρηματική αξία ή όφελος. Με την ελαχιστοποίηση τη πιθανότητας μη ανίχνευσης της απάτης ή της εσφαλμένης σήμανσης ως απάτη μπορεί να επιτευχθεί με αρκετά υψηλή πιθανότητα μία σωστή διάγνωση. Παραθέτουμε στη συνέχεια ορισμένους τεχνικούς όρους και κάποια θέματα που λαμβάνονται υπ' όψιν κατά την αξιολόγηση ενός συστήματος ανίχνευσης απάτης :

- Ρυθμός εσφαλμένων σημάνσεων απάτης (false alarm ή false positive rate) ονομάζουμε το ποσοστό των νόμιμων συναλλαγών που εσφαλμένα χαρακτηρίστηκαν ως παράνομες.
- Ρυθμός εσφαλμένων αρνητικών (false negative rate) το ποσοστό τω απατών που λανθασμένα χαρακτηρίστηκαν ως νόμιμες.
- Ρυθμός ανίχνευσης της απάτης (fraud catching rate, detection accuracy rate, true positive rate) ονομάζουμε το ποσοστό των παράνομων συναλλαγών που εντοπίστηκαν.

Στην ανίχνευση απάτης το κόστος εσφαλμένης ταξινόμησης δεν είναι το ίδιο για τις δύο περιπτώσεις και μπορεί να αλλάζει με το χρόνο. Συνήθως είναι πιο δαπανηρό να αναγνωρίσουμε μια παράνομη συναλλαγή ως νόμιμη από το αντίθετο (Phua et al., 2005). Σε ένα σύστημα ανίχνευσης απάτης είναι σημαντικό να ορίζονται με προσοχή οι μετρικές για τον έλεγχο της απόδοσής του (Kou et al., 2004). Ένα τυπικό σύστημα επιχειρεί να μεγιστοποιήσει το ρυθμό ανίχνευσης απάτης και να ελαχιστοποιήσει το ρυθμό εσφαλμένων σημάνσεων συναγερμού, λαμβάνοντας υπ' όψιν και το κόστος ελέγχου των συναλλαγών που κρίθηκαν ύποπτες. Επιπλέον οι μετρικές μπορούν να περιλαμβάνουν ανάλογα με την εφαρμογή και τις πολιτικές αναζήτησης μιας απάτης τα εξής: πόσο γρήγορα ανακαλύπτεται η απάτη, πόσα είδη

απάτης μπορεί να ανακαλύψει, το εάν η ανίχνευση γίνεται σε πραγματικό χρόνο ή όχι.

Το γεγονός ότι η απάτη είναι συνήθως σε χαμηλά ποσοστά δημιουργεί και επιπλέον προβλήματα κατανόησης σχετικά με το ποιο εργαλείο εντοπισμού είναι αποτελεσματικό που φαίνονται καλύτερα στο παρακάτω παράδειγμα.. Συμπερασματικά ενδεχομένως να έχουμε τη δυνατότητα να μειώσουμε την απάτη σε όποιο επίπεδο θέλουμε, εις βάρος της αύξησης των εξόδων που απαιτούνται για τη διαδικασία της ανίχνευσης όμως πρέπει επίσης να λάβουμε υπ'όψιν και την ενδεχόμενη δυσαρέσκεια που μπορεί να προκαλέσουμε είτε αυτή μεταφράζεται σε χρόνο είτε σε ποιότητα ή κόστος υπηρεσιών, επιλέγοντας μια συμβιβαστική λύση που ικανοποιεί τις ανάγκες της εφαρμογής και είναι σύμφωνη με την πολιτική προστασίας από τις απάτες που ακολουθείται.

2.5 Τεχνικές ανίχνευσης της απάτης

Ένα τυπικό σύστημα ανίχνευσης απάτης επεξεργάζεται κάθε νέα περίπτωση και επιστρέφει ένα αποτέλεσμα χαρακτηρίζοντάς την ως πλαστή ή νόμιμη. Αυτό το αποτέλεσμα που παίρνουμε ως έξοδο από το μηχανισμό ανίχνευσης μπορεί να δίνεται και μέσω ενός σκορ του πόσο ύποπτη για απάτη είναι μια περίπτωση ή μιας βαθμολογίας του κατά πόσο αυτή η περίπτωση μοιάζει με τις άλλες απάτες. Υπάρχουν διάφορες μέθοδοι, κάθε μια με τις αντίστοιχες τεχνικές και τους αλγορίθμους που τις υλοποιούν, τις οποίες μπορούμε να χρησιμοποιήσουμε για την ανίχνευση της απάτης. Μπορούμε να διακρίνουμε τις μεθόδους σε δύο μεγάλες κατηγορίες: τις εποπτευόμενες (supervised) και τις μη εποπτευόμενες (unsupervised) μεθόδους.

Οι *εποπτευόμενες μέθοδοι* εφαρμόζονται σε δεδομένα που έχουν προηγουμένως χαρακτηριστεί οι εγγραφές τους. Έχουμε δηλαδή ένα σύνολο δοκιμαστικό και η ανάλυση ξεκινά γνωρίζοντας ποιες εγγραφές αφορούν περιπτώσεις απάτης και ποιες όχι. Στις εποπτευόμενες μεθόδους χρησιμοποιούνται δείγματα των οποίων οι διαθέσιμες εγγραφές έχουν χαρακτηριστεί τόσο ως ύποπτες όσο και ως μη ύποπτες. Συνεπώς, αφού δημιουργηθεί ένα μοντέλο που χρησιμοποίει τα δεδομένα εκπαίδευσης, ελέγχονται και οι νέες εγγραφές διαχωρίζοντας τις σε ύποπτες ή νόμιμες εγγραφές. Φυσικά, κάποιος θα πρέπει να είναι σίγουρος για την σωστή ταξινόμηση των στοιχείων του μοντέλου εκπαίδευσης, γιατί αυτή είναι και η βάση του μοντέλου.

Για το λόγο αυτό, στην εφαρμογή, των εποπτευομένων μεθόδων υπάρχουν θέματα τα οποία αφορούν στη χρήση μιας εκ των προτέρων γνώσης που οφείλει να λάβει κανείς υπ' όψιν. Πρώτα απ' όλα η γνώση αυτή ενδέχεται να είναι αρκετά δαπανηρή και δύσκολο να αποκτηθεί (Brockett et al., 2002). Τα δεδομένα μιας τέτοιας ανάλυσης υπόκεινται σε κάποιο σχετικό σφάλμα, αφού ενδεχόμενα ορισμένες εγγραφές μπορεί να έχουν χαρακτηριστεί ως περιπτώσεις απάτης ενώ δεν είναι ή

ορισμένες περιπτώσεις απάτης να έχουν λανθασμένα καταγραφεί ως νόμιμες. Επίσης είναι δυνατόν ορισμένες περιπτώσεις απάτης να μην έχουν αποκαλυφθεί. Ορισμένα τέτοια ζητήματα σθεναρότητας των τεχνικών θίγονται από τους Provost και Fawcett (2001).

Αντίθετα, οι μη εποπτευόμενες μέθοδοι χρησιμοποιούνται στις περιπτώσεις όπου δεν υπάρχει προηγούμενη γνώση για τα στοιχεία των δεδομένων. Οι μέθοδοι δηλαδή αυτές δεν χρησιμοποιούν κάποιο χαρακτηριστικό για το διαχωρισμό των δεδομένων σε περιπτώσεις απάτης ή νόμιμης συμπεριφοράς. Τα δεδομένα εδώ συνδυάζονται χωρίς τη χρήση τέτοιων διαχωρισμών και δίνουν ως αποτέλεσμα ύποπτα σκορ. Οι τεχνικές που χρησιμοποιούνται προέρχονται από την ανακάλυψη παράτυπων σημείων. Συνήθως μοντελοποιούν τα δεδομένα και έπειτα προσπαθούν να εντοπίσουν αυτά που απέχουν περισσότερο από την "κανονική" συμπεριφορά.

Ένα σημαντικό θέμα για τις μη εποπτευόμενες μεθόδους στην περίπτωση ειδικά της ανίχνευσης απάτης όπου οι βάσεις δεδομένων συνεχώς εξελίσσονται είναι η μη ομογένεια που μπορεί να εμφανίζουν τα δεδομένα ως προς το χρόνο. Λόγου χάρη ένας τραπεζικός λογαριασμός δεν έχει την ίδια συμπεριφορά καθώς ο ιδιοκτήτης του λογαριασμού μεγαλώνει ηλικιακά. Συνήθως παρατηρείται βελτίωση της οικονομικής του κατάστασης και συχνά αλλάξει καταναλωτικές συνήθειες, πραγματοποιώντας αγορές που δεν έκανε στο παρελθόν ή μετακινώντας μεγαλύτερα ποσά. Ένας τέτοιος λογαριασμός που μοντελοποιείται συνολικά χωρίς να λαμβάνεται υπ' όψιν η επίδραση του χρόνου μπορεί να εμφανιστεί αδικαιολόγητα ως ύποπτος απάτης.

Στην ενότητα που ακολουθεί γίνεται μια επισκόπηση εποπτευόμενων και μη εποπτευόμενων τεχνικών που έχουν χρησιμοποιηθεί στην περίπτωση της ανίχνευσης της απάτης. Στην επισκόπηση αυτή λαμβάνονται υπ'όψιν μόνο τεχνικές από το πεδίο της εξόρυξης δεδομένων (data mining) και όχι παραδοσιακές τεχνικές ανάλυσης δεδομένων. Επιπλέον, είναι περιορισμένη στην αναφορά και μόνο των τεχνικών που έχουν χρησιμοποιηθεί, χωρίς να διευκρινίζονται οι υποθέσεις που έχει θεωρήσει ο εκάστοτε συγγραφέας. Εκτενέστερες πληροφορίες παρέχονται από τους Phua et al. (2005) και τους Bolton και Hand (2002). Αρχικά, θα γίνει αναφορά στις εποπτευόμενες μεθόδους και στη συνέχεια στις μη εποπτευόμενες μεθόδους.

2.6 Εφαρμογές εποπτευόμενων μεθόδων

Η χρήση των εποπτευόμενων μεθόδων στην ανίχνευση απάτης έχει γίνει αντικείμενο μελέτης σε αρκετές έρευνες. Μια αρκετά διαδεδομένη περίπτωση είναι αυτή των νευρωνικών δικτύων (neural networks). Οι μελέτες των Barson et al. (1996), Fanning και Cogger (1998) και των Green και Choi (1997) χρησιμοποιούν την τεχνολογία των νευρωνικών δικτύων για την αντιμετώπιση της απάτης σε δίκτυα κινητής τηλεφωνίας. Οι Lin et al. (2003) εφάρμοσαν ένα ασαφές νευρωνικό δίκτυο στον τομέα της απάτης χρηματοοικονομικών αναφορών. Ένα συνδυασμό νευρωνικών

δικτύων και κανόνων χρησιμοποιούν οι Brause et al. (1999) και οι Estevez et al. (2006). Ειδικότερα, οι Brause et al. χρησιμοποίησαν παραδοσιακούς κανόνες συσχέτησης (traditional association rules) ενώ οι Estevez et al. ασαφής κανόνες. Στην περίπτωση της ανίχνευσης απάτης σε πιστωτικές κάρτες, οι Maes et al.(2002) εφάρμοσαν ένα Bayesian εκπαιδευόμενο νευρωνικό δίκτυ. Το ίδιο μοντέλο εφάρμοσαν οι Ezawa και Norton (1996) σε απάτες στο χώρο των τηλεπικοινωνιών, ενώ οι Viaene et al. (2005) στο πεδίο της ασφάλισης αυτοκίνητων.

Οι Major and Riedinger (2002) πρότειναν ένα σύστημα βασισμένο στην υβριδική γνώση/στατιστική, (hybrid knowledge/statistical-based system), όπου η γνώση των εμπειρογνωμόνων συνδυάζεται με στατιστικές μεθόδους ως εργαλείο για την ανίχνευση απάτης στις ασφάλειες υγείας. Επιπλέον παράδειγμα για το συνδυασμό διαφορετικών τεχνικών αποτελεί η εργασία των Fawcett και Provost (1997).

Οι Yang και Hwang (2006) παρουσίασαν ένα άλλο πλαίσιο μελέτης για την ανίχνευση της άπατης στο χώρο της υγείας βασισμένο στην ιδέα των κλινικών μονοπατιών (*clinical pathways*), όπου αρχικά ανακαλύπτονται τα διάφορα πρότυπα συμπεριφοράς και στη συνέχεια αναλύονται.

Τα ασαφή έμπειρα συστήματα αποτελούν άλλο ένα κομμάτι τεχνικών το οποίο έχει δοκιμαστεί σε διάφορες μελέτες. Παράδειγμα τέτοιων ερευνών είναι αυτές των Pathak et al. (2003), Bordoni et al. (2001) και των Deshmukh και Talluru (1998). Οι Stolfo et al. (2001) και Lee και Xiang (2001) έδωσαν στη δημοσιότητα ένα πλαίσιο μελέτης γνωστό ως MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection).

Οι Cahill et al. (2000) σχεδίασαν ένα σύστημα, βασισμένο σε δεδομένα ύποπτων κλήσεων με σκοπό την ανίχνευση απάτης στις τηλεπικοινωνίες. Η ανάλυση κανόνων εκμάθησης (Rule-learning) και η ανάλυση των δέντρων απόφασης έχει, επίσης, εφαρμοστεί από διάφορους ερευνητές όπως οι Shao et al. (2002), ο Fan (2004), οι Bonchi et al. (1999) και οι Rosset et al. (1999).

Η ανάλυση σύνδεσης (link analysis) έχει ως στόχο την ανακάλυψη και ανάλυση των σχέσεων ανάμεσα στις μεταβλητές (Gentle, 2002) και μπορεί να χρησιμοποιηθεί με απώτερο σκοπό τη σύνδεση τους με πράξεις απάτης. Οι Cortes et al. (2001) αναφέρουν ότι απατεώνες συνδέονται μεταξύ τους. Στις τηλεπικοινωνίες λόγου χάρη οι απατεώνες τηλεφωνούν ο ένας τον άλλο ή τηλεφωνούν στους ίδιους αριθμούς. Οι Nigrini και Mittermaier (1997) και ο Nigrini (1999) χρησιμοποίησαν τον νόμο του Benford[1] (Hill, 1995) για την ανίχνευση απάτης σε λογιστικά δεδομένα.

[1] ότι η κατανομή των πρώτων σημαντικών ψηφίων θα έχει ασυμπτωτικά μια συγκεκριμένη μορφή

2.7 Εφαρμογές μη εποπτευόμενων μεθόδων

Η χρήση των μη εποπτευόμενων μεθόδων για την ανίχνευση απάτης δεν έχει διερευνηθεί τόσο συστηματικά όσο οι εποπτευόμενες μέθοδοι. Οι Bolton και Hand (2001) παρακολούθησαν τη συμπεριφορά κατά το πέρασμα του χρόνου μέσω της peer group ανάλυσης. Η peer group ανάλυση εντοπίζει μεμονωμένα στοιχεία τα οποία αρχίζουν να συμπεριφέρονται με ένα διαφορετικό τρόπο από τα στοιχεία που προηγουμένως είχαν παρόμοια συμπεριφορά. Ένα άλλο εργαλείο το οποίο ανέπτυξαν οι Bolton και Hand για την ανίχνευση απάτης είναι η break point analysis. Σε αντίθεση με την peer group analysis, η break point analysis λειτουργεί σχετικά με ένα επίπεδο λογαριασμού. Ένα σημείο καμπής είναι μια παρατήρηση όπου ανιχνεύεται η ύποπτη συμπεριφορά ενός συγκεκριμένου λογαριασμού. Τα δυο αυτά εργαλεία εφαρμόζονται για την αξιολόγηση των κινήσεων που γίνονται σε λογαριασμούς πιστωτικών καρτών.

Οι Murad και Pinkas (1999) εστίασαν την προσοχή τους στις αλλαγές της συμπεριφοράς για το σκοπό της ανίχνευσης απάτης και παρουσίασαν το προφίλ τριών επίπεδων (three-level-profiling). Όπως και η break point analysis των Bolton και Hand, η μέθοδος three-level-profiling λειτούργει σε ένα επίπεδο λογαριασμού και χαρακτηρίζει κάθε αξιόλογη απόκλιση από την κανονική συμπεριφορά ως ένδειξη απάτης. Για να συμβεί αυτό, δημιουργούνται «κανονικά» προφίλ, σε τρία επίπεδα, βασισμένα στα δεδομένα που δεν έχουν εμφανιστεί ενδεχόμενα απάτης. Στην περίπτωση αυτή, είναι προτιμότερο να χρησιμοποιηθεί ο όρος ήμι-εποπτευόμενη μέθοδος αντί για μη εποπτευόμενη. Η three-level-profiling εφαρμόστηκε στις άπατες στις τηλεπικοινωνίες. Το 2001 οι Burge και Shawe-Taylor χρησιμοποίησαν τα χαρακτηριστικά συμπεριφοράς με σκοπό την ανίχνευση απάτης στις τηλεπικοινωνίες. Ωστόσο, όταν γίνεται χρήση ενός επαναλαμβανόμενου νευρωνικού δικτύου (recurrent neural network), τότε εφαρμόζονται οι μη-εποπτευόμενες μέθοδοι. Για την κατασκευή των προφίλ λαμβάνονται υπόψη δυο χρονικά διαστήματα, οδηγώντας, έτσι, σε ένα τρέχον προφίλ συμπεριφοράς (current behavior profile, CBP) και σε ένα ιστορικό προφίλ συμπεριφοράς (behavior profile history, BPH) για κάθε περίπτωση. Στο επόμενο βήμα, γίνεται χρήση της απόστασης Hellinger για να συγκριθούν οι δυο κατανομές πιθανότητας και κάθε περίπτωση (τηλεφωνική κλήση) βαθμολογείται με ένα σκορ για το πόσο ύποπτη είναι.

Ένα σύντομο άρθρο των Cox et al. (1997) συνδυάζει τις ανθρώπινες ικανότητες αναγνώρισης προτύπων με τα αυτοματοποιημένα αλγοριθμικά συστήματα. Στην εργασία τους, οι πληροφορίες παρουσιάζονται οπτικά από διεπαφές συγκεκριμένων τομέων (domain-specific interfaces). Η ιδέα είναι ότι το ανθρώπινο οπτικό σύστημα είναι αρκετά δυναμικό και μπορεί εύκολα να προσαρμοστεί στις συνεχώς μεταβαλλόμενες τεχνικές που χρησιμοποιούνται από τους απατεώνες. Από την άλλη μεριά, οι μηχανές παρουσιάζουν το πλεονέκτημα της πολύ μεγαλύτερης υπολογιστικής ικανότητας, κατάλληλη για επαναλαμβανόμενες εργασίες ρουτίνας.

Επιπλέον, οι Yamanishi et al. (2004) χρησιμοποίησαν τον μη εποπτευόμενο αλγόριθμο SmartSifter με τη χρήση της απόστασης Hellinger για την ανακάλυψη παράτυπων σημείων σε ιατρικά δεδομένα, ενώ οι Brockett et al. (2002) πρότειναν την ανάλυση κύριων συνιστωσών σε ασφαλιστικά δεδομένα αυτοκινήτων.

Οι εργασίες των Bolton και Hand, Murad και Pinkas (1999), Burge και Shawe-Taylor (2001) και των Cox et al. (1997), συμπληρώνουν τη λίστα των σημαντικότερων εργασιών που αφορούν τις μη εποπτευόμενες μεθόδους στην ανίχνευση απάτης.

Παρά το γεγονός ότι η παραπάνω ανασκόπηση δεν είναι εξαντλητική, είναι σαφές ότι οι μη εποπτευόμενες μέθοδοι στην ανίχνευση απάτης συνεχώς κερδίζουν έδαφος.

2.8 Περιγραφή διαφόρων ειδών απάτης

Σε αυτή την ενότητα θα περιγράψουμε διάφορες χαρακτηριστικές μορφές απάτης. Η επιλογή των συγκεκριμένων μορφών έγινε με γνώμονα το βαθμό διάδοσης που αυτές εμφανίζουν και με προσπάθεια να καλυφθεί ένα μεγάλο μέρος του φάσματος στο οποίο συμβαίνουν οι απάτες. Τέλος αναφέρουμε επιγραμματικά ορισμένες ακόμη απάτες με στόχο την απόκτηση μιας όσο το δυνατόν πιο ολοκληρωμένης εικόνας, παραθέτοντας και μια ενδεικτική βιβλιογραφία.

2.8.1 Απάτη στις τηλεπικοινωνίες

Τα τελευταία χρόνια παρατηρείται άνθιση της βιομηχανίας των τηλεπικοινωνιών. Αυτή η γιγαντιαία εξάπλωση οφείλεται κυρίως στην πρόοδο της τεχνολογίας των κινητών τηλεφώνων που έχει επιτρέψει την παραγωγή τους με οικονομικά αποδεκτό κόστος. Η ανάπτυξη αυτή έχει μεγαλώσει τον αριθμό των χρηστών αλλά μαζί τους αυξήθηκε και η συχνότητα των περιπτώσεων στο πεδίο αυτό.

Τηλεπικοινωνιακή απάτη είναι οποιαδήποτε ενέργεια μέσω της οποίας κάποιος αποκτά πρόσβαση σε τηλεπικοινωνιακές υπηρεσίες για τις οποίες δεν έχει σκοπό να πληρώσει (Carberry, 2001). Βάσει αυτού του ορισμού η τηλεπικοινωνιακή απάτη περιλαμβάνει ενέργειες τόσο των εσωτερικών όσο και εξωτερικών χρηστών ενός δικτύου και μπορεί να ανιχνευθεί μόνον αφού έχει συμβεί. Επίσης, σήμερα, ο όρος τηλεπικοινωνίες είναι πιο ευρύς από ποτέ. Περιλαμβάνει συστήματα τόσο σταθερής όσο και κινητής πρόσβασης, συστήματα που βασίζονται σε παραδοσιακές τηλεπικοινωνιακές τεχνολογίες (PSTN, ISDN) αλλά και συστήματα που χρησιμοποιούν ως πλατφόρμα τους το διαδίκτυο. Η πληθώρα τεχνολογιών πρόσβασης και προσφερόμενων υπηρεσιών αλλά και ο ρυθμός εισόδου νέων

εταιρειών στην αγορά με αυξημένες ανάγκες προσέλκυσης πελατών, καθιστά τον εντοπισμό απάτης ακόμα πιο δύσκολο.

Αρκετά άρθρα δίνουν διαφορετικά ποσά ή ποσοστά που περιγράφουν την έκταση της απάτης. Εκτιμήσεις αναφέρουν ότι το 3% με 5% των εσόδων από τις τηλεπικοινωνίες παγκοσμίως χάνεται εξαιτίας της απάτης. Αυτό αναλογεί σε $55 δισεκατομμύρια δολάρια κάθε χρόνο. Μάλιστα οι νέες εταιρίες τηλεπικοινωνιών έχουν μεγαλύτερες απώλειες που σε ποσοστά μπορεί να ανέρχονται ακόμα και σε 20% των εσόδων τους (Cahill et al., 2002). Το εύρος των τιμών αυτών έγκειται στο γεγονός ότι αποτελούν εκτιμήσεις και έτσι δικαιολογείται μια μεταβλητότητα.

