ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ, ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΔΙΚΤΥΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ ΜΕ ΘΕΜΑ:

« ΔΙΑΤΗΡΗΣΗ ΤΗΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ ΚΑΤΑ ΤΗΝ ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ »

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ
ΒΕΡΥΚΙΟΣ ΒΑΣΙΛΕΙΟΣ

ΕΚΠΟΝΗΣΗ
ΚΑΤΣΑΡΟΥ ΑΛΙΚΗ

ΒΟΛΟΣ 2008

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Βερύκιο Βασίλειο για την εκπόνηση της διπλωματικής μου εργασίας αλλά και για την πολύτιμη βοήθειά του κατά τη διάρκεια της συγγραφής. Επίσης θερμά θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα κ. Γκουλαλά - Διβάνη Άρη για την συνεχή καθοδήγηση και βοήθεια που μου παρείχε .

Τέλος, ευχαριστώ την οικογένειά μου για τη συμπαράσταση και τη στήριξη τους καθώς επίσης και την συμφοιτήτρια και καλή μου φίλη Παππά Κατερίνα για την πολύτιμη βοήθεια της.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Στις μέρες μας ζούμε σε μια εποχή τεράστιας τεχνολογικής εξέλιξης. Η πρόοδος της τεχνολογίας είναι ορατή σε όλους σχεδόν τους τομείς, πόσο μάλλον στην επιστήμη των υπολογιστών. Ένα από τα θετικά επακόλουθα αυτής, είναι η δυνατότητα συλλογής και διατήρησης μεγάλης ποσότητας πληροφοριών σε ηλεκτρονική μορφή. Η κατοχή όμως τέτοιων βάσεων δεδομένων, δεν θα είχε κανένα νόημα αν δεν ήταν εφικτό να εφαρμοστούν πάνω σε αυτές, τεχνικές και αλγόριθμοι για την ανάλυση και επεξεργασία τους. Η αυτοματοποιημένη εξαγωγή συμπερασμάτων από ηλεκτρονικά δεδομένα υλοποιείται από ένα πλήθος τεχνικών και αλγορίθμων που στο σύνολο τους αποτελούν τον τομέα της εξόρυξης γνώσης.

Όπως συνεπάγεται από τα παραπάνω, η εξέλιξη στον τομέα της εξόρυξης γνώσης υπήρξε σημαντική τα τελευταία χρόνια. Εκτός όμως από τις βελτιώσεις και τα πλεονεκτήματα από την αυτοματοποιημένη αυτή εξαγωγή συμπερασμάτων, προκύπτουν και σοβαρά ζητήματα εμπιστευτικότητας των δεδομένων. Με τα ζητήματα αυτά θα ασχοληθούμε και στην παρούσα εργασία. Πιο συγκεκριμένα στην εργασία μελετάται το πρόβλημα της διατήρησης της ιδιωτικότητας κατά την εξόρυξη γνώσης από βάσεις δεδομένων με χρήση της διαδικασίας της κατηγοριοποίησης. Σκοπός της εργασίας είναι αρχικά να γίνει κατανοητό το πρόβλημα. Στο κυρίως μέρος παρουσιάζονται κάποιες προσεγγίσεις που αφορούν την επίλυση του εν λόγω προβλήματος με ιδιαίτερη έμφαση στην διατήρηση της εμπιστευτικότητας με απόκρυψη κανόνων. Τέλος, εισάγεται ένας νέος αλγόριθμος, που στοχεύει στην διατήρηση της ιδιωτικότητας των εξαγόμενων προτύπων από την διαδικασία της κατηγοριοποίησης, με την ελεγχόμενη τροποποίηση κάποιων εγγραφών. Ο αλγόριθμος δοκιμάστηκε σε πειραματικά δεδομένα στα πλαίσια της εργασίας και επιτυγχάνει την απόκρυψη των ευαίσθητων κανόνων.

.

1. ΕΙΣΑΓΩΓΗ

Η εξέλιξη της τεχνολογίας στην σύγχρονη κοινωνία έχει σαν αποτέλεσμα η ανάπτυξη και η διατήρηση βάσεων δεδομένων για ποικίλους λόγους, να αποτελεί πλέον κανόνα. Ο αριθμός των οργανισμών (π.χ. δημόσιοι οργανισμοί, τράπεζες) και επιχειρήσεων (π.χ. σουπερμάρκετ, διαφημιστικές εταιρείες) που διατηρούν στοιχεία σε τέτοια μορφή αυξάνεται συνεχώς. Για παράδειγμα, τα σουπερμάρκετ καταγράφουν σε καθημερινή βάση τις αγορές των καταναλωτών σε βάσεις δεδομένων για την εξαγωγή συσχετίσεων μεταξύ των διάφορων προϊόντων. Ωστόσο, παρά τα πολλά πλεονεκτήματα που μπορεί να προσφέρει η διατήρηση τέτοιων βάσεων δεδομένων, σε ορισμένες περιπτώσεις προκύπτουν σημαντικά προβλήματα εμπιστευτικότητας.

Ο σκοπός για τον οποίο διατηρείται ο τεράστιος αυτός όγκος δεδομένων είναι κυρίως για να αναλυθούν και να χρησιμοποιηθούν για την εξαγωγή κάποιων συμπερασμάτων. Τις περισσότερες φορές η συλλογή των στοιχείων γίνεται με συναίνεση των εμπλεκομένων ατόμων και ο εκάστοτε οργανισμός που κάνει την καταχώρηση παρέχει κάποιου είδους εγγύηση ότι τα δεδομένα θα παραμείνουν εμπιστευτικά. Πολλές φορές όμως, τα δεδομένα μπορεί αργότερα να χρησιμοποιηθούν για δευτερεύοντες σκοπούς ή ακόμα και να πουληθούν σε άλλους οργανισμούς. Υπάρχει λοιπόν ο κίνδυνος τόσο της έκθεσης των ευαίσθητων προσωπικών δεδομένων που μπορεί να καταγράφονται σε τέτοιες συλλογές δεδομένων, όσο και της εξαγωγής προτύπων από τα δεδομένα αυτά, που πιθανώς να έπρεπε να κρατηθούν εμπιστευτικά.

Τα ευαίσθητα δεδομένα μπορεί να περιλαμβάνουν προσωπικά στοιχεία των εμπλεκομένων ατόμων όπως ονοματεπώνυμο, διεύθυνση, αριθμό ταυτότητας. Τα ευαίσθητα πρότυπα είναι συμπεράσματα που μπορούν να εξαχθούν από τα δεδομένα και τα οποία οδηγούν σε έμμεση διαρροή πληροφορίας. Για παράδειγμα, πρότυπα που οδηγούν σε έμμεση αναγνώριση της ταυτότητας ενός ατόμου. Επομένως, μπορούμε να θεωρήσουμε δυο εκδοχές του προβλήματος της διατήρησης της εμπιστευτικότητας κατά τις διαδικασίες εξόρυξης γνώσης. Η πρώτη αφορά τα ευαίσθητα προσωπικά στοιχεία που μπορεί να διατηρούνται στις βάσεις δεδομένων και η δεύτερη αφορά τα διάφορα πρότυπα που μπορούν να εξαχθούν από το σύνολο των δεδομένων.

Στην παρούσα εργασία θα ασχοληθούμε με το πρόβλημα της αξιοπιστίας των δεδομένων στην δεύτερη αυτή εκδοχή του. Για παράδειγμα, ένας οργανισμός θέλει να

δώσει ένα τμήμα της βάσης δεδομένων που διατηρεί σε μια τρίτη μη έμπιστη πηγή προκειμένου αυτό να αναλυθεί. Ο οργανισμός πρέπει να εξασφαλίσει ότι από τα δεδομένα αυτά δεν θα εξαχθούν ευαίσθητα πρότυπα. Για να μπορεί λοιπόν να υπάρξει μια τέτοια εγγύηση ότι τα ευαίσθητα πρότυπα παραμένουν σε κάθε περίπτωση προστατευμένα, έχει υλοποιηθεί ένα πλήθος προσεγγίσεων. Οι διάφορες προσεγγίσεις πρέπει να προστατεύουν επιτυχώς τα ευαίσθητα δεδομένα, ενώ παράλληλα να επιτρέπουν την εφαρμογή τεχνικών εξόρυξης γνώσης και σε κάποιες περιπτώσεις και την στατιστική ανάλυση των δεδομένων. Επίσης, είναι καίριας σημασίας να μπορούν οι προτεινόμενες λύσεις να εφαρμοστούν σε βάσεις ανεξαρτήτως του μεγέθους τους.

Καθώς η εξόρυξη γνώσης υλοποιείται από ένα πλήθος τεχνικών, έχουν προταθεί και υλοποιηθεί προσεγγίσεις, που προσπαθούν να καλύψουν σε όλο του το εύρος τον τομέα της εξόρυξης γνώσης. Στην παρούσα εργασία θα ασχοληθούμε με την διατήρηση της ιδιωτικότητας μόνο όσον αφορά την τεχνική της κατηγοριοποίησης. Πιο συγκεκριμένα, θα ασχοληθούμε με την απόκρυψη ευαίσθητων κανόνων κατηγοριοποίησης που μπορούν να εξαχθούν από μια βάση δεδομένων. Η δομή της εργασίας ακολουθεί την εξής γραμμή. Στο Κεφάλαιο 2 γίνεται ορισμός του προβλήματος της προστασίας της ιδιωτικότητας κατά την κατηγοριοποίηση δεδομένων. Δίνονται επίσης κάποια μέτρα για την αξιολόγηση της απόδοσης των αλγορίθμων που παρουσιάζονται και για την αξιολόγηση του αλγορίθμου που υλοποιείται στο κυρίως μέρος της εργασίας. Τέλος, παρουσιάζεται ένας αριθμός αλγορίθμων που έχουν υλοποιηθεί για να δώσουν λύση στο πρόβλημα. Στο Κεφάλαιο 3 γίνεται μια ακριβής παρουσίαση της νέας τεχνικής που εισάγει η παρούσα εργασία, του Αλγόριθμου Ελάχιστης Τροποποίησης, που αφορά την απόκρυψη ευαίσθητων κανόνων κατηγοριοποίησης με τροποποίηση κάποιων από τις εγγραφές της βάσης δεδομένων. Στο Κεφάλαιο 4 γίνεται η αξιολόγηση της προτεινόμενης προσέγγισης και παρατίθενται τα αποτελέσματα της εφαρμογής της σε πειραματικά δεδομένα. Τέλος, το Κεφάλαιο 5 παρέχει μια γενική ανασκόπηση της προτεινόμενης λύσης αλλά και της εργασίας γενικότερα.

2. ΠΡΟΣΤΑΣΙΑ ΤΗΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ ΚΑΤΑ ΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

Το ζήτημα της προστασίας της ιδιωτικότητας, αφορά όλες τις τεχνικές εξόρυξης γνώσης. Στα πλαίσια της εργασίας θα αναφερθούμε μόνο στην τεχνική της κατηγοριοποίησης. Για την καλύτερη παρουσίαση και κατανόηση της υπόλοιπης εργασίας, πριν παρουσιάσουμε την σχετική βιβλιογραφία, παρατίθενται κάποιες βασικές έννοιες και ορισμοί. Επίσης, δίνεται μια σύντομη περιγραφή των αλγορίθμων κατηγοριοποίησης C4.5 και RIPPER καθώς και τα μέτρα αξιολόγησης που θα χρησιμοποιηθούν στο υπόλοιπο της εργασίας.

2.1 Βασικές Έννοιες

Η προστασία των ευαίσθητων εξαγόμενων προτύπων, στα πλαίσια της εργασίας, γίνεται με απόκρυψη από το αρχικό σύνολο κανόνων κατηγοριοποίησης της βάσης δεδομένων, των κανόνων που θεωρούνται ευαίσθητοι.

Ορισμός 2.1.1: Μια βάση δεδομένων D είναι ένα σύνολο της μορφής {T, A, V, f}, όπου $T=\{t_1, t_2, .., t_n\}$ ένα πεπερασμένο σύνολο εγγραφών, A ένα πεπερασμένο σύνολο γνωρισμάτων τέτοιο ώστε για κάθε γνώρισμα $a \epsilon A$ μπορούμε να συσχετίσουμε ένα σύνολο τιμών V_α, V το σύνολο των τιμών των γνωρισμάτων και f μια συνάρτηση $T \times A \rightarrow V_\alpha$ που επιστρέφει την τιμή ενός γνωρίσματος για μια εγγραφή.

Με χρήση ειδικών αλγορίθμων μπορεί να εξαχθεί από μια βάση το σύνολο των κανόνων που κατηγοριοποιούν τις εγγραφές της σε διακριτές κλάσεις – κατηγορίες. Δηλαδή με χρήση κάποιου αλγορίθμου κατηγοριοποίησης που εφαρμόζεται στο σύνολο της βάσης δεδομένων, οι εγγραφές της ομαδοποιούνται ανάλογα με τις τιμές των γνωρισμάτων τους. Για κάθε τέτοια «ομάδα» εγγραφών εξάγεται ένας κανόνας που να την περιγράφει. Δίνουμε τον ορισμό ενός κανόνα κατηγοριοποίησης καθώς και της ίδιας της διαδικασίας κατηγοριοποίησης με χρήση κανόνων.

Ορισμός 2.1.2: Ένας κανόνας κατηγοριοποίησης R_i που προκύπτει από μια βάση δεδομένων D, είναι μια πρόταση της μορφής $<A_1=a_1>$ and $<A_2=a_2>$ and…and $<A_n=a_n>$

→ c, όπου A_1, A_2,…, A_n είναι γνωρίσματα της βάσης (διακριτά ή συνεχή), a_1, a_2,.., a_n είναι τιμές γνωρισμάτων και c μια τιμή του γνωρίσματος κλάσης C. Ο ακέραιος n καλείται και μήκος του κανόνα R_i.

Ορισμός 2.1.3: Το πρόβλημα της κατηγοριοποίησης με χρήση κανόνων ορίζεται ως η αυτόματη εύρεση του συνόλου (R_1, R_2, …,R_k) των κανόνων της μορφής <A_1=a_1> and <A_2=a_2> and…and <A_n=a_n> → c που κατηγοριοποιούν το σύνολο των εγγραφών της βάσης.

2.2 Ορισμός του Προβλήματος

Στόχος των αλγορίθμων προστασίας της ιδιωτικότητας, στα πλαίσια της τεχνικής της κατηγοριοποίησης, είναι η προστασία των ευαίσθητων προτύπων που μπορεί να προκύψουν κατά την κατηγοριοποίηση των δεδομένων. Για παράδειγμα έστω μια χρηματοπιστωτική εταιρεία που θέλει να κοινοποιήσει μια βάση δεδομένων με εγκεκριμένες και μη αιτήσεις έκδοσης πιστωτικών καρτών και τα ονόματα των αιτούντων, σε ένα τραπεζικό ίδρυμα παροχής στεγαστικών δανείων. Κάθε εγγραφή στην βάση αντιστοιχεί σε έναν αιτούντα. Τα γνωρίσματα που περιλαμβάνονται σε κάθε εγγραφή είναι τα εξής: οικονομική κατάσταση, αριθμός ετών στην παρούσα θέση εργασίας, φύλο, εισοδηματική κλίμακα, διεύθυνση κατοικίας και ηλικία. Το γνώρισμα-κλάση είναι η έγκριση ή όχι της έκδοσης πιστωτικής κάρτας. Η εταιρείας παροχής στεγαστικών δανείων που θα παραλάβει την βάση σκοπεύει να τη χρησιμοποιήσει για τη δημιουργία ενός μοντέλου κατηγοριοποίησης για τους μελλοντικούς πελάτες της. Εφόσον οι δυο επιχειρήσεις «αντιλαμβάνονται» αλλιώς τα στοιχεία της βάσης θεωρητικά δεν υπάρχει κίνδυνος από την κοινοποίηση της βάσης. Ωστόσο, από το σύνολο των δεδομένων μπορούν εξαχθούν ευαίσθητα πρότυπα από τα οποία να επωφεληθεί η εταιρεία στεγαστικών δανείων. Είναι δυνατό να χρησιμοποιήσει της εγγραφές της βάσης για να καθορίσει μελλοντικές ομάδες πελατών ή ακόμα και μεμονωμένα άτομα θέτοντας απλά ως γνώρισμα-κλάση τον ταχυδρομικό κώδικα της διεύθυνσης κατοικίας.

Το πρόβλημα που επιδιώκουμε να λύσουμε ορίζεται στην συνέχεια.

Ορισμός 2.2.1: Έχουμε μια βάση δεδομένων D με ένα σύνολο κατηγοριών – κλάσεων C και ένα σύνολο κανόνων R εξαγόμενων από αυτή. Αν κάποιοι από τους κανόνες της βάσης R' χαρακτηριστούν ως ευαίσθητοι, πρέπει να βρεθεί μια νέα βάση D' από την οποία να είναι δυνατόν να εξαχθούν μόνο οι μη ευαίσθητοι κανόνες R-R'.

Έστω δηλαδή μια βάση δεδομένων, από την οποία εξάγεται με την διαδικασία της κατηγοριοποίησης ένα σύνολο κανόνων. Κάποιοι από τους κανόνες θεωρούνται εμπιστευτικοί και πρέπει να προστατευτούν. Πρωταρχικός σκοπός είναι η απόκρυψη όλων των ευαίσθητων κανόνων, εν συνεχεία να χαθούν όσο το δυνατό λιγότεροι μη ευαίσθητοι κανόνες (false drop rules) και τέλος να δημιουργηθούν όσο το δυνατό λιγότεροι κανόνες φαντάσματα(ghost rules). Οι έννοιες αυτές ορίζονται αναλυτικά στην ενότητα 2.4. Στόχος δηλαδή είναι η κατασκευή μια νέας βάσης δεδομένων με όσο το δυνατόν όμοια χαρακτηριστικά(αριθμός εγγραφών, αριθμός γνωρισμάτων, στατιστικά στοιχεία γνωρισμάτων, αριθμός αγνώστων τιμών) με την αρχική βάση από την οποία όμως να μην είναι δυνατό να εξαχθούν οι ευαίσθητοι κανόνες.

2.3 Επίλυση του Προβλήματος

Γενικά, το θέμα της προστασίας της ιδιωτικότητας κατά την εξόρυξη γνώσης από δεδομένα μπορεί να προσεγγιστεί με δυο τρόπους. Είτε με βάση την στατιστική, είτε με βάση την κρυπτογραφία. Στην πρώτη κατηγορία, στόχος των αλγορίθμων είναι να προστατέψουν τα δεδομένα εφαρμόζοντας κάποιου είδους τροποποίηση (όπως για παράδειγμα γενίκευση, προσθήκη θορύβου). Η τροποποίηση γίνεται με τέτοιο τρόπο ώστε να διατηρούνται τα χαρακτηριστικά της βάσης. Να μην μεταβάλλονται δηλαδή χαρακτηριστικά όπως ο αριθμός των εγγραφών της βάσης δεδομένων, ο αριθμός των γνωρισμάτων, το είδος των γνωρισμάτων (διακριτά, συνεχή) και τα στατιστικά στοιχεία των τιμών γνωρισμάτων(π.χ. μέση τιμή, διακύμανση, ποσοστό αγνώστων τιμών). Στην δεύτερη κατηγορία τα δεδομένα προστατεύονται με χρήση κρυπτογραφικών τεχνικών. Στα ευαίσθητα δεδομένα της βάσης εφαρμόζεται κάποιος κρυπτογραφικός αλγόριθμος ώστε να μην είναι δυνατή η ανάκτηση τους. Ωστόσο, οι αλγόριθμοι που βασίζονται σε κρυπτογραφικές τεχνικές απαιτούν κατά κανόνα μεγαλύτερο χρόνο εκτέλεσης και είναι πιο δύσκολο να εφαρμοστούν σε μεγάλο όγκο δεδομένο. Στην εργασία παρουσιάζονται

αλγόριθμοι που βασίζονται στην στατιστική, ενώ και η υλοποίηση που προτείνουμε αφορά ένα αλγόριθμο τροποποίησης επίσης της πρώτης κατηγορίας.

Ο κεντρικός άξονας πάνω στον οποίο κινούνται οι προσεγγίσεις στον τομέα της απόκρυψης/προστασίας ευαίσθητης γνώσης, είναι η δημιουργία όσο το δυνατόν ακριβέστερων μοντέλων κατηγοριοποίησης που όμως προστατεύουν τα ευαίσθητα δεδομένα. Η λογική στην οποία στηρίζονται είναι ότι τα πρότυπα δεν προκύπτουν από ξεχωριστές εγγραφές της βάσης αλλά από τα γενικά χαρακτηριστικά της. Αν τα χαρακτηριστικά αυτά διατηρηθούν τότε η βάση δεδομένων δεν χάνει τη χρησιμότητα της. Η αξιολόγηση των εν λόγω προσεγγίσεων θα πρέπει να γίνεται λαμβάνοντας υπόψη την διατήρηση τόσο της εμπιστευτικότητας των δεδομένων όσο και της χρησιμότητας της βάσης.

2.4 Μέτρα Αξιολόγησης

Η αξιολόγηση της απόδοσης των διαφόρων προσεγγίσεων που καλούνται να δώσουν λύση στο πρόβλημα της προστασίας των δεδομένων κατά την εφαρμογή τεχνικών εξόρυξης γνώσης είναι εξίσου σημαντική [1]. Είναι βασικό οι προτεινόμενοι αλγόριθμοι να έχουν ρεαλιστικούς χρόνους εκτέλεσης, να διατηρούν τη λειτουργικότητα της βάσης δεδομένων, να προστατεύουν αποτελεσματικά τα ευαίσθητα δεδομένα και τέλος να είναι δυνατή η αντίσταση τους σε διαφορετικούς αλγόριθμους εξόρυξης. Όσο καλύτερες είναι οι επιδόσεις τους όσον αφορά τα παραπάνω κριτήρια, τόσο πιο αποτελεσματικός και επομένως προτιμότερος είναι ο αλγόριθμος.

2.4.1 Μέτρα Διασφάλισης της Ποιότητας της Παραγόμενης Βάσης

Για την μέτρηση της αποτελεσματικότητας και της ποιότητας των αλγορίθμων όσον αφορά την διατήρηση της ποιότητας της βάσης δεδομένων, τα μέτρα που έχουν προταθεί μπορούν να χωριστούν σε δυο μεγάλες κατηγορίες.

- Γενικά μέτρα. Στην κατηγορία αυτή εντάσσονται τα μέτρα που καταγράφουν το ποσοστό τροποποίησης της αρχικής βάσης δεδομένων σε σχέση με την βάση που τελικά θα δημοσιοποιηθεί.

- Μέτρα προσανατολισμένα στην κάθε ξεχωριστή εργασία εξόρυξης. Τα μέτρα που ανήκουν στην κατηγορία αυτή λαμβάνουν υπόψη τους την εκάστοτε διαδικασία εξόρυξης που λαμβάνει χώρα και πιστοποιούν την ποιότητα του αλγορίθμου ανάλογα με την ακρίβεια των αποτελεσμάτων. Όσο πιο ακριβή είναι τα αποτελέσματα τόσο λιγότερο θεωρείται ότι χάνεται η λειτουργικότητα της βάσης.

2.4.2 Μέτρα Προστασίας της Ιδιωτικότητας κατά την Εξόρυξη Κανόνων Κατηγοριοποίησης

Οι αλγόριθμοι που στοχεύουν στην προστασία των δεδομένων κατά την εφαρμογή τεχνικών κατηγοριοποίησης είναι επιθυμητό να διατηρούν τα δεδομένα σε τέτοιο επίπεδο, ώστε να είναι μεν δυνατή η δημιουργία ενός όσο το δυνατόν ακριβέστερου κατηγοριοποιητή αλλά συγχρόνως να επιτυγχάνεται η προστασία των ευαίσθητων δεδομένων που μπορεί να περιλαμβάνονται.

Για τον έλεγχο της αποτελεσματικότητας του Αλγορίθμου Ελάχιστης Τροποποίησης που θα παρουσιάσουμε, θα χρησιμοποιήσουμε για να μετρήσουμε τις «παρενέργειες» που υφίσταται το αρχικό σύνολο κανόνων τα εξής δυο μέτρα. Τον αριθμό των απολεσθέντων μη ευαίσθητων κανόνων (false drop rules) που εμφανίζονται και τον αριθμό των κανόνων «φαντασμάτων» (ghost rules).

Ορισμός 2.4.2.1: Με τον όρο απολεσθέντες μη ευαίσθητοι κανόνες (false drop rules) καλούνται οι μη-ευαίσθητοι κανόνες που υπήρχαν στο σύνολο κανόνων κατηγοριοποίησης της αρχικής βάσης δεδομένων αλλά μετά την εφαρμογή του αλγορίθμου δεν εμφανίζονται στο νέο σύνολο κανόνων της τροποποιημένης βάσης.

Ορισμός 2.4.2.2: Με τον όρο κανόνες «φαντάσματα» (ghost rules) καλούνται οι μη-ευαίσθητοι κανόνες που δεν εμφανίζονται στο σύνολο κανόνων κατηγοριοποίησης της αρχικής βάσης δεδομένων αλλά μετά την εφαρμογή του αλγορίθμου εμφανίζονται στο νέο σύνολο κανόνων της τροποποιημένης βάσης.

Στις περισσότερες από τις μελέτες που παρουσιάστηκαν, εισάγονται νέα μέτρα για την αξιολόγηση των αλγορίθμων που περιγράφουν. Τα μέτρα αυτά είναι προσαρμοσμένα στην φύση της κάθε προσέγγισης, ώστε να μπορέσουν να αξιολογήσουν όσο το δυνατόν αποτελεσματικότερα τον κάθε αλγόριθμο.

Η επιλογή των αλγορίθμων που θα χρησιμοποιηθούν σε κάθε περίπτωση για την διασφάλιση των δεδομένων, πρέπει να γίνεται με βάση την απόδοση τους στα κριτήρια που θεωρούνται κρίσιμα στην παρούσα περίπτωση. Με τον τρόπο αυτό επιτυγχάνεται το καλύτερο αποτέλεσμα όσον αφορά το εκάστοτε πρόβλημα διατήρησης της εμπιστευτικότητας κατά την εξαγωγή δεδομένων με χρήση της διαδικασίας της κατηγοριοποίησης.

2.5 Αλγόριθμοι Εξόρυξης Κανόνων Κατηγοριοποίησης

Τόσο στην σχετική βιβλιογραφία όσο και στα πειράματα του Κεφαλαίου 4 χρησιμοποιούνται οι αλγόριθμοι κατηγοριοποίησης C4.5[2], [3], [5] και RIPPER [4], [5]. Παρατίθεται μια σύντομη περιγραφή τους.

2.5.1 C4.5

Ο C4.5 είναι ένας αλγόριθμος δένδρου απόφασης. Κατασκευάζει δηλαδή ένα δένδρο απόφασης με βάση το οποίο γίνεται η κατηγοριοποίηση των εγγραφών.

Ορισμός 2.5.1: Έχουμε μια βάση δεδομένων D με ένα σύνολο γνωρισμάτων $A=\{A_1, A_2, \ldots, A_n\}$ και ένα σύνολο κατηγοριών – κλάσεων $C=\{C_1, C_2, \ldots, C_m\}$. Ένα δέντρο απόφασης ή κατηγοριοποίησης είναι ένα δέντρο που σχετίζεται με το D και έχει τις εξής ιδιότητες. Κάθε εσωτερικός κόμβος παίρνει το όνομα του από ένα γνώρισμα A_i. Κάθε ακμή παίρνει το όνομα μιας τιμής από το γνώρισμα του πατέρα – κόμβου. Και τέλος κάθε φύλλο έχει ως όνομα μια κλάση – κατηγορία C_j.

Η στρατηγική που εκτελείται από τον C4.5 για την κατασκευή του δέντρου είναι η επιλογή κάθε φορά ως γνωρίσματος διάσπασης αυτό με το μεγαλύτερο κέρδος πληροφορίας.

Ορισμός 2.5.2: Δεδομένης μιας κατάστασης μιας βάσης δεδομένων D, η εντοπία H(D) βρίσκει την ποσότητα της τάξης ή της έλλειψης αυτής σε αυτήν την κατάσταση. Έστω οι πιθανότητες p_1, p_2,…,p_S όπου $\sum^{S}_{i=1}p=1$και plog(1/p) η αναμενόμενη πληροφορία με βάση την πιθανότητα ενός περιστατικού, η εντροπία ορίζεται ως:

$$H(D)=H(p_1, p_2,\ldots,p_S)= \sum^{S}_{i=1}(p_i \log(1/p_i))$$

Η εντροπία μπορεί να πάρει τιμές που κυμαίνονται μεταξύ του 0 και του 1. Η έλλειψη «αταξίας» σε ένα σύνολο μεγιστοποιεί το κέρδος πληροφορίας.

Ορισμός 2.5.3: Το κέρδος πληροφορίας είναι η διαφορά στην εντροπία κατά την μετάβαση από μια κατάσταση σε μια άλλη με την επιλογή ενός γνωρίσματος διάσπασης.

$$IG(Ex,α) = H(Ex) - H(Ex|α)$$

Όπου με IG συμβολίζεται το κέρδος πληροφορίας, με H(Ex) η αρχική εντροπία και με H(Ex|α) η εντροπία που προκύπτει αν χρησιμοποιηθεί ως γνώρισμα διάσπασης το α.

Ο C4.5 μπορεί να εφαρμοστεί τόσο σε διακριτά όσο και συνεχή δεδομένα. Επίσης, μπορεί να χειριστεί ελλιπή δεδομένα. Τέλος, στα πλαίσια του αλγορίθμου μπορούν να εφαρμοστούν διαδικασίες κλαδέματος (αντικατάστασης υποδέντρου, ανύψωσης υποδέντρου) δίνοντας σαν αποτέλεσμα ένα μικρότερου μεγέθους δένδρο κατηγοριοποίησης. Το αποτέλεσμα της κατηγοριοποίησης μπορεί να ληφθεί και σε μορφή συνόλου κανόνων, μια δυνατότητα που θα αξιοποιήσουμε και στο πειραματικό μέρος της εργασίας.

2.5.2 RIPPER

Ο RIPPER είναι ένας αλγόριθμος εξαγωγής κανόνων από ένα ήδη κατηγοριοποιημένο σύνολο δεδομένων εκπαίδευσης . Η δημιουργία των κανόνων γίνεται όπως και στον C4.5 με βάση το κέρδος πληροφορίας. Ο RIPPER δημιουργεί διαδοχικά τους κανόνες μέχρι να οδηγηθεί στο τελικό σύνολο κανόνων κατηγοριοποίησης. Σε κάθε κανόνα προστίθενται συνθήκες με βάση ένα μέτρο που στηρίζεται στην έννοια του κέρδους πληροφορίας. Μπορεί επίσης να εφαρμοστεί σε συνεχή δεδομένα. Τέλος, έχει

πολύ καλή απόδοση από άποψη χρόνου εκτέλεσης σε σχέση με άλλους αλγορίθμους εξαγωγής κανόνων κατηγοριοποίησης ακόμα και για μεγάλες βάσεις δεδομένων.

Στο υπόλοιπο του κεφαλαίου παρατίθενται ορισμένες προηγούμενες μελέτες που καλούνται να λύσουν το ζήτημα της διατήρησης της ιδιωτικότητας κατά την κατηγοριοποίηση γενικά αλλά και την εξαγωγή κανόνων κατηγοριοποίησης ειδικότερα

2.6 Σχετική Βιβλιογραφία

Η κατηγοριοποίηση δεδομένων αποτελεί μια από τις πιο διαδεδομένες τεχνικές εξόρυξης γνώσης. Για την διασφάλιση της εμπιστευτικότητας των δεδομένων κατά την κατηγοριοποίηση, είτε πρόκειται για ευαίσθητα προσωπικά στοιχεία είτε για εξαγόμενα πρότυπα έχουν υλοποιηθεί ποικίλες προσεγγίσεις. Στην συνέχεια παρουσιάζονται ενδεικτικά κάποιες μελέτες που επιδιώκουν την διασφάλιση του εν λόγω ζητήματος.

2.6.1 Αλγόριθμοι Προστασίας της Ιδιωτικότητας Κατά την Κατηγοριοποίηση

Η κατηγοριοποίηση των δεδομένων μιας βάσης μπορεί να γίνει με διάφορες μεθόδους όπως είναι τα δέντρα απόφασης (decision trees), οι κανόνες απόφασης (decision rules), τα νευρωνικά δίκτυα (neural networks), ταξινομητές βασισμένοι σε στατιστικές μεθόδους (Bayesian classifiers, Παλινδρόμηση) και τα μέτρα απόστασης . Έχουν προταθεί και υλοποιηθεί αλγόριθμοι που επιλύουν το πρόβλημα της διατήρησης της ιδιωτικότητας των δεδομένων για όλες τις παραπάνω μεθόδους.

Μια αξιόλογη προσέγγιση για την προστασία της εμπιστευτικότητας των δεδομένων κατά την διαδικασία της κατηγοριοποίησης με χρήση δέντρων απόφασης προτείνουν και οι Jaideep Vaidyal και Chris Clifton [8]. Ο αλγόριθμος που υλοποιείται αποτελεί ουσιαστικά μια ασφαλή έκδοση του ID3 όσον αφορά την διατήρηση της ιδιωτικότητας. Σε αντίθεση με τις προσεγγίσεις των οι LiWu Chang και James Tracy [6] και [7] , ο αλγόριθμος της μελέτης [8] επιτυγχάνει την προστασία των ευαίσθητων δεδομένων.