Η τηλεπικοινωνιακή απάτη έχει μερικά χαρακτηριστικά που την κάνουν ιδιαίτερα ελκυστική στους επίδοξους απατεώνες. Το κύριο χαρακτηριστικό είναι ότι ο κίνδυνος εντοπισμού είναι μικρός. Αυτό συμβαίνει επειδή όλες οι ενέργειες γίνονται από μακριά και η τοπολογία και το μέγεθος των δικτύων κάνουν τη διαδικασία εντοπισμού χρονοβόρα και ακριβή. Από την άλλη είτε δεν απαιτείται καθόλου εξοπλισμός ή αυτός είναι πολύ φτηνός. Για παράδειγμα, απλά η γνώση ενός κωδικού πρόσβασης, που μπορεί να αποκτηθεί και με μεθόδους κοινωνικής μηχανικής (social engineering), καθιστά εφικτή την εκτέλεση απάτης. Τέλος, το προϊόν της επίθεσης, δηλαδή ένα τηλεφώνημα, είναι άμεσα μετατρέψιμο σε χρήμα. Η παράνομη πώληση του παράνομα αποκτηθέντα χρόνου πρόσβασης σε τηλεπικοινωνιακά συστήματα είναι μεγάλη επιχείρηση (Forest, 1996).

Αναφέρονται τέσσερις κύριες υποκατηγορίες τηλεπικοινωνιακής απάτης, η τεχνική απάτη (technical fraud), η απάτη συμβολαίου (contractual fraud), η απάτη με μορφή πειρατείας (hacking fraud) και η διαδικαστική απάτη (procedural fraud) (Carberry, 2001).

Η *τεχνική απάτη* αφορά σε επιθέσεις σε αδύναμα σημεία ενός συστήματος. Απαιτεί ειδικές τεχνικές γνώσεις από τον επιτιθέμενο. Από τη στιγμή που ένα αδύναμο σημείο ενός συστήματος αποκαλυφθεί η γνώση διαχέεται γρήγορα και σε εκείνους που δεν έχουν την κατάλληλη τεχνική εμπειρία. Παράδειγμα τεχνικής απάτης είναι ο κλωνισμός των κινητών τηλεφώνων. Είναι χαρακτηριστικό παράδειγμα τεχνολογίας η οποία όταν βρίσκεται στα πρώτα στάδια ανάπτυξής της παρουσιάζει κενά ασφάλειας που μπορεί να χρησιμοποιήσει κάποιος παράνομα. Μια από τις μεθόδους ανίχνευσής της είναι ο έλεγχος για ταυτόχρονα τηλεφωνήματα ή για τηλεφωνήματα που γίνονται από περιοχές τόσο απομακρυσμένες ώστε να είναι αδύνατο να μεταφερθεί ο χρήστης από την μια στην άλλη μέσα στο χρονικό διάστημα που χωρίζει τις κλήσεις. Η μέθοδος είναι γνωστή ως παγίδα ταχύτητας (velocity trap).

Η *απάτη συμβολαίου ή συνδρομής* αναφέρεται σε εκείνες τις περιπτώσεις που κάποιος γίνεται συνδρομητής σε μια υπηρεσία, συνήθως με πλαστή ταυτότητα, για την οποία δεν προτίθεται να πληρώσει. Η ανίχνευση της απάτης μπορεί να γίνει μόνο μετά τη λήξη της περιόδου εξόφλησης του σχετικού λογαριασμού. Στην κατηγορία

αυτή εκτός από τις περιπτώσεις ατόμων που γίνονται συνδρομητές κινητών και δεν πληρώνουν ποτέ μέχρι τη διακοπή του συμβολαίου, ανήκουν και σύγχρονες μορφές απάτης που εκμεταλλεύονται τους οικονομικούς διακανονισμούς μεταξύ παρόχων καθώς και τις θεσμικές και διαδικαστικές διαφορές που ελέγχουν το καθεστώς αδειοδότησης από χώρα σε χώρα. Οι διακρατικές συμφωνίες καθορίζουν ποσοστά επί του κόστους των κλήσεων τα οποία αποδίδονται στον φορέα τερματισμού της κλήσης. Αν αυτός είναι ένας εικονικός φορέας σε μια χώρα με χαλαρό σύστημα αδειοδότησης μπορεί να δημιουργήσει κέρδος από τον τερματισμό εικονικών κλήσεων που προέρχονται από το εξωτερικό της. Η απάτη συμβολαίου είναι η πιο «μοντέρνα» και ταχύτατα εξελισσόμενη περίπτωση απάτης (Lane και Brodley, 1997)

Η *απάτη πειρατείας* αναφέρεται σε εκείνες τις περιπτώσεις όπου ο απατεώνας αποκτά πρόσβαση στο δίκτυο μιας εταιρείας και πουλά πόρους της εταιρείας σε τρίτους. Χαρακτηριστική περίπτωση είναι ο εντοπισμός των εξερχόμενων γραμμών ενός δικτύου και η παροχή πρόσβασης σε ακριβές κλήσεις μέσω αυτών σε τρίτους. Αντίστοιχη είναι η περίπτωση πώλησης υπολογιστικής ισχύος από τους υπολογιστές ενός οργανισμού. Μια συχνή περίπτωση μικρο-απάτης είναι η εξής. Τα ιδιωτικά δίκτυα εταιρειών και οργανισμών συνήθως επιτρέπουν στο προσωπικό τους τη δωρεάν πρόσβαση σε αστικά τηλεφωνήματα. Ο εργαζόμενος καλεί δύο φίλους ή συγγενείς και χρησιμοποιώντας την υπηρεσία της τηλε-διάσκεψης τους βάζει σε συνομιλία χρεώνοντας την εταιρεία. Μια άλλη περίπτωση είναι εκείνη που ο υπάλληλος χρησιμοποιεί την παροχή δωρεάν κλήσεων καθώς και τον υπολογιστή του γραφείου του ώστε να έχει δωρεάν πρόσβαση στο διαδίκτυο από το σπίτι του.

Η *διαδικαστική απάτη* σχετίζεται με επιθέσεις στις αδυναμίες των διαδικασιών της επιχείρησης με στόχο την πρόσβαση στο σύστημα. Έτσι, κάποιος μπορεί να παραποιήσει παραστατικά ενεργοποίησης λογαριασμών και να καρπωθεί την υπηρεσία κατά τη διάρκεια του χρόνου επεξεργασίας τους.

Στη βιβλιογραφική αναφορά των Moreau et al. (1997) παρουσιάζονται δώδεκα διαφορετικές περιπτώσεις τηλεπικοινωνιακής απάτης. Στο ίδιο άρθρο παρουσιάζεται και το νομικό καθεστώς που αφορά τόσο στην ίδια την απάτη όσο και στα συστήματα ανίχνευσής της. Η προστασία του απορρήτου των επικοινωνιών και της έρευνας καθώς και η προστασία της ιδιωτικότητας των χρηστών είναι τα σημαντικότερα στοιχεία που αναφέρονται. Η εμπειρία έχει δείξει πως υπάρχουν και άλλες παραλλαγές ή συνδυασμοί τους. Ένα παράδειγμα είναι όταν κάποιος γίνει συνδρομητής σε μια υπηρεσία χωρίς να σκοπεύει να πληρώσει (απάτη συμβολαίου) και μετά πουλάει την υπηρεσία σε τρίτους (απάτη πειρατείας).

Αντίθετα με ό,τι θα περίμενε κανείς, οι προπληρωμένες υπηρεσίες είναι επίσης ένας σημαντικός στόχος των απατεώνων. Σε πρώτη ματιά οι προπληρωμένες υπηρεσίες, π.χ. καρτοκινητά, εξασφαλίζουν στην εταιρεία πάροχο πως τα τέλη χρήσης της υπηρεσίας έχουν πληρωθεί πριν τη χρήση της. Η υπηρεσίες αυτές έχουν, όμως, ένα σημαντικό δέλεαρ για τον επίδοξο απατεώνα, την ανωνυμία του χρήστη.

Αν κάποιος μπορεί να εισβάλει στους μεταγωγείς του παροχέα και να επέμβει στις ρυθμίσεις, τότε έχει τη δυνατότητα να προσθέτει πιστωτικές μονάδες σε μια προπληρωμένη υπηρεσία, να εισάγει συγκεκριμένους προορισμούς κλήσης σε λίστες με αριθμούς πρώτης ανάγκης (που συνήθως δεν χρεώνονται), ή ακόμα να αντιστρέψει τη διαδικασία αφαίρεσης μονάδων που γίνεται μετά από κάθε κλήση και να κάνει την προπληρωμένη υπηρεσία να συμπεριφέρεται σαν υπηρεσία συμβολαίου.

Τα δεδομένα που παράγονται από τα τηλεπικοινωνιακά δίκτυα είναι τεράστια και μπορούν να είναι της τάξεως των gigabyte ανά μέρα, έτσι οι τεχνικές εξόρυξης δεδομένων είναι ιδιαίτερα δημοφιλείς. Ένας επιπλέον λόγος είναι ότι συναντούμε και εδώ ποικιλία στο είδος των δεδομένων: ημερομηνία και ώρα, διάρκεια κλήσης, είδος κλήσης, γεωγραφική προέλευση, γεωγραφικός προορισμός, συνόψεις που αφορούν σε χρεώσεις κτλ. Χαρακτηριστικό παράδειγμα αποτελεί η βάση δεδομένων της εταιρίας AT&T που περιέχει 350 εκατομμύρια προφίλ χρηστών και επεξεργάζεται 275 εκατομμύρια κλήσεις την εβδομάδα (Cortes και Pregibon, 1998).

Η Ένωση για τον Έλεγχο Τηλεπικοινωνιακής Απάτης (Communications Fraud Control Association - CFCA), με έδρα το Φοίνιξ των ΗΠΑ, ανακοίνωσε τα αποτελέσματα μιας παγκόσμιας έρευνας που διεξήγαγε το 2003. Σύμφωνα με αυτή, οι απώλειες των εταιρειών εξαιτίας περιπτώσεων απάτης εκτιμώνται στην περιοχή των 35 – 40 εκατομμυρίων δολαρίων ΗΠΑ. Επανάληψη της έρευνας το 2005 έδειξε αύξηση αυτών των μεγεθών της τάξης του 52%. Τα ποσά αντιστοιχούν στο 5% των εσόδων των εταιρειών. Το 85% των εταιρειών που συμμετείχαν ανέφεραν μεγέθυνση του προβλήματος, ενώ το 65% ανέφερε περιστατικά απάτης στις ίδιες. Τέλος, όπως αναφέρεται στην έρευνα, το 47,5% των περιπτώσεων απάτης αφορούσαν σε απάτες συμβολαίου, κλοπής ταυτότητας και συνέβησαν σε ιδιωτικά τηλεφωνικά κέντρα και συστήματα ταχυδρομείου φωνής (Adomavicious και Tuzhilin, 1999) Στην ίδια έρευνα η τηλεπικοινωνιακή απάτη συνδέεται στενά με τη χρηματοδότηση τρομοκρατικών οργανώσεων ανά τον κόσμο.

Τέλος, τα εργαλεία που χρησιμοποιούνται είναι η αναγνώριση παράτυπων σημείων και η εποπτευόμενη ταξινόμηση είτε μέσω κανόνων είτε μέσω δέντρων απόφασης. Τα συστήματα που βασίζονται σε κανόνες αναζητούν τηλεφωνήματα μεγάλης διάρκειας ή μεγάλου κόστους, επικαλυπτόμενα, ή αναζητούν τη χρήση του ίδιου τηλεφώνου σε διαφορετικά μέρη σε κοντινά χρονικά σημεία. Σε άλλες περιπτώσεις στατιστικές συνόψεις που ονομάζονται προφίλ συγκρίνονται με σταθερές που ορίζονται από τους ειδικούς ή από εφαρμογές μεθόδων εποπτευόμενης εκμάθησης. Επίσης τα νευρωνικά δίκτυα έχουν ευρεία χρήση και σε αυτή την περίπτωση με παραδείγματα αυτά των Taniguchi et al. (1998) και Hollmen (2000). Η ανάλυση συνδέσεων επίσης χρησιμοποιείται στην ανίχνευση απάτης και στις τηλεπικοινωνίες (Cortes et al., 2001). Ακόμα μέθοδοι οπτικοποίησης που έχουν αναπτυχθεί για την εξόρυξη δεδομένων χρησιμοποιούνται και στην ανίχνευση απάτης στις τηλεπικοινωνίες (Cox et al., 1997). Το 1999, οι Fawcett και Provost, χρησιμοποίησαν ένα συνδυασμό μεθόδων για να ανιχνεύσουν απάτη. Κατασκεύασαν

ένα σύστημα εκμάθησης κανόνων για να εντοπίσουν δείκτες απάτης μέσα από σχεσιακές βάσεις δεδομένων. Οι δείκτες χρησιμοποιήθηκαν για την κατασκευή προτύπων τα οποία τροφοδοτούνται σε ένα σύστημα που συνδυάζει πρότυπα συμπεριφοράς. Οι έξοδοι αυτού του συστήματος συνδυάζονται μεταξύ τους μέσω ενός εκπαιδευόμενου γραμμικού μοντέλου ώστε να παράγονται συναγερμοί. Στο ίδιο συνέδριο, όπου παρουσιάστηκε η προηγούμενη εργασία, ο Rosset και οι συνεργάτες του (1999) αναφέρουν ενθαρρυντικά αποτελέσματα από τη χρήση κανόνων που εξάγονται με μια παραλλαγή του αλγορίθμου C4.5.

2.8.2 Απάτες Πιστωτικών Καρτών

Μια από τις περισσότερο κοινές απάτες αφορά στη χρήση πιστωτικών καρτών, όπου βαθμός εξάπλωσης της δεν είναι δυνατόν να υπολογιστεί με ακρίβεια. Στη συνέχεια θα περιγράφει το γενικότερο πλαίσιο της συγκεκριμένης απάτης. Τα τραπεζικά ιδρύματα που εκδίδουν πιστωτικές κάρτες, είναι διστακτικά στην ανακοίνωση των εν λόγω μεγεθών έτσι δεν έχουμε σαφή εικόνα για την έκταση που έχει αυτή η μορφή απάτης. Ακόμη και οι σχετικές ανακοινώσεις δεν είναι απαραίτητο να περιέχουν τα πραγματικά νούμερα και αυτό διότι υπάρχει ο κίνδυνος η ανακοίνωση των μεγεθών να μειώσει το βαθμό εμπιστοσύνης των πελατών προς την τράπεζα. Αυτό μεταφράζεται σε απώλεια των συνολικών εσόδων της τράπεζας. Ενδεικτικές είναι οι τιμές και τα ποσοστά απάτης που προκύπτουν από προηγούμενες αναφορές. Η έγκυρη οικονομική εφημερίδα Financial Times το 2000 έγραψε ότι για κάθε £100 περίπου 13p είναι οι ζημίες από τις απάτες πιστωτικών καρτών. Η Association for Payment Clearing Services (APACS) του Λονδίνου ανακοίνωσε ότι οι συνολικές απώλειες από απάτες πιστωτικών καρτών ανέρχονταν: το έτος 1998 σε £135 εκατομμύρια, το 1999 σε £188, το 2000 σε £293 και το 2001 σε £374. Από τα στοιχεία της APACS γίνεται φανερό ότι η απάτη μέσω πιστωτικών καρτών είναι ένα φαινόμενο σε έξαρση τουλάχιστον στο Ηνωμένο Βασίλειο, χωρίς όμως να υπάρχουν λόγοι να αποκλείσουμε μια αντίστοιχη συμπεριφορά παγκοσμίως.

Οι περιπτώσεις απάτης στις πιστωτικές κάρτες, προκαλούνται εξαιτίας της απώλειας, κλοπής, πλαστογραφίας ή της μη παραλαβής των καρτών από το νόμιμο δικαιούχο. Για τον εντοπισμό τέτοιων μορφών απάτης έχουν χρησιμοποιηθεί η διακριτική ανάλυση (Richardson, 1997), και τα νευρωνικά δίκτυα πολλαπλών επιπέδων (Ghosh and Reilly, 1994).

Μία από τις ιδιαιτερότητες του προβλήματος εντοπισμού παράνομων συναλλαγών στην περίπτωση των πιστωτικών καρτών, έγκειται στην ανάγκη εντοπισμού των παράνομων συναλλαγών τη στιγμή ακριβώς που συμβαίνουν, ώστε να εμποδιστούν πριν ολοκληρωθεί η συναλλαγή. Το πλαίσιο, με βάση το οποίο θα διεκπεραιωθεί ο έλεγχος νομιμότητας της συναλλαγής, καθορίζεται από το μικρό διαθέσιμο χρόνο για το χαρακτηρισμό της συναλλαγής ως πιθανά μη νόμιμης, σε συνδυασμό με τον τεράστιο καθημερινό όγκο του συνόλου των συναλλαγών τις

οποίες καλείται να διαχειριστεί ένα χρηματοπιστωτικό ίδρυμα. Αναφέρονται μάλιστα μελέτες όπου η έρευνα εντοπίζει με επιτυχία παράνομες συναλλαγές χωρίς χρήση ιστορικών δεδομένων συναλλαγών του συγκεκριμένου πελάτη, αλλά μόνο με τη αξιολόγηση της τρέχουσας συναλλαγής, καθώς και των πιο πρόσφατων συναλλαγών (Doronsoro et al, 1997). Το γεγονός ότι ο αριθμός των μη νόμιμων συναλλαγών είναι εξαιρετικά μικρός σε σχέση με τις νόμιμες, κάνει τον εντοπισμό τους ιδιαίτερα δύσκολο.

Μελέτες δεδομένων από τη χρήση πιστωτικών καρτών, υποδεικνύουν ότι η απάτη στην περιοχή αυτή είναι δυνατό να εντοπιστεί με αξιόλογο βαθμό βεβαιότητας, εάν αντιμετωπισθεί με τεχνικές ταξινόμησης. Στην περίπτωση αυτή κάθε συναλλαγή ταξινομείται ως πιθανά νόμιμη ή μη νόμιμη. Επιπροσθέτως, λόγω της γοργής εξάπλωσης του ηλεκτρονικού εμπορίου διεθνώς, προτείνονται τεχνικές κλιμακωτών αλγορίθμων και παράλληλης επεξεργασίας, με στόχο την έγκαιρη ανάλυση του τεράστιου όγκου δεδομένων (Chan and Stolfo, 1993).

Επί πλέον υπάρχουν ετησίως απώλειες δισεκατομμυρίων δολαρίων για τις εταιρίες παροχής πιστωτικών καρτών, εξαιτίας πραγματικής αδυναμίας ορισμένων πελατών τους να εκπληρώσουν τις υποχρεώσεις τους, δηλαδή λόγω προσωπικής χρεοκοπίας. Όπως αναφέρεται (Donato et al, 1998), ο συνδυασμός διαφορετικών τεχνικών, όπως δέντρα απόφασης και νευρωνικά δίκτυα, είναι δυνατό να οδηγήσει στη δημιουργία ενός συστήματος υποστήριξης λήψης αποφάσεων για το έλεγχο των συναλλαγών των πελατών ενός παροχέα σε πραγματικό χρόνο.

Τέλος, υπάρχει μία κατηγορία απάτης στις πιστωτικές κάρτες, η οποία συχνά αναφέρεται και ως κατάχρηση. Στην περίπτωση αυτή ο χρήστης της πιστωτικής κάρτας προβαίνει σε αγορές με πίστωση, τις οποίες όμως δεν προτίθεται να εξοφλήσει. Σε ορισμένες περιπτώσεις πρόκειται για μια προσχεδιασμένη δραστηριότητα και αυτό προϊδεάζει για πιθανή αλλαγή συμπεριφοράς του χρήστη σε αντιπαραβολή με τη συνήθη συμπεριφορά του. Το ειδικό αυτό πρόβλημα απέχει στην αντιμετώπισή του από το ευρύτερο πρόβλημα απάτης με πιστωτικές κάρτες και έχει περισσότερα κοινά χαρακτηριστικά με το πρόβλημα αφερεγγυότητας πελατών τηλεπικοινωνιακών οργανισμών.

Για την ανακάλυψη της απάτης πιστωτικών καρτών μπορούν να χρησιμοποιηθούν μέθοδοι που αναζητούν αλλαγές στα μοτίβα συναλλαγών καθώς και μέθοδοι που αναζητούν συγκεκριμένα μοτίβα αγορών τα οποία θεωρούνται ύποπτα. Μοντελοποιώντας ατομικά την προηγούμενη συμπεριφορά κάθε πελάτη ή συγκρίνοντας την με μια αναμενόμενη συμπεριφορά ή ακόμη με την αναζήτηση ορισμένων συμπεριφορών που σχετίζονται με περιπτώσεις απάτης αλλά και μέσω των εποπτευομένων μοντέλων μπορούμε να ανακαλύψουμε την απάτη. Για παράδειγμα υπάρχουν κάτοχοι πιστωτικών καρτών που αγοράζουν ορισμένα προϊόντα (π.χ. βενζίνη) με μία κάρτα και κάνουν αγορές καταναλωτικών αγαθών (π.χ. supermarket) με διαφορετική κάρτα. Επίσης οι νέοι κάτοχοι κάρτας είναι γενικά πιο φειδωλοί στη

χρήση της. Οι Hand και Blunt (2001) αναφέρουν ότι οι αθροιστικές χρεώσεις μιας πιστωτικής κάρτας είναι αξιοσημείωτα γραμμικές έτσι ένα απότομο άλμα ή μια αλλαγή στην κλίση είναι ενδείξεις για μια πιθανή απάτη.

Η βιβλιογραφία που είναι διαθέσιμη πάνω στην ανίχνευση απάτης κάθε άλλο παρά εκτενής θα μπορούσε να χαρακτηριστεί. Άλλα ακόμη και η υπάρχουσα δεν φαίνεται να έχει στόχο την περιγραφή μεθόδων ανίχνευσης απάτης αλλά την περιγραφή νέων εργαλείων ανάλυσης με εφαρμογή πάνω στην ανίχνευση απάτης. Οι Bolton και Hand σχολιάζουν ότι εξαιτίας του γεγονότος ότι η ανάλυση παράτυπων σημείων βασίζεται πολύ στο περιβάλλον που γίνεται η εφαρμογή και δεν μπορεί εύκολα να γενικευτεί, στις εργασίες συναντούμε κυρίως εποπτευόμενες διαδικασίες (Bolton και Hand, 2002). Συγκεκριμένα οι κανόνες συσχέτισης και τα νευρωνικά δίκτυα όπως οι εργασίες των Ghosh και Reilly (1994) και Dorronsoro et al. (1997) έχουν προσελκύσει το ενδιαφέρον των περισσότερων επιστημόνων. Στην εργασία των Stolfo et al. (2000) είδαμε την κατασκευή ενός ρεαλιστικού μετα-ταξινομητή με βάση το κόστος.

2.8.3 Ανίχνευση εισβολών σε υπολογιστικά συστήματα

Ο στόχος της ανίχνευσης εισβολών (intrusion detection) σε υπολογιστικά συστήματα είναι η ανίχνευση μη εξουσιοδοτημένης πρόσβασης σε υπολογιστικά συστήματα. Οι προσεγγίσεις στο πρόβλημα κατατάσσονται σε δύο γενικές κατηγορίες μεθόδων, την ανίχνευση ανωμαλιών (anomaly detection) και την ανίχνευση κατάχρησης (misuse detection). Με την ανίχνευση ανωμαλιών διερευνάται η ύπαρξη αποκλίσεων από τη συνηθισμένη χρήση του συστήματος. Κατά την ανίχνευση κατάχρησης συγκρίνονται τα μοτίβα χρήσης του συστήματος με γνωστές τεχνικές υπονόμευσης της ασφάλειας υπολογιστικών συστημάτων (Kumar, 1995) Από τις πρώτες εργασίες στο αντικείμενο είναι εκείνες του Anderson (1980) και του Denning (1987). Στις εργασίες αυτές, που έθεσαν τα θεμέλια για τις έρευνες που ακολούθησαν, χρησιμοποιούνται οι εγγραφές ελέγχου και τα αρχεία ημερολογίου των υπολογιστικών συστημάτων καθώς και η κίνηση του δικτύου. Ο Lunt (1990) συνδυάζει τεχνικές που ανήκουν και στις δύο κύριες κατηγορίες στοχεύοντας στην αναίρεση των μειονεκτημάτων της κάθε μιας.