Πιο αναλυτικά, ο αλγόριθμος που παρουσιάζεται δημιουργεί ένα ID3 δέντρο κατηγοριοποίησης που διατηρεί όμως την εμπιστευτικότητα των ευαίσθητων δεδομένων. Η προσέγγιση αφορά καθέτως κατανεμημένα δεδομένα. Στον συγκεκριμένο αλγόριθμο, θεωρείται ότι κάθε τοποθεσία γνωρίζει μόνο τα γνωρίσματα που διατηρεί και ότι το

γνώρισμα-κλάση διατηρείται σε μια μόνο τοποθεσία. Ένα από τα μειονεκτήματα προηγούμενων αλγορίθμων ήταν ότι το γνώρισμα κλάση έπρεπε να είναι γνωστό σε όλες τις τοποθεσίες. Επομένως, ο εν μελέτη αλγόριθμος αποτελεί βελτίωση των προηγούμενων προσεγγίσεων αφού μπορεί να υλοποιήσει και αυτή την περίπτωση και μάλιστα με καλύτερη απόδοση.

Η τοποθέτηση μιας συναλλαγής σε μια κατηγορία γίνεται με τον ίδιο τρόπο όπως και στον κλασικό ID3 με την μόνη διαφορά ότι οι κόμβοι όπως και τα γνωρίσματα είναι κατανεμημένοι. Αρχικά, με την χρήση της μιας βοηθητικής συνάρτησης υπολογίζεται η κλάση πλειοψηφίας, δηλαδή η κλάση που κατηγοριοποιεί την πλειοψηφία των εγγραφών, καθώς και η κατανομή των εγγραφών στις διάφορες κλάσεις. Η συνάρτηση αυτή δίνει επίσης την δυνατότητα στην κάθε τοποθεσία να υπολογίσει τον αριθμό των εγγραφών της που ικανοποιούν τους εκάστοτε περιορισμούς του υπό κατασκευή δέντρου, ώστε να είναι δυνατόν να γίνουν οι παραπάνω υπολογισμοί. Αν όλες οι συναλλαγές που ικανοποιούν τους περιορισμούς του συγκεκριμένου σταδίου ανήκουν στην ίδια κλάση τότε έχουμε ένα κόμβο φύλλο στο δέντρο κατηγοριοποίησης. Σε αντίθετη περίπτωση υπολογίζεται το καλύτερο γνώρισμα για διάσπαση ακριβώς όπως και στον κλασικό ID3. Ο κόμβος του δέντρου κάθε φορά μπορεί να είναι μια από τις τοποθεσίες όπου φυλάσσονται τα γνωρίσματα. Με επανάληψη της παραπάνω διαδικασίας παίρνουμε το δέντρο κατηγοριοποίησης.

Η κατηγοριοποίηση ξεκινά από τον κόμβο που την ζήτησε και ο έλεγχος περνά με μια απλή δεικτοδότηση από τοποθεσία σε τοποθεσία. Με τον τρόπο αυτό επιτυγχάνεται η διατήρηση της εμπιστευτικότητας των δεδομένων αφού κάθε τοποθεσία γνωρίζει τις τιμές της συναλλαγής για τους κόμβους που διατηρούνται σε αυτήν. Δεν γνωρίζει ωστόσο τις τιμές των γνωρισμάτων της συναλλαγής για κανέναν από τους άλλους κόμβους που διατηρούνται σε άλλες τοποθεσίες.

Για παράδειγμα, έστω ότι έχουμε δυο τοποθεσίες στις οποίες κρατούνται διαφορετικές μετεωρολογικές μετρήσεις. Στην πρώτη συλλέγονται μετρήσεις που αφορούν την υγρασία της ατμόσφαιρας και την ένταση των ανέμων ενώ στην δεύτερη συλλέγονται δεδομένα που αφορούν την διακύμανση της θερμοκρασίας και τις μεταβολές στα σύννεφα. Στην δεύτερη τοποθεσία διατηρείται επίσης το γνώρισμα κλάση που μπορεί να πάρει τις τιμές ΝΑΙ και ΟΧΙ. Έστω ότι θέλουμε να χρησιμοποιήσουμε τις

παραπάνω μετρήσεις για να απαντήσουμε στο ερώτημα «αν η σημερινή μέρα είναι καλή για τένις». Καμία από τις δυο τοποθεσίες δεν προτίθεται να αποκαλύψει τα δεδομένα της στην άλλη. Ωστόσο και οι δυο επιθυμούν να συνεργαστούν ώστε να μπορέσουν να δώσουν μια ικανοποιητική απάντηση στο παραπάνω ερώτημα. Το δέντρο κατηγοριοποίησης που απαιτείται μπορεί να προκύψει με εφαρμογή του παραπάνω αλγορίθμου. ■

Δεν έχει γίνει πραγματική υλοποίηση του αλγορίθμου. Ωστόσο στην μελέτη παρουσιάζεται και μια σειρά προτάσεων που αποδεικνύουν την αποτελεσματικότητα της παραπάνω προσέγγισης όσον αφορά την διατήρηση της εμπιστευτικότητας των δεδομένων που κρατούνται στις διάφορες τοποθεσίες. Τέλος, έχει υλοποιηθεί μια αντίστοιχη ασφαλής έκδοση του ID3 που αφορά δεδομένα με οριζόντια κατανομή.

Με την προστασία της ιδιωτικότητας κατά την κατηγοριοποίηση με χρήση μέτρων απόστασης ασχολούνται οι Keke Chen και Ling Liu στο [9]. Εισάγουν μια τροποποίηση των δεδομένων που όμως γίνεται έτσι, ώστε να μην μεταβάλλεται η «απόσταση» μεταξύ τους. Με τον τρόπο αυτό επιτυγχάνουν την διατήρηση της αποδοτικότητας των κατηγοριοποιητών που στηρίζονται στην συγκεκριμένη τεχνική κατηγοριοποίησης, επιλύοντας όμως ταυτόχρονα και το πρόβλημα της ιδιωτικότητας των στοιχείων της βάσης.

Σε αντίθεση με τους υπόλοιπους αλγορίθμους, όπου έπρεπε να βρεθεί μια ισορροπία ανάμεσα στην προστασία των δεδομένων και στην διατήρηση της λειτουργικότητας της βάσης δεδομένων στην συγκεκριμένη προσέγγιση τα δυο αυτά μεγέθη δεν αλληλεξαρτώνται. Αυτό συμβαίνει γιατί ο αλγόριθμος εισάγει μια περιστροφική τροποποίηση των δεδομένων που όμως αφού δεν μεταβάλλει την απόσταση έχει μηδενική απώλεια πληροφορίας για ορισμένες ομάδες κατηγοριοποιητών (KNN, SVM, Perceptrons). Παρακάτω μελετάμε την εφαρμογή του αλγορίθμου για την κατηγοριοποίηση που βασίζεται στην μέθοδο των K πλησιέστερων γειτόνων (KNN – K nearest neighbors).

Ό αλγόριθμος που παρουσιάζεται χρησιμοποιεί μια περιστροφική τροποποίηση των δεδομένων που όμως δεν μεταβάλει τα γεωμετρικά γνωρίσματα της βάσης. Για την κατανόηση της λειτουργίας του αλγορίθμου παρατίθενται μερικές βασικές έννοιες.

Θεωρούμε το σύνολο δεδομένων X στη μορφή $X=[x_1,x_2,\ldots x_N]$ με N γραμμές και d στήλες όπου x_i είναι μια πλειάδα της βάσης. Κάθε πλειάδα ανήκει σε μια προκαθορισμένη κλάση που ορίζεται από το γνώρισμα κλάση y το οποίο δεν θεωρείται εμπιστευτικό.

Ως μια περιστροφική τροποποίηση ενός συνόλου δεδομένων X μπορούμε να ορίσουμε την συνάρτηση g(X) = RX όπου R είναι ένας πίνακας περιστροφής d×d. Μια βασική ιδιότητα ενός πίνακα περιστροφής είναι η ορθοκανονικότητα. Δηλαδή $R^TR=RR^T=I$.

Το βασικό σε έναν περιστροφικό μετασχηματισμό είναι η διατήρηση του μήκους. Αν ορίσουμε ως μήκος του διανύσματος x το $\|x\|=x^Tx$ τότε από τον ορισμό του πίνακα περιστροφής έχουμε $\|Rx\| =\|R\| \|x\|= R^TR \|x\|=I \|x\|=\|x\|$. Επομένως η περιστροφή διατηρεί και την Ευκλείδεια απόσταση μεταξύ δύο σημείων x και y αφού $\|R(x-y)\|=\|R\| \|x-y\|=\|x-y\|$. Μπορούμε λοιπόν λαμβάνοντας υπόψη τα παραπάνω και τον ορισμό την KNN κατηγοριοποίησης να κατανοήσουμε γιατί δεν επηρεάζεται η απόδοση της εν λόγω τεχνικής.

Στην μελέτη εισάγεται επίσης ένα νέο μέτρο για την μέτρηση της ποιότητας της ασφάλειας των δεδομένων που καλείται Φ και ορίζεται ως Φ = (p,w). Όπου με $p=(p_1,p_2,\ldots p_d)$ συμβολίζουμε το διάνυσμα που μετρά την εμπιστευτικότητα των γνωρισμάτων και με $w=(w_1,w_2,\ldots w_d)$ τα βάρη εμπιστευτικότητας που αντιστοιχούν στο κάθε γνώρισμα. Με την χρήση του μέτρου αυτού μπορούμε να αξιολογήσουμε τον βαθμό ασφάλειας που παρέχεται από μια τροποποίηση και συνεπώς να καθορίζουμε την τροποποίηση που παρέχει τον μεγαλύτερο βαθμό προστασίας των δεδομένων.

Ο αλγόριθμος αποδεικνύεται μέσα από πειραματικές μετρήσεις εξαιρετικά αποτελεσματικός. Όχι μόνο εγγυάται την προστασία των δεδομένων αλλά δεν έχει και καμία απώλεια πληροφορίας όσον αφορά την μοντελοποίηση των δεδομένων. Η παραπάνω τεχνική, της περιστροφικής τροποποίησης μπορεί να εφαρμοστεί και για την προστασία δεδομένων κατά την κατηγοριοποίηση με χρήση νευρωνικών δικτύων.

Με μια άλλη τεχνική κατηγοριοποίησης, την παλινδρόμηση, ασχολούνται οι Wenliang Du, Yunghsiang S.Han και Shigang Chen [10]. Προσπαθούν να εισάγουν την έννοια της ασφάλειας κατά την κατηγοριοποίηση δεδομένων με χρήση γραμμικής

παλινδρόμησης. Θεωρούν δυο ανεξάρτητους κατόχους Α και Β, καθένας από τους οποίους διαθέτει κάποια στοιχεία που όμως δεν είναι διατεθειμένος να τα αποκαλύψει ούτε στον άλλο ούτε όμως και σε κάποια τρίτη έμπιστη πηγή. Οι Α και Β λοιπόν θέλουν να εφαρμόσουν την τεχνική της γραμμικής παλινδρόμησης στο σύνολο των δεδομένων τους χωρίς όμως να χρειαστεί να τα αποκαλύψουν.

Πιο συγκεκριμένα το πρόβλημα της προστασίας της ιδιωτικότητας κατά την εφαρμογή της τεχνικής της γραμμικής παλινδρόμησης όπως περιγράφεται στην εν λόγω μελέτη μπορεί να οριστεί ως εξής. Έστω δυο ιδιοκτήτες βάσεων δεδομένων, από τους οποίους ο ένας διατηρεί μια βάση Α και ο άλλος μια βάση Β που τις γνωρίζουν μόνο αυτοί. Έστω V_{ai} τα δεδομένα που γνωρίζει μόνο ο Α και V_{bi} τα δεδομένα που γνωρίζει μόνο ο Β. Η διαδικασία της γραμμικής παλινδρόμησης θα εφαρμοστεί στο σύνολο δεδομένων M = (A:B:Y) και θα δείξει πώς μεταβάλλονται οι τιμές της ανεξάρτητης μεταβλητής Y σε σχέση με τις τιμές των Α και Β.

Για να επιλύσουν το παραπάνω πρόβλημα οι Wenliang Du και Yunghsiang S.Han εισάγουν ένα νέο μοντέλο ασφαλείας που προσεγγίζει το πρωτόκολλο SMC (Secure Multi-party Computation) [11]. Ωστόσο το πρωτόκολλο που εισάγεται, προκειμένου να γίνει πιο αποδοτικό δεν επιτυγχάνει το ίδιο επίπεδο ασφάλειας. Δηλαδή, μπορεί ένα τμήμα των δεδομένων κάποιου εκ των συμβαλλομένων να αποκαλυφθεί. Έχει προβλεφθεί όμως, το ποσοστό της αποκάλυψης να είναι τέτοιο ώστε να μην μπορεί να γίνει οποιαδήποτε χρήση των δεδομένων που έχουν υποκλαπεί. Επίσης εισάγουν μια σειρά συμπληρωματικών πρωτοκόλλων (AB Protocol, $(A+B)^{-1}$ Protocol, Matrix Product II Protocol) για τους ασφαλείς υπολογισμούς των ενδιάμεσων σταδίων που απαιτούνται για την ολοκλήρωση του αλγορίθμου.

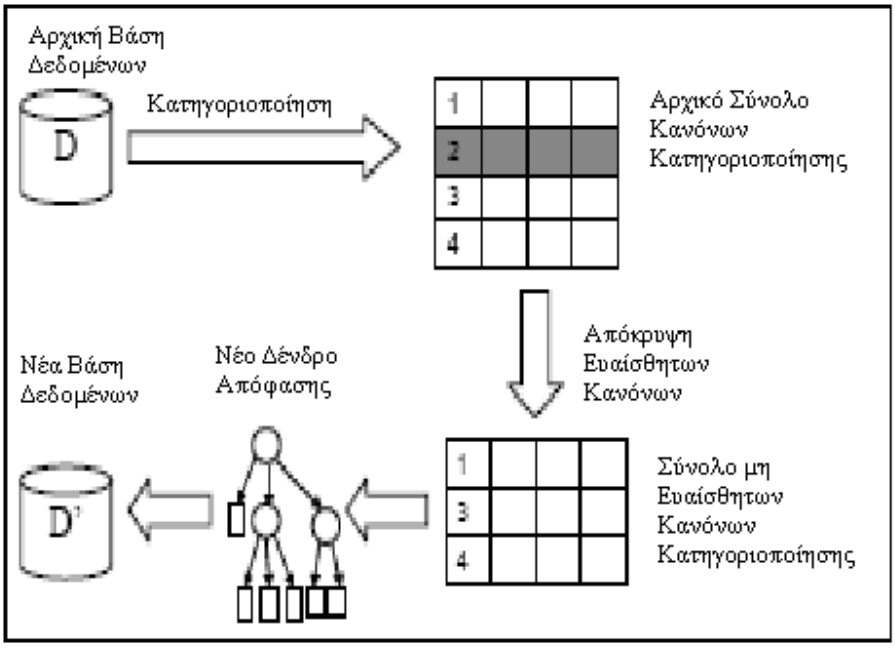
Με την εφαρμογή των παραπάνω επιτυγχάνεται η ασφαλής εφαρμογή της διαδικασίας της παλινδρόμησης μεταξύ δυο κατόχων από τους οποίους κανένας δεν είναι διατεθειμένος να εμπιστευτεί τον άλλο. Στα μειονεκτήματα της προτεινόμενης μεθόδου συγκαταλέγεται το γεγονός ότι η προσέγγιση περιορίζεται στις δυο πλευρές όχι σε ένα κατανεμημένο περιβάλλον με περισσότερους από δυο χρήστες.

2.6.2 Απόκρυψη Κανόνων Κατηγοριοποίησης

Στην ενότητα αυτή παρουσιάζονται κάποιες προσεγγίσεις που αφορούν αποκλειστικά στην απόκρυψη κανόνων κατηγοριοποίησης. Οι υλοποιήσεις του επόμενου κεφαλαίου στηρίζονται στις μελέτες της παρούσας ενότητας

Με την προστασία ευαίσθητων προτύπων που μπορεί να προκύψουν κατά την κατηγοριοποίηση με χρήση κανόνων ασχολούνται οι Juggapong Natwichai, Xue Li και Maria E.Orlowska [12]. Ο ανακατασκευαστικός αλγόριθμος (Α.Α.) που παρουσιάζεται φτιάχνει ξανά την βάση ακολουθώντας συγκεκριμένα βήματα που περιγράφονται παρακάτω. Αν και δεν διατηρεί την μορφή της αρχικής βάσης καταφέρνει να αποκρύψει αποτελεσματικά τους κανόνες που πρέπει να παραμείνουν εμπιστευτικοί.