Αρκετές εργασίες χρησιμοποιούν νευρωνικά δίκτυα για την ανίχνευση εισβολών. Σε άλλες επιδιώκεται η ταξινόμηση της συμπεριφοράς σε κανονική ή σε εισβολή (Tan, 1995) ενώ σε άλλες επιδιώκεται η εκμάθηση της συμπεριφοράς των χρηστών ώστε να αναγνωρίζεται η νομιμότητα της (Ryan, Ling, και Miikkulainen, 1997). Στην εργασία των Lane και Brodley (1997) συγκρίνεται η τρέχουσα συμπεριφορά του χρήστη με την ιστορική του συμπεριφορά και προτείνονται τρόποι και μέτρα σύγκρισης. Η χρήση στατιστικών μοντέλων και πολυμεταβλητών ταξινομητών με εφαρμογή σε ασύρματα ad-hoc δίκτυα προτείνεται στην εργασία των Manikopoulos και Papavassiliou (2002) ενώ η ομαδοποίηση της συμπεριφοράς του

κάθε χρήστη προτείνεται στην εργασία των Oh και Lee (2003). Είναι χαρακτηριστικό ότι σε αυτή την προσέγγιση, όπως και σε προηγούμενες χρησιμοποιούνται πάρα πολλές, συγκεκριμένα 86, παράμετροι συμπεριφοράς.

Ένα θέμα που υποκινεί έντονη ερευνητική δραστηριότητα τα τελευταία χρόνια είναι η κατασκευή έξυπνων συστημάτων για τον εντοπισμό ενοχλητικών ηλεκτρονικών μηνυμάτων (spam). Οι κυριότερες τακτικές προσέγγισης του προβλήματος είναι με πιθανοτικά μοντέλα και νευρωνικά δίκτυα εμπρόσθετης τροφοδότησης (Yoshida et al., 2004). Η περιοχή έχει να επιδείξει και μια από τις πιο καινοτόμες προτάσεις για την αντιμετώπιση της απάτης η οποία χρησιμοποιεί στοιχεία από τη θεωρία παιγνίων (Dalvi et al., 2004).

Σε πρώτη ανάγνωση το πρόβλημα της ανίχνευσης εισβολών σε υπολογιστικά συστήματα φαίνεται να μοιάζει πολύ με το πρόβλημα της ανίχνευσης απάτης σε τηλεπικοινωνιακά δίκτυα. Παρόλα αυτά, όπως φαίνεται και από τα συμπεράσματα της εργασίας των Fawcett και Provost (1999) δεν πρέπει να θεωρείται δεδομένη η επιτυχής μεταφορά συστημάτων από τη μια περιοχή στην άλλη.

2.8.4 Το πρόβλημα της απάτης στον τομέα της υγείας

Στον τομέα της υγείας πολυάριθμα παραδείγματα εξαπάτησης των ασφαλιστικών φορέων από ιδιώτες και μη είναι διαθέσιμα. Το 1999 επιβεβαιώθηκε, κατόπιν ερευνών, ότι το οργανωμένο έγκλημα εμπλέκεται άμεσα πλέον στις απάτες στο χώρο της υγείας. Συνήθεις πρακτικές είναι π.χ. πλαστή υπερχρέωση δημόσιων και ιδιωτικών ασφαλιστικών φορέων από ιατρούς για υποτιθέμενα παρεχόμενες πολύπλοκες (up coding) και δαπανηρές ιατρικές διαδικασίες ή συσκευές και μηχανήματα (π.χ. μηχανήματα τεχνητής υποστήριξης), συνταγογραφήσεις επωνύμων φαρμάκων ενώ χρησιμοποιούνται φάρμακα «φασόν» κατά τη θεραπεία, χρέωση για υπηρεσίες που ποτέ δεν παρασχέθηκαν, πλαστή εικόνα της κατάστασης του ασθενούς, παραποίηση της ταυτότητας του ασθενούς για ασφαλιστική κάλυψη άλλου ασθενούς που δεν διαθέτει ασφαλιστική ικανότητα, σκόπιμη απόκρυψη ή παραποίηση πληροφοριών που αφορούν την κλινική εικόνα ή το ιστορικό του ασθενούς, ψευδή δήλωση ατυχήματος, ψευδή δήλωση στοιχείων του ασθενούς, παραποίηση του χρόνου παροχής ιατρικών υπηρεσιών, παραπομπή για μη αναγκαίες εξετάσεις, χωριστή χρέωση για υπηρεσίες που φυσιολογικά καλύπτονται από μία χρέωση, διπλή χρέωση για την ίδια παρεχόμενη υπηρεσία, χρήση κωδικού που δεν αντιστοιχεί (ανακόλουθη κωδικοποίηση) στην ιατρική πράξη που παρασχέθηκε, πλαστά έγγραφα για υποτιθέμενη ανάγκη διακομιδής του ασφαλισμένου από και προς το νοσοκομείο με ασθενοφόρο, κ.ά. Συνήθης στόχος απάτης είναι ακόμα και άτομα με ανίατες ή σε τελικό στάδιο ασθένειες, τα οποία, πέφτουν θύματα ιατρών που με μη συμβατικές μεθόδους θεραπείας εκθέτουν τη ζωή τους σε άμεσο κίνδυνο αφενός και αφετέρου υπερχρεώνουν τους ασφαλιστικούς τους φορείς για υπηρεσίες ανακόλουθες της κατάστασης των ασφαλισμένων.

Η σκόπιμη υποβολή ψευδών αιτημάτων αποζημιώσεων σε ιδιωτικούς ή δημόσιους ασφαλιστικούς φορείς τείνει να είναι αυξανόμενης συχνότητας φαινόμενο διεθνώς. Οι συνηθέστερες μορφές απάτης συνίστανται σε λανθασμένη αναφορά διάγνωσης ή ιατρικής πράξης (43%), με σκοπό τη μεγιστοποίηση των πληρωμών των αιτημάτων αποζημίωσης, ενώ το 34% αφορά τιμολόγηση παροχής ιατρικών υπηρεσιών που ποτέ δεν παρασχέθηκαν. Ανακριβή αιτήματα αποζημιώσεων που υποβάλλονται όχι σκόπιμα μπορεί επίσης να δημιουργήσουν προβλήματα και να οδηγήσουν σε ανάγκη λήψης νομικών μέτρων. Η επιτυχία των μέτρων κατά της απάτης έγκειται σε δύο στοιχεία:

• τους διαθέσιμους πόρους που προκύπτουν από κέρδη των ίδιων ασφαλιστικών φορέων από την καταπολέμηση της απάτης και

• τη μεγάλη προτεραιότητα που δίνεται στην ανίχνευση της απάτης πλέον από τους νομοθέτες και την κοινωνία γενικότερα.

Τέλος, για να αναδείξουμε το ποσό σημαντική είναι η ανίχνευση της άπατης και στο χώρο της υγείας θα ήταν αξιόλογη η αναφορά σε κάποια ποσοτικά στοιχεία. Η ευαισθητοποίηση του κοινού και η επαγρύπνηση μέσω ενημέρωσης είναι κάτι που ειδικά στον τομέα της υγείας πρέπει οπωσδήποτε να γίνει, διότι εκτός από την οικονομία, εμπλέκεται άμεσα και η ίδια η ανθρώπινη ζωή. Στις ΗΠΑ ο NHCAA είναι ο φορέας πρόληψης, ανίχνευσης, έρευνας και προσαγωγής εγκλήματος στον τομέα της υγείας, υποστηριζόμενος από τα υπουργεία υγείας και δικαιοσύνης. Στις ΗΠΑ μετά το 1996 η καταπολέμηση της απάτης στον τομέα της υγείας έγινε η δεύτερη προτεραιότητα νόμου μετά το βίαιο έγκλημα, ιδιαίτερα όταν έγινε αντιληπτό ότι οι απώλειες στο κράτος ανέρχονταν στο 3% των ετήσιων δαπανών για την υγεία. Το 1994 έγινε εθνικό ζήτημα λόγω της επίδρασής του στο εσωτερικό εμπόριο. Το 1999 περίπου το 40% των ασφαλιστικών φορέων είχαν ειδικές ερευνητικές μονάδες προς ανίχνευση της απάτης, ενώ το 2001 είχαν σχεδόν διπλασιαστεί. Για κάθε 1 εκατομμύριο ($) δολάρια επένδυσης σε προσωπικό απασχολούμενο στην ανίχνευση της απάτης, είχαν κέρδος περίπου 27 εκατομμύρια δολάρια $. Αν και αυτό το κέρδος συχνά φαίνεται να μειώνεται λόγω των νεώτερων και ολοένα περιπλοκότερων μορφών απάτης , ωστόσο εξακολουθεί να υφίσταται, αν και μικρότερο.

Σύμφωνα με την υπηρεσία καταμέτρησης απάτης του Αγγλικού συστήματος δημόσιας υγείας, οι δαπάνες της ΕΕ για τη δημόσια υγειονομική περίθαλψη υπολογίζονται στα 1 τρισεκατομμύριο euro ετησίως. Η απάτη αποτελεί σχεδόν το 3% των δημόσιων δαπανών υγειονομικής περίθαλψης στην Αγγλία και από 3% έως 10% των ετήσιων δαπανών (ιδιωτικών και δημόσιων) στις ΗΠΑ. Η εμπειρία της Αγγλίας έχει καταδείξει τον οικονομικό αντίκτυπο της καταπολέμησης της απάτης: από την ίδρυση της υπηρεσίας καταμέτρησης απάτης οι απώλειες λόγω απάτης μειώθηκαν κατά 45% περίπου.

2.8.5 Άλλα είδη απάτης

Ένα άλλο είδος απάτης που με τη ευρεία εξάπλωση και χρήση του διαδικτύου γίνεται ολοένα και πιο προσφιλές κυρίως σε φοιτητές είναι η λογοκλοπή (plagiarism). Λογοκλοπή είναι η χρησιμοποίηση του έργου ή μέρους του έργου κάποιου και η οικειοποίηση αυτού. Παρατηρείται λοιπόν το φαινόμενο φοιτητές να αντιγράφουν άρθρα ή εργασίες άλλων και να τα παρουσιάζουν ως δικά τους στα πλαίσια ενός μαθήματος ή μιας πτυχιακής εργασίας. Η εύκολη πρόσβαση και διαχείριση των πληροφοριακών πόρων σε ηλεκτρονική μορφή ενισχύει τα κρούσματα λογοκλοπής.

Το πρόβλημα απασχολεί από καιρό αρκετά πανεπιστήμια του εξωτερικού τα οποία έχουν αναπτύξει δικά τους συστήματα και βάσεις ώστε να ελέγχουν την τιμιότητα των εργασιών. Σε αυτά μπορεί να κανείς να εισάγει ένα κείμενο και το σύστημα συγκρίνοντάς το με τη δική του βάση από άρθρα (και προηγούμενες εργασίες) να δώσει ένα στατιστικό μέτρο της πρωτοτυπίας του κειμένου. Ενδεικτικά αναφέρουμε την ιστοσελίδα http://www.plagiarism.org στην οποία μπορεί κανείς να βρει μια περιγραφή του θέματος καθώς και συνδέσμους για αρκετά συστήματα που βοηθούν στην ανίχνευση αυτής της απάτης.

Η ανίχνευση ύποπτων ή μη κανονικών συναλλαγών που μπορεί να πραγματοποιηθεί μέσω τεχνικών ΕΔ και να βοηθήσει στην αντιμετώπιση άλλων παράνομων ενεργειών όπως το ξέπλυμα χρήματος (Senator et al., 1995) και διάφορες απάτες λογιστικής φύσεως όπως οι δηλώσεις τελωνείων (Shao et al., 2002), η φοροδιαφυγή (Bonchi et al., 1999) ή οι χρηματιστηριακές απάτες (Kirkos et al., 2007). Η ανάλυση συνδέσεων ή άλλες εποπτευόμενες μέθοδοι μπορούν να αναγνωρίσουν ύποπτα μοτίβα συναλλαγών (συνεχής καταθέσεις μικρών ποσών, συναλλαγές με ύποπτους κατόχους ή επιχειρήσεις) μέσα από τεράστιους όγκους δεδομένων (Senator et al., 1995). Η ανίχνευση απάτης μπορεί να βρει εφαρμογές ακόμη και στις πιο απίθανες περιοχές όπως για παράδειγμα στην αντιτρομοκρατία και την επιδημιολογία (Heino και Toivonen, 2003; Rath et al., 2003) αλλά και στον αθλητισμό όπου οι Robinson et al. (1995) και Smith (1997) μελέτησαν αν ορισμένοι εξαιρετικοί χρόνοι αθλητών ήταν εκτός αυτών που ήταν αναμενόμενοι.

ΚΕΦΑΛΑΙΟ 3

«ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ»

3.1 Εισαγωγή στην εξόρυξη δεδομένων

Στην σύγχρονη εποχή της πληροφορίας είναι ιδιαίτερα σημαντικό, αν όχι απαραίτητο, να υπάρχει ένα εργαλείο για την ανάλυση και την ερμηνεία της τεράστιας ποσότητας δεδομένων που είναι καταχωρημένα σε αρχεία, βάσεις δεδομένων και άλλα μέσα αποθήκευσης, με στόχο την εξαγωγή της γνώσης που θα βοηθήσει στην ουσιαστική και απρόσκοπτη διαδικασία λήψης αποφάσεων, σε μια σειρά από προβλήματα της καθημερινής ζωής. Η ύπαρξη ενός τέτοιου εργαλείου καθίσταται ακόμα πιο σημαντική αν ληφθεί υπόψη ότι ο όγκος των αποθηκευμένων δεδομένων διπλασιάζεται κάθε 3 χρόνια (Lyman και Hal, 2003). Ένα τέτοιο εργαλείο είναι η εξόρυξη γνώσης από δεδομένα (data mining). Η εξόρυξη γνώσης από δεδομένα ή απλούστερα εξόρυξη δεδομένων (ΕΔ), είναι μια νέα και δυναμική τεχνολογία. Ο τομέας της εξόρυξης δεδομένων (ΕΔ) αποτελεί αντικείμενο μελέτης από πολλούς ερευνητές και μηχανικούς ειδικά τα τελευταία χρόνια με τη ραγδαία αύξηση του όγκου της πληροφορίας. Η έρευνα σε αυτό τον τομέα έχει προχωρήσει αρκετά, έχουν γίνει σημαντικά βήματα και έχουν εξαχθεί πολλά συμπεράσματα.

Γενικά ο όρος ΕΔ αναφέρεται σε υψηλού επιπέδου εφαρμογές, μεθόδους και παρόμοια εργαλεία, που χρησιμοποιούνται για να παρουσιάσουν και να αναλύσουν δεδομένα σε πεδία λήψης αποφάσεων. Η βασική ιδέα πίσω από τον όρο εξόρυξη δεδομένων είναι η ανεύρεση εκείνης της μη προφανούς λύσης η οποία δίνει την δυνατότητα εξαγωγής χρήσιμων και ουσιαστικών συμπερασμάτων. Η όλη διαδικασία βασίζεται στη χρησιμοποίηση αλγορίθμων οι οποίοι αναζητούν κανόνες μεταξύ των μεταβλητών των δεδομένων, και έπειτα καταχωρούν τα δεδομένα σε νέες βάσεις δεδομένων. Οι αλγόριθμοι αυτοί είναι τα συστατικά της διαδικασίας η οποία βρίσκει συσχετισμούς ή κανόνες μέσα από τεράστιες βάσεις αποθηκευμένων δεδομένων / πληροφοριών.

3.2 Ορισμός της Εξόρυξης Δεδομένων

Υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων (ΕΔ). Ωστόσο, ο ορισμός των Hand et al. (2001) θεωρείται αξιόλογος. Σύμφωνα, λοιπόν, με αυτόν τον ορισμό Εξόρυξη Δεδομένων ονομάζεται :

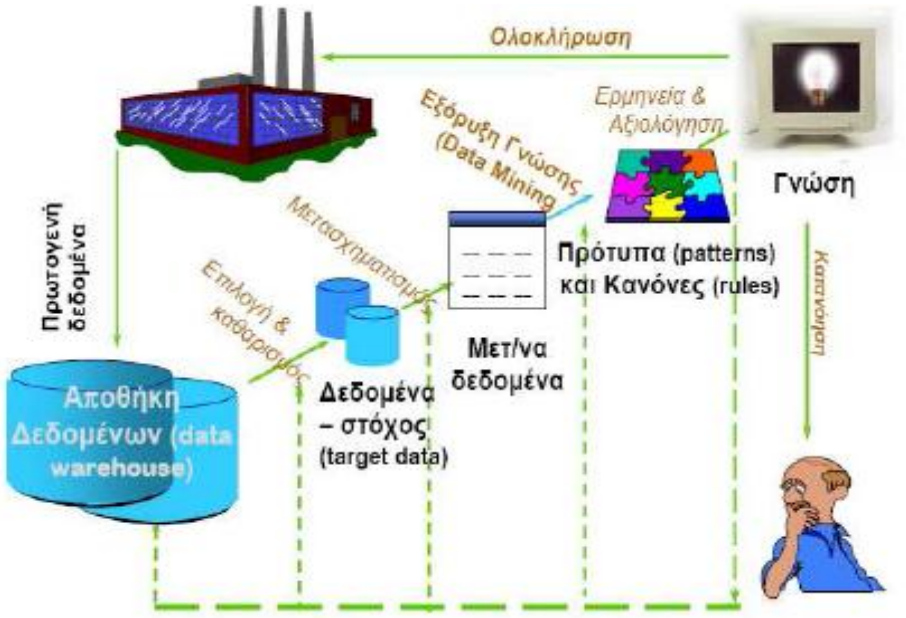
"Η ανάλυση (συχνά μεγάλων) παρατηρούμενων συνόλων δεδομένων για την αναζήτηση ανύποπτων σχέσεων και η σύνοψη των δεδομένων με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στον ιδιοκτήτη των δεδομένων".

Η δήλωση των σχέσεων και η σύνοψη των στοιχείων στην οποία αναφέρεται ο ορισμός αυτός, συχνά αναφέρεται ως μοντέλο ή πρότυπο. Βασικοί στόχοι της ΕΔ είναι η περιγραφή και η πρόβλεψη. Δηλαδή, η αναγνώριση των προτύπων που επικρατούν σε ένα μεγάλο σύνολο δεδομένων και η δημιουργία προβλέψεων όσον αφορά τη μελλοντική αξία ή συμπεριφορά κάποιων μεταβλητών. Η αναγνώριση των προτύπων γίνεται μέσω γραμμικών εξισώσεων, κανόνων, διάκρισης σε κατηγορίες, απόδοσης γραφημάτων και δενδρικών δομών, καθώς και επαναλαμβανόμενων προτύπων σε μορφή χρονοσειρών.

Θεμελιώδες στοιχείο του ορισμού της ΕΔ είναι η δομή που τελικά αναζητούμε να είναι χρήσιμη και να περιγράφει το φαινόμενο με πρωτότυπο τρόπο. Η περίπτωση της ανάλυσης μιας βάσης δεδομένων και η εύρεση της ήδη υπάρχουσας γνώσης δεν αποτελεί στόχο των αναλυτών. Παρότι πρόκειται για ένα σημαντικό θέμα στην ΕΔ, αλλά λίγοι αλγόριθμοι λαμβάνουν υπ' όψιν τους την εκ των προτέρων γνώση του χρήστη για τα δεδομένα (Benoit, 2002).

Ένα σημαντικό σημείο, στο οποίο πρέπει να σταθούμε, είναι ότι ο παραπάνω ορισμός αναφέρεται σε παρατηρούμενα (observable) δεδομένα και όχι σε εμπειρικά ή πειραματικά (experimental). Τα δεδομένα σε μια εφαρμογή ΕΔ προέρχονται από την απλή καταγραφή ιδιοτήτων και όχι από την προσεκτική επιλογή τους μέσω ενός πειράματος. Συχνά ο στόχος αυτής της συλλογής των δεδομένων είναι άλλος από αυτόν της ανάλυσής τους. Για το λόγο αυτό, η ΕΔ αναφέρεται συχνά ως «δευτερογενής» ανάλυση δεδομένων, αφού στην ουσία ασχολείται με δεδομένα που έχουν ήδη συλλεχθεί.

Εξάλλου, ο ουσιαστικός σκοπός της ΕΔ δεν είναι η ανάπτυξη στρατηγικής ως προς τη συλλογή δεδομένων και τη δημιουργία μοντέλων, έτσι όπως γίνεται στη στατιστική.



Σχήμα 3.1 :Σχηματική αναπαράσταση του ορισμού της ΕΔ

3.3 Ιστορική εξέλιξη ΕΔ

Στη δεκαετία του 1950 η δουλειά των μαθηματικών και των επιστημόνων των υπολογιστών συνδυάστηκε για να δημιουργήσουν τον κλάδο της τεχνητής νοημοσύνης και της μηχανικής μάθησης (Buchanan, 2006).

Κατά τη δεκαετία του 1960, η τεχνητή νοημοσύνη και ο κλάδος της στατιστικής εφάρμοσαν νέους αλγορίθμους όπως η ανάλυση παλινδρόμησης, ο υπολογισμός της μέγιστης πιθανότητας, τα νευρωνικά δίκτυα καθώς και τα γραμμικά μοντέλα κατηγοριοποίησης (Dunham, 2003). Στην διάρκεια της ίδιας επίσης δεκαετίας, έγινε για πρώτη φορά, αναφορά στον όρο εξόρυξη δεδομένων (ΕΔ), αλλά χρησιμοποιήθηκε κυρίως υποτιμητικά για να περιγράψει την πρακτική επεξεργασία των δεδομένων με σκοπό την αναζήτηση προτύπων, η οποία δεν είχε στατιστική σημαντικότητα (Fayyad et al., 1996).

Επίσης, στα έτη της δεκαετίας του 1960, το πεδίο της ανάκτησης της πληροφορίας συνείσφερε στις τεχνικές κατηγοριοποίησης και σε παρόμοια μέτρα. Αυτήν την περίοδο οι τεχνικές αυτές εφαρμόστηκαν σε κείμενα αλλά

χρησιμοποιήθηκαν αργότερα με την εξόρυξη δεδομένων σε βάσεις δεδομένων και σε άλλες μεγάλες πήγες δεδομένων (Dunham, 2003). Τα συστήματα βάσεων δεδομένων εστιάζουν στα ερωτήματα και στην επεξεργασία δοσοληψίας δομημένων δεδομένων, όπου η ανάκτηση της πληροφορίας σχετίζεται με τον οργανισμό και την ανάκτηση πληροφορίας από μεγάλο αριθμό εγγραφών κειμένου (Han και Kamber, 2001). Στα τέλη της δεκαετίας του 1960, η ανάκτηση πληροφορίας και τα συστήματα βάσεων δεδομένων εφαρμόζονταν παράλληλα.

Το 1971, ο Gerard Salton δημοσιοποίησε την καινοτόμα δουλειά του για τα συστήματα ανάκτησης πληροφοριών διαμορφώνοντας μια νέα προσέγγιση στην ανάκτηση πληροφορίας η οποία χρησιμοποιούσε το μοντέλο αλγεβρικού διανύσματος (VSM). Το μοντέλο θα αποδειχθεί ότι είναι το κλειδί για την εξόρυξη δεδομένων (Dunham, 2003).

Στις αρχές της δεκαετίας του 1990, ο όρος «ανακάλυψη γνώσης σε βάσεις δεδομένων» (Knowlekdge Discovery in Databases, KDD) χρησιμοποιήθηκε για πρώτη φορά και δημιουργήθηκε το πρώτο εργαστήριο KDD (Fayyad, 1996). Το μεγάλο μέγεθος των διαθέσιμων δεδομένων δημιούργησε την ανάγκη για νέες τεχνικές για τη διαχείριση αυτής της πληροφοριας που ήταν κρυμμένη σε τεράστιες βάσεις δεδομένων. Παράλληλα, με την ανάπτυξη των αποθηκών δεδομένων βρεθήκαν στο προσκήνιο οι βάσεις δεδομένων OLAP, τα συστήματα στήριξης αποφάσεων, ο μετασχηματισμός δεδομένων και οι αλγόριθμοι κανόνων συσχέτισης (Dunham, 2003, Han και Kamber, 2001).

Κατά την δεκαετία 1990, η ΕΔ εξελίχτηκε από μια ενδιαφέρουσα τεχνολογία σε μια πρότυπη πρακτική από τις εταιρείες. Αυτό επετεύχθη επειδή μειώθηκε το κόστος της αποθήκευσης των δεδομένων, το κόστος επεξεργασίας ανέβηκε και τα πλεονεκτήματα της ΕΔ έγιναν πιο εμφανή. Οι επιχειρήσεις ξεκίνησαν να χρησιμοποιούν την ΕΔ για να μπορούν να διευθύνουν όλες τις φάσεις του κύκλου ζωής πελάτη, συμπεριλαμβανόμενου την προσπάθεια απόκτησης νέων πελατών, την αύξηση εσόδων από τους υπάρχοντες πελάτες και τη διατήρηση καλών πελατών (Two Crows, 1999).

Στη συνέχεια ακολουθεί ένας πίνακας όπου παρουσιάζονται συνοπτικά τα βήματα της εξέλιξης της ΕΔ.

Βήματα Εξέλιξης	Υποβοηθητική Τεχνολογία	Χαρακτηριστικά
Συλλογή Δεδομένων (Data Collection), ‘60s	Υπολογιστές, Δίσκοι, Κασέτες	Επεξεργασία ανακεφαλαιωτικών και στατικών δεδομένων
Διαχείριση Δεδομένων (Data Management), ‘70s	Βάσεις δεδομένων και σχεσιακές βάσεις δεδομένων	Δημιουργία σχεσιακού σχήματος
Πρόσβαση Δεδομένων (Data Access), ‘80s	Σχεσιακές βάσεις δεδομένων, SQL	Επεξεργασία ανακεφαλαιωτικών και δυναμικών δεδομένων σε επίπεδο εγγραφών
Αποθήκες Δεδομένων (Data Warehousing), ‘90s	OLAP, Αποθήκες δεδομένων, πολυδιάστατες βάσεις δεδομένων	Επεξεργασία ανακεφαλαιωτικών και δυναμικών δεδομένων σε πολλαπλά επίπεδα
Εξόρυξη Δεδομένων (Data Mining), σήμερα	Εξελιγμένοι αλγόριθμοι, πολύ-επεξεργαστικά συστήματα, μαζικές βάσεις δεδομένων	Επεξεργασία ανακεφαλαιωτικών δυναμικών δεδομένων σε επίπεδο εγγραφών

Πίνακας 3.1 : Τα βήματα της εξέλιξης της ΕΔ (Πηγή : An Introduction to Data Mining, Pilot Software, 1998)

3.4 Τα προβλήματα που οδήγησαν στην ΕΔ

Είναι γεγονός πως η υπολογιστική ισχύς των υπολογιστών διπλασιάζεται κάθε 18 μήνες. Επίσης η χωρητικότητα δεδομένων διπλασιάζεται κάθε 12 εβδομάδες (Porter, 1998). Το αποτέλεσμα είναι μια διαφορά στις δύο τάσεις η οποία αυξάνεται εκθετικά και καλείται το κενό ή χάσμα δεδομένων (data gap) ή ο νόμος της αποθήκευσης (storage law) (Fayyad, 2002). Είναι γεγονός πως το κενό μεταξύ της απόδοσης του υλικού και της ποσότητας των δεδομένων που θέλουμε να επεξεργαστούμε είναι ένα σημαντικό πρόβλημα.

Οι τυπικοί αλγόριθμοι που διαχειρίζονται πολύ λιγότερα δεδομένα αντιμετωπίζουν προβλήματα από τη στιγμή που το υλικό δεν μπορεί να καλύψει το κενό από τον όγκο των δεδομένων. Για παράδειγμα ένας αλγόριθμος ταξινόμησης που λειτουργεί ορθά με λίγα megabytes δεδομένων θα μπορούσε να έχει προβλήματα απόδοσης αν εφαρμοστεί σε gigabytes δεδομένων. Στην πραγματικότητα υπάρχουν συγκεκριμένα προβλήματα με τους κλασικούς αλγορίθμους τα οποία θα παρουσιαστούν συνοπτικά παρακάτω.

Το βασικό πρόβλημα των κλασικών αλγορίθμων είναι ο χρόνος εκτέλεσης. Ωστόσο, μπορεί να υπάρχουν άλλοι λόγοι που έχουν ως αποτέλεσμα μεγάλους χρόνους εκτέλεσης.

Από τη στιγμή που μιλάμε για μεγάλο όγκο δεδομένων δεν μπορούμε να θεωρήσουμε πως θα χωρέσει ολόκληρος στη μνήμη RAM του υπολογιστή αλλά ούτε μπορούμε να θεωρούμε πως έχουμε διαθέσιμο άπειρο χώρο βοηθητικής μνήμης. Ο αλγόριθμος είναι πιθανό να μην λειτουργήσει ορθά για μεγέθη δεδομένων μεγαλύτερα από το μέγεθος της μνήμης RAM, εάν υποθέσουμε ότι ο αλγόριθμος που έχει σχεδιαστεί θα τοποθετεί όλα τα δεδομένα στην μνήμη RAM έτσι ώστε να είναι γρήγορη και άμεση η προσπέλαση των δεδομένων. Ένα επιπλέον μειονέκτημα είναι η έλλειψη του απαιτούμενου χώρου αποθήκευσης στη βοηθητική μνήμη έτσι ώστε να μπορούμε να έχουμε δύο ή τρία αντίγραφα των δεδομένων μας Συμπερασματικά, χρειαζόμαστε αλγόριθμους που λαμβάνουν υπ’ όψιν όλες αυτές τις παραμέτρους.

Επιπρόσθετα, τη δεκαετία του ‘90 πρόεκυψε άλλο ένα σημαντικό πρόβλημα καθώς τα δεδομένα που προέρχονταν από εταιρίες ή επιστημονικούς οργανισμούς αυξάνονταν πολύ γρήγορα. Νέες βάσεις δεδομένων και συστήματα αποθήκευσης, σε συνεργασία με υπερσύγχρονα συστήματα συλλογής δεδομένων ξεκίνησαν να συγκεντρώνουν όλο και πιο πολλά δεδομένα μέρα με τη μέρα. Γι’ αυτό το λόγο χρειάστηκε ένα μοντέλο που να μπορεί να σε οδηγήσει από τα απλά δεδομένα στη χρήσιμη πληροφορία. Το μοντέλο αυτό έπρεπε να τηρεί τους ακόλουθους κανόνες:

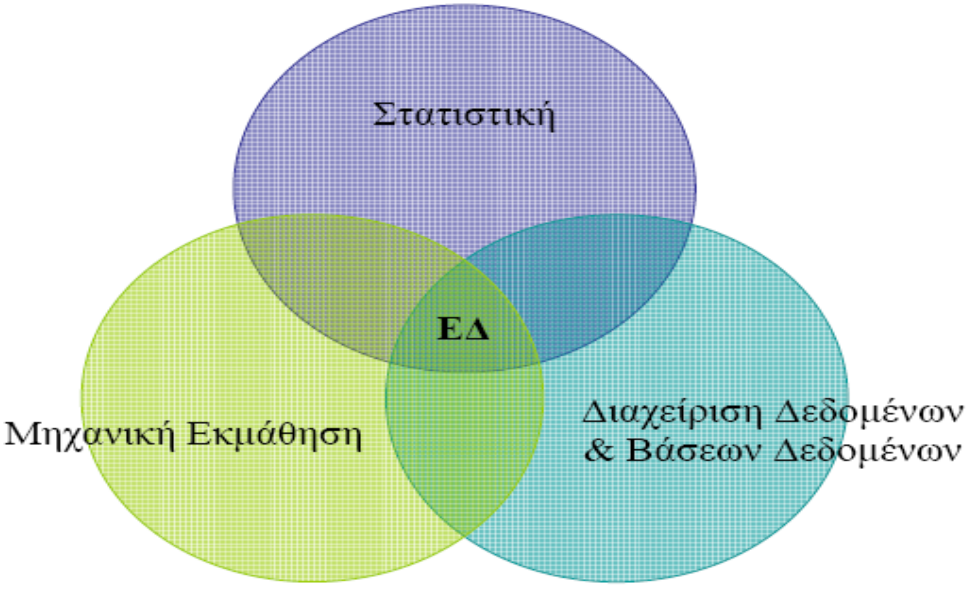
- Να μπορεί να διαχειριστεί μεγάλο όγκο πληροφορίας.
- Να κατέχει μηχανισμούς ώστε να εξάγει χρήσιμη πληροφορία από αυτά τα δεδομένα.

Στην περίπτωση που το χρησιμοποιούμενο μοντέλο ικανοποιεί αυτούς τους δυο κανόνες, τότε είναι σε θέση να ανακαλύψει την χρήσιμη πληροφορία από τα διαθέσιμα δεδομένα που πρόκειται να επεξεργαστούμε.

Η ανάγκη επίλυσης των προβλημάτων που αναφέρθηκαν παραπάνω οδήγησαν την επιστημονική κοινότητα στην ανάγκη εύρεσης ενός νέου τομέα ικανού να ανταπεξέλθει τα προβλήματα, οδηγώντας τους έτσι στην εξόρυξη δεδομένων.

3.5 Εξόρυξη δεδομένων και στατιστική

Παρόλο που με μια πρώτη ματιά η ΕΔ και η στατιστική φαίνεται να ενδιαφέρονται και οι δυο για την ανακάλυψη δομών και την εξαγωγή πληροφορίας από τα δεδομένα και κατά συνέπεια να έχουν κοινό στόχο (Glymour et al., 1997), υπάρχει μια σημαντική διαφορά μεταξύ τους (Hand et al., 2001). Η διαφοροποίηση της ΕΔ έγκειται στο ότι χρησιμοποιεί εκτός από τη στατιστική, ιδέες, εργαλεία και μεθόδους από τις βάσεις δεδομένων και από τη μηχανική εκμάθηση (βλέπε Σχήμα 3.2).



Σχήμα 3.2 : Γραφική απεικόνιση της σχέσης της ΕΔ με τη στατιστική

3.6 Διαχωρισμός των μεθόδων εξόρυξης δεδομένων

Οι μέθοδοι της ΕΔ μπορούν να χωριστούν σε τέσσερα βασικά μέρη. Αρχικά, όμως, πρέπει να αναφέρουμε ότι υπάρχουν δύο βασικοί τύποι ΕΔ: η επαλήθευση και η ανακάλυψη. Στον πρώτο τύπο, γίνεται επαλήθευση των υποθέσεων του χρήστη από

το σύστημα ενώ, στο δεύτερο τύπο, το σύστημα βρίσκει νέους κανόνες και πρότυπα μέσα από αυτόνομες διαδικασίες.

Οι μέθοδοι επαλήθευσης ασχολούνται με την εκτίμηση μιας υπόθεσης που προτείνεται από μια εξωτερική πηγή. Αυτές οι μέθοδοι περιλαμβάνουν περισσότερο παραδοσιακές μεθόδους από τη στατιστική (όπως έλεγχος καλής προσαρμογής, έλεγχος υποθέσεων, ανάλυση διακύμανσης) και σχετίζονται λιγότερο με την ΕΔ απ'ότι οι μέθοδοι ανακάλυψης, οι οποίες εντοπίζουν αυτόματα πρότυπα στα δεδομένα. Το στάδιο της ανακάλυψης χωρίζεται στην περιγραφή και την πρόβλεψη (Maimon and Rokach, 2005) που είναι και οι δύο στόχοι της ΕΔ, όπως αναφέρθηκε και στην εισαγωγή.

Τα τέσσερα μέρη ή ομάδες εργασιών της ΕΔ (data mining tasks), σύμφωνα με τους Hand et al. (2001) και τους Tan et al. (2005), είναι **η περιγραφική μοντελοποίηση** (descriptive modeling), **η προβλεπτική** (predictive modeling), **η ανάλυση συνάφειας** (association analysis) και **η ανίχνευση ανωμαλιών** (anomaly detection).

Ο στόχος ενός **μοντέλου περιγραφής** είναι να γίνει περιγραφή όλου του συνόλου δεδομένων ή της διαδικασίας που παράγει τα δεδομένα. Η σημαντικότερη εφαρμογή των περιγραφικών μοντέλων είναι η ομαδοποίηση, η οποία επιχειρεί να βρει ομάδες παρατηρήσεων που είναι κοντά μεταξύ τους ως προς τα χαρακτηριστικά που περιλαμβάνουν. Οι μέθοδοι περιγραφής και ειδικά η ομαδοποίηση είναι πολύ χρήσιμες σε πελατοκεντρικά συστήματα που βασίζονται στο CRM (Customer Relationship Management), καθώς έτσι μπορούν να εντοπιστούν ομάδες πελατών που αναμένεται να έχουν όμοια συμπεριφορά.

Η κατασκευή ενός **μοντέλου πρόβλεψης** στοχεύει στη δυνατότητα πρόγνωσης της τιμής μιας μεταβλητής (απόκριση) μέσα από τις τιμές άλλων μεταβλητών (επεξηγηματικές) που είναι γνωστές. Εάν η μεταβλητή απόκρισης είναι (ή μπορεί να θεωρηθεί) κατηγορική, τότε είμαστε σε θέση να εφαρμόσουμε μια μέθοδο ταξινόμησης. Όμως, αν έχουμε συνεχή απόκριση, τότε προχωράμε σε παλινδρόμηση. Επιθυμώντας να διευκρινήσουμε τη διαφορά μεταξύ ταξινόμησης και κατηγορίες, πρέπει να σημειώσουμε ότι στην ταξινόμηση επιχειρείται η περιγραφή μιας λειτουργίας που αντιστοιχεί (ταξινομεί) ένα στοιχείο σε μια εκ των κατηγοριών οι οποίες είναι ήδη προκαθορισμένες.

Η **ανάλυση συνάφειας** χρησιμοποιείται ώστε να ανακαλυφθούν πρότυπα που περιγράφουν χαρακτηριστικά συνάφειας μεταξύ των δεδομένων. Τα πρότυπα αυτά απεικονίζονται συνήθως στα πλαίσια κανόνων συνεπαγωγής (implication rules) ή υποομάδων των χαρακτηριστικών. Η χαρακτηριστικότερη εφαρμογή και η αιτία από την οποία ξεκίνησαν οι κανόνες συνάφειας (ή κανόνες «if – then») είναι η «ανάλυση του καλαθιού αγοράς» (market basket analysis). Άλλες εφαρμογές πραγματοποιούνται στην προώθηση προϊόντων ή στην τοποθέτησή τους στα ράφια καταστημάτων, στη διαχείριση αποθεμάτων κ.λπ. Στους κανόνες συνάφειας δίνεται και εκτίμηση για το πόσο πιθανό να συμβεί αυτή η σχέση αιτίας – αποτελέσματος,

Τέλος, στην **ανίχνευση ανωμαλιών** ανήκουν εργασίες εντοπισμού παρατηρήσεων των οποίων τα χαρακτηριστικά διαφέρουν σημαντικά από αυτά του υπόλοιπου συνόλου δεδομένων (έκτροπες παρατηρήσεις ή outliers). Στόχος είναι η υψηλού επιπέδου ανίχνευση πιθανών ανωμαλιών, διατηρώντας όμως χαμηλά ποσοστά λανθασμένης προειδοποίησης. Ως εφαρμογή μπορούμε να αναφέρουμε τον προσδιορισμό απειλής στην έγκριση δανείων ή πιστωτικών καρτών από μια τράπεζα.

3.7 Αλγόριθμοι εξόρυξης δεδομένων

Οι αλγόριθμοι εξόρυξης δεδομένων είναι πολλοί και σε αυτή την ενότητα θα παρουσιαστούν σε κατηγορίες οι πιο σημαντικοί από αυτούς. Οι κατηγορίες στις οποίες θα τους συναντήσουμε είναι οι παρακάτω:

- Κατηγοριοποίηση
- Ομαδοποίηση
- Κανόνες συσχέτισης
- Πρότυπα ακολουθιών
- Παλινδρόμηση
- Δέντρα απόφασης

Οι παραπάνω κατηγορίες χωρίς αμφιβολία αναπαριστούν όλη τη περιοχή των αλγορίθμων που χρησιμοποιούνται στον τομέα αυτό. Τα τελευταία χρόνια η ερευνητική κοινότητα δίνει ιδιαίτερη έμφαση στη βελτίωση υπαρχόντων τεχνικών και δημιουργία νέων για να αντιμετωπιστούν τα προβλήματα που τείθονται σε αυτές τις κατηγορίες που θα αναλύσουμε παρακάτω.

3.7.1 Κατηγοριοποίηση

Η κατηγοριοποίηση (classification) αποτελεί μια από τις βασικές εργασίες (tasks) στην εξόρυξη δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προκατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί (να ανατεθεί σε κάποια από τις κατηγορίες).

Στις περισσότερες περιπτώσεις, υπάρχει ένας περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες.).

3.7.2 Ομαδοποίηση

Μια από τις πιο σημαντικές διεργασίες στη διαδικασία εξόρυξης γνώσης είναι ομαδοποίηση (clustering). Πρόκειται για μια μεθοδολογία ανακάλυψης συστάδων και κατανομών ή προτύπων που παρουσιάζουν ενδιαφέρον στα υπό μελέτη δεδομένα. Ως ομάδα ορίζεται μια συλλογή αντικειμένων από τα δεδομένα, με βάση τη μεταξύ τους ομοιότητα. Οι απαρχές της ανάλυσης ομαδοποίησης εντοπίζονται από τα δεκαετία του 1960 (Anderberg, 1973). Σήμερα, οι διαδικασίες τη κατηγορίες θεωρούνται απαραίτητο συστατικό για πολλές εφαρμογές ετερόκλητων κλάδων, όπως η αναγνώριση προτύπων, η ανάλυση δεδομένων, η επεξεργασία εικόνας, η έρευνα αγοράς και άλλες.

Η διαδικασία της ομαδοποίησης μπορεί να οδηγήσει σε διαφορετικές τμηματοποιήσεις ενός συνόλου δεδομένων, ανάλογα με το κριτήριο κατηγορίες που χρησιμοποιείται. Τα βασικά βήματα της διαδικασίας είναι **η επιλογή των κατάλληλων χαρακτηριστικών** (attributes) στα οποία πρόκειται να εφαρμοστεί η

ομαδοποίηση, **η επιλογή ενός αλγορίθμου** που οδηγεί στον καθορισμό ενός καλού σχήματος κατηγορίες, **η επικύρωση** των αποτελεσμάτων και εν τέλει η ερμηνεία τους.

Ο αλγόριθμος που επιλέγεται καθορίζεται από το **μέτρο εγγύτητας** που προσδιορίζει πόσο «όμοια» είναι δύο αντικείμενα (τα χαρακτηριστικότερα μέτρα απόστασης/ομοιότητας δίνονται από τους Johnson and Wichern, 1998) και το **κριτήριο κατηγορίες**, το οποίο μπορεί να εκφραστεί μέσω μιας συνάρτησης ή κάποιου τύπου κανόνων, ή να ληφθεί υπόψη ο τύπος συστάδων που αναμένεται να εμφανιστούν στο σύνολο δεδομένων

3.7.3 Κανόνες Συσχέτισης

Η εξαγωγή κανόνων συσχέτισης (association rules) θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων. Έχει προσελκύσει μεγάλο ενδιαφέρον γιατί παρέχουν έναν συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Αυτοί οι συσχετισμοί παρουσιάζονται στην ακόλουθη μορφή $A \rightarrow B$ όπου το A και το B αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα.

3.7.4 Πρότυπα Ακολουθιών

Η εξόρυξη πρότυπων ακολουθιών (sequential patterns) είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικών με το χρόνο ή άλλες ακολουθίες. Ο περισσότερες μελέτες στα πρότυπα ακολουθιών επικεντρώνονται στα συμβολικά πρότυπα. Ο χρήστης εδώ μπορεί να προσδιορίσει τους περιορισμούς στα είδη των προτύπων ακολουθιών που εξάγονται με την παροχή των προσχεδίων προτύπων (template patterns) υπό μορφή σειριακών επεισοδίων, παράλληλων επεισοδίων ή κανονικών εκφράσεων. Παραδείγματα προτύπων ακολουθιών έχουμε στην καθημερινή μας ζωή όπως τα κείμενα, οι μουσικές νότες, τα δεδομένα του καιρού και οι ακολουθίες του DNA.

3.7.5 Παλινδρόμηση

Η παλινδρόμηση (regression) είναι θέμα το οποίο έχει μελετηθεί πολύ στην στατιστική. Κύριος σκοπός εδώ είναι η πρόβλεψη τη τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν. Συνήθως χρησιμοποιούμε ένα μοντέλο για την μεταβλητή. Η παλινδρόμηση καλύπτει ένα μεγάλο τμήμα του τομέα της εξόρυξης δεδομένων που έχει να κάνει με προβλέψεις.

3.7.6 Δέντρα Απόφασης

Τα δέντρα απόφασης (decision trees) έχουν μελετηθεί αρκετά σαν ένα ζήτημα μηχανικής μάθησης. Για να γίνει κατανοητό, ας υποθέσουμε ότι έχουμε ένα σύνολο εγγραφών και καθεμία από αυτές έχει μια λίστα χαρακτηριστικών. Ένα δέντρο απόφασης στο σύνολο των εγγραφών είναι ένα δέντρο όπου σε κάθε κόμβο του (που δεν είναι φύλλο) υπάρχει ένα ερώτημα που αναφέρεται στα χαρακτηριστικά των εγγραφών και κάθε ερώτημα καταλήγει σε ένα συγκεκριμένο παιδί ενός κόμβου. Τα φύλλα του δηλώνουν τις κλάσεις. Έτσι ένα δέντρο απόφασης εκτελεί κατηγοριοποίηση χρησιμοποιώντας ερωτήματα σχετικά με τα χαρακτηριστικά των εγγραφών. Οι εφαρμογές που χρησιμοποιούν δέντρα απόφασης είναι παρόμοιες με αυτές που κάνουν κατηγοριοποίηση (Mitchell, 1997).

3.8 Εφαρμογές εξόρυξης δεδομένων

Σε αυτή την ενότητα θα παρουσιάσουμε τις βασικές περιοχές εφαρμογής του τομέα της εξόρυξης δεδομένων.

- Παγκόσμιος ιστός

Ο τομέας της εξόρυξης δεδομένων είχε άμεση εφαρμογή με επιτυχία στο διαδίκτυο. Το πιο δημοφιλές παράδειγμα εξόρυξης δεδομένων στο διαδίκτυο είναι η Google (www.google.com). Για να γίνει πιο κατανοητή η σημαντικότητα της συνεισφοράς αυτής θα πρέπει να αντιληφθούμε πως ο όγκος της πληροφορίας που υπάρχει μέχρι τώρα στο διαδίκτυο είναι αδύνατο να μετρηθεί με ακρίβεια.