Η βασική ιδέα πάνω στην οποία στηρίζεται ένας ανακατασκευαστικός αλγόριθμος είναι η εξής. Η αρχική βάση κατηγοριοποιείται με την χρήση ενός αλγορίθμου κατηγοριοποίησης και το αποτέλεσμα είναι ένα σύνολο κανόνων κατηγοριοποίησης. Στην συνέχεια ο ιδιοκτήτης της βάσης δεδομένων αποφασίζει ποιούς από τους κανόνες επιθυμεί να αποκρύψει. Οι υπόλοιποι κανόνες, αυτοί δηλαδή που σύμφωνα με τον κάτοχο της βάσης δεν αποκαλύπτουν ευαίσθητα πρότυπα, χρησιμοποιούνται για την κατασκευή ενός δέντρου απόφασης. Τέλος, το δέντρο που προέκυψε από το προηγούμενο βήμα χρησιμοποιείται για την ανακατασκευή μιας νέας βάσης με τον ίδιο αριθμό εγγραφών και τα ίδια γνωρίσματα που όμως δεν περιλαμβάνει τους ευαίσθητους κανόνες. Μια σχηματική αναπαράσταση της λογικής του ανακατασκευαστικού αλγορίθμου φαίνεται στο σχήμα 1.



Σχήμα 1. Λογική Ανακατασκευαστικού Αλγορίθμου.

Ο αλγόριθμος που παρουσιάζεται στην μελέτη χρησιμοποιεί τους αλγόριθμους RIPPER και C4.5 για την εξαγωγή των κανόνων κατηγοριοποίησης από την αρχική βάση. Έπειτα αποκρύπτονται οι κανόνες που πρέπει να παραμείνουν εμπιστευτικοί και κατασκευάζεται το δέντρο από τους υπόλοιπους. Το δέντρο κατασκευάζεται ως εξής. Για κάθε έναν από τους εναπομείναντες κανόνες υπολογίζεται ο αριθμός των εγγραφών της αρχικής βάσης που κατηγοριοποιεί. Επίσης, υπολογίζεται το gain ratio των γνωρισμάτων. Στην συνέχεια οι κανόνες εισάγονται στο δέντρο ξεκινώντας από αυτόν που κατηγοριοποιεί τις περισσότερες εγγραφές. Τα γνωρίσματα του εκάστοτε κανόνα εισάγονται κατά φθίνουσα τιμή gain ratio.

Για να γίνει πιο κατανοητή η διαδικασία κατασκευής του δέντρου ακολουθεί ένα παράδειγμα.

Παράδειγμα 1

Έστω ότι έχουμε τα 6 δυαδικά γνωρίσματα A, B, C, D, E, F με τις τιμές gain ratio που φαίνονται στον πίνακα 1. Το γνώρισμα κλάση μπορεί να πάρει τις τιμές +/- και το σύνολο των εγγραφών που διατηρούνται στην βάση είναι 800.

Πίνακας 1. Τιμές gain ratio των γνωρισμάτων.

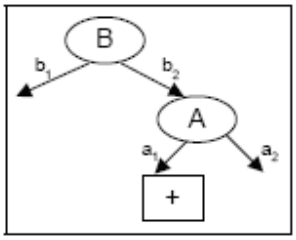
ΓΝΩΡΙΣΜΑ	GAIN RATIO
A	0.150
B	0.153
C	0.080
D	0.070
E	0.098
F	0.072

Από την βάση εξάγεται το σύνολο κανόνων κατηγοριοποίησης που φαίνεται στον πίνακα 2. Στον πίνακα 2 φαίνεται επίσης και ο αριθμός των εγγραφών που κατηγοριοποιεί ο κάθε κανόνας:

Πίνακας 2. Κανόνες κατηγοριοποίησης.

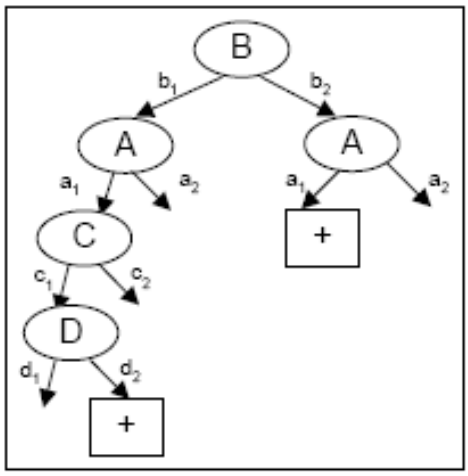
ΑΡΙΣΤΕΡΟ ΜΕΡΟΣ ΚΑΝΟΝΑ	ΚΛΑΣΗ	ΑΡΙΘΜΟΣ ΕΓΓΡΑΦΩΝ
(A=a1) και (B=b2)	+	143
(A=a1) και (C=c1) και (D=d2)	+	47
(B=b1) και (E=e2)	+	21
(E=e1) και (F=f1)	+	82
Εξ ορισμού	-	407

Έστω ότι θέλουμε να αποκρύψουμε τον κανόνα (E=e1) και (F=f1)=>+ . Ο κανόνας που κατηγοριοποιεί τις περισσότερες εγγραφές είναι ο πρώτος δηλαδή ο (A=a1) και (B=b2)=>+ . Επομένως θα είναι ο πρώτος που θα εισαχθεί στο δέντρο. Τα γνώρισμα που περιλαμβάνει στο αριστερό μέλος του είναι τα A και B από τα οποία το B έχει μεγαλύτερο gain ratio άρα θα είναι το πρώτο γνώρισμα που θα χρησιμοποιηθεί. Το υπό κατασκευή δέντρο που προκύπτει φαίνεται παρακάτω (σχήμα 2):



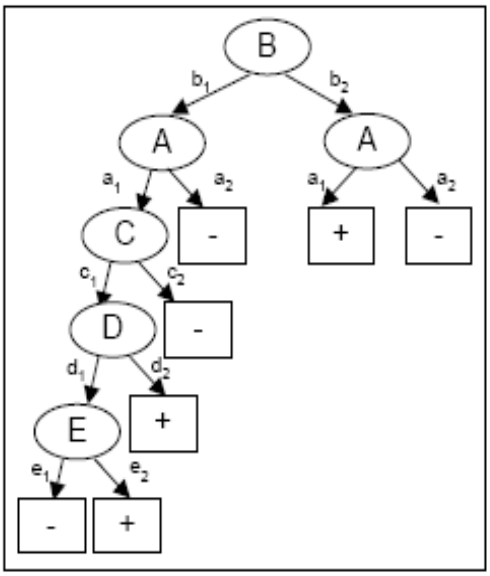
Σχήμα 2 . Εισαγωγή 1ου κανόνα στο δέντρο απόφασης.

Ο αμέσως επόμενος κανόνας που θα εισαχθεί στο δέντρο είναι ο (A=a1) και (C=c1) και (D=d2)=>+ . Τα γνωρίσματά του θα εισαχθούν με την εξής σειρά: A, C, D. Το δέντρο με την εισαγωγή του νέου κανόνα διαμορφώνεται όπως φαίνεται στο σχήμα 3.



Σχήμα 3. Εισαγωγή 2ου κανόνα στο δέντρο απόφασης

Τέλος, εισάγεται και ο κανόνας (B=b1) και (E=e2)=>+ με τα γνωρίσματα του στην εξής σειρά: B, E. Το τελικό δέντρο από το οποίο θα γίνει η ανακατασκευή της βάσης φαίνεται στο σχήμα 4.

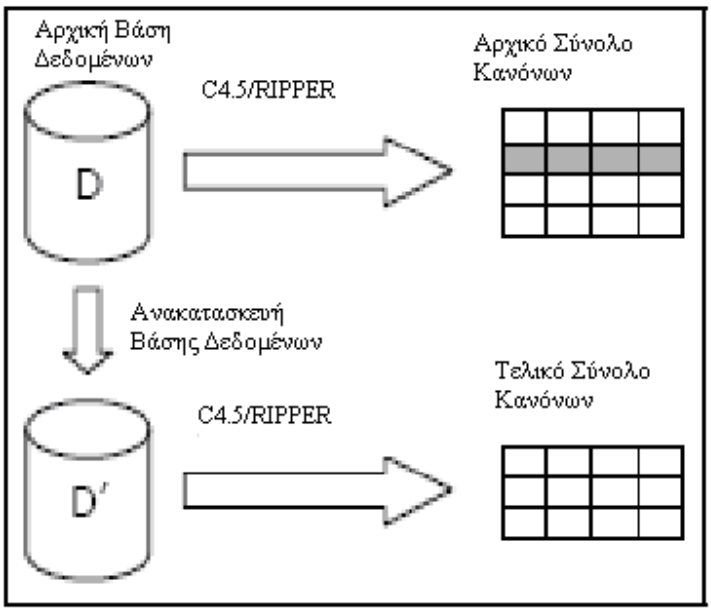


Σχήμα 4. Τελικό δέντρο απόφασης.

■

Αφού κατασκευαστεί το δέντρο από τους μη ευαίσθητους κανόνες όπως περιγράφτηκε από την παραπάνω διαδικασία, ξεκινά η ανακατασκευή της βάσης. Για λόγους ευκολίας στην υλοποίηση χρησιμοποιείται ομοιόμορφη κατανομή για τις τιμές των γνωρισμάτων στις διάφορες εγγραφές. Επομένως η τελική βάση που παίρνουμε έχει τον ίδιο αριθμό εγγραφών και τις ίδιες τιμές γνωρισμάτων. Ωστόσο αν εφαρμόσουμε πάλι τους αλγόριθμους RIPPER και C4.5 δεν μας επιστρέφονται οι ευαίσθητοι κανόνες.

Σύμφωνα με τα πειράματα που διεκπεραίωσαν οι συγγραφείς, ο αλγόριθμος αποδείχτηκε εξαιρετικά αποτελεσματικός στην προστασία ευαίσθητων προτύπων ενώ ταυτόχρονα οι παρενέργειες που παρατηρούνται στην βάση είναι σπάνιες και διατηρούνται στο ελάχιστο. Στο σχήμα 5 φαίνεται παραστατικά η διαδικασία που ακολουθήθηκε κατά τον πειραματισμό..



Σχήμα 5. Αξιολόγηση Αλγορίθμου.

Επίσης, έχει υλοποιηθεί και παρουσιαστεί σε προηγούμενη μελέτη [13] από τους ίδιους συγγραφείς ο παραπάνω αλγόριθμος με μια διαφοροποίηση στην κατασκευή του δένδρου απόφασης. Τα γνωρίσματα του κάθε κανόνα που εισάγεται στο δέντρο δεν επιλέγονται με φθίνουσα τιμή gain ratio αλλά επιλέγεται κάθε φορά το λιγότερο κοινό γνώρισμα (Least Common). Δηλαδή το γνώρισμα που εμφανίζεται τις λιγότερες φορές στο σύνολο των κανόνων. Τα υπόλοιπα βήματα του αλγορίθμου είναι όμοια με αυτόν που παρουσιάστηκε παραπάνω. Τα αποτελέσματα στην Least Common υλοποίηση ήταν λιγότερο ικανοποιητικά.

Μια ακόμη μελέτη παρουσιάζεται από τους Juggapong Natwichai, Maria E.Orlowska και Xingzhi Sun [14], όπου η απόκρυψη των ευαίσθητων κανόνων γίνεται με διαγραφή τμήματος ή του συνόλου των εγγραφών που τους υποστηρίζουν.

Για την αποδοτικότερη και πιο αποτελεσματική λειτουργία του αλγορίθμου οι συγγραφείς εισάγουν την έννοια του κανόνα κατηγοριοποίησης κανονικής μορφής. Στηριζόμενοι στον ορισμό αυτό, καταφέρνουν κάθε φορά που διαγράφεται μια εγγραφή από την βάση, να παίρνουν το νέο σύνολο κανόνων κατηγοριοποίησης χωρίς να χρειάζεται να εφαρμοστεί εκ νέου ο αλγόριθμος κατηγοριοποίησης.

Για την κατανόηση του αλγορίθμου χρησιμοποιείται μια γεωμετρική περιγραφή στην οποία κάθε εγγραφή αναπαρίσταται με ένα σημείο σε ένα κ – διάστατο χώρο, η διάσταση κ του οποίου καθορίζεται από τον αριθμό των γνωρισμάτων της βάσης. Κάθε κανόνας που εξάγεται από την βάση δεδομένων αναπαρίσταται με έναν από τους παρακάτω τρείς τρόπους. Είτε με ένα σημείο είτε με μια ευθεία είτε με ένα επίπεδο.

Ένα σύνολο κανόνων κατηγοριοποίησης για να είναι σε κανονική μορφή πρέπει να είναι ακριβές και όσο πιο γενικό γίνεται. Δηλαδή πρέπει να κατηγοριοποιεί με ακρίβεια όλες τις εγγραφές της βάσης αλλά επίσης να μην υπάρχει ένα ζεύγος γνωρίσματος – τιμής σε κάποιο κανόνα που αν αφαιρεθεί ο κανόνας θα εξακολουθεί να κατηγοριοποιεί τις ίδιες εγγραφές. Ένα παράδειγμα δίνεται στην συνέχεια.

Παράδειγμα 2

Έστω η παρακάτω βάση δεδομένων, η οποία αποτελείται από τα γνωρίσματα A_1, A_2, A_3 που μπορούν να πάρουν τις τιμές {0, 1} και το γνώρισμα κλάση C που μπορεί να πάρει τις τιμές {1, 2, 3}. Η βάση περιλαμβάνει 3 εγγραφές.

Πίνακας 3. Σύνολο εγγραφών της βάσης (Παράδειγμα 4).

Αριθμός Εγγραφής	**A_1**	**A_2**	**A_3**	**C**
1	0	0	0	1
2	0	0	1	1
3	0	1	0	1

Από τις παραπάνω εγγραφές μπορούν να προκύψουν τα παρακάτω διαφορετικά σύνολα κανόνων .

Σύνολο Α

1. $(A_1, 0) \rightarrow 1$

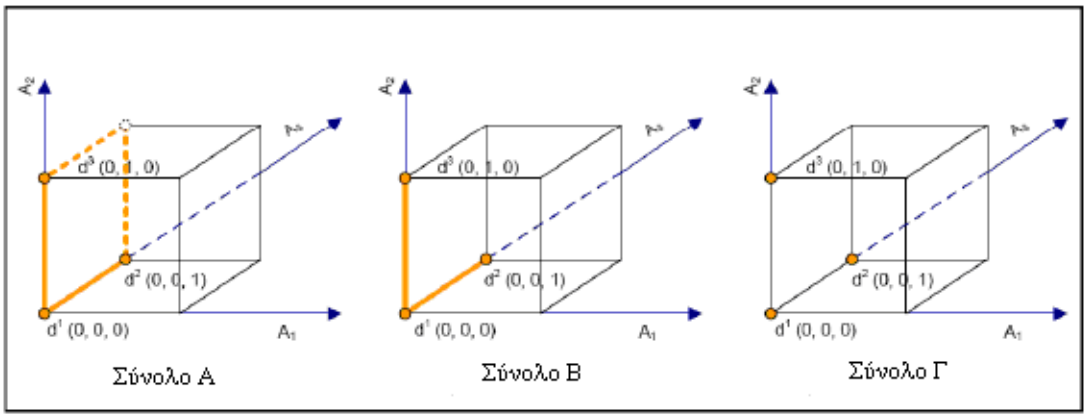
Σύνολο Β

1. $(A_1, 0)$ and $(A_3, 0) \rightarrow 1$
2. $(A_1, 0)$ and $(A_2, 0) \rightarrow 1$

Σύνολο Γ

1. $(A_1, 0)$ and $(A_2, 0)$ and $(A_3, 0) \rightarrow 1$
2. $(A_1, 0)$ and $(A_2, 0)$ and $(A_3, 1) \rightarrow 1$
3. $(A_1, 0)$ and $(A_2, 1)$ and $(A_3, 0) \rightarrow 1$

Στο σχήμα 6 φαίνεται η γεωμετρική αναπαράσταση για το κάθε σύνολο κανόνων.



Σχήμα 6. Γεωμετρική αναπαράσταση του συνόλου κανόνων.

Ωστόσο από τα παραπάνω σύνολα κανόνων μόνο το δεύτερο είναι σε κανονική μορφή. Όπως φαίνεται και από το σχήμα 6, το πρώτο σύνολο αναπαρίσταται με ένα επίπεδο που όμως περιλαμβάνει και το σημείο (0, 1, 1) επομένως δεν είναι αρκετά ακριβές, ενώ το σύνολο Γ αναπαρίσταται με τρία σημεία τα οποία μπορεί να

περιγράφουν με ακρίβεια της εγγραφές, όμως θα μπορούσε να περιλαμβάνει πιο γενικούς κανόνες. ■

Ο αλγόριθμος που υλοποιείται στην μελέτη είναι ο ακόλουθος. Αρχικά, εξάγεται από την αρχική βάση δεδομένων το σύνολο κανόνων σε κανονική μορφή. Για κάθε κανόνα υπάρχει ένας βαθμός εμπιστοσύνης και υποστήριξης. Το ζητούμενο είναι να προστατευτούν οι ευαίσθητοι κανόνες μειώνοντας την υποστήριξη τους. Για κάθε ευαίσθητο κανόνα ο αλγόριθμος διαγράφει διαδοχικά εγγραφές που τον υποστηρίζουν και ανανεώνει το σύνολο των κανόνων φέρνοντας τους ξανά αν χρειαστεί σε κανονική μορφή. Η διαδικασία σταματά όταν κανένας ευαίσθητος κανόνας δεν εμφανίζεται πλέον στο σύνολο κανόνων. Οι εγγραφές που διαγράφονται επιλέγονται έτσι ώστε να ελαχιστοποιείται η επίδραση στους μη ευαίσθητους κανόνες.