Η Google και γενικά ο τομέας της εξόρυξης δεδομένων στο διαδίκτυο έχουν σήμερα τεράστια επιτυχία γιατί έχουν εκπληρώσει δυο σημαντικούς στόχους. Πρώτα, μπορούν να κάνουν αναζήτηση (με κάθε ερώτημα) σε τόσα πολλά δεδομένα σε πολύ σύντομο χρόνο. Δεύτερον, μπορούν να επιστρέψουν σε κάθε ερώτημα τα πρώτα αποτελέσματα που είναι πιο χρήσιμα. Έτσι τελικά ο χρήστης λαμβάνει γρήγορα και εύκολα μόνο την ουσιώδη πληροφορία που θέλει.

- Επιστήμη

Αλγόριθμοι εξόρυξης δεδομένων χρησιμοποιούνται ευρέως σε εφαρμογές από διάφορους άλλους επιστημονικούς τομείς. Ένα αξιοσημείωτο παράδειγμα είναι το SKYCAT (Fayyad, 1996) ένα σύστημα εξόρυξης δεδομένων που αναλαμβάνει

ανάλυση και κατηγοριοποίηση χωρικών αντικειμένων. Αυτό που είναι αξιοσημείωτο, είναι πως το SKYCAT εκτελεί αλγόριθμους για την ανίχνευση αντικειμένων από εικόνες.

- Μάρκετινγκ

Μια κατηγορία πολύ γνωστών εφαρμογών εξόρυξης δεδομένων είναι αυτές του μάρκετινγκ. Αυτό είναι αναμενόμενο μιας και μεγάλες εταιρίες χρησιμοποιούν μεγάλα συστήματα διαχείρισης δεδομένων για να διαχειρίζονται μεγάλο αριθμό πελατών και οικονομικών στοιχείων. Τα τελευταία χρόνια οι τάσεις του μάρκετινγκ ορίζουν μια πολιτική έρευνας των αναγκών των πελατών. Αναζητούν απαντήσεις σε ερωτήματα όπως, τι είναι αυτό που θέλουν οι πελάτες, ποιες είναι οι ανάγκες τους κ.α.

Ο τομέας της εξόρυξης δεδομένων έχει συνεισφέρει σημαντικά σε αυτή την κατεύθυνση από την ανάλυση δεδομένων μια επιχείρησης και την εξαγωγή χρήσιμων συμπερασμάτων για την συμπεριφορά των πελατών.

- Χρηματοοικονομικά

Πολυάριθμες χρηματιστηριακές εταιρίες χρησιμοποιούν τεχνικές εξόρυξης δεδομένων έτσι ώστε να μπορούν να γνωρίζουν που να επενδύσουν. Στην πραγματικότητα μια μεγάλη μερίδα έρευνας στο τομέα εξόρυξης δεδομένων έχει γίνει έχοντας ως αφετηρία χρηματιστηριακές και χρηματοοικονομικές εφαρμογές. Μια άλλη χρήση των τεχνικών εξόρυξης δεδομένων είναι οι εφαρμογές εξόρυξης δεδομένων από κείμενα. Για παράδειγμα αλγόριθμοι που εξάγουν χρήσιμη πληροφορία από μη δομημένα κείμενα, έτσι ώστε να προβλεφθούν οι τάσεις σε μετοχές (Wurthrich et al., 1999).

- Πρόληψη και Ασφάλεια

Η εξόρυξη δεδομένων έχει με επιτυχία εφαρμοστεί και στην πρόληψη και αποφυγή διάφορων τύπων απάτης. Από την αναγνώριση κακόβουλων ενεργειών σε συναλλαγές κάποιος μπορεί να αντιληφθεί συναλλαγές που μπορεί να σχετίζονται με οικονομικές παρανομίες ή άλλου είδους απάτες. Ένα παράδειγμα συστήματος είναι το FAIS (Financial Crimes Enforcement Network (FINCEN) AI system)(Senator et al., 1995).

Ωστόσο τα τελευταία χρόνια, όπως βλέπουμε και ακούμε, υπάρχει μια τάση για πρόληψη σε κακόβουλες ενέργειες. Οι κινήσεις μας σε δημόσιους χώρους καταγράφεται όπως και αυτές που έχουν να κάνουν με τον παγκόσμιο ιστό. Για παράδειγμα μια πρόσφατη εφαρμογή μπορούσε να αναγνωρίζει ανώμαλα πρότυπα χρησιμοποιώντας κανόνες σε δεδομένα νοσοκομείων έτσι ώστε να αναγνωρίζει, σε πραγματικό χρόνο, εμφάνιση ασθενειών.

ΚΕΦΑΛΑΙΟ 4

«ΕΦΑΡΜΟΓΗ ΔΕΔΟΜΕΝΩΝ»

4.1 Το πρόγραμμα WEKA

4.1.1 Γενικά

Το Weka είναι ένα περιβάλλον ανάπτυξης εφαρμογών μηχανικής μάθησης και εξόρυξης γνώσης, το οποίο αναπτύχθηκε από το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Η ανάπτυξη του ξεκίνησε το 1999 με σκοπό να καλύψει τις πειραματικές ανάγκες του εργαστηρίου Μηχανικής Μάθησης του Πανεπιστημίου του Waikato. Το όνομα του Weka αποτελεί ακρωνύμιο της φράσης Waikato Environment for Knowledge Analysis. Το πρόγραμμα Weka είναι γραμμένο στη γλώσσα προγραμματισμού Java, ώστε να μπορεί να χρησιμοποιηθεί με όσο το δυνατόν περισσότερα λειτουργικά συστήματα, και διατίθεται ελεύθερα στο Διαδίκτυο, συνοδευόμενο από κατάλληλα εγχειρίδια. Σήμερα αποτελεί ένα από τα δημοφιλέστερα εργαλεία μηχανικής μάθησης και χρησιμοποιείται στην έρευνα, στην εκπαίδευση και σε εφαρμογές.

Το Weka υλοποιεί τις περισσότερες τεχνικές και αλγορίθμους που υπάρχουν στη θεωρία, παρέχει ένα εύχρηστο γραφικό περιβάλλον για τον χειρισμό τους, καθώς και εργαλεία για οπτική παρουσίαση δεδομένων, προ-επεξεργασία δεδομένων και ανάλυση αποτελεσμάτων.

4.1.2 Μέτρα αξιολόγησης

- ✓ ***Πίνακας ταξινόμησης***

Στον παρακάτω πίνακα (Πίνακας 4.1) εμφανίζεται ο πίνακας ταξινόμησης (confusion matrix) που αναπαριστά τις τέσσερις δυνατές εκβάσεις μιας προσπάθειας δυαδικής ταξινόμησης. Στην κύρια διαγώνιο του πίνακα οι συμβολισμοί TP (True Positive) και TN (True Negative) αντιστοιχούν στις περιπτώσεις ορθής ταξινόμησης ενός θετικού παραδείγματος στην κλάση των θετικών και ενός αρνητικού στην κλάση των αρνητικών. Οι άλλες δύο περιπτώσεις αναφέρονται σε εσφαλμένες ταξινομήσεις,

που συμμετέχουν στον υπολογισμό του συνολικού σφάλματος του ταξινομητή. Η περίπτωση FN (False Negative), που αντιστοιχεί στην εσφαλμένη ταξινόμηση ενός θετικού παραδείγματος ως αρνητικό, αναφέρεται στη βιβλιογραφία ως Σφάλμα Τύπου I (Error Type I ή ε_I). Η αντίθετη περίπτωση FP (False Positive) αποτελεί το Σφάλμα Τύπου II – ε_{II}. Ας σημειωθεί επίσης ότι το άθροισμα TP + FN ισούται με το σύνολο των θετικών παραδειγμάτων που χρησιμοποιήθηκαν κατά τον έλεγχο του ταξινομητή, και κατ' επέκταση το FP + TN με το πλήθος των αρνητικών που είτε ταξινομήθηκαν σωστά ή εσφαλμένα.

	Positive Class	Negative Class
Classified as Positive	**TP**	**FP**
Classified as Negative	**$FN_{(\varepsilon_I)}$**	**TN**

Πίνακας 4.1: Πίνακας ταξινόμησης ενός προβλήματος ταξινόμησης δύο κλάσεων

- ✓ ***Ακρίβεια***

Ως ακρίβεια (accuracy) ορίζουμε την αναλογία όλων των προβλέψεων που ήταν σωστές:

$$\boldsymbol{accuracy} = \frac{\#\boldsymbol{\sigma\omega\sigma\tau\acute{\omega}\nu\ \pi\rho o\beta\lambda\acute{\epsilon}\psi\epsilon\omega\nu}}{\#\boldsymbol{\pi\rho o\beta\lambda\acute{\epsilon}\psi\epsilon\omega\nu}} \qquad (4.1)$$

Για ένα πρόβλημα ταξινόμησης δύο κλάσεων, η ποσότητα εκφράζεται με την βοήθεια του πλήθους των σωστών και εσφαλμένων ταξινομήσεων ενός συστήματος ως εξής:

$$\boldsymbol{accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \quad (4.2)$$

- ✓ ***Ορθότητα***

Ως ορθότητα (precision–degree of soundness) ορίζουμε τη δεσμευμένη πιθανότητα αν ταυτίζεται η κλάση που προβλέπει ένας ταξινομητής για ένα στιγμιότυπο με την πραγματική του κλάση:

$$\mathbf{precision_c} = \frac{\#\boldsymbol{\sigma\omega\sigma\tau\acute{\omega}\nu\ \pi\rho o\beta\lambda\acute{\epsilon}\psi\epsilon\omega\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\varsigma\ c}}{\#\boldsymbol{\pi\rho o\beta\lambda\acute{\epsilon}\psi\epsilon\omega\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\varsigma\ c}} \qquad (4.3)$$

Για ένα πρόβλημα ταξινόμησης δύο κλάσεων έχουμε:

$$\mathbf{precision_P} = \frac{\mathbf{TP}}{\mathbf{TP+FP}}, \quad \mathbf{precision_N} = \frac{\mathbf{TN}}{\mathbf{TN+FN}} \qquad (4.4)$$

✓ ***Ανάκληση***

Ως ανάκληση (recall – degree of completeness) ορίζουμε τη δεσμευμένη πιθανότητα αν ένα στιγμιότυπο ανήκει σε μια κλάση, έστω c, αυτή να αναγνωριστεί σωστά από τον ταξινομητή:

$$\mathbf{recall_C} = \frac{\#\textbf{σωστών προβλέψεων κλάσης c}}{\#\textbf{δεδομένων κλάσης c}} \qquad (4.5)$$

Για ένα πρόβλημα ταξινόμησης δύο κλάσεων έχουμε:

$$\boldsymbol{recall_P} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}, \quad \boldsymbol{recall_N} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \qquad (4.6)$$

✓ ***Μέτρο F***

Στην πράξη οι δύο παραπάνω μετρικές δεν μπορούν να εκτιμηθούν χωριστά, καθώς παρέχουν μια αλληλοσυμπληρούμενη εικόνα της αποτελεσματικότητας ενός ταξινομητή. Ένα μέτρο που τα συνδυάζει είναι η συνάρτηση F (F–score), που ορίζεται ως:

$$\mathbf{F_c} = \frac{\mathbf{2{*}precision_c{*}recall_c}}{\mathbf{precision_c{+}recall_c}} \qquad (4.7)$$

✓ ***Εμβαδόν κάτω από την καμπύλη ROC***

Η καμπύλη ROC (Receiver Operating Characteristic) είναι η γραφική παράσταση του ποσοστού των σωστών ταξινομήσεων των θετικών παραδειγμάτων, $\mathbf{Pr}[\boldsymbol{f(X)} > \boldsymbol{\theta}|\boldsymbol{Y} = \mathbf{1}]$, ως προς το ρυθμό των εσφαλμένων ταξινομήσεων των θετικών παραδειγμάτων, $\mathbf{Pr}[\boldsymbol{f(X)} > \boldsymbol{\theta}|\boldsymbol{Y} = -\mathbf{1}]$, καθώς μεταβάλλουμε το κατώφλι απόφασης θ. Η καμπύλη ROC αποτελεί ένα δισδιάστατο μέτρο αξιολόγησης ενός ταξινομητή, ενώ ένα άλλο μέτρο αξιολόγησης που χρησιμοποιείται πιο συχνά είναι το εμβαδό κάτω από την καμπύλη (area under curve – AUC). Όταν το εμβαδό είναι ίσο με 1 τότε ο ταξινομητής επιτυγχάνει τέλεια ακρίβεια, αν το κατώφλι απόφασης επιλεχθεί σωστά, ενώ ένας ταξινομητής που ταξινομεί τα δεδομένα τυχαία έχει AUC ίσο με 0,5.

4.1.3 Προ-επεξεργασία δεδομένων

Η εξόρυξη γνώσης αφορά κυρίως στις μεθοδολογίες για την εξαγωγή των προτύπων από μεγάλες βάσεις δεδομένων. Δεδομένου ότι ένα σύστημα εξόρυξης γνώσης μπορεί να παράγει κάτω από διαφορετικές υποθέσεις χιλιάδες ή εκατομμύρια πρότυπα, προκύπτουν κάποια ερωτήματα σχετικά με την ποιότητα των αποτελεσμάτων εξόρυξης γνώσης, όπως ποια από τα πρότυπα που έχουν εξαχθεί από ένα σύνολο δεδομένων παρουσιάζουν ενδιαφέρον και ποια από αυτά αντιπροσωπεύουν γνώση.

Γενικά, ένα πρότυπο είναι ενδιαφέρον εάν γίνεται εύκολα κατανοητό, είναι έγκυρο, νέο και χρήσιμο στο μέλλον. Επίσης, ένα πρότυπο μπορεί να θεωρηθεί ενδιαφέρον εάν κάποιες υποθέσεις που έχει θέσει κάποιος χρήστης. Σε κάθε περίπτωση ένα ενδιαφέρον πρότυπο αντιπροσωπεύει γνώση.

Η ποιότητα των προτύπων εξαρτάται και από την ποιότητα των στοιχείων που έχουν αναλυθεί και από την ποιότητα των αποτελεσμάτων εξόρυξης γνώσης. Κατά συνέπεια, διάφορες τεχνικές έχουν αναπτυχθεί στοχεύοντας στην αξιολόγηση και την προετοιμασία των στοιχείων που χρησιμοποιούνται ως είσοδος στη διαδικασία εξόρυξης γνώσης.

Πολλές φορές τα δεδομένα, όπως είναι διαθέσιμα από τις πραγματικές εφαρμογές, στερούνται ποιότητας. Οι χρήστες των βάσεων δεδομένων συχνά αναφέρουν λάθη, εμφάνιση ασυνήθιστων τιμών και ασυνέπειες στα αποθηκευμένα δεδομένα. Για το λόγο αυτό είναι συνηθισμένο σε πραγματικές εφαρμογές τα υπό ανάλυση δεδομένα να είναι:

- Ημιτελή: Να παρουσιάζουν δηλαδή έλλειψη κάποιων τιμών γνωρισμάτων, έλλειψη ορισμένων γνωρισμάτων που ίσως έχουν ενδιαφέρον, ή να περιέχουν μόνο αθροιστικά δεδομένα.
- Θόρυβος: Περιέχουν λάθη ή ακραίες τιμές.
- Ασυνεπή: Περιέχουν αποκλίσεις στις κωδικοποιήσεις που χρησιμοποιούνται για να ταξινομήσουμε τα δεδομένα ή στα ονόματα που χρησιμοποιούνται για να αναφερθούμε στα ίδια δεδομένα.

Με βάση ένα σύνολο δεδομένων που στερείται ποιότητας τα αποτελέσματα της διαδικασίας εξόρυξης γνώσης αναπόφευκτα τείνουν να είναι ανακριβή και να μην παρουσιάζουν κάποιο ενδιαφέρον για τον ειδικό του χώρου. Για το λόγο αυτό, η προ-επεξεργασία των δεδομένων είναι ένα σημαντικό βήμα στη διαδικασία ανακάλυψης γνώσης. Η εφαρμογή των τεχνικών προ-επεξεργασίας δεδομένων πριν από το βήμα της εξόρυξης δεδομένων θα μπορούσε να συμβάλλει στη βελτίωση της ποιότητας των υπό ανάλυση δεδομένων και συνεπώς της ακρίβειας και της αποτελεσματικότητας των διαδικασιών που ακολουθούν της εξόρυξης δεδομένων.

Υπάρχει ένας αριθμός από τεχνικές προ-επεξεργασίας δεδομένων που στοχεύει ουσιαστικά στη βελτίωση της γενικής ποιότητας των εξαγομένων προτύπων. Οι πιο ευρύτατα χρησιμοποιούμενες τεχνικές μπορούν να συνοψιστούν στις ακόλουθες :

- Καθαρισμός δεδομένων, ο οποίος μπορεί να εφαρμοστεί για να αφαιρεθεί ο θόρυβος και να διορθωθούν τυχόν ασυνέπειες στα δεδομένα.
- Μετασχηματισμός δεδομένων : Μια κοινή τεχνική μετασχηματισμού είναι η κανονικοποίηση.
- Μείωση δεδομένων : Χρησιμοποιείται προκειμένου να μειωθεί το μέγεθος στοιχείων με τη συνάθροιση, εξαλείφοντας τα περιττά χαρακτηριστικά.

4.1.4 Καθαρισμός Δεδομένων

Τα πραγματικά δεδομένα είναι συνήθως ατελή, περιέχουν θόρυβο και δεν έχουν συνοχή. Ο καθαρισμός τους περιλαμβάνει όλες τις διαδικασίες της συμπλήρωσης των ατελών δεδομένων, της εξομάλυνσης του θορύβου και του εντοπισμού των ακραίων τιμών.

Ελλιπείς τιμές

Αρκετά συχνά εμφανίζονται περιπτώσεις εγγραφών με ελλιπή δεδομένα (missing data). Διάφορες μέθοδοι έχουν κατά καιρούς προταθεί στη βιβλιογραφία για την αντικατάσταση των ελλιπών τιμών όπως : 1. αντικατάσταση της τιμής της ιδιότητας με σταθερή τιμή, 2. αντικατάσταση με την τιμή της ιδιότητας που εμφανίζει τη μεγαλύτερη συχνότητα ή με τη μέση τιμή της ιδιότητας, 3. διαγραφή όλων των παραδειγμάτων στα οποία εμφανίζεται έστω και μία ελλιπής τιμή.

Δεδομένα με θόρυβο

Ως θόρυβο ορίζουμε το τυχαίο σφάλμα ή διακύμανση σε μια μετρήσιμη μεταβλητή. Συνήθεις αιτίες δημιουργίας θορύβου στα δεδομένα είναι:

α) Προβλήματα κατά τη συλλογή, εισαγωγή ή μετάδοση των δεδομένων

β) Τεχνολογικοί περιορισμοί ή προβληματικά όργανα

χ) Χρήση μη συμβατής ονοματολογίας δεδομένων

δ) Διπλότυπες εγγραφές

Ο τρόπος αντιμετώπισης του θορύβου είναι η εξομάλυνσή του. Οι πιο βασικές μέθοδοι εξομάλυνσης αναφέρονται συνοπτικά παρακάτω:

- Κατάτμηση: Εξομάλυνση της τιμής ενός γνωρίσματος με βάση την τιμή των γειτόνων του. Στην αρχή γίνεται ταξινόμηση και διαχωρισμός των δεδομένων σε ισομήκη τμήματα και στη συνέχεια εφαρμόζεται η ανάλογη μέθοδος εξομάλυνσης (smooth by bin means, smooth by bin median, smooth by bin boundaries, κτλ).
- Ομαδοποίηση: Εύρεση και απομάκρυνση των ακραίων τιμών, δηλαδή των μεταβλητών με τιμές με μεγάλη απόκλιση από το σύνολο των υπολοίπων τιμών.
- Συνδυασμένη επίβλεψη από τον υπολογιστή και από τον χρήστη: Εύρεση "ύποπτων" τιμών από τον υπολογιστή και έλεγχος από τον χρήστη για την εγκυρότητα των τιμών αυτών.
- Παλινδρόμηση: Εξομάλυνση των δεδομένων με τη χρήση συναρτήσεων παλινδρόμησης.

4.1.5 Μετασχηματισμός Δεδομένων

Τα δεδομένα μετασχηματίζονται με σκοπό να διευκολύνουν την εξόρυξη από τα δεδομένα και να παρέχουν πιο κατανοητά αποτελέσματα. Δεδομένα από

διαφορετικές πηγές μετατρέπονται σε μια κοινή μορφή που επιτρέπει την επεξεργασία τους. Επιπλέον, πολλοί αλγόριθμοι εξόρυξης απαιτούν συγκεκριμένες δομές δεδομένων με αποτέλεσμα να επιβάλλεται η προσαρμογή των αρχικών δεδομένων σε αυτές τις δομές.

Παραδείγματα μετασχηματισμών που μπορούν να γίνουν είναι τα εξής:

- Επιλογή ή συγχώνευση χαρακτηριστικών, ώστε να μειωθεί η πολυπλοκότητα των δεδομένων.
- Αντικατάσταση ενός χαρακτηριστικού από άλλο.
- Μετατροπή συνεχών τιμών σε διακριτές τιμές (διακριτοποίηση).
- Απομάκρυνση σπάνια εμφανιζόμενων ακραίων τιμών, όπως είναι τα ακραία σημεία (outliers).
- Μετασχηματισμός με εφαρμογή κάποιας συνάρτησης (π.χ. λογαριθμικής) στις τιμές ενός χαρακτηριστικού.

.

Όλες οι παραπάνω τεχνικές διευκολύνουν τη διαδικασία εξόρυξης, είτε μειώνοντας τον αριθμό των χαρακτηριστικών (dimensionality reduction), είτε μειώνοντας τον αριθμό των τιμών που παίρνει ένα χαρακτηριστικό (variability reduction).

4.1.6 Μετα-επεξεργασία

Σε αυτό το βήμα λαμβάνουν χώρα διαδικασίες όπως η ερμηνεία, η αξιολόγηση και γενικότερα η διαχείριση των αποτελεσμάτων. Διάφορες τεχνικές οπτικής αναπαράστασης (visualization) μπορούν να χρησιμοποιηθούν για την παρουσίαση των δεδομένων. Η κατανόηση της χρησιμότητας των αποτελεσμάτων εξαρτάται σε μεγάλο βαθμό από τον τρόπο παρουσίασής τους. Οι παραπάνω τεχνικές, δίνουν στο χρήστη τη δυνατότητα να συνοψίζει και να εξάγει πιο πολύπλοκα συμπεράσματα από ότι με μαθηματικές βασισμένες σε κείμενο περιγραφές των αποτελεσμάτων. Πολλές φορές η γνώση που προκύπτει καταγράφεται σε μια βάση γνώσης (knowledge base), οπότε ενδέχεται να παρουσιαστεί η ανάγκη επίλυσης συγκρούσεων με προϋπάρχουσα γνώση. Συνήθως τα αποτελέσματα μετά από ένα κύκλο της διαδικασίας δίνουν ερέθισμα για νέες αναζητήσεις με αποτέλεσμα την επανάληψη ολόκληρης της διαδικασίας.