Προς το παρόν η ελαχιστοποίηση της επίδρασης της διαγραφής στους υπόλοιπους κανόνες είναι ο μόνος παράγοντας που λαμβάνεται υπόψη στην υλοποίηση. Στα μελλοντικά σχέδια αναφέρεται η επέκταση του αλγορίθμου να περιλαμβάνει διπλότυπες εγγραφές, καθώς επίσης και η μείωση των παρενεργειών όσον αφορά μη ευαίσθητους κανόνες που μπορεί να εμφανίζονται στην νέα βάση ή να χάνονται από την παλιά.

Τέλος, μια ακόμα μελέτη που αφορά στην προστασία προτύπων που μπορεί να προκύψουν από την κατηγοριοποίηση δεδομένων παρουσιάζεται από τους Zahidul Islam και Ljiljana Brankovic [15]. Η μέθοδος βασίζεται στην προσθήκη θορύβου στα δεδομένα και επιτυγχάνει την απόκρυψη των ευαίσθητων προτύπων χωρίς όμως να επηρεάζει τις στατιστικές παραμέτρους της βάσης, διατηρώντας έτσι τη χρησιμότητα της όσον αφορά στατιστικές αναλύσεις.

Η μελέτη αφορά βάσεις δεδομένων που αποτελούνται από αριθμητικά γνωρίσματα και ένα μόνο κατηγορικό γνώρισμα, το γνώρισμα κλάση. Ο θόρυβος εισάγεται τόσο σε ευαίσθητα όσο και σε μη ευαίσθητα γνωρίσματα καθιστώντας δυσκολότερο το διαχωρισμό τους και ενισχύοντας την παρεχόμενη ασφάλεια. Επιπλέον, ο αλγόριθμος μπορεί να επεκταθεί με τέτοιο τρόπο ώστε να διατηρεί τις συσχετίσεις των γνωρισμάτων και να διατηρεί τη λειτουργικότητα της εκάστοτε βάσης δεδομένων και για κατηγοριοποίηση και για στατιστική ανάλυση.

Για την εφαρμογή του αλγορίθμου αρχικά λαμβάνεται το δέντρο κατηγοριοποίησης με χρήση κάποιου αλγορίθμου. Επίσης, εισάγονται οι έννοιες των LINAs και LIAs που διατηρούνται για κάθε φύλλο. Τα LINAs είναι τα γνωρίσματα που εξετάζονται κατά την διαπέραση ενός μονοπατιού από τη ρίζα μέχρι το φύλλο ενώ τα LIAs είναι τα γνωρίσματα που δεν εξετάζονται. Ο διαχωρισμός δεν αφορά το γνώρισμα κλάση.

Ο αλγόριθμος μπορεί να διαχωριστεί σε τρία στάδια. Στο πρώτο στάδιο γίνεται η προσθήκη θορύβου στα LINAs κάθε φύλλου της αρχικής βάσης. Στο δεύτερο στάδιο γίνεται η προσθήκη θορύβου στα LIAs κάθε φύλλου επίσης της αρχικής βάσης. Τέλος, στο τρίτο στάδιο γίνεται η προσθήκη θορύβου για το γνώρισμα κλάση. Η προσθήκη θορύβου στο κάθε στάδιο γίνεται με διαφορετική μέθοδο. Τα πειραματικά αποτελέσματα που παρουσιάζονται στην μελέτη, έδειξαν ότι η απόκρυψη των ευαίσθητων προτύπων γίνεται επιτυχώς.

Οι τελευταίες αυτές μελέτες που παρουσιάστηκαν αποτελούν ουσιαστικά το σύνολο των προσπαθειών που αφορούν την προστασία ευαίσθητων προτύπων. Στο επόμενο κεφάλαιο εισάγεται μια προσέγγιση που αφορά επίσης στο ίδιο θέμα.

3. ΠΡΟΣΕΓΓΙΣΕΙΣ ΑΠΟΚΡΥΨΗΣ ΚΑΝΟΝΩΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

3.1 Βασική Προσέγγιση (Αλγόριθμος Ελάχιστης Τροποποίησης)

Στο κεφάλαιο αυτό, παρουσιάζουμε έναν αλγόριθμο τροποποίησης που στοχεύει στην απόκρυψη των ευαίσθητων κανόνων. Ο Αλγόριθμος Ελάχιστης Τροποποίησης που εισάγουμε στηρίζεται στην αντιγραφή μέρους των εγγραφών της αρχικής βάσης. Η λογική στην οποία στηρίζεται και παρατίθεται αναλυτικά παρακάτω.

Σχήμα 7. Βασική Δομή Αλγορίθμου

Αρχικά με χρήση ενός αλγορίθμου κατηγοριοποίησης (C4.5, RIPPER) λαμβάνονται οι κανόνες κατηγοριοποίησης της βάσης δεδομένων. Στην συνέχεια, ο ιδιοκτήτης της βάσης αποφασίζει ποιους από τους κανόνες θεωρεί ευαίσθητους και θέλει να κρύψει. Για κάθε έναν από τους υπόλοιπους μη ευαίσθητους κανόνες αντιγράφονται

από την αρχική βάση οι εγγραφές που τους υποστηρίζουν. Οι εγγραφές ελέγχονται να υποστηρίζουν έναν μόνο κανόνα κάθε φορά. Σε αντίθετη περίπτωση εισάγεται στην εγγραφή που παρουσιάζει την διαρροή πληροφορίας μια τροποποίηση προκειμένου να υποστηρίζει μόνο έναν κανόνα.

Η τροποποίηση αυτή αφορά στην αλλαγή τιμών σε ένα γνωρίσμα της εγγραφής. Το γνώρισμα που αλλάζει τιμή είναι ένα από αυτά που δημιουργούν το πρόβλημα, δηλαδή αυτά που εμφανίζονται με την ίδια τιμή και σε άλλο κανόνα. Για να μην αυξηθεί η υποστήριξη σε κάποια άλλη τιμή γνωρίσματος και συνεπώς να υπάρχει ο κίνδυνος για την εμφάνιση κανόνων «φαντασμάτων», το γνώρισμα που τροποποιείται παίρνει την τιμή εκείνη από το σύνολο τιμών του, που εμφανίζεται σπανιότερα στις μέχρι εκείνη τη στιγμή αντιγραμμένες εγγραφές.

Για να είναι εφικτή η αλλαγή αυτή στις τιμές των γνωρισμάτων, διατηρείται μια λίστα που περιλαμβάνει όλα τα γνωρίσματα και τις τιμές τους. Για κάθε τιμή γνωρίσματος διατηρείται ένας μετρητής. Αρχικά, όλοι οι μετρητές έχουν τιμή μηδέν. Κάθε φορά που μια εγγραφή αντιγράφεται ενημερώνει την λίστα για τις τιμές των γνωρισμάτων από τις οποίες αποτελείται. Όταν μια εγγραφή βρεθεί ότι υποστηρίζει και κάποιο άλλο κανόνα, για το γνώρισμα που συμβάλλει στην διαρροή πληροφορίας, γίνεται αναζήτηση στη λίστα και επιστρέφεται η τιμή που έχει χρησιμοποιηθεί τις λιγότερες φορές στις μέχρι τώρα εγγραφές. Αυτή η τιμή θα αντικαταστήσει την αρχική. Για την καλύτερη κατανόηση της διαδικασίας δίνεται ένα παράδειγμα.

Παράδειγμα 3

Έστω μια βάση δεδομένων, με 4 γνωρίσματα, τα A, B, C, D και ένα δυαδικό γνώρισμα κλάση που παίρνει τις τιμές { 0, 1}. Το γνώρισμα A μπορεί να πάρει τις τιμές {a1, a2, a3}, το B τις τιμές {b1, b2, b3}, το C τις τιμές {c1, c2, c3} και το D τις {d1, d2, d3}. Θεωρούμε ότι στην λίστα, οι μετρητές των τιμών των γνωρισμάτων έχουν τις τιμές που φαίνονται στους πίνακες που ακολουθούν.

Πίνακας 4. Τιμές μετρητή λίστας για το γνώρισμα Α.

	Τιμή Γνωρίσματος	Τιμή Μετρητή
Γνώρισμα Α	a1	1
	a2	3
	a3	6
Γνώρισμα Β	b1	7
	b2	3
	b3	0
Γνώρισμα C	c1	2
	c2	3
	c3	5
Γνώρισμα D	d1	3
	d2	3
	d3	4

Έστω ότι από το παρακάτω σύνολο εγγραφών που υποστηρίζουν τον κανόνα X: <A=a1> AND <D=d1> => 1 που εξετάζουμε, η εγγραφή 4 υποστηρίζει και τον κανόνα Y: < C=c3 >=> 1.

1. a1, b1, c1, d1, 1
2. a1, b2, c2, d1, 1
3. a1, b1, c2, d1, 1
4. a1, b2, c3, d1, 1

Σύμφωνα με τον προτεινόμενο αλγόριθμο η τιμή του γνωρίσματος C θα αντικατασταθεί με την τιμή του γνωρίσματος C που έχει χρησιμοποιηθεί το λιγότερο μέχρι τώρα. Δηλαδή, στην συγκεκριμένη περίπτωση η τιμή c3 θα αντικατασταθεί από την c1. ■

Με τον ελεγχόμενο αυτό τρόπο οι αλλαγές που γίνονται στις εγγραφές και συνεπώς οι πιθανότητες για εμφάνιση παρενεργειών περιορίζονται στο ελάχιστο. Στην συνέχεια η περιγραφή του σε μορφή ψευδοκώδικα.

ΑΛΓΟΡΙΘΜΟΣ ΕΛΑΧΙΣΤΗΣ ΤΡΟΠΟΠΟΙΗΣΗΣ (MINIMUM MODIFICATION ALGORITHM)

Είσοδος: Αρχική Βάση Δεδομένων D, Σύνολο μη ευαίσθητων κανόνων κατηγοριοποίησης R-R', Σύνολο ευαίσθητων κανόνων κατηγοριοποίησης R' .

Έξοδος: Νέα τροποποιημένη Βάση Δεδομένων D'.

D'={}.

Για κάθε κανόνα R_i Є R-R'

{

Βήμα 1. Βρες εκείνες τις εγγραφές t_i Є T της D που υποστηρίζουν τον κανόνα. Επέστρεψε το σύνολο των εγγραφών T_i που υποστηρίζουν τον κανόνα.

Βήμα 2. Για κάθε τιμή γνωρίσματος κάθε εγγραφής στο T_i αύξησε κατά 1 τον αντίστοιχο μετρητή την λίστα L.

Βήμα 3. Για κάθε εγγραφή t_i Є T_i έλεγξε αν υποστηρίζει κάποιον άλλο μη ευαίσθητο κανόνα, R_j ≠ R_i , R_j Є R-R'. Αν υποστηρίζει, προχώρα στο βήμα 3.1. Διαφορετικά πήγαινε στο βήμα 4.

Βήμα 3.1. Για την εγγραφή t_i που βρέθηκε στο προηγούμενο βήμα, βρες τα γνωρίσματα A_m Є t_i που εμφανίζονται στον R_j .

Βήμα 3.1.1. Από τα γνωρίσματα που βρέθηκαν στο βήμα 3.1 επέλεξε ένα που δεν εμφανίζεται και στον R_i και αντικατέστησε την τιμή του στην θεωρούμενη εγγραφή με την τιμή αυτή που έχει τον μικρότερο αριθμό μετρητή σύμφωνα με την λίστα L.

Βήμα 3.1.2. Ενημέρωσε με βάση τις αλλαγές την λίστα L.

Βήμα 4. Για κάθε εγγραφή t_i Є T_i έλεγξε αν υποστηρίζει κάποιον ευαίσθητο κανόνα, R_n Є R_S. Αν υποστηρίζει, προχώρα στο βήμα 4.1. Διαφορετικά πήγαινε στο βήμα 5.

Βήμα 4.1. Για την εγγραφή που βρέθηκε στο προηγούμενο βήμα, βρες τα γνωρίσματα A_m Є t_i που εμφανίζονται στον R_n .

Βήμα 4.1.1. Από τα γνωρίσματα που βρέθηκαν στο βήμα 4.1 επέλεξε ένα που δεν εμφανίζεται και στον R_i και αντικατέστησε την τιμή του στην θεωρούμενη εγγραφή με την τιμή αυτή που έχει τον μικρότερο αριθμό μετρητή σύμφωνα με την λίστα L.

Βήμα 4.1.2. Ενημέρωσε με βάση τις αλλαγές την λίστα L.

Βήμα 5. Πρόσθεσε κυκλικά τόσες από τις εγγραφές T_i στο T' όσες προσδιορίζονται από το $_{τον}$ αριθμό των εγγραφών που κατηγοριοποιεί ο κανόνας στην αρχική βάση.

}

Βήμα 6. Επέστρεψε την νέα τροποποιημένη βάση D' .

3.2 Συμπληρωματικοί Αλγόριθμοι

Για να υπάρχει ένα μέτρο σύγκρισης της αποδοτικότητας του αλγορίθμου που παρουσιάστηκε, υλοποιήσαμε και έναν αριθμό αλγορίθμων που αφορούν επίσης την απόκρυψη κανόνων. Πρόκειται για δυο τροποποιήσεις του Αλγορίθμου Ελάχιστης Τροποποίησης καθώς και για τροποποιήσεις των προσεγγίσεων που παρουσιάζονται στα [12] και [13].

3.2.1 Παραλλαγές Αλγόριθμου Ελάχιστης Τροποποίησης

Η πρώτη υλοποίηση αφορά μια παραλλαγή του Αλγόριθμου Ελάχιστης Τροποποίησης. Στην παραλλαγή αυτή, που καλείται Αλγόριθμος Διαγραφής (Α.Δ.)όσες από τις εγγραφές ενός κανόνα βρίσκονται να υποστηρίζουν κάποιον άλλο κανόνα (ευαίσθητο ή μη) διαγράφονται εντελώς από το νέο σύνολο εγγραφών. Πιο αναλυτικά, αφού ανακτηθεί το σύνολο κανόνων κατηγοριοποίησης με χρήση κάποιου κατάλληλου αλγορίθμου (C4.5, RIPPER) , καθορίζονται οι ευαίσθητοι προς απόκρυψη κανόνες. Για κάθε έναν από τους μη ευαίσθητους κανόνες αντιγράφεται το σύνολο των εγγραφών που τον υποστηρίζουν από την αρχική βάση δεδομένων. Στην συνέχεια για κάθε εγγραφή ελέγχεται εάν υποστηρίζει κάποιον άλλο ευαίσθητο ή μη ευαίσθητο κανόνα. Σε περίπτωση που κάτι τέτοιο ισχύει η εγγραφή διαγράφεται. Το τελικό σύνολο των εγγραφών που παράγεται με τον τρόπο αυτό για τον κάθε κανόνα χρησιμοποιείται όσες φορές χρειάζεται προκειμένου να δημιουργηθεί για κάθε κανόνα ίσος αριθμός εγγραφών με αυτόν που κατηγοριοποιεί από την αρχική βάση. Με τον τρόπο αυτό ο τελικός αριθμός εγγραφών που δημιουργείται από τον Αλγόριθμο Διαγραφής για την νέα βάση δεδομένων είναι ίσος με αυτόν της αρχικής βάσης.

Η δεύτερη παραλλαγή του Αλγορίθμου Ελάχιστης Τροποποίησης είναι ο Αλγόριθμος Διαγραφής - Ελάχιστης Τροποποίησης. Πρόκειται για ένα συνδυασμό των δύο αλγορίθμων. Όπως και προηγουμένως, αφού ανακτηθεί το σύνολο κανόνων κατηγοριοποίησης με χρήση κάποιου κατάλληλου αλγορίθμου (C4.5, RIPPER) , καθορίζονται οι ευαίσθητοι προς απόκρυψη κανόνες. Για κάθε έναν από τους μη ευαίσθητους κανόνες αντιγράφεται πάλι το σύνολο των εγγραφών που τον υποστηρίζουν από την αρχική βάση δεδομένων. Στην συνέχεια για κάθε εγγραφή ελέγχεται εάν

υποστηρίζει κάποιον άλλο ευαίσθητο ή μη ευαίσθητο κανόνα. Αυτή την φορά όμως σε περίπτωση που κάτι τέτοιο ισχύει, ανάλογα με τον αριθμό των εγγραφών που ανακτούνται από την αρχική βάση και υποστηρίζουν τον κανόνα γίνεται και η αντίστοιχη τροποποίηση. Αν ο αριθμός των εγγραφών που ανακτώνται είναι μεγαλύτερος από τον αριθμό των εγγραφών που κατηγοριοποιεί ο κανόνας στη αρχική βάση, τότε οι εγγραφές, που υποστηρίζουν κάποιον άλλο κανόνα (ευαίσθητο ή μη), διαγράφονται. Αν ο αριθμός των εγγραφών που ανακτήθηκαν είναι μικρότερος ή ίσος από τον αριθμό των εγγραφών που κατηγοριοποιεί ο κανόνας στη αρχική βάση τότε οι εγγραφές που υποστηρίζουν κάποιον άλλο κανόνα τροποποιούνται σύμφωνα με τον αλγόριθμο ελάχιστης τροποποίησης. Το τελικό σύνολο των εγγραφών για τον κάθε κανόνα παράγεται πάλι με τον ίδιο τρόπο όπως στον Αλγόριθμο Διαγραφής ώστε η τελική βάση να έχει τον ίδιο αριθμό εγγραφών με την αρχική.

3.2.2 Παραλλαγές Αλγορίθμων Βιβλιογραφίας

Ο πρώτος αλγόριθμος της βιβλιογραφίας του οποίου εισάγουμε μια τροποποίηση, είναι ένας απλός ευριστικός αλγόριθμος που αναφέρεται στο [6], ο οποίος αποκρύπτει τους ευαίσθητους κανόνες διαγράφοντας εγγραφές από το σύνολο υποστήριξης τους.

Αρχικά παίρνουμε με χρήση κάποιου αλγορίθμου (C4.5, RIPPER) το σύνολο των κανόνων κατηγοριοποίησης. Στην συνέχεια, καθορίζονται οι ευαίσθητοι κανόνες που πρέπει να κρύψουμε και το πλήθος των εγγραφών που θα τροποποιηθούν. Η τροποποίηση μπορεί να γίνει είτε σε όλες είτε στις μισές από τις εγγραφές που υποστηρίζουν τον προς απόκρυψη κανόνα. Για κάθε ευαίσθητο κανόνα λοιπόν, η τιμή του γνωρίσματος κλάση αντικαθίσταται στο προκαθορισμένο πλήθος εγγραφών με την άλλη τιμή που μπορεί να πάρει.

Ένας ψευδοκώδικας του ευριστικού αλγορίθμου (Ε.Α.) παρατίθεται στην συνέχεια, προκειμένου να γίνει πιο κατανοητός ο τρόπος λειτουργίας του.

ΕΥΡΙΣΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ (HEURISTIC ALGORITHM)

Είσοδος: Αρχική Βάση Δεδομένων D, Σύνολο ευαίσθητων κανόνων κατηγοριοποίησης R-R', παράμετρος x για την επιλογή του αριθμού των εγγραφών που θα τροποποιηθούν.
Έξοδος: Νέα Βάση Δεδομένων D'.

Για κάθε κανόνα $R_i \epsilon$R-R'
{

Βήμα 1. Βρες εκείνες τις εγγραφές t_i Є T που υποστηρίζουν τον κανόνα.

Βήμα 2. Επέστρεψε το σύνολο των εγγραφών T_i που υποστηρίζουν τον κανόνα.

Βήμα 3. Έλεγξε την τιμή του x. Αν x=1 προχώρα στο επόμενο βήμα. Διαφορετικά, αν x=2 πήγαινε στο βήμα 5.

Βήμα 4. Άλλαξε την τιμή του γνωρίσματος κλάση σε όλες τις εγγραφές T_i.

Βήμα 5. Άλλαξε την τιμή του γνωρίσματος κλάση στις πρώτες μισές εγγραφές του Ti.

}

Βήμα 6. Επέστρεψε την νέα τροποποιημένη βάση D'.

Η δεύτερη βοηθητική υλοποίηση αφορά τον ανακατασκευαστικό αλγόριθμο που παρουσιάζεται στο [13]. Όπως και ο αρχικός αλγόριθμος έτσι και η εκδοχή που παρουσιάζεται στην συνέχεια αφορά κατηγορικές βάσεις δεδομένων με δυαδικά γνωρίσματα κλάσεις. Στα δεδομένα μπορεί να περιλαμβάνονται και άγνωστες τιμές (?).

Ο νέος αλγόριθμος ακολουθεί σε γενικές γραμμές την ίδια λογική. Αρχικά λαμβάνονται οι κανόνες κατηγοριοποίησης με χρήση των αλγορίθμων C4.5 ή RIPPER. Στην συνέχεια επιλέγονται οι ευαίσθητοι κανόνες που πρέπει να παραμείνουν κρυφοί. Για την δημιουργία του δέντρου, οι μη ευαίσθητοι κανόνες εισάγονται πάλι με φθίνοντα αριθμό εγγραφών που κατηγοριοποιούν, ενώ τα γνωρίσματα του κάθε κανόνα εισάγονται με φθίνουσα τιμή κέρδους πληροφορίας (gain ratio). Η διαφορά είναι ότι για κάθε μονοπάτι του δέντρου διατηρείται ακριβώς ο αριθμός των εγγραφών που κατηγοριοποιεί.

Ο αριθμός αυτός αντιστοιχεί στον αριθμό των εγγραφών που κατηγοριοποιεί ο εκάστοτε κανόνας που εισάγεται και δημιουργεί το μονοπάτι.

Επίσης, ο κόμβος του δέντρου κατηγοριοποίησης στον οποίο θα γίνει η εισαγωγή του επόμενου κανόνα επιλέγεται με βάση τα παρακάτω. Αν υπάρχει ήδη κόμβος με το ίδιο όνομα γνωρίσματος, ελέγχεται καταρχήν αν υπάρχει τμήμα του κανόνα που να εμφανίζεται ήδη στο δέντρο και αν μπορεί το υπόλοιπο του κανόνα να εισαχθεί εκεί. Σε κάθε άλλη περίπτωση ως αρχικός κόμβος του νέου μονοπατιού, επιλέγεται ο πλησιέστερος κόμβος στην ρίζα του δέντρου που διαθέτει ελεύθερες ακμές. Με τον τρόπο αυτό καταφέρνουμε να ελαχιστοποιήσουμε τον αριθμό των γνωρισμάτων που θα εμφανιστούν στον κανόνα αλλά δεν θα ανήκουν σε αυτόν, περιορίζοντας έτσι και τον αριθμό των κανόνων που μπορεί να εμφανιστούν (false drop rules).

Τέλος, όσον αφορά πάντα την δημιουργία του δέντρου κατηγοριοποίησης από το οποίο θα παραχθεί η νέα βάση, κάθε φορά που εισάγεται ένας κανόνας , γίνεται αναζήτηση σε όλα τα μονοπάτια για την εμφάνιση κάποιου ευαίσθητου κανόνα. Σε περίπτωση που βρεθεί ότι σε κάποιο μονοπάτι σχηματίζεται ένας ευαίσθητος κανόνας, οι αλλαγές που έχουν γίνει στο δέντρο στο παρόν στάδιο αναιρούνται και ο κανόνας δεν εισάγεται. Με τον τρόπο αυτό ελέγχεται ότι στην νέα βάση δεν θα εμφανίζονται ευαίσθητοι κανόνες που θα έπρεπε να έχουν κρυφτεί.

Στο στάδιο της ανακατασκευής της βάσης, δηλαδή της δημιουργίας των εγγραφών της νέας βάσης εισάγονται επίσης κάποιες διαφοροποιήσεις. Δεν χρησιμοποιείται ομοιόμορφη κατανομή για την παραγωγή των εγγραφών. Οι εγγραφές δημιουργούνται δίνοντας τιμή αρχικά στα γνωρίσματα που ανήκουν στο κάθε μονοπάτι και στην συνέχεια συμπληρώνοντας τις υπόλοιπες τιμές με βάση τα στατιστικά για τις τιμές του κάθε γνωρίσματος στην αρχική βάση (stratification).

Για να γίνει πιο κατανοητός ο τρόπος λειτουργίας του αλγορίθμου δίνεται η υλοποίηση του σε μορφή ψευδοκώδικα.

ΑΝΑΚΑΤΑΣΚΕΥΑΣΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ (RECONSTRUCTION ALGORITHM)

Είσοδος: Αρχική Βάση Δεδομένων D, Σύνολο ευαίσθητων κανόνων κατηγοριοποίησης R-R', Σύνολο μη ευαίσθητων κανόνων κατηγοριοποίησης R ταξινομημένων κατά φθίνοντα αριθμό εγγραφών που κατηγοριοποιούν.

Έξοδος: Νέα Βάση Δεδομένων D'.

Για κάθε κανόνα R_i ϵR-R'

{

Βήμα 1. Πρόσθεσε κάθε γνώρισμα a_i Є R_i με gain ratio(a_i) ≥ gain ratio(a_j), a_j, a_i Є R_i στο δέντρο κατηγοριοποίησης.

Βήμα 2. Βρες όλα τα μονοπάτια που προκύπτουν και έλεγξε ότι δεν δημιουργείται κανένας ευαίσθητος κανόνας R_k ϵR-R'. Αν δημιουργείται πήγαινε στο επόμενο βήμα. Διαφορετικά συνέχισε στον επόμενο κανόνα.

Βήμα 3. Αναίρεσε όλες τις αλλαγές που έγιναν στο δέντρο για τον κανόνα αυτό.

}

Για κάθε μονοπάτι του δέντρου p_i

{

Για όσες εγγραφές t_i πρέπει να δημιουργηθούν για αυτό

{

Βήμα 4. Δώσε τιμή στα γνωρίσματα που παίρνουν τιμή στο μονοπάτι.

Βήμα 5. Δώσε με stratification τιμή στα υπόλοιπα γνωρίσματα

}

}

Βήμα 6. Επέστρεψε την τελική βάση D'.

Με τον ίδιο τρόπο, υλοποιείται και ο δεύτερος ανακατασκευαστικός αλγόριθμος, με τη μόνη διαφορά ότι τα γνωρίσματα του κάθε κανόνα εισάγονται στο δέντρο ξεκινώντας από αυτό που εμφανίζεται λιγότερο στο σύνολο των κανόνων.

Όλες οι παραπάνω συμπληρωματικές υλοποιήσεις θα χρησιμοποιηθούν στις πειραματικές δοκιμές του επόμενου κεφαλαίου.

4. ΠΕΙΡΑΜΑΤΑ – ΕΦΑΡΜΟΓΕΣ ΑΛΓΟΡΙΘΜΩΝ

Στην ενότητα αύτη παρουσιάζονται τα αποτελέσματα της εφαρμογής του αλγορίθμου Ελάχιστης Τροποποίησης και των παραλλαγών της ενότητας 3.2 σε τρείς πραγματικές βάσεις δεδομένων. Πρόκειται για τις Mushrooms, Kr-vs-kp και Voting Records από το UCI Repository. Για τα πειράματα που πραγματοποιήσαμε χρησιμοποιήσαμε την υλοποίηση του RIPPER που υπάρχει στο WEKA (JRip). Πρόκειται για κατηγορικές βάσεις δεδομένων, με δυαδικά γνωρίσματα κλάσης. Περισσότερες λεπτομέρειες φαίνονται στον πίνακα 8 που ακολουθεί.

Πίνακας 5. Στοιχεία Πειραματικών Δεδομένων.

	Βάσεις Δεδομένων		
Στοιχεία	Mushroom	Kr-vs-kp	Vote
Αριθμός Εγγραφών	8124	3196	435
Αριθμός Γνωρισμάτων	22	37	16
Αριθμός Κανόνων με χρήση RIPPER με χρήση κλαδέματος	9	16	4
Αριθμός Κανόνων με χρήση RIPPER χωρίς χρήση κλαδέματος	8	18	10

Στα πειραματικά δεδομένα εφαρμόστηκαν εκτός από τον Αλγόριθμο Ελάχιστης Τροποποίησης (Α.Ε.Τ.) και τις παραλλαγές του, Αλγόριθμος Διαγραφής (Α.Δ.) και Αλγόριθμος Διαγραφής – Ελάχιστης Τροποποίησης (Α.Δ. – Ε.Τ.) και ο Ευριστικός Αλγόριθμος (Ε.Α.). Επίσης, εφαρμόζεται σε κάποια από τα πειραματικά δεδομένα και ο Ανακατασκευαστικός Αλγόριθμος (Α.Α.) και στις δυο εκδοχές του (διάσπαση με βάση το Gain Ratio και με βάση το Least Common γνώρισμα).

Οι πειραματικές δοκιμές των αλγορίθμων έγιναν για διάφορα σενάρια απόκρυψης κανόνων που φαίνονται στου πίνακες 6 - 20. Σε κάθε πίνακα αναφέρεται το όνομα της βάσης στην οποία εφαρμόζονται οι αλγόριθμοι, ο αλγόριθμος που

χρησιμοποιείται για την παραγωγή του αρχικού και τελικού συνόλου κατηγοριοποίησης, ο αρχικός αριθμός κανόνων κατηγοριοποίησης, ο αριθμός των ευαίσθητων κανόνων που πρέπει να κρυφτούν, ο αριθμός των ευαίσθητων κανόνων που χάνονται (αριθμός απολεσθέντων κανόνων, ο αριθμός των κανόνων «φαντασμάτων», ο αριθμός των ευαίσθητων κανόνων που μπορεί να εμφανιστούν στο τελικό σύνολο κανόνων και τέλος ο αριθμός των κανόνων στο τελικό σύνολο κανόνων κατηγοριοποίησης. Πριν από κάθε πίνακα αποτελεσμάτων δίνονται οι εκάστοτε λεπτομέρειες. Επίσης μετά από κάθε πίνακα ακολουθεί μια σύντομη ανάλυση των αποτελεσμάτων που παρουσιάζονται.

Στον πίνακα 6 φαίνεται η εφαρμογή των αλγορίθμων στην βάση Mushroom. Το αρχικό και τελικό σύνολο κανόνων κατηγοριοποίησης αποτελείται από 8 κανόνες και ανακτήθηκε με χρήση του αλγορίθμου RIPPER χωρίς χρήση κλαδέματος. Οι αλγόριθμοι εφαρμόζονται για απόκρυψη ενός κανόνα.

Πίνακας 6. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς κλάδεμα. Απόκρυψη ενός κανόνα.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	1	1	0	7
Α.Δ. – Ε.Τ.	1	1	0	7
Α.Ε.Τ	1	1	0	7
Α.Α. (Gain Ratio)	5	34	0	36
Ε.Α.(Όλες)	1	1	0	7
Ε.Α.(Μισές)	1	2	0	8
Α.Α(Λιγότερο κοινό γνώρισμα)	5	43	0	45

Στην εφαρμογή γίνεται απόκρυψη ενός μόνο κανόνα από το αρχικό σύνολο. Η απόδοση των αλγορίθμων Διαγραφής, Διαγραφής – Ελάχιστης Τροποποίησης, Ελάχιστης Τροποποίησης και του Ευριστικού είναι καλή με λίγες παρενέργειες ενώ δεν εμφανίζεται στην νέα βάση κανένας ευαίσθητος κανόνας. Αντίθετα, ο Ανακατασκευαστικός

αλγόριθμος και στις δυο εκδοχές του (gain ratio και least common) εισάγει πολλές παρενέργειες (αριθμός απολεσθέντων κανόνων και κανόνων «φαντασμάτων»).

Στον πίνακα 7 φαίνεται η εφαρμογή των αλγορίθμων πάλι στην βάση Mushroom. Το αρχικό και τελικό σύνολο κανόνων κατηγοριοποίησης ανακτήθηκε με χρήση του αλγορίθμου RIPPER χωρίς χρήση κλαδέματος. Αυτή τη φορά το πείραμα εκτελείται με απόκρυψη 2 κανόνων από το αρχικό σύνολο 8 κανόνων.

Πίνακας 7. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς κλάδεμα. Απόκρυψη δυο κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	0	1	0	7
Α.Δ. – Ε.Τ.	0	0	0	6
Α.Ε.Τ	0	0	0	6
Α.Α. (Gain Ratio)	4	52	0	54
Ε.Α.(Όλες)	0	0	0	6
Ε.Α.(Μισές)	0	6	0	12
Α.Α(Λιγότερο κοινό γνώρισμα)	4	49	0	51

Όπως φαίνεται στο σύνολο των εφαρμογών η απόδοση των αλγορίθμων Διαγραφής, Διαγραφής – Ελάχιστης Τροποποίησης, Ελάχιστης Τροποποίησης και του Ευριστικού είναι εξαιρετικά ικανοποιητική. Διατηρεί τις παρενέργειες σε πολύ χαμηλά επίπεδα ενώ κανένας ευαίσθητος κανόνας δεν μπορεί να εξαχθεί από το νέο σύνολο κανόνων. Ωστόσο οι δυο ανακατασκευαστικοί αλγόριθμοι εισάγουν παρενέργειες τόσο με τη μορφή «ψευδών» κανόνων όσο και με τη μορφή κανόνων «φαντασμάτων».

Στους υπόλοιπους πίνακες φαίνονται τα αποτελέσματα των υπόλοιπων πειραματικών δοκιμών οι οποίες συνεχίζονται μόνο για τους αλγόριθμους Διαγραφής, Διαγραφής – Ελάχιστης Τροποποίησης, Ελάχιστης Τροποποίησης και του Ευριστικού(στις δυο εκδοχές του) τα αποτελέσματα των οποίων έχουν αξία σύγκρισης.