4.2 Το πρόβλημα μελέτης

4.2.1 Ο χώρος της ασφάλισης αυτοκίνητων

Οι ασφαλιστικές απάτες με ψευδείς ισχυρισμούς αποζημίωσης στον κλάδο της ασφάλισης αυτοκινήτου πλήττουν περισσότερο από ποτέ ολόκληρο τον κλάδο των ασφαλίσεων. Δυστυχώς, το φαινόμενο των εικονικών δηλώσεων ζημιών από ασφαλισμένους με στόχο την εξαπάτηση και την αποκόμιση αποζημιώσεων ύψους

χιλιάδων ευρώ, παρουσιάζει έξαρση. Παρά το γεγονός ότι δεν υπάρχουν στατιστικά στοιχεία για την έκταση που έχει πάρει μέχρι σήμερα η ασφαλιστική απάτη, δεκάδες είναι οι περιπτώσεις που γίνονται αντιληπτές σε όλες σχεδόν τις εταιρίες του κλάδου, με αποτέλεσμα τα τελευταία χρόνια το ποσοστό περιπτώσεων άπατης να έχει αυξηθεί δραματικά. Όπως προκύπτει από εκτιμήσεις, καθώς δεν υπάρχουν μετρήσιμα στοιχεία, στο σύνολο της αγοράς, για όλους τους κλάδους ασφάλισης δηλαδή, υπολογίζεται ότι σε ετήσια βάση καταβάλλονται γύρω στα 200 εκατ. ευρώ σε περιπτώσεις στημένων υποθέσεων. Από το ποσό αυτό γύρω στα 80 εκατ. ευρώ αφορούν τον κλάδο αυτοκινήτων.

4.2.2 Δεδομένα

Όπως έχει ήδη αναφερθεί σε προηγούμενο κεφάλαιο βασικό μελέτης της εργασίας αυτής είναι η ανίχνευση άπατης σε δεδομένα προερχόμενα από το χώρο της ασφάλισης αυτοκίνητων στον οποίο τα διαθέσιμα δεδομένα είναι αρκετά περιορισμένα. Συγκεκριμένα, στην εργασία αυτή, τα δεδομένα που χρησιμοποιούνται παρέχονται από την Angoss KnowledgeSeeker software(αναλύονται στην εργασία του Pyle, 1999).

Το σύνολο των δεδομένων περιέχει 11338 περιπτώσεις από τον Ιανουάριο 1994 έως το Δεκέμβριο 1995, τα οποία αποτελούν τα λεγόμενα ‘δεδομένα εκπαίδευσης’ και 4083 περιπτώσεις από τον Ιανουάριο 1996 έως το Δεκέμβριο 1996 που αποτελούν τα score data.

Το αρχικό αρχείο δεδομένων περιλαμβάνει συνολικά 31 μεταβλητές από τις οποίες οι 6 είναι αριθμητικές και οι υπόλοιπες 25 είναι κατηγορικές. Σε γενικές γραμμές, η ποιότητα των δεδομένων είναι αρκετά καλή, αν και υπάρχουν κάποια προβλήματα, όπως για παράδειγμα στη μεταβλητή make, που αφορά τη μάρκα του αυτοκίνητου που εμπλέκεται στη υπόθεση, υπάρχουν συνολικά 19 πιθανές τιμές από τις οποίες 5 από αυτές καταλαμβάνουν το 90% σχεδόν του συνόλου των τιμών.

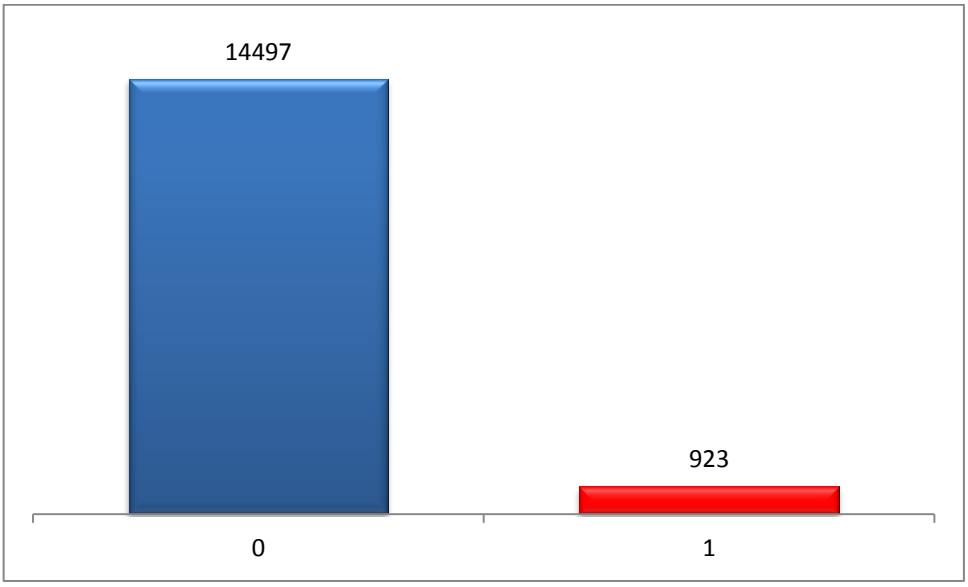
Οι μεταβλητές που τελικά θα χρησιμοποιηθούν στη μελέτη αυτή παρουσιάζονται στον πίνακα 4.2.

VARIABLE	ΤΥΠΟΣ ΜΕΤΑΒΛΗΤΗΣ
Month	Κατηγορική
Age	Αριθμητική
Policytype	Κατηγορική
Ageofpolicyholder	Αριθμητική
Monthclaimed	Κατηγορική
PastNoofclaims	Αριθμητική
Dayofweek	Κατηγορική
Fault	Κατηγορική
Vehiclecategory	Κατηγορική
Basepolicy	Κατηγορική
Adresschangeclaim	Κατηγορική
Make	Κατηγορική
Agenttype	Κατηγορική
Policynumber	Αριθμητική
Fraudfound	Κατηγορική
Sex	Κατηγορική
Accidentarea	Κατηγορική
Numberofsuppliments	Κατηγορική
vehicleprice	Αριθμητική
Deductible	Αριθμητική
Year	Αριθμητική
Ageofvehicle	Αριθμητική

Πίνακας 4.2 : Μεταβλητές του προβλήματος μελέτης

Συνολικά, τελικά, θα χρησιμοποιηθούν 22 μεταβλητές εκ των οποίων οι 14 είναι κατηγορικές και οι υπόλοιπες 8 είναι αριθμητικές. Η εξαρτημένη μεταβλητή

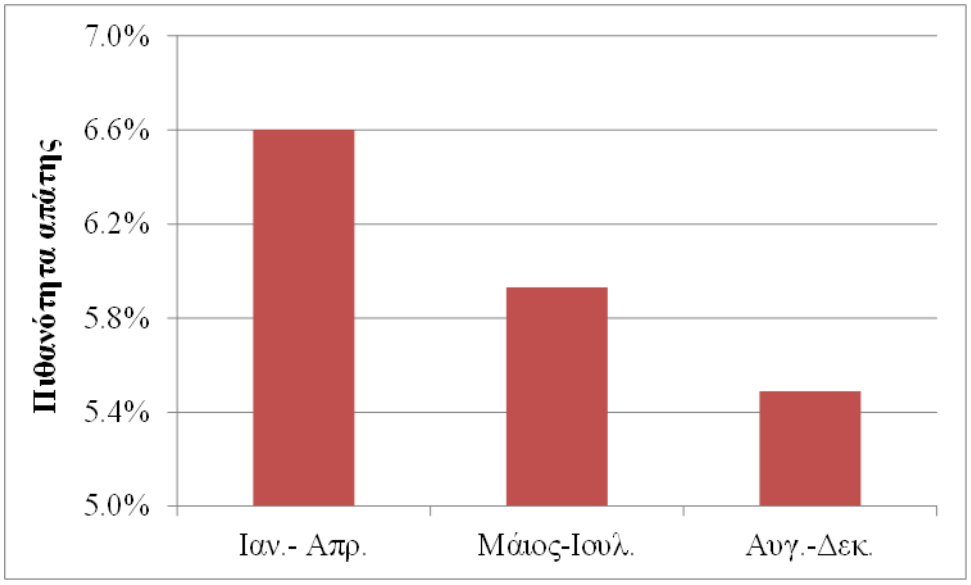
είναι το χαρακτηριστικό fraudfound και λαμβάνει τιμές 0 και 1. Πρόκειται για την κατηγορία στην οποία εντάσσεται κάθε στιγμιότυπο του αρχείο των δεδομένων και κατηγοριοποιεί τις περιπτώσεις με βάση το γεγονός εάν εντοπίστηκε απάτη ή όχι. Με 0 ορίζεται η περίπτωση «μη απάτης» και με 1 οι περιπτώσεις που εμφανίζουν απάτη. Από τα 15420 στιγμιότυπα, τα 14497 (94%) κατατάσσονται στην κατηγορία «μη απάτης», ενώ τα υπόλοιπα 923 (6%) στην κατηγορία «απάτης» (σχήμα 4.1). Όσον αφορά το σύνολο των μεταβλητών, αριθμητικές είναι οι : age, year, ageofvehicle, deductible, vehicleprice, policynumber, pastnoofclaims και η ageofpolicyholder, ενώ οι υπόλοιπες αποτελούν τις κατηγορικές μεταβλητές του προβλήματος.



Σχήμα 4.1 : Μεταβλητή fraudfound

Όλα τα δεδομένα έχουν κωδικοποιηθεί με αριθμητικές τιμές έτσι ώστε τα δεδομένα εισόδου σε όλους τους υπό εξέταση αλγορίθμους να είναι ίδια και να παρέχουν συγκρίσιμα αποτελέσματα.

Η ταυτόχρονη απεικόνιση της σχέσης ενός χαρακτηριστικού με όλα τα υπόλοιπα χαρακτηριστικά είναι δυνατή και εξαιρετικά χρήσιμο εργαλείο. Στα Σχήματα 4.2-4.4 φαίνεται η σχέση της πιθανότητας απάτης με κάποια από τα υπόλοιπα χαρακτηριστικά. Πιο συγκεκριμένα, η μεταβλητή month, για παράδειγμα, έχει κωδικοποιηθεί σε τρεις τιμές 1,2 και 3 όπου η τιμή 1 αντιπροσωπεύει τους μήνες Ιανουάριο, Φεβρουάριο, Μάρτιο και Απρίλιο και αντιστοιχεί στο 35% του συνόλου των περιπτώσεων (5317 σε σύνολο 15420 περιπτώσεων), η τιμή 2 αντιστοιχεί στους μήνες Μάιο, Ιούνιο και Ιούλιο, κατέχοντας το 25% του συνόλου και τέλος η τιμή 3 στους μήνες Αύγουστο, Σεπτέμβριο, Οκτώβριο, Νοέμβριο και Δεκέμβριο με το ποσοστό συμμετοχής του να φτάνει το 40% (Σχήμα 4.2).



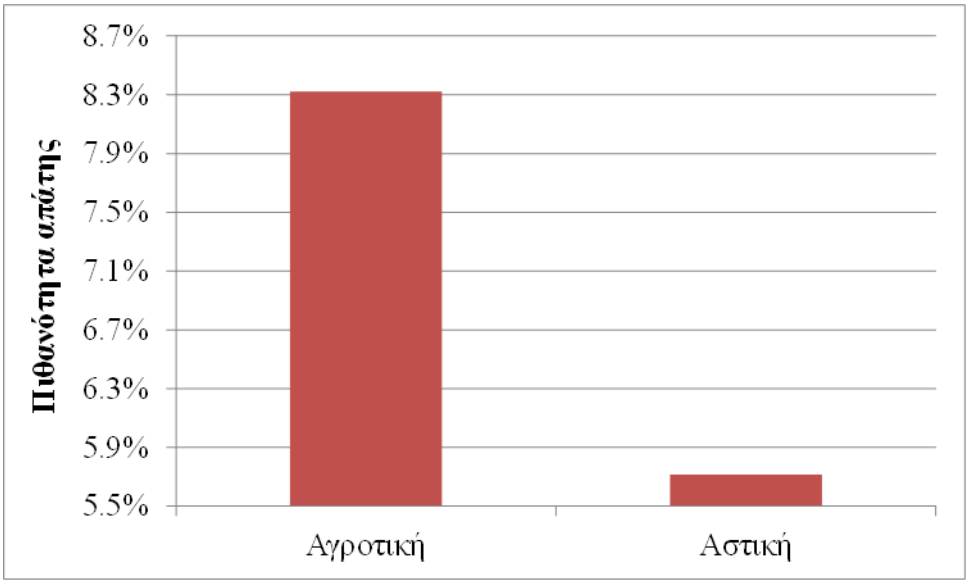
Σχήμα 4.2 : Πιθανότητα απάτης για το χαρακτηριστικό month

Άλλο παράδειγμα αποτελεί η μεταβλητή sex, η όποια δηλώνει το φύλο του ασφαλιζόμενου και λαμβάνει δυο τιμές - 1, 2 - αντιπροσωπεύοντας το φύλο άνδρας και γυναίκα αντίστοιχα. Στο σύνολο δεδομένων που χρησιμοποιείται η τιμή 1 αντιστοιχεί στις 13000 των περιπτώσεων, λαμβάνοντας ένα υψηλό ποσοστό κοντά στο 85% του συνόλου (Σχήμα 4.3).

Σχήμα 4.3 : Πιθανότητα απάτης για το χαρακτηριστικό sex

Τέλος, η μεταβλητή accidentarea, η όποια δηλώνει το τόπο όπου εξελίχθηκε το ατύχημα λαμβάνει δυο τιμές - 1, 2 – και αντιπροσωπεύει την αγροτική και την αστική περιοχή αντίστοιχα. Στο σύνολο δεδομένων που χρησιμοποιείται η τιμή 1

αντιστοιχεί στις 1598 των περιπτώσεων ενώ η τιμή 2 στις υπόλοιπες 13822 (Σχήμα 4.4)



Σχήμα 4.4 : Πιθανότητα απάτης για το χαρακτηριστικό accidentarea

4.3 Εφαρμογή Αλγορίθμων

Στο παραπάνω σύνολο δεδομένων εφαρμόστηκαν με τη βοήθεια του προγράμματος Weka, διάφοροι δημοφιλείς αλγόριθμοι εξόρυξης γνώσης.

4.3.1 Μηχανές διανυσμάτων υποστήριξης

Γενικά

Οι μηχανές διανυσμάτων υποστήριξης (support vector machines –SVM, (Cortes.C. and Vapnik V.N.,1995) έχουν αρκετές αξιοπρόσεκτες ιδιότητες που άλλοι αλγόριθμοι μάθησης δεν έχουν, όπως η μεγιστοποίηση του περιθωρίου ταξινόμησης (maximization of margin) και ο μη γραμμικός ταξινομητής μετασχηματισμός του χώρου εισόδου (input space) στο χώρο των χαρακτηριστικών (feature space) χρησιμοποιώντας μεθόδους πυρήνων (kernel methods). Ένας γραμμικός ταξινομητής SVM είναι ένα υπερεπίπεδο (hyperplane) που διαχωρίζει ένα σύνολο θετικών παραδειγμάτων από ένα σύνολο αρνητικών, μεγιστοποιώντας το περιθώριο στο χώρο των χαρακτηριστικών, την απόσταση δηλαδή του υπερεπιπέδου από τα πλησιέστερα θετικά ή αρνητικά παραδείγματα.

Τα παραδείγματα που βρίσκονται πάνω στο περιθώριο καλούνται διανύσματα υποστήριξης. Για τα προβλήματα που δεν είναι γραμμικά διαχωρίσιμα, μπορούν να

χρησιμοποιηθούν μέθοδοι πυρήνων που μετασχηματίζουν ένα μη γραμμικό χώρο εισόδου σε ένα γραμμικό χώρο χαρακτηριστικών. Επίσης, στις περιπτώσεις όπου τα σημεία δεν είναι γραμμικά διαχωρίσιμα, ο αλγόριθμος SVM έχει μια παράμετρο, C, η οποία επηρεάζει το πλήθος των δεδομένων εκπαίδευσης που θα βρίσκονται στη λάθος μεριά του υπερεπιπέδου.

Οι πιο συνηθισμένοι τύποι πυρήνων για το μετασχηματισμό του χώρου είναι οι:

I. Πολυωνυμικός πυρήνας (polynomiakernel): $k(x, y) = (x * y)^d$
II. Πυρήνας ακτινικής βάσης (radial basis function kernel, RBF): $k(x, y) = \exp(-\gamma\|x - y\|)^d$
III. Σιγμοειδής πυρήνας (sigmoid kernel): k (x, y) = $\tanh(\kappa x * y + \theta)$
IV. Γραμμικός πυρήνας (linear kernel) : $k(x, y) = x * y$

Παρουσίαση αποτελεσμάτων

Για την εφαρμογή του αλγορίθμου στην παρούσα εργασία θα γίνει χρήση δυο από των τύπων πυρήνα για το μετασχηματισμό του χώρου που αναφέρθηκαν παραπάνω, ο πυρήνας ακτινικής βάσης (RBF) και ο γραμμικός πυρήνας.

Σε αυτό το σημείο πρέπει να σημειωθεί ότι για όλες τις περιπτώσεις αλγορίθμων που χρησιμοποιήθηκαν, έγιναν αρκετές επαναλήψεις έτσι ώστε το TP rate και το FP rate για τις δυο κλάσεις να συγκλίνουν όσο το δυνατό περισσότερο. Η παράμετρος που υπέστη τροποποίηση για την επίτευξη ορθών αποτελεσμάτων είναι ο πίνακας κόστους.

i. *Πυρήνας RBF*

Με δεδομένο ότι το πλήθος των περιπτώσεων απάτης στο δείγμα είναι πολύ μικρότερο από το πλήθος των περιπτώσεων μη απάτης, η ανάπτυξη και αξιολόγηση κατάλληλων μοντέλων πρόβλεψης πρέπει να προσαρμοστεί ώστε να ληφθεί υπόψη το στοιχείο αυτό. Για αυτό το λόγο τα σφάλματα ταξινόμησης κατά την ανάπτυξη κάθε μοντέλου σταθμίζονται κατάλληλα, ώστε να επιτευχθεί μια ικανοποιητική ισορροπία στις ακρίβειες ταξινόμησης για κάθε κατηγορία περιπτώσεων. Αυτό επιτυγχάνεται με τη χρήση ενός πίνακα κόστους των εσφαλμένων ταξινομήσεων. Για την περίπτωση αυτή, ο πίνακας κόστους που επιλέχτηκε παρουσιάζεται στον Πίνακα 4.3. Σύμφωνα με τον πίνακα αυτόν, η εσφαλμένη ταξινόμηση μιας περίπτωσης απάτης στην κατηγορία των περιπτώσεων μη απάτης θεωρείται 35 φορές σημαντικότερη σε σχέση με την εσφαλμένη ταξινόμηση μιας περίπτωσης μη απάτης στην κατηγορία των περιπτώσεων απάτης.

	Μη απάτη	Απάτη
Μη απάτη	0	1
Απάτη	35	0

Πίνακας 4.3 : Πίνακας κόστους για την περίπτωση SVM - *radial basis function*

Στον πίνακα 4.4 παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εφαρμογή του αλγορίθμου. Όπως μπορούμε να διαπιστώσουμε από τη μελέτη του πίνακα το ποσοστό των περιπτώσεων ορθής ταξινόμησης ενός θετικού παραδείγματος στην κλάση των θετικών (TP RATE) για την κλάση 0 ισούται με 0,78 ενώ το ποσοστό ορθών ταξινομήσεων ενός θετικού παραδείγματος ως αρνητικό (FP RATE) για την κλάση 1 ισούται με 0,22.

CLASS	TP rate	FP rate	Precision	Recall	F-measure	ROC area
0	0.78	0.351	0.972	0.78	0.865	*0.714*
1	*0.649*	*0.22*	*0.158*	*0.649*	*0.254*	*0.714*

Πίνακας 4.4: Πίνακας αποτελεσμάτων για τον ταξινομητή SVM - Πυρήνας RBF

Όπως έχουμε ήδη αναφέρει όταν το εμβαδό κάτω από την καμπύλη ROC είναι ίσο με 1 τότε ο ταξινομητής επιτυγχάνει τέλεια ακρίβεια. Η τιμή που προκύπτει από την εκτέλεση του αλγορίθμου SVM- *με τον πυρήνα RBF* είναι ίση με 0,7143, γεγονός που σημαίνει ότι ο ταξινομητής επιφέρει ικανοποιητικά αποτελέσματα.

ii. *Γραμμικός πυρήνας*

Ο πίνακας κόστους που επιλέχθηκε έχει διαφορετικά στοιχεία σε σχέση με τη προηγούμενη περίπτωση του αλγορίθμου και είναι αυτός που φαίνεται παρακάτω :

	Μη απάτη	Απάτη
Μη απάτη	0	1
Απάτη	9,65	0

Πίνακας 4.5 : Πίνακας κόστους για τον ταξινομητή SVM *–με γραμμικό πυρήνα*

Η εφαρμογή του αλγορίθμου στην περίπτωση αυτή, όπως προκύπτει από τη μελέτη του πίνακα 4.6, έχει σαν αποτέλεσμα το ποσοστό των περιπτώσεων ορθής ταξινόμησης ενός θετικού παραδείγματος στην κλάση των θετικών (TP RATE) για την κλάση 0 και για την κλάση 1 να είναι πολύ κοντά και ισούνται με 0,731 και 0,699 αντίστοιχα. Επιπλέον, ενδιαφέρον παρουσιάζει και το μετρό αξιολόγησης ορθότητα (precision), δηλαδή, η πιθανότητα να ταυτίζεται η κλάση που προβλέπει ένας ταξινομητής για ένα στιγμιότυπο με την πραγματική του κλάση, η οποία ισούται με 0,974 για την κλάση 0 και με 0,142 για την κλάση 1.

CLASS	TP rate	FP rate	Precision	Recall	F-measure	ROC area
0	0.731	0.301	0.974	0.731	0.835	*0.715*
1	*0.699*	*0.269*	*0.142*	*0.699*	*0.236*	*0.715*

Πίνακας 4.6 : Πίνακας αποτελεσμάτων για ταξινομητή SVM *με το γραμμικό πυρήνα*

Το εμβαδό κάτω από την καμπύλη ROC είναι 0.715, γεγονός που σημαίνει ότι ο ταξινομητής επιφέρει ικανοποιητικά αποτελέσματα.

4.3.2 Bagging

Γενικά

Ο όρος bagging είναι συντομογραφία του bootstrap aggregating (Breiman L., 1996). Bootstrap είναι η μέθοδος που χρησιμοποιείται για την αναπαραγωγή των παραδειγμάτων εκπαίδευσης στην περίπτωση όπου το σύνολο εκπαίδευσης είναι μικρό. Δεδομένου ότι το συνολικό πλήθος των παραδειγμάτων εκπαίδευσης είναι n, κάθε νέο σύνολο εκπαίδευσης παράγεται επιλέγοντας n φορές ένα παράδειγμα από το αρχικό σύνολο. Η επιλογή γίνεται με επανατοποθέτηση σύμφωνα με την ομοιόμορφη κατανομή. Με αυτόν τον τρόπο κάποια παραδείγματα εκπαίδευσης μπορεί να εμφανίζονται περισσότερες φορές και άλλα να μην εμφανίζονται καθόλου στο σύνολο εκπαίδευσης (περίπου το 36.8% των παραδειγμάτων - Breiman., 1996). Κάθε νέο τέτοιο σύνολο χρησιμοποιείται ως είσοδος στον αλγόριθμο μάθησης. Με αυτόν τον τρόπο δημιουργείται μια σειρά διαφορετικών μοντέλων τα οποία χρησιμοποιούνται για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής συνδυάζοντας τις επί μέρους προβλέψεις όλων των μοντέλων. Η μέθοδος είναι ιδιαίτερα αποδοτική όταν χρησιμοποιούνται ασταθείς ταξινομητές βάσης με μεγάλη διακύμανση (ταξινομητές οι οποίοι μεταβάλλονται αρκετά σε μικρές διαταραχές του συνόλου εκπαίδευσης). Η μέθοδος δεν είναι επιρρεπής σε φαινόμενα υπερεκπαίδευσης καθώς αυξάνεται το πλήθος των παραγόμενων υποθέσεων.