Στον πίνακα 8 φαίνεται η εφαρμογή των αλγορίθμων πάλι στην βάση Mushroom. Το αρχικό και τελικό σύνολο κανόνων κατηγοριοποίησης ανακτήθηκε με χρήση του αλγορίθμου RIPPER χωρίς χρήση κλαδέματος. Αυτή τη φορά το πείραμα εκτελείται με απόκρυψη 3 κανόνων από το αρχικό σύνολο 8 κανόνων.

Πίνακας 8. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς κλάδεμα. Απόκρυψη τριών κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	0	1	0	6
Α.Δ. – Ε.Τ.	0	0	0	5
Α.Ε.Τ	0	0	0	5
Ε.Α.(Όλες)	0	0	0	5
Ε.Α.(Μισές)	0	0	0	5

Στον πίνακα 9 φαίνεται η εφαρμογή των αλγορίθμων για την βάση Mushroom. με απόκρυψη 4 κανόνων από το αρχικό σύνολο 8 κανόνων.

Πίνακας 8. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς κλάδεμα. Απόκρυψη τεσσάρων κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	1	2	0	5
Α.Δ. – Ε.Τ.	1	1	0	4
Α.Ε.Τ	1	1	0	4
Ε.Α.(Όλες)	2	2	0	4
Ε.Α.(Μισές)	1	29	0	32

Στους πίνακες 9 - που ακολουθούν η εφαρμογή των αλγορίθμων στην βάση Mushroom με απόκρυψη ενός ,δυο, τριών και τεσσάρων κανόνων αντίστοιχα. Το αρχικό και το τελικό σύνολο κανόνων κατηγοριοποίησης παράγεται πάλι με χρήση του

αλγόριθμου RIPPER αυτή την φορά όμως με χρήση κλαδέματος. Το αρχικό σύνολο αποτελείται από 9 κανόνες.

Πίνακας 9. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη ενός κανόνα.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	0	1	0	9
Α.Δ. – Ε.Τ.	0	1	0	9
Α.Ε.Τ	0	0	0	8
Ε.Α.(Όλες)	0	1	0	9
Ε.Α.(Μισές)	1	1	0	9

Πίνακας 10. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη δυο κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	0	0	0	7
Α.Δ. – Ε.Τ.	0	0	0	7
Α.Ε.Τ	0	0	0	7
Ε.Α.(Όλες)	1	0	0	6
Ε.Α.(Μισές)	0	1	0	8

Πίνακας 11. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη τριών κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	0	0	0	6
Α.Δ. – Ε.Τ.	0	0	1	7
Α.Ε.Τ	0	1	1	8
Ε.Α.(Όλες)	1	2	1	7
Ε.Α.(Μισές)	0	1	1	8

Πίνακας 12. Εφαρμογή Αλγορίθμων για την βάση Mushroom. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη τεσσάρων κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	0	0	0	5
Α.Δ. – Ε.Τ.	0	0	0	5
Α.Ε.Τ	0	0	0	5
Ε.Α.(Όλες)	0	0	1	6
Ε.Α.(Μισές)	0	4	1	10

Ακόμα και με απόκρυψη περισσότερων κανόνων, όπως φαίνεται από τους παραπάνω πίνακες η απόδοση των αλγορίθμων Διαγραφής, Διαγραφής – Ελάχιστης Τροποποίησης, Ελάχιστης Τροποποίησης και του Ευριστικού είναι καλή με λίγες παρενέργειες ενώ δεν εμφανίζεται στην νέα βάση κανένας ευαίσθητος κανόνας.

Στους πίνακες 13 -16 φαίνονται τα αποτελέσματα των πειραματικών δοκιμών για την βάση Vote. Το αρχικό σύνολο κανόνων κατηγοριοποίησης εξάγεται με χρήση του αλγορίθμου RIPPER χωρίς χρήση κλαδέματος και αποτελείται από 10 κανόνες. Οι πίνακες αντιστοιχούν σε σενάρια απόκρυψης ενός, δύο, τριών και τεσσάρων κανόνων αντίστοιχα

Πίνακας 6. Εφαρμογή Αλγορίθμων για την βάση Vote. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη ενός κανόνα.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	4	5	0	9
Α.Δ. – Ε.Τ.	1	2	0	10
Α.Ε.Τ	1	2	0	10
Ε.Α.(Όλες)	4	9	1	14
Ε.Α.(Μισές)	5	8	1	12

Πίνακας 7. Εφαρμογή Αλγορίθμων για την βάση Vote. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη δυο κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	5	6	0	9
Α.Δ. – Ε.Τ.	3	4	0	8
Α.Ε.Τ	3	3	0	7
Ε.Α.(Όλες)	4	9	0	12
Ε.Α.(Μισές)	3	12	0	16

Πίνακας 15. Εφαρμογή Αλγορίθμων για την βάση Vote. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη τριών κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	3	3	0	7
Α.Δ. – Ε.Τ.	3	3	0	7
Α.Ε.Τ	2	2	0	7
Ε.Α.(Όλες)	4	10	1	13
Ε.Α.(Μισές)	4	11	2	16

Πίνακας 16. Εφαρμογή Αλγορίθμων για την βάση Vote. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη τεσσάρων κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	3	4	0	7
Α.Δ. – Ε.Τ.	3	4	0	7
Α.Ε.Τ	3	5	0	8
Ε.Α.(Όλες)	3	11	2	16
Ε.Α.(Μισές)	3	12	2	17

Στους επόμενους πίνακες 17 -20 φαίνονται τα αποτελέσματα των πειραματικών δοκιμών επίσης για την βάση Vote. Το αρχικό σύνολο κανόνων κατηγοριοποίησης εξάγεται με χρήση του αλγορίθμου RIPPER αυτή την φορά με χρήση κλαδέματος και αποτελείται από 4 κανόνες. Οι πίνακες αντιστοιχούν σε σενάρια απόκρυψης ενός και δύο κανόνων αντίστοιχα

Πίνακας 17. Εφαρμογή Αλγορίθμων για την βάση Vote. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη ενός κανόνα.

	Αριθμός Απολεσθέντων Κανόνων	**Αριθμός Κανόνων «Φαντασμάτων»**	**Αριθμός Ευαίσθητων Κανόνων**	**Αριθμός Τελικών Κανόνων**
Α.Δ.	2	1	0	2
Α.Δ. – Ε.Τ.	2	2	0	3
Α.Ε.Τ	2	2	0	3
Ε.Α.(Όλες)	2	1	0	2
Ε.Α.(Μισές)	1	1	1	4

Πίνακας 18. Εφαρμογή Αλγορίθμων για την βάση Vote. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη δυο κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	1	1	0	2
Α.Δ. – Ε.Τ.	0	0	0	2
Α.Ε.Τ	0	0	0	2
Ε.Α.(Όλες)	1	3	1	5
Ε.Α.(Μισές)	1	1	0	2

Τέλος, στους πίνακες 19 - 22 και 23 - 26 φαίνονται τα αποτελέσματα των πειραματικών δοκιμών την βάση Kr-vs-kp. Το αρχικό σύνολο κανόνων κατηγοριοποίησης εξάγεται με χρήση του αλγορίθμου RIPPER χωρίς και με χρήση κλαδέματος και αποτελείται από 16 και 18 κανόνες αντίστοιχα. Τα πειράματα εκτελούνται για την απόκρυψη ενός, δύο, τριών και τεσσάρων κανόνων.

Πίνακας 19. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη ενός κανόνα.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	11	8	0	15
Α.Δ. – Ε.Τ.	8	11	0	19
Α.Ε.Τ	9	11	0	18
Ε.Α.(Όλες)	14	27	0	30
Ε.Α.(Μισές)	14	43	1	47

Πίνακας 20. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη δυο κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	12	8	0	13
Α.Δ. – Ε.Τ.	8	8	0	17
Α.Ε.Τ	11	13	0	18
Ε.Α.(Όλες)	14	29	1	32
Ε.Α.(Μισές)	14	54	1	57

Πίνακας 21. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη τριών κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	10	6	0	11
Α.Δ. – Ε.Τ.	7	8	0	16
Α.Ε.Τ	9	11	0	17
Ε.Α.(Όλες)	13	32	1	35
Ε.Α.(Μισές)	13	57	1	60

Πίνακας 22. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη τεσσάρων κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	9	4	0	9
Α.Δ. – Ε.Τ.	5	6	0	15
Α.Ε.Τ	5	7	0	16
Ε.Α.(Όλες)	11	37	1	40
Ε.Α.(Μισές)	11	64	1	68

Πίνακας 23. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη ενός κανόνα.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	7	4	0	12
Α.Δ. – Ε.Τ.	4	6	0	16
Α.Ε.Τ	7	6	0	14
Ε.Α.(Όλες)	13	14	0	16
Ε.Α.(Μισές)	11	26	1	30

Πίνακας 24. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη δυο κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	8	6	0	12
Α.Δ. – Ε.Τ.	5	6	0	14
Α.Ε.Τ	7	7	0	14
Ε.Α.(Όλες)	12	15	0	17
Ε.Α.(Μισές)	11	28	1	32

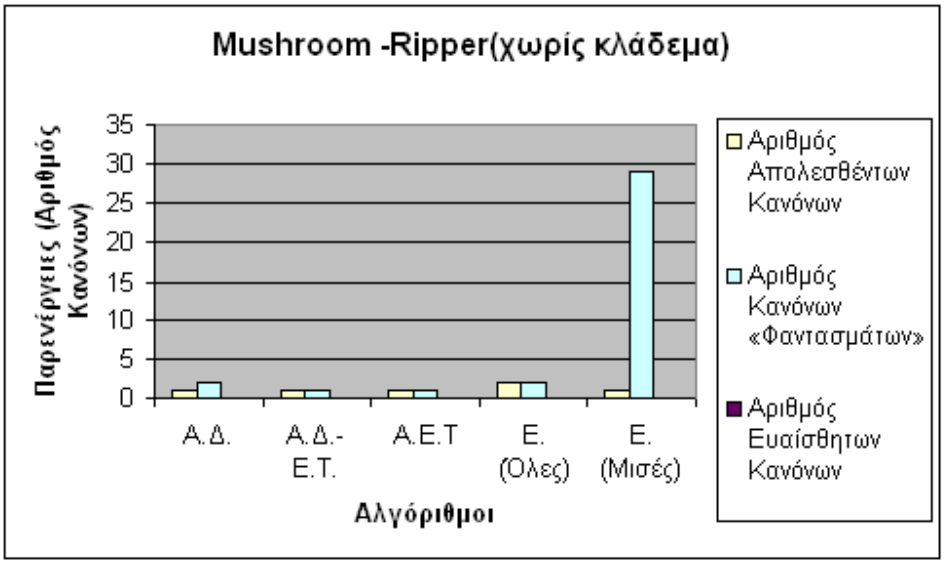
Πίνακας 25. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER με χρήση κλαδέματος. Απόκρυψη τριών κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	7	3	0	9
Α.Δ. – Ε.Τ.	0	0	0	13
Α.Ε.Τ	4	4	0	13
Ε.Α.(Όλες)	12	14	0	15
Ε.Α.(Μισές)	12	35	0	36

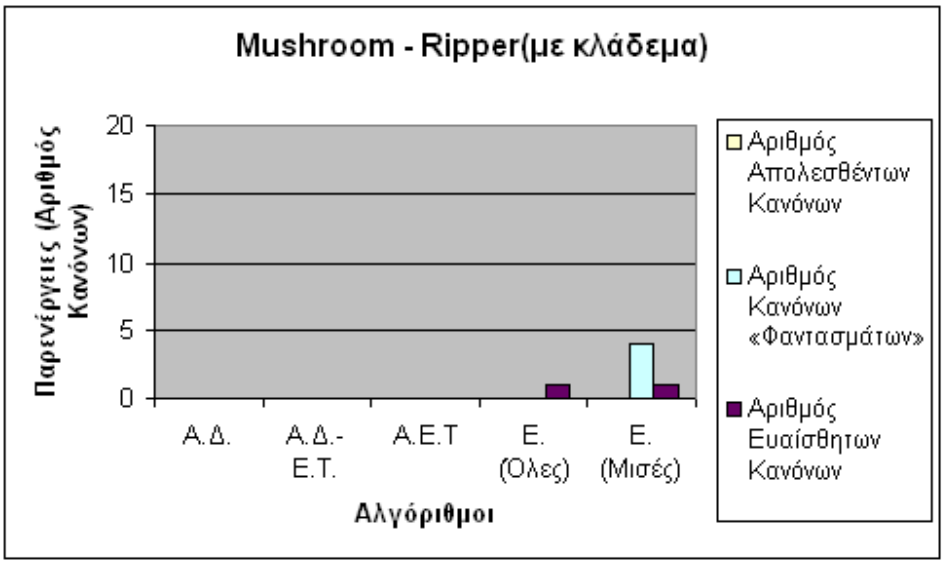
Πίνακας 26. Εφαρμογή Αλγορίθμων για την βάση kr-vs-kp. Οι κανόνες εξάγονται με χρήση του RIPPER χωρίς χρήση κλαδέματος. Απόκρυψη τεσσάρων κανόνων.

	Αριθμός Απολεσθέντων Κανόνων	Αριθμός Κανόνων «Φαντασμάτων»	Αριθμός Ευαίσθητων Κανόνων	Αριθμός Τελικών Κανόνων
Α.Δ.	6	1	0	7
Α.Δ. – Ε.Τ.	2	2	0	12
Α.Ε.Τ	1	6	0	13
Ε.Α.(Όλες)	11	7	0	8
Ε.Α.(Μισές)	9	26	0	29

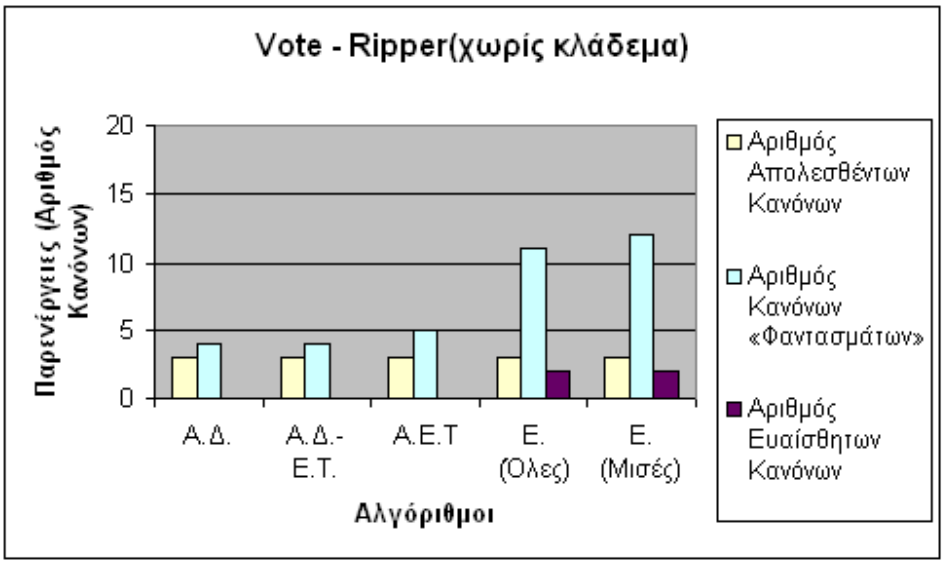
Για καλύτερη σύγκριση των αποτελεσμάτων των αλγορίθμων για την κάθε βάση δεδομένων, απεικονίζονται στην συνέχεια με την μορφή γραφημάτων τα χειρότερα σενάρια απόκρυψης κανόνων για τις τρεις βάσεις δεδομένων (σχήματα 7 - 12).



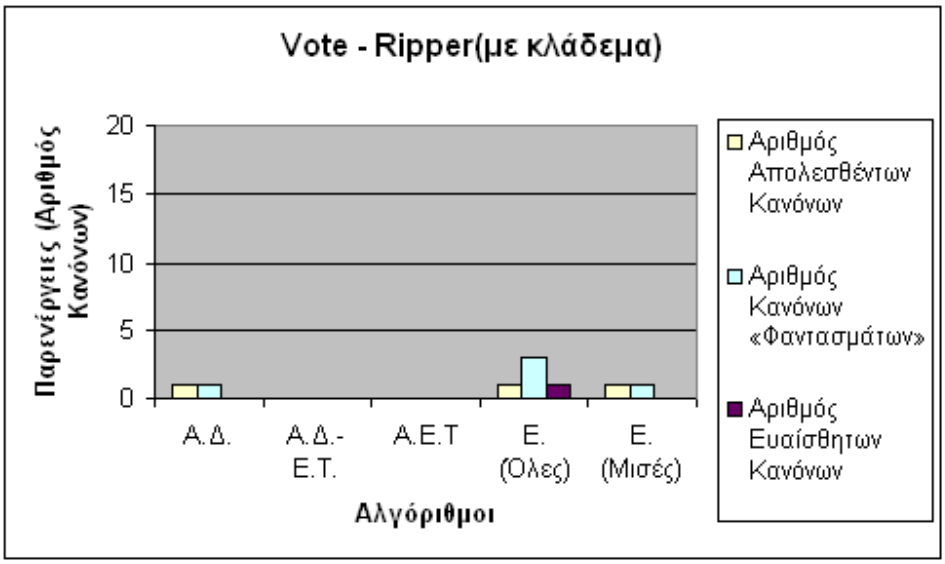
Σχήμα 7. Γραφική παράσταση για το σενάριο απόκρυψης τεσσάρων κανόνων για τη βάση Mushroom. Το αρχικό και τελικό σύνολο κανόνων εξάγονται με χρήση του Ripper χωρίς κλάδεμα.