Παρουσίαση αποτελεσμάτων

Ο πίνακας κόστους για τον αλγόριθμο bagging στον όποιο καταλήξαμε ύστερα από αρκετές επαναλήψεις παρουσιάζεται στον πίνακα 4.7 :

	Μη απάτη	Απάτη
Μη απάτη	0	1
Απάτη	47	0

Πίνακας 4. 7: Πίνακας κόστους για τον ταξινομητή bagging

Στον πίνακα 4.8 παρουσιάζονται τα αποτελέσματα της εφαρμογής του αλγορίθμου bagging. Σύμφωνα με τα δεδομένα που προκύπτουν, το ποσοστό ορθών ταξινομήσεων ενός θετικού παραδείγματος ως αρνητικό (FP RATE) για την κλάση 0 ισούται με 0.271 και για την κλάση 1 με 0.216. Το μέτρο αξιολόγησης ανάκληση (recall) αποτελεί ένα ακόμα σημαντικό μέτρο για την περιγραφή των αποτελεσμάτων και δηλώνει την πιθανότητα που υπάρχει να αναγνωρίσει σωστά ο ταξινομητής εάν ένα στιγμιότυπο ανήκει στην κλάση 0 ή 1. Το ποσοστό που προκύπτει είναι αρκετά υψηλό και για τις δυο κλάσεις 0 και 1 του προβλήματος μελέτης και ανέρχεται σε 0.784 και 0.729 αντίστοιχα.

CLASS	TP rate	FP rate	Precision	Recall	F-measure	ROC area
0	0.784	0.271	0.978	0.784	0.871	*0.842*
1	*0.781*	*0.216*	*0.177*	*0.729*	*0.285*	*0.842*

Πίνακας 4.8 : Πίνακας αποτελεσμάτων για τον ταξινομητή bagging

4.3.3. Τυχαία Δάση (Random Forests)

Γενικά

Τα τυχαία δάση (Breiman, 2001) αποτελούν μια ειδική κατηγορία των συνδυαστικών μεθόδων ταξινόμησης η οποία χρησιμοποιεί για ταξινομητές δέντρα απόφασης (Quinlan, 1986). Για την κατασκευή ενός δέντρου απόφασης ανατίθεται αρχικά στη ρίζα του το σύνολο των δειγμάτων εκπαίδευσης. Κάθε ενδιάμεσος κόμβος περιέχει υποσύνολο των δειγμάτων το οποίο μέσω της εφαρμογής ενός κατάλληλου ελέγχου διαχωρίζεται σε δυο ή περισσότερα μικρότερα υποσύνολα (παιδιά) στο επόμενο επίπεδο. Ο έλεγχος συνήθως αφορά ένα υποσύνολο των χαρακτηριστικών των δειγμάτων εκπαίδευσης. Η επιλογή του καλύτερου διαχωρισμού γίνεται σύμφωνα με ένα κατάλληλο μέτρο όπως π.χ. εντροπία. Τα δέντρα του δάσους αναπτύσσονται στο μέγιστο μέγεθός τους, χωρίς κλάδεμα. Η μέθοδος Bagging χρησιμοποιώντας για ταξινομητές δέντρα απόφασης αποτελεί μια ειδική κατηγορία των Random Forests. Σε αυτή την περίπτωση η τυχαιότητα ενσωματώνεται στο μοντέλο και μέσω της τυχαίας επιλογής N παραδειγμάτων εκπαίδευσης, με επανατοποθέτηση, από το αρχικό σύνολο εκπαίδευσης.

Η διαδικασία ταξινόμησης «άγνωστων» παραδειγμάτων πραγματοποιείται μέσω της διάσχισης των δέντρων του δάσους ξεκινώντας από τη ρίζα και καταλήγοντας σε ένα από τα φύλλα του δέντρου και στη συνέχεια συνδυάζοντας τις προβλέψεις των ταξινομητών σύμφωνα με ένα πλειοψηφικό σύστημα ψηφοφορίας (majority voting scheme). Κάθε παράδειγμα ανατίθεται στην κατηγορία με τη μεγαλύτερη συχνότητα.

Παρουσίαση αποτελεσμάτων

Ο πίνακας κόστους για τον αλγόριθμο randomforest στον όποιο καταλήξαμε ύστερα από αρκετές επαναλήψεις παρουσιάζεται στον πίνακα 4.9 :

	Μη απάτη	Απάτη
Μη απάτη	0	1
Απάτη	210	0

Πίνακας 4. 9: Πίνακας κόστους για τον ταξινομητή randomforest

Τα αποτελέσματα που προκύπτουν από την υλοποίηση του αλγορίθμου randomforest είναι αρκετά ικανοποιητικά εφόσον το FP RATE και το TP RATE για τις δυο κλάσεις είναι αντίστοιχα αρκετά κοντά. Πιο συγκεκριμένα, το FP RATE για την κλάση 0 είναι ίσο με 0.764 ενώ για την κλάση 1 αγγίζει το 0.758 (πίνακας 4.10). Η ικανοποιητική εφαρμογή του αλγορίθμου απεικονίζεται και από την τιμή που προκύπτει για το εμβαδό κάτω από την καμπύλη ROC που είναι ισο με 0.838 (σχήμα 4.5).

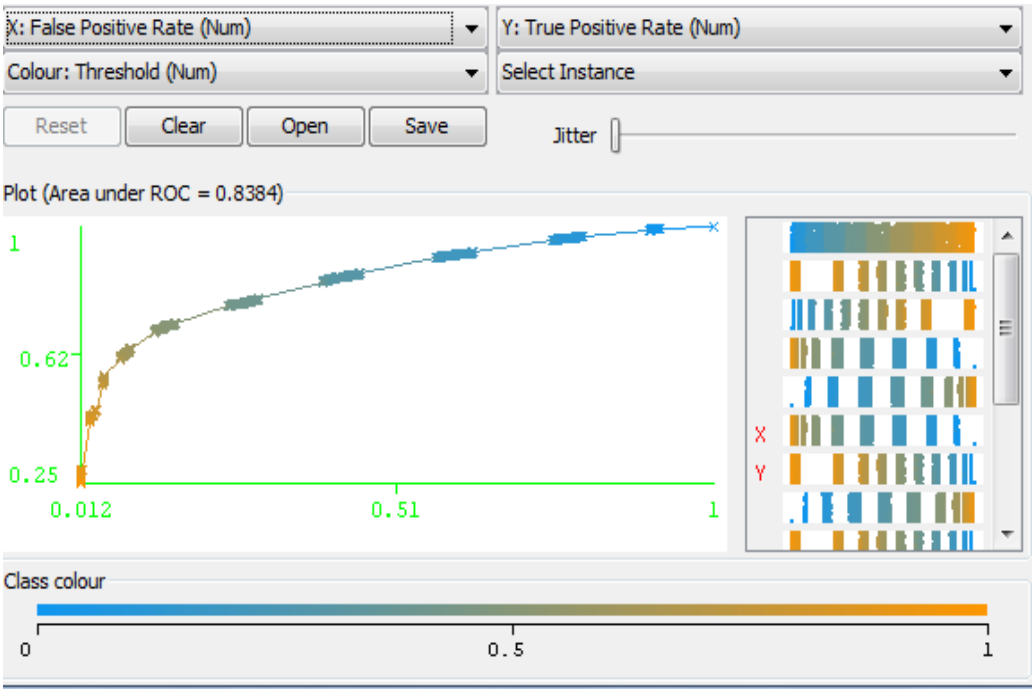
Σημαντικό μέτρο αξιολόγησης αποτελεί και ο πίνακας ταξινόμησης (confusion matrix). Από τον πίνακα 4.11 καταλήγουμε στο συμπέρασμα ότι οι περιπτώσεις που αντιστοιχούν σε ορθή ταξινόμηση ενός θετικού παραδείγματος στην κλάση των θετικών, δηλαδή οι περιπτώσεις μη άπατης που ταξινομήθηκαν επιτυχώς στην κλάση μη απάτης, είναι ίσο 11072 σε ένα σύνολο 15420 περιπτώσεων και ενός αρνητικού στην κλάση των αρνητικών ισούται με 700. Συνολικά, δηλαδή, οι σωστά ταξινομημένες περιπτώσεις φτάνουν τις 11772, και το ποσοστό επιτυχίας του αλγορίθμου ισούται με 76,34%.

CLASS	TP rate	FP rate	Precision	Recall	F-measure	ROC area
0	0.764	0.242	0.98	0.764	0.859	*0.838*
1	*0.758*	*0.236*	*0.17*	*0.758*	*0.277*	*0.838*

Πίνακας 4.10: Πίνακας αποτελεσμάτων για τον ταξινομητή randomforest

	Μη απάτη	Απάτη
Μη απάτη	11072	3425
Απάτη	223	700

Πίνακας 4.11 : Πίνακας ταξινόμησης για τον ταξινομητή randomforest



Σχήμα 4.5 : Εμβαδό κάτω από την καμπύλη ROC για τον ταξινομητή randomforest

4.3.4 Adaboost

Γενικά

Ο αλγόριθμος adaboost, συντομογραφία του Adaptive Boosting, είναι ένας αλγόριθμος μηχανικής μάθησης, που διατυπώθηκε για πρώτη φορά από τους Freund και Schapire (1995). Πρόκειται για ένα μετά-αλγόριθμο και μπορεί να χρησιμοποιηθεί σε συνδυασμό με άλλους αλγορίθμους μάθησης για να βελτιωθεί η απόδοση τους. Ο αλγόριθμος adaboost είναι ευαίσθητος σε δεδομένα με θορυβο και σε ακραίες τιμές. Οι ταξινομητές που χρησιμοποιεί μπορεί να είναι αδύναμοι, δηλαδή, να εμφανίζουν ένα σημαντικό ποσοστό σφάλματος, ωστόσο όσο η εκτέλεσή τους δεν είναι τυχαία (καταλήγοντας σε ένα αποτέλεσμα με ποσοστό λάθους 0,5 για τη δυαδική κατάταξη), το τελικό μοντέλο θα βελτιωθεί. Ακόμα και οι ταξινομητές που εμφανίζουν ποσοστό σφάλματος υψηλότερο από ό, τι θα αναμενόταν από μια τυχαία ταξινόμηση είναι χρήσιμοι, δεδομένου ότι θα έχουν αρνητικούς συντελεστές στον τελικό γραμμικό συνδυασμό των ταξινομητών.

Ο αλγόριθμος adaboost, σε κάθε γύρο, παράγει και καλεί ένα νέο αδύναμο ταξινομητή t = 1…..T. Για κάθε κλήση, η κατανομή των βαρών D(t) ενημερώνεται, υποδεικνύοντας τη σπουδαιότητα των παραδειγμάτων στο σύνολο των δεδομένων για την ταξινόμηση. Σε κάθε γύρο, οι συντελεστές στάθμισης κάθε εσφαλμένου παραδείγματος αυξάνονται, ενώ μειώνονται οι συντελεστές στάθμισης κάθε ορθώς ταξινομημένου παραδείγματος, έτσι ώστε ο νέος ταξινομητής να εστιάσει στα παραδείγματα που έχουν μέχρι στιγμής ξεφύγει από σωστή ταξινόμηση.

Παρουσίαση αποτελεσμάτων

Για την περίπτωση αυτή, ο πίνακας κόστους (πίνακας 4.12) που επιλέχτηκε παρουσιάζεται παρακάτω :

	Μη απάτη	Απάτη
Μη απάτη	0	1
Απάτη	117	0

Πίνακας 4.12 : Πίνακας κόστους για τον ταξινομητή adaboost

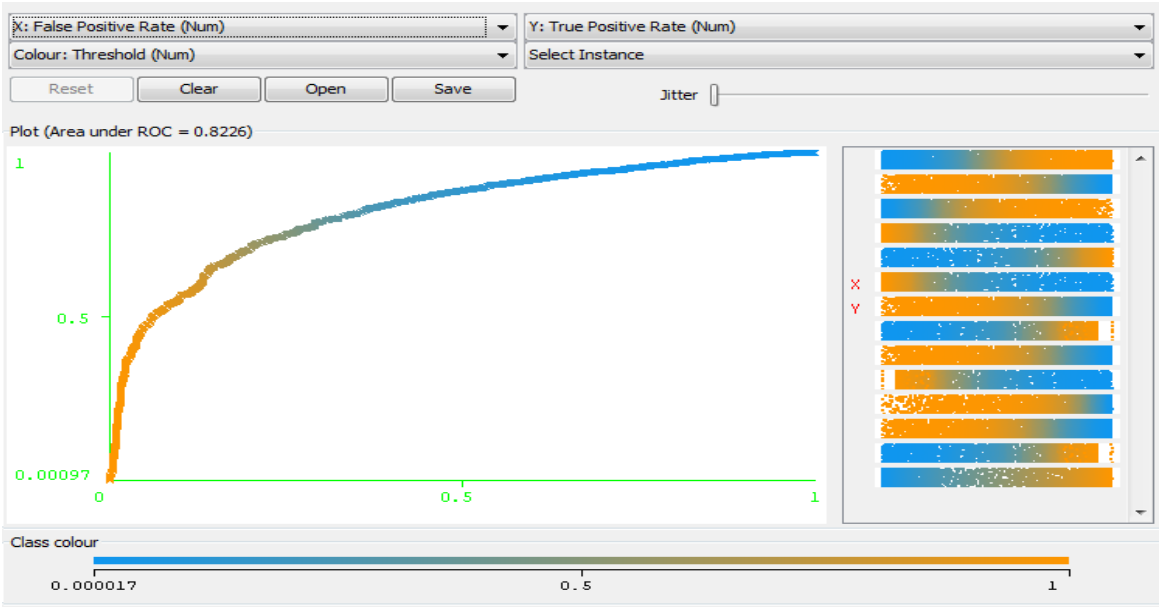
Στον πίνακα 4.13 παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εφαρμογή του αλγορίθμου. Όπως μπορούμε να διαπιστώσουμε από τη μελέτη του πίνακα το ποσοστό των περιπτώσεων ορθής ταξινόμησης ενός θετικού παραδείγματος στην κλάση των θετικών (TP RATE) για την κλάση 0 ισούται με 0,746 ενώ το ποσοστό ορθών ταξινομήσεων ενός θετικού παραδείγματος ως αρνητικό (FP RATE) για την ίδια κλάση ισούται με 0,248.

CLASS	TP rate	FP rate	Precision	Recall	F-measure	ROC area
0	0.746	0.248	0.979	0.746	0.847	*0.823*
1	*0.752*	*0.254*	*0.159*	*0.752*	*0.262*	*0.823*

Πίνακας 4.13: Πίνακας αποτελεσμάτων για τον ταξινομητή adaboost

Οι τιμές που προκύπτουν για τα μέτρα αξιολόγησης ανάκληση και ορθότητα είναι αρκετά ικανοποιητικές και δηλώνουν την επιτυχία του αλγορίθμου. Για την περίπτωση της κλάσης 0 οι τιμές που αντιστοιχούν στην ανάκλαση και την ορθότητα είναι 0.746 και 0.979.

Επιπλέον, ένα ιδιαίτερα ενδιαφέρον αποτέλεσμα προκύπτει για το εμβαδό κάτω από την καμπύλη ROC το όποιο είναι ίσο με 0.823 και θα χαρακτηρίζεται αρκετά υψηλό (σχήμα 4.6).



Σχήμα 4.6 : Εμβαδό κάτω από την καμπύλη ROC για τον ταξινομητή randomforest

4.3.5 Λογιστική παλινδρόμηση

Γενικά

Η λογιστική παλινδρόμηση είναι μια μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης (multivariate statistical analysis) που χρησιμοποιεί ένα σύνολο ανεξαρτήτων μεταβλητών για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής.

Η λογιστική παλινδρόμηση (Logistic Regression) είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη της ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή ενός συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης. Εκτός από την πρόβλεψη ένα μοντέλο λογιστικής παλινδρόμησης δίνει τη δυνατότητα να εκτιμήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής.

Στη λογιστική παλινδρόμηση, σε αντίθεση με την πολλαπλή παλινδρόμηση είναι δυνατό να χρησιμοποιηθούν ως εξαρτημένες μεταβλητές εκτός από αναλογικές αριθμητικές μεταβλητές και κατηγορικές μεταβλητές.

Η πιο διαδεδομένη βιβλιογραφικά έκφραση της λογιστικής παλινδρόμησης είναι : $$\ln\left(\frac{p(x|t=1)}{p(x|t=0)}\right) = b + w^T x \ .$$

Το δεξί μέρος της εξίσωσης δημιουργείται από ένα γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που συμμετέχουν στο μοντέλο παλινδρόμησης και το

αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής εκφράζοντας την πιθανότητα του συμβάντος του γεγονότος.

Παρουσίαση αποτελεσμάτων

Για την περίπτωση αυτή, ο πίνακας κόστους (πίνακας 4.12) που επιλέχτηκε παρουσιάζεται παρακάτω :

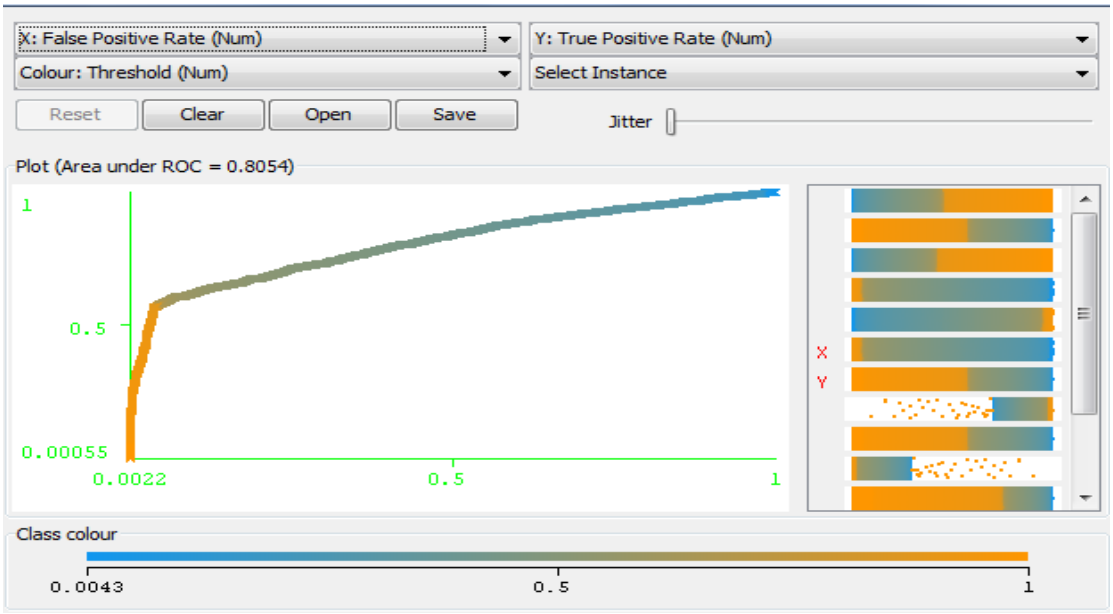
	Μη απάτη	Απάτη
Μη απάτη	0	1
Απάτη	10	0

Πίνακας 4.14 : Πίνακας κόστους για τον ταξινομητή λογιστικής παλινδρόμησης

Η εφαρμογή της λογιστικής παλινδρόμησης επιφέρει ικανοποιητικά αποτελέσματα εφόσον το FP RATE και το TP RATE για τις δυο κλάσεις είναι αντίστοιχα αρκετά κοντά. Πιο συγκεκριμένα, το FP RATE για την κλάση 0 είναι ίσο με 0.256 ενώ για την κλάση 1 αγγίζει το 0.287 (πίνακας 4.15). Τέλος, το σχήμα 4.7 απεικονίζει τη τιμή που λαμβάνει το εμβαδό κάτω από την καμπύλη ROC, η τιμή του οποίου είναι ίση με 0.805.

CLASS	TP rate	FP rate	Precision	Recall	F-measure	ROC area
0	0.713	0.256	0.978	0.713	0.824	*0.805*
1	*0.744*	*0.287*	*0.142*	*0.744*	*0.238*	*0.805*

Πίνακας 4.15: Πίνακας αποτελεσμάτων για τον ταξινομητή λογιστικής παλινδρόμησης



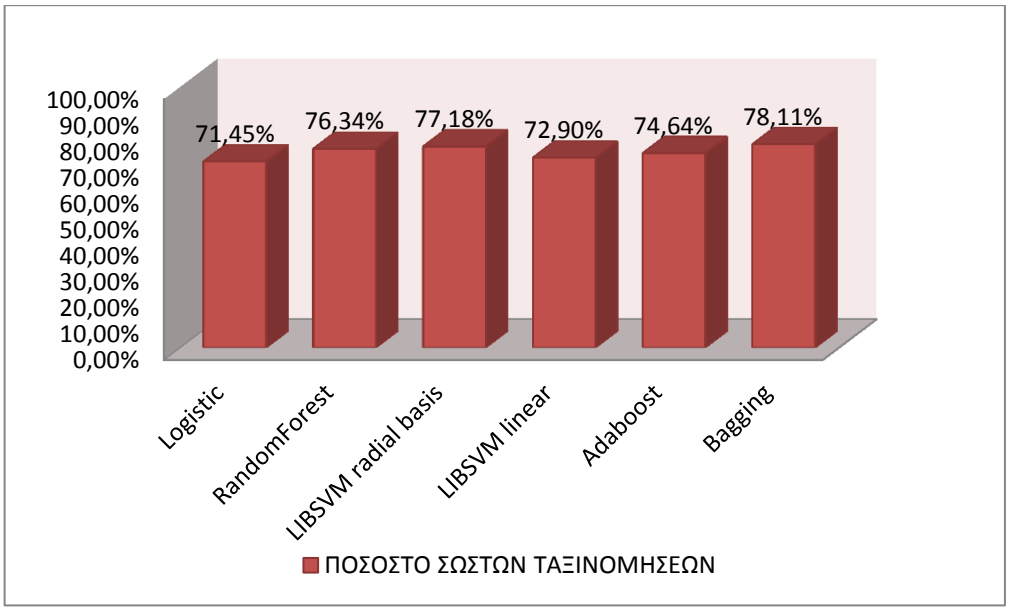
Σχήμα 4.7 : Εμβαδό κάτω από την καμπύλη ROC για τον ταξινομητή λογιστικής παλινδρόμησης

4.4 Σύγκριση αποτελεσμάτων

Στον πίνακα 4.16 και στην γραφική αναπαράσταση των στοιχείων του πίνακα στο σχήμα 4.8 φαίνεται η συγκριτική απόδοση των αλγορίθμων που δοκιμάστηκαν. Από το γράφημα, διαφαίνεται πως η ακρίβεια των μοντέλων που προέκυψαν κυμαίνεται από 71,45% έως και 78,11%. Η ακρίβεια τους, επομένως, μπορεί να θεωρηθεί αρκετά ικανοποιητική. Το μοντέλο που παρουσιάζει τη μεγαλύτερη ακρίβεια είναι αυτό που προκύπτει από την εφαρμογή του αλγορίθμου bagging, ενώ τη μικρότερη αυτό που προκύπτει από τη λογιστική παλινδρόμηση.

ΤΥΠΟΣ ΑΛΓΟΡΙΘΜΟΥ	ΠΟΣΟΣΤΟ ΣΩΣΤΩΝ ΤΑΞΙΝΟΜΗΣΕΩΝ
Logistic	71,45%
RandomForest	76,34%
LIBSVM radial basis	77,18%
LIBSVM linear	72,90%
Adaboost	74,64%
Bagging	78,11%

Πίνακας 4.16 : Απόδοση αλγορίθμων



Σχήμα 4.8 : Συγκριτική γραφική αναπαράσταση αλγορίθμων

Στη συνέχεια παρουσιάζεται ο πίνακας 4.17 στον οποίο συνοψίζονται τα αποτελέσματα που προκύπτουν από την εφαρμογή των αλγορίθμων.

		ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ					
		TP rate	FP rate	Precision	Recall	F-measure	ROC area
ΤΥΠΟΣ ΑΛΓΟΡΙΘΜΟΥ	Logistic	0.713	0.256	0.978	0.713	0.824	0.805
	RandomForest	0.764	0.242	0.98	0.764	0.859	0.838
	LIBSVM radial	0.78	0.351	0.972	0.78	0.865	0.714
	LIBSVM linear	0.731	0.301	0.974	0.731	0.835	0.715
	Adaboost	0.746	0.248	0.979	0.746	0.847	0.823
	Bagging	0.784	0.271	0.978	0.784	0.871	0.842

Πίνακας 4.17 : Αναλυτική σύγκριση αποτελεσμάτων

Η αξιολόγηση με βάσει μόνο την ακρίβεια ταξινόμησης και τα στοιχεία που δίνει ο πίνακας ταξινόμησης δεν αρκεί για την ολοκληρωμένη ανάλυση της προβλεπτικής ικανότητας των αλγορίθμων, γιατί τα στοιχεία αυτά βασίζονται σε συγκεκριμένες υποθέσεις για τη σημασία και το κόστος των εσφαλμένων ταξινομήσεων. Το εμβαδό κάτω από την καμπύλη ROC παρέχει μια ολοκληρωμένη εικόνα για διαφορετικά κόστη εσφαλμένων ταξινομήσεων.

Στο σημείο αυτό αξίζει να αναφέρουμε πως η διαδικασία για την εξαγωγή των διάφορων μοντέλων, εφόσον προηγουμένως έχουμε ορίσει τις επιθυμητές παραμέτρους, είναι μια σύντομη διαδικασία μέσω του προγράμματος Weka, αφού οι περισσότεροι αλγόριθμοι δεν απαιτούν παρά μόνο λίγα δευτερόλεπτα για τη δημιουργία του μοντέλου. Ο χρόνος που απαιτείται για τη δημιουργία του μοντέλου για τους περισσότερου αλγορίθμους κυμαίνεται από 0,40 έως 10 δευτερόλεπτα. Το μοντέλο που χρειάστηκε περισσότερο χρόνο είναι αυτό που πρόεκυψε από την εφαρμογή των μηχανών διανυσμάτων απόφασης και συγκεκριμένα η περίπτωση με το γραμμικό πυρήνα (128,39 sec).

Τέλος, αξίζει να τονιστεί ότι τα αποτελέσματα που προκύπτουν από την εφαρμογή των σύνθετων ταξινομητών (bagging, randomforest, adaboost) είναι αρκετά καλύτερα και στη συνέχεια ακολουθούν τα αποτελέσματα των μηχανών διανυσμάτων υποστήριξης και της λογιστικής παλινδρόμησης.

ΚΕΦΑΛΑΙΟ 5

«ΣΥΜΠΕΡΑΣΜΑΤΑ»

Σύμφωνα με τα επιχειρήματα των ασφαλιστικών φορέων, το κόστος ελέγχου της απάτης είναι ιδιαίτερα υψηλό, δεδομένου ότι στην Ελλάδα δεν υπάρχει ασφαλές σύστημα πρόβλεψης.

Αν δεχθούμε ότι το επιχείρημα των ασφαλιστικών φορέων περί υψηλού κινδύνου των ασφαλίσεων είναι σωστό, θα πρέπει επίσης να σημειώσουμε ότι όλοι οι πελάτες δεν είναι αφερέγγυοι. Ωστόσο, το υψηλό κόστος βαρύνει το σύνολο των πελατών και όχι μόνο τους ασυνεπείς.

Στα πλαίσια της παρούσας εργασίας εφαρμόστηκαν διάφοροι αλγόριθμοι εξόρυξης δεδομένων σε ιστορικά δεδομένα, που περιείχαν στοιχεία πελατών με ασφάλιση στον κλάδο των αυτοκίνητων, και εξήχθησαν μοντέλα που μπορούν να χρησιμοποιηθούν για την αξιολόγηση της φερεγγυότητας του υποψήφιου ως προς την ορθή χορήγηση αποζημίωσης σε περίπτωση ζημιάς. Η ακρίβεια των μοντέλων που προέκυψαν κυμαίνεται από 72% έως και 78%. Επομένως, η ακρίβεια τους μπορεί να θεωρηθεί αρκετά ικανοποιητική και κατά συνέπεια θα μπορούσαν τα μοντέλα αυτά να χρησιμοποιηθούν από διάφορους ελεγκτικούς οργανισμούς για την ταξινόμηση των υποψήφιων πελατών ανάλογα με τη φερεγγυότητά τους κατά τη διαδικασία έγκρισης χορήγησης αποζημίωσης

Μελλοντικά, θα μπορούσαν να γίνουν τα παρακάτω βήματα:

1. Προεπεξεργασία των δεδομένων : Επιλογή των κατάλληλων χαρακτηριστικών και μετασχηματισμός τους με διάφορους τρόπους, δηλαδή χρησιμοποιώντας διάφορα φίλτρα, ανάλογα με τους αλγορίθμους που θα επιλεγούν για να χρησιμοποιηθούν στη συνέχεια.

2. Δημιουργία των μοντέλων ταξινόμησης/ πρόβλεψης : Εκτέλεση εκτενών πειραμάτων με εφαρμογή διαφόρων αλγορίθμων εξόρυξης δεδομένων και συνδυασμού αυτών, ώστε να προκύψουν τα διάφορα μοντέλα πρόβλεψης. Τα μοντέλα αυτά πρέπει να αξιολογηθούν από κάποιον ειδικό του τομέα εφαρμογής και

σε περίπτωση ανακάλυψης γνώσης για το συγκεκριμένο πρόβλημα, είναι απαραίτητη η ερμηνεία των αποτελεσμάτων. Επιπλέον, είναι αναγκαίο να ερευνηθεί αν οι χρησιμοποιούμενες τεχνικές είναι cost sensitive, αν δηλαδή οι μεταβλητές έχουν διαφορετικό κόστος υπολογισμού και ανάκτησης.

Το καλύτερο, βέβαια, θα ήταν να συγκεντρωθούν πρωτογενή δεδομένα, που να ανταποκρίνονται στην ελληνική πραγματικότητα και κατόπιν να εφαρμοστούν τα παραπάνω βήματα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Adomavicious, G., and Tuzhilin, A., "User Profiling in Personalization Applications through Rule Discovery and Validation," Proceedings of the ACM Fifth International Conference on Data Mining and Knowledge Discovery (KDD 99), 377 -381, 1999.

Anderberg M.R. Cluster Analysis for Applications, Academic Press, 1973.

Anderson D., Frivold T., and Valdes A. Next-generation Intrusion Detection Expert System (NIDES) A Summary, Technical Report SRII-CSL-95-97, 1995.

Anderson, J. P. Computer security threat monitoring and surveillance. Technical report, James P. Anderson Co., 1980.

Andritsos P. Scalable Clustering of Categorical Data and Applications, PhD Thesis, University of Toronto, Department of Computer Science, 2004.

Barson, P., S. Field, N. Davey, G. McAskie, and R. Frank. The detection of fraud in mobile phone networks. Working paper University of Hertfordshire, Hatfield, 1996.

Bauer E., Kohavi R., "An empirical comparison of voting classification algorithms: bagging, boosting and variants", Machine Learning, Vol. 36, pp. 1545- 1588, 1999.

Benoit G., Data Mining., Annual Review of Information Science and Technology (ARIST), 36, 265-310, 2002.

Bhargava B., Zhong Y., & Lu Y., Fraud Formalisation and Detection. Proc. Of DaWaK2003, 330-339, 2003.

Bolton R. and D. Hand Statistical fraud detection: A review. Statistical Science 17 (3), 235-255, 2002.

Bolton, R. & Hand, D., Unsupervised Profiling Methods for Fraud Detection. Credit Scoring and Credit Control VII, 2001.

Bonchi F., Giannotti F., Mainetto G., and Pedreschi D., A classifcation based methodology for planning audit strategies in fraud detection. In KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA. ACM Press, 1999.

Bordoni, S., R. Emilia, and G. Facchinetti, Insurance fraud evaluation – a fuzzy expert system. In FUZZ-IEEE, pp. 1491-1494, 2001.

Brause R., Langsdorf, T. & Hepp, M., Neural Data Mining for Credit Card Fraud Detection. Proc. of 11th IEEE International Conference on Tools with Artificial Intelligence, 1999.

Breiman L., “Bagging predictors”, Machine Learning, Vol. 24, pp. 124-140, 1996.

Breiman L., “Random forests”, Machine Learning, 45(1), pp. 5-32, 2001.

Brockett P. L., Xia, X. and Derrig R. A. Using Kohonen’s self-organising feature map to uncover automobile bodily injury claims fraud. The Journal of Risk and Insurance, 245–274., 1998.

Brockett, P., Derrig, R., Golden, L., Levine, A. & Alpert, M., Fraud Classification using Principal Component Analysis of RIDITs. Journal of Risk and Insurance 69(3): 341-371, 2002.

Burge P. and Shawe-Taylor J., An unsupervised neural network approach to profling the behavior of mobile phone users to use in fraud detection. Journal of Parallel and Distributed Computing 61, 915-925, 2001.

Cahill, D. Lambert, J.C. Pinheiro, D.X. Sun, Detecting Fraud in the Real World, Technical Report, Lucent Technologies, 2000

Cahill, M. H., D. Lambert, J. C. Pinheiro, and D. X. Sun. Detecting Fraud in the Real World. In Handbook of Massive Datasets, Kluwer Academic Publishers, pp 911–930, 2002.

Carberry S. “Techniques for Plan Recognition”. User Modeling and User Adapted Interaction, vol. 11, pp 31-48, 2001.

Chan P. and Stolfo S. Toward Parallel and Distributed Learning by Meta-learning, In Working Notes AAAI Work. Knowledge Discovery in Databases, p227-240., 1993.

Chan P., Fan W., Prodromidis A. & Stolfo S., Distributed Data Mining in Credit Card Fraud Detection. IEEE Intelligent Systems 14: 67-74, 1999.

Cortes C. & Pregibon D., Signature-Based Methods for Data Streams. Data Mining and Knowledge Discovery 5: 167-182, 2001.

Cortes C. and Vapnik V.N., “Support-vectors Networks”, Machine learning, 273-297 1995.

Cortes, C., & Pregibon, D., Giga-mining. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (pp. 174-178). New York, NY: AAAI Press, 1998

Cox K., Eick S., and Wills G., Viual data mining: Recognizing telephone calling fraud. Data Mining and Knowledge Discovery 1, 225-231, 1997.

Dalvi N., Domingos P., Mausam, Sanghai P. & Verma, D. Adversarial Classification. Proc.of SIGKDD04, 99-108, 2004.

Davia H. R.,Coggins P., Wideman J., and Kastantin J., Accountant's Guide to Fraud Detection and Control (2 ed.), John Wiley & Sons, 2000.

Denning D. E., Cyberspace attacks and countermeasures. In Internet Besieged (D. E. Denning and P. J. Denning, eds.) 29–55. ACM Press, New York, 1997.

Deshmukh, A. and L. Talluru, A rule based fuzzy reasoning system for assessing the risk of management fraud. Journal of Intelligent Systems in Accounting, Finance & Management 7 (4), 223-241, 1998.

Donato M., Floriana E., Francesca A. L., Learning Recursive Theories with ATRE. ECAI: 435-439, 1998.

Dorronsoro J., Ginel, F., Sanchez, C. & Cruz, C., Neural Fraud Detection in Credit Card Operations. IEEE Transactions on Neural Networks 8(4): 827-834, 1997.

Dunham M. Data mining introduction and advanced topics, New Jersey, Prentice Hall, 2003.

Estevez P., Held C., and Perez C., Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. Expert Systems with Applications 31, 337-344, 2006.

Ezawa K. & Norton S., Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts. IEEE Expert October: 45-51, 1996.

Fan, W., Systematic data selection to mine concept-drifting data streams. Proceedings of SIGKDD04, 128-137, 2004.

Fanning K. and Cogger K., Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management 7, 21-41, 1998.

Fawcett T. and Phua C., Fraud Detection Bibliography Available from: http://liinwww.ira.uka.de/bibliography/Ai/fraud.detection.html/, 2005.

Fawcett, T. & Provost, F., Adaptive Fraud Detection. Data Mining and Knowledge Discovery 1(3): 291-316, 1997.

Fawcett, T. and Provost, F., Activity monitoring: Noticing interesting changes in behavior. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 53–62. ACM Press, New York, 1999.

Fayyad U.M, Uthurusamy R.. Evolving data mining into solutions for insights Communications of the ACM, 45(8):28-31, August 2002.

Fayyad U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

Forest, S. Hofmeyr, and A. Somayaji, “Computer Immunology,” Communications of the ACM, 1996.

Ganti V., Gehrke, J. and Ramakrishnan, R. CACTUS: Clustering Categorical Data Using Summaries, In Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 73-83, ACM Press, 1999.

Gentle, J. E., Elements of Computational Statistics, Springer, 2002.

Ghosh S. & Reilly, D., Credit Card Fraud Detection with a Neural Network. Proc. of 27th Hawaii International Conference on Systems Science 3: 621-630, 1994.

Gibson D., Kleinberg, J. and Raghavan, P. Clustering Categorical Data: An Approach Based on Dynamical Systems, In Proceedings of the 24th International Conference on Very Lerge Data Bases, 311-322, Morgan Kaufmann, 1998.

Glymour C., D. Madigan, D. Pregibon, and P. Smyth, Statistical Themes and Lessons for Data Mining, Data Mining and Knowledge Discovery, 1, 11-28, 1997.

Green, B. P., and J. H. Choi., Assessing the risk of management fraud through neural network technology. Auditing: A Journal of Practice & Theory 16 (Spring): 14–28, 1997.

Guha S., Rastogi R. and Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes. Information Systems, 25, 345-366, 2000.

Han J. and Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufman Publishers, 2001.

Hand D., Mannila, H. and Smyth, P. Principles of Data Mining, MIT Press, Cambridge., 2001.

Hassibi, K. , Detecting pay ment card fraud with neural networks. In Business Applications of Neural Networks (P. J. G. Lisboa, A. Vellido and B. Edisbury, eds.). World Scientific, Singapore, 2000.

Heino J. & Toivonen H., Automated Detection of Epidemics from Usage Logs of a Physicians Reference Database. Proc. of PKDD2003, 180-191, 2003.

Hill T., Base-invariance implies Benford's law, Proc. Amer. Math. Soc., 123:3 (March 1995) 887—895, 1995.

Hollmen J. & Tresp V. Call-based Fraud detection in Mobile Communication Networks using a Hierarchical Regime-Switching Model. Proc. Of Advances in Neural Information Processing Systems, 1998.

Hollmen, J., User Profiling and Classification for fraud detection in mobile communication networks, PhD Dissertation, Helsinki University of Technology,

Finland, 2000. Insurance Information Institute. Facts and Statistics on Auto Insurance, NY, USA, 2003.

James N. Porter, Disk Drives' Evolution, at the 100th Anniversary Conference on Magneti Recording and Information Storage, Santa Clara University, December, 1998.

Johnson R.A. and Wichern D.W, Applied Multivariate Statistical Analysis, Prentice Hall, 1998.

Kirkos E., Spathis C. and Manolopoulos Y., Data Mining Techniques for the Detection of Fraudulent Financial Statements, Expert Systems with Applications, 32,

Kononenko I., Kukar M., "Machine Learning and Data Mining: Introduction to Principles and Algorithms", Horwood, 2007.

Kou Y., Lu C. T., Sirwongwattana S., and Huang Y. P, A Survey of Fraud Detection Techniques, the 2004 IEEE International Conference on Networking, Sensing and Control, 2, 749-754, 2004.

Kumar, S. Classification and detection of computer intrusions. Ph. D. thesis, Purdue University. 1995.

Lane T. & Brodley C., An Empirical Study of Two Approaches to Sequence Learning for Anomaly Detection. Machine Learning 51:73-107, 2003.

Lane, T. AND Brodley, C. E., Sequence matching and learning in anomaly detection for computer security. In Proceedings of the AAAI-97 Workshop on AI Approaches to Fraud Detection and Risk Management (AAAI-97). 43-49, 1997.

Lee W. & Xiang D., Information-theoretic Measures for Anomaly Detection. Proc. of 2001 IEEE Symposium on Security and Privacy, 2001.

Leonard K. J., Detecting credit card fraud using expert systems. Computers and Industrial Engineering 25 103–106, 1993.

Lin J., Hwang M. & Becker, J., A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. Managerial Auditing Journal 18(8): 657-665, 2003.

Lunt, T. F. "IDES: An intelligent system for detecting intruders". In Proceedings of the Symposium on Computer Security (CS'90), Rome, Italy, pp. 110–121, 1990.

Lyman P. and Hal R. V., "How Much Information", www2.sims.berkeley.edu/research/projects/ how-much-info, 2003

MacQueen J.B. Some Methods for Classification and Analysis of Multivariate Observations, In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability, Vol.I: Statistics, 281-297, 1967

Maes S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B., Credit Card Fraud Detection using Bayesian and Neural Networks. Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies, 2002.

Maimon O. and Rokah L. Data Mining and Knowledge Discovery Handbook, Springer., 2005.

Major J. & Riedinger D., EFD: A Hybrid Knowledge/Statistical-based system for the Detection of Fraud. Journal of Risk and Insurance 69(3): 309-324, 2002.

Marques de Sá, J. P. Pattern Recognition: Concepts, Methods and Applications, Springer., 2001.

Mitchell T, M,, Machine Learning, McGraw-Hill Higher Education, 1997.

Moore Law, http://www.intel.com/technology/mooreslaw/index.htm

Moreau Y. & Vandewalle, J., Detection of Mobile Phone Fraud Using Supervised Neural Networks: A First Prototype. Proc. of 1997 International Conference on Artificial Neural Networks, 1065-1070, 1997.

Murad U. & Pinkas G., Unsupervised Profiling for Identifying Superimposed Fraud. Proc. of PKDD99, 1999.

Nigrini M., Fraud Detection: I Have Got Your Number. Journal of Accountancy 187: 79-83, 1999.

Nigrini, M. J., and L. J. Mittermaier. The use of Benford's Law as an aid in analytical procedures. Auditing: A Journal of Practice and Theory 16 (Fall): 52-67, 1997.

Oh, S. H., and W. S. Lee. "An anomaly intrusion detection method by clustering normal user behavior". Computers & Security, vol. 22, no. 7, pp 596 –612, 2003.

Pathak J., Vidyarthi N. & Summers S, A Fuzzy-based Algorithm for Auditors to Detect Element of Fraud in Settled Insurance Claims, Odette School of Business Administration, 2003.

Phua C, Alahakoon D, Lee V, Minority report in fraud detection: classification of skewed data. SIGKEE Explorations 6 (1):50–59, 2004.

Phua, C., Lee, V., Smith-Miles, K. and Gayler, R., A Comprehensive Survey of Data Mining-based Fraud Detection Research. Clayton School of Information Technology, Monash University, 2005.

Piatetsky-Shapiro G., Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, AI Magazine, 11 (5), 68-70, 1991.

Provost, F. and T. Fawcett, Robust Classication for Imprecise Environments'. Machine Learning 42(3), 203-231, 2001.

Pyle D., Data Preparation for Data Mining, Morgan Kaufmann Publishers, San Francisco, USA, 1999.

Quinlan J. R., “Induction of decision trees”, Machine Learning, 1(1), pp. 81-106, 1986.

Rath T., Carreras M. & Sebastiani P., Automated Detection of Influenza Epidemics with Hidden Markov Models. Proc. of IDA2003, 521-532, 2003.

Richardson R., Neural Networks Compared to Statis-tical Techniques, Proceedings of the 1997 IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, New York, N.Y., USA, 89-95, 1997.

Robinson M. E. and Tawn J. A., Statistics for Exceptional Athletics Records, Applied Statistics, 44, 499-511, 1995.

Rosset S., Murad U., Neumann E., Idan Y. & Pinkas G., Discovery of Fraud Rules for Telecommunications - Challenges and Solutions. Proc. of SIGKDD99, 409-413, 1999.

Ryan J., Lin, M. and Miikkukainen R., Intrusion detection with neural networks. In AAAI Workshop on AI Approaches to Fraud Detection and Risk Management 72–79. AAAI Press, Menlo Park, CA., 1997.

Salton, G., The SMART retrieval system: Experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall, 1971.

Senator T., Goldberg H., Wooton J., Cottini M., Khan U., Klinger C., Llamas W., Marrone M. & Wong R., The Financial Crimes Enforcement Network AI System (FAIS). AAAI 16(4): Winter, 21-39, 1995.

Shao H., Zhao H. & Chang G., Applying Data Mining to Detect Fraud Behaviour in Customs Declaration. Proc. of 1st International Conference on Machine Learning and Cybernetics, 1241-1244., 2002.

Stolfo S., Lee W., Chan P., Fan W. & Eskin E., Data Mining-Based Intrusion Detectors: An Overview of the Columbia IDS Project. SIGMOD Record 30(4): 5-14, 2001.

Tan P.N., Steinbach M., Kumar V., “Introduction to Data Mining”, Addison-Wesley, May 2005.

Taniguchi M., Haft M., Hollmen J., and Tresp V., Fraud detection in communication networks using neural and probabilistic methods, in: Proc. 1998 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 2, pp. 1241–1244 , 1998.

Thomas B., Clergue J., Schaad A., and Dacier M., A comparison of conventional and online fraud, SAP Research and Institut Eurecom, Sophia Antipolis, France, 2004.

Viaene, S., Derrig, R. & Dedene, G., A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. IEEE Transactions on Knowledge and Data Engineering 16(5): 612-620, 2005.

Williams, G., Evolutionary Hot Spots Data Mining: AnArchitecture for Exploring for Interesting Discoveries. Proc. of PAKDD99, 1999,

Witten H. and Frank E. ,Data Mining Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann Series in Data Management Systems, Morgan Kaufman Publishers, 2005.

Wüthrich B., Permunetilleke D., Leung S. Cho V., Zhang J., Lam W., Daily Prediction of Major Stock Indices from Textual WWW Data. In Knowledge Discovery From Distributed And Textual Data (1999)

Yamanishi K., Takeuchi J., Williams G. & Milne P., On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. Data Mining and Knowledge Discovery 8: 275-300, 2004.

Yang, W.-S. and S.-Y. Hwang, A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications 31, 56-68, 2006.

Yoshida K., Adachi F., Washio T., Motoda H., Homma T., Nakashima A., Fujikawa H. & Yamazaki K. Density Based Spam Detector. Proc. of SIGKDD04, 486-493, 2004.

Zhang M., Salerno J. & Yu P., Applying Data Mining in Investigating Money Laundering Crimes. Proc. of SIGKDD03, 747-752, 2003.