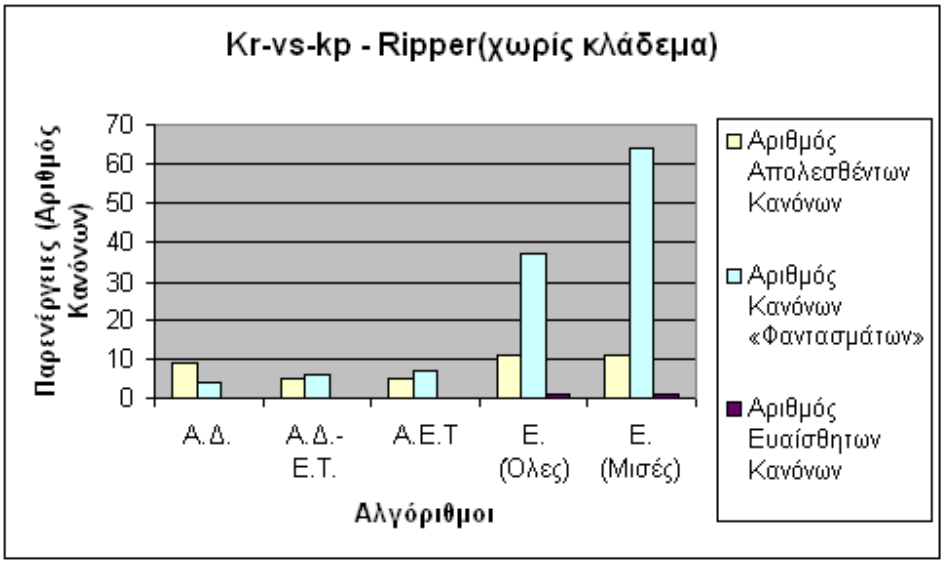
Σχήμα 8. Γραφική παράσταση για το σενάριο απόκρυψης τεσσάρων κανόνων για τη βάση Mushroom. Το αρχικό και τελικό σύνολο κανόνων εξάγονται με χρήση του Ripper με κλάδεμα.



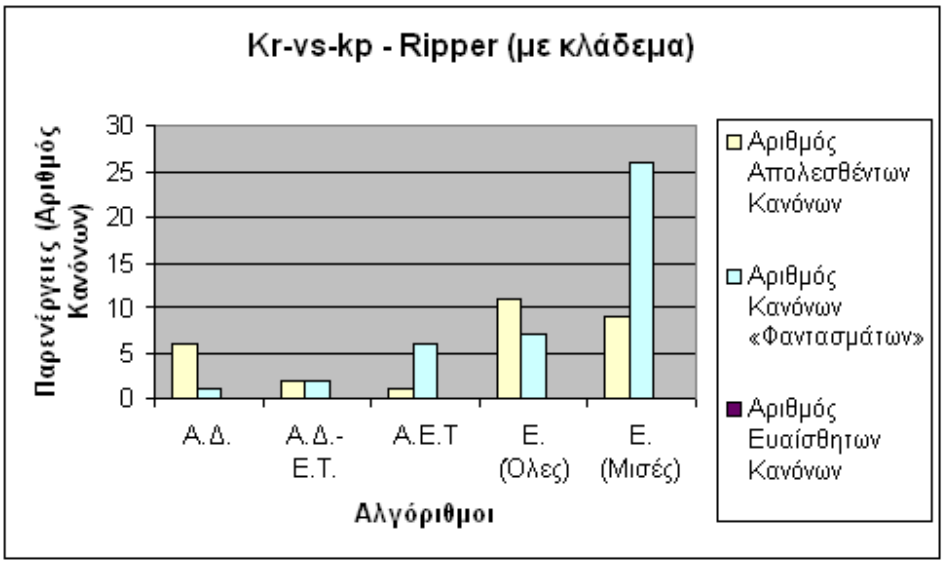
Σχήμα 9. Γραφική παράσταση για το σενάριο απόκρυψης τεσσάρων κανόνων για τη βάση Vote. Το αρχικό και τελικό σύνολο κανόνων εξάγονται με χρήση του Ripper χωρίς κλάδεμα.



Σχήμα 10. Γραφική παράσταση για το σενάριο απόκρυψης δυο κανόνων για τη βάση Vote. Το αρχικό και τελικό σύνολο κανόνων εξάγονται με χρήση του Ripper με κλάδεμα.



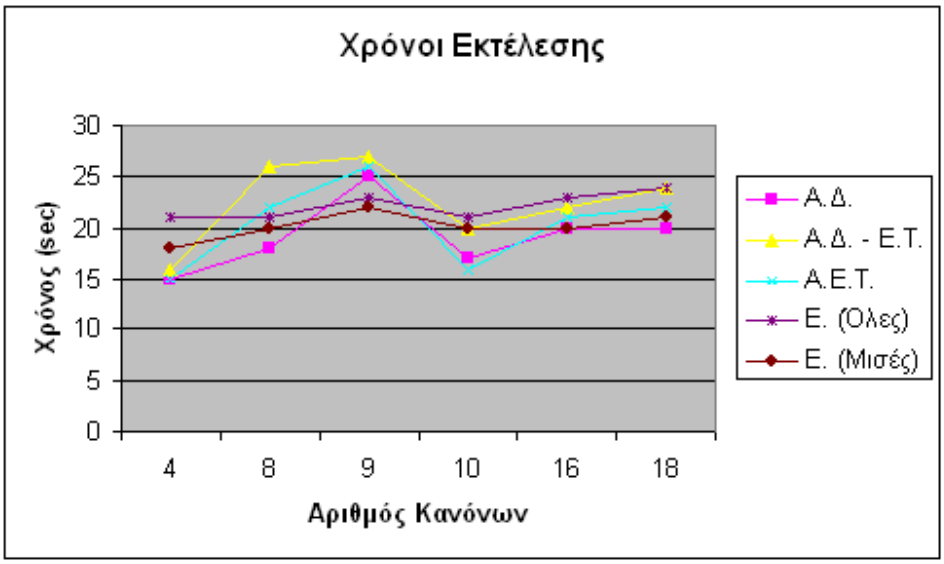
Σχήμα 10. Γραφική παράσταση για το σενάριο απόκρυψης τεσσάρων κανόνων για τη βάση Kr-vs-kp. Το αρχικό και τελικό σύνολο κανόνων εξάγονται με χρήση του Ripper χωρίς κλάδεμα.



Σχήμα 10. Γραφική παράσταση για το σενάριο απόκρυψης τεσσάρων κανόνων για τη βάση Kr-vs-kp. Το αρχικό και τελικό σύνολο κανόνων εξάγονται με χρήση του Ripper με κλάδεμα.

4.1 Χρονική Πολυπλοκότητα

Ο χρόνος εκτέλεση των αλγορίθμων που χρησιμοποιήθηκαν για τα πειράματα φαίνεται στο επόμενο σχήμα (σχήμα 7). Για τον σχεδιασμό της γραφικής παράστασης μετρήθηκε ο χρόνος εκτέλεσης του κάθε αλγόριθμου για το κάθε σενάριο απόκρυψης κανόνων που παρουσιάστηκε προηγουμένως. Στην συνέχεια για τον κάθε αλγόριθμο υπολογίστηκε ένας μέσος όρος του χρόνου εκτέλεσης ο οποίος και χρησιμοποιείται. Η γραφική παράσταση γίνεται συναρτήσει του αριθμού κανόνων που εμφανίζονται στο αρχικό σύνολο κανόνων κατηγοριοποίησης και τους οποίους πρέπει να επεξεργαστεί ο κάθε αλγόριθμος.



Σχήμα 11. Γραφική παράσταση των χρόνων εκτέλεσης των αλγορίθμων συναρτήσει του αριθμού κανόνων στο αρχικό σύνολο κανόνων κατηγοριοποίησης.

Όπως φαίνεται και από το γράφημα και οι τρείς αλγόριθμοι έχουν μικρές διαφορές στους χρόνους εκτέλεσης. Επίσης ο χρόνος δεν παρουσιάζει μεγάλη αύξηση κατά την αύξηση του αριθμού των κανόνων στο αρχικό σύνολο κανόνων. Ενώ ο αριθμός των κανόνων διπλασιάζεται (από 8 σε 16) ο χρόνος εκτέλεσης παρουσιάζει μια πολύ μικρή αύξηση της τάξης των 5 sec. Επίσης, κατά την αύξηση του αριθμού κανόνων οι χρόνοι εκτέλεσης των Α.Ε.Τ., Α.Δ.Τ., Α.Δ –Ε.Τ. είναι καλύτεροι από του Ευριστικού Αλγορίθμου.

4.2 Συμπεράσματα

Όπως φαίνεται και από τις πειραματικές εφαρμογές της προηγούμενης ενότητας, ο Αλγόριθμος Ελάχιστης Τροποποίησης και οι παραλλαγές του (Αλγόριθμος Διαγραφής και Αλγόριθμος Διαγραφής – Ελάχιστης Τροποποίησης) έχουν καλύτερη απόδοση από τον ευριστικό αλγόριθμο και στις δυο εκδοχές του. Επίσης, ο Αλγόριθμος Ελάχιστης Τροποποίησης φαίνεται να έχει καλύτερη απόδοση όταν εφαρμόζεται σε σύνολα κανόνων με μεγάλη υποστήριξη. Δηλαδή όταν τα σύνολα κανόνων κατηγοριοποίησης αποτελούνται από κανόνες που υποστηρίζονται από μεγάλο αριθμό εγγραφών. Αυτό

συμβαίνει γιατί τα σύνολα αυτά δεν είναι τόσο ευαίσθητα σε μικρές αλλαγές των τιμών των γνωρισμάτων των εγγραφών. Έτσι ενώ η απόδοση του αλγορίθμου είναι πολύ καλή για την βάση mushroom, χειροτερεύει για την βάση Vote και ακόμα περισσότερο για την βάση Kr-vs-kp. Επίσης, βασικό ρόλο παίζει ο αριθμός των κανόνων στο αρχικό σύνολο κατηγοριοποίησης καθώς και ο αριθμός των προς απόκρυψη κανόνων. Τέλος, ένα σημαντικό στοιχείο είναι ότι στο σύνολο των πειραμάτων δεν εμφανίζονται στο νέο – τελικό σύνολο κανόνων κατηγοριοποίησης κανόνες που είχαν χαρακτηριστεί ως ευαίσθητοι.

Οι παραλλαγές του αλγορίθμου, δηλαδή ο Αλγόριθμος Διαγραφής και ο Αλγόριθμος Διαγραφής – Ελάχιστης Τροποποίησης, ακολουθούν την απόδοση του Αλγόριθμου Ελάχιστης Τροποποίησης. Εισάγουν τις ανάλογες παρενέργειες και καταφέρνουν επίσης να αποκρύψουν επιτυχώς τους ευαίσθητους κανόνες.

Οι δυο ανακατασκευαστικοί αλγόριθμοι, με διάσπαση με βάση το gain ratio και το λιγότερο κοινό γνώρισμα, εισάγουν πολύ μεγάλες παρενέργειες, και από πλευράς «ψευδών» κανόνων και από πλευράς κανόνων «φαντασμάτων». Η νέα βάση που παράγεται δεν έχει ομοιότητες με την αρχική και το νέο σύνολο κανόνων περιλαμβάνει ελάχιστους από τους αρχικούς κανόνες. Αυτό οφείλεται σε μεγάλο βαθμό στον τρόπο ανακατασκευής της βάσης. Ο αριθμός των γνωρισμάτων των οποίων οι τιμές καθορίζονται στα πλαίσια των κανόνων – μονοπατιών των δέντρων είναι μικρός και στα υπόλοιπα γνωρίσματα οι τιμές δίνονται με χρήση στατιστικών μέτρων (stratification). Συνεπώς, οι εγγραφές που δημιουργούνται έχουν μεγάλες διαφορές σε σχέση με τις εγγραφές της αρχικής βάσης δεδομένων.

Γενικά, ο Αλγόριθμος Ελάχιστης Τροποποίησης επιτυγχάνει σε κάθε περίπτωση την απόκρυψη των ευαίσθητων κανόνων, έχει μικρή χρονική και χωρική πολυπλοκότητα. Είναι βασισμένος σε μια απλή λογική και έχει πολύ καλά αποτελέσματα σε βάσεις με σύνολα κανόνων κατηγοριοποίησης μεγάλης υποστήριξης.

5. ΕΠΙΛΟΓΟΣ

"You have zero privacy. Get over it."
Scott McNealy, 1999

Δυστυχώς, η παραπάνω άποψη εκφράζει σε μεγάλο βαθμό την κατάσταση στις μέρες μας. Η εξέλιξη της τεχνολογίας δημιουργεί τεράστια προβλήματα ιδιωτικότητας. Η αυξανόμενη χρήση υπολογιστών και δικτύων, το μειωμένο κόστος αποθήκευσης και διατήρησης δεδομένων και η δυνατότητα ανάλυσης και επεξεργασίας μεγάλου όγκου δεδομένων δημιουργούν προβλήματα ασφάλειας. Τα προβλήματα αυτά μπορεί να αφορούν τόσο τα προσωπικά στοιχεία που μπορεί να δοθούν από μεμονωμένα άτομα όσο και ευαίσθητα πρότυπα που μπορεί να εξαχθούν από το σύνολο των δεδομένων.

Στην παρούσα εργασία μελετήθηκε το πρόβλημα της ιδιωτικότητας όπως αυτό διαμορφώνεται κατά την εξόρυξη γνώσης από δεδομένα με την τεχνική της κατηγοριοποίησης. Για την επίλυση του προβλήματος έχει προταθεί πληθώρα προσεγγίσεων που μπορούν να οργανωθούν με πολλούς διαφορετικούς τρόπους. Στην εργασία παρουσιάστηκε ένας αριθμός αλγορίθμων που στοχεύουν στην επίλυση του προβλήματος καθώς επίσης και πειραματικές εφαρμογές κάποιων υλοποιήσεων.

Από τους αλγορίθμους που μελετήθηκαν παραπάνω κάποιοι επιτυγχάνουν την ασφαλή εφαρμογή της κατηγοριοποίησης ή σε άλλες περιπτώσεις την ασφαλή κοινοποίηση των εξαγόμενων συμπερασμάτων. Κάποιοι άλλοι ωστόσο, ενώ καταφέρνουν να διατηρήσουν την λειτουργικότητα της βάσης, δεν επιτυγχάνουν την προστασία των δεδομένων.

Σημαντικά θέματα για μελλοντική μελέτη όσον αφορά τους αλγορίθμους που παρουσιάζονται στην εργασία είναι η υλοποίηση κάποιων προσεγγίσεων που παραμένουν σε θεωρητικό επίπεδο αλλά και η ανάπτυξη κάποιων άλλων ώστε να είναι δυνατή η εφαρμογή τους σε μεγαλύτερης κλίμακας προβλήματα.

Εν κατακλείδι, το ζήτημα της ιδιωτικότητας των δεδομένων καλό είναι να επιλύεται με μεθόδους προσανατολισμένες στην εκάστοτε περίπτωση. Για παράδειγμα αν έχουμε ένα κατανεμημένο περιβάλλον, ποιος αλγόριθμος κατηγοριοποίησης θα

χρησιμοποιηθεί. Πρόκειται για ένα κρίσιμο ζήτημα που θα συνεχίσει να απασχολεί τις ερευνητικές μελέτες και που λόγω της εξέλιξης των τεχνικών εξόρυξης γνώσης θα γίνεται συνεχώς πιο δυσεπίλυτο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Vassilios S.Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis. State-of-the-art in Privacy Preserving Data Mining. 2001.

[2] http://www.rulequest.com/Personal/

[3] Ron Kohavi, Ross Quinlan. Decision Tree Discovery. 1999.

[4] http://www.cs.cmu.edu/~wcohen/

[5] Margaret H.Dunham. Data Mining Introductory and Advanced Topics. 55-228. 2004.

[6] LiWu Chang, James Tracy. Multi-Dimensional Inference and Confidential Data Protection with Decision Tree Methods. 2002.

[7] LiWu Chang, James Tracy. Parsimonious Downgrading and Decision Trees Applied to the Inference Problem. 1998.

[8] Jaideep Vaidyal, Chris Clifton. Privacy-Preserving Decision Trees over Vertically Partitioned Data. 2005.

[9] Keke Chen, Ling Liu. A Random Rotation Perturbation Approach to Privacy Preserving Data Classification. 2005.

[10] Wenliang Du, Yunghsiang S.Han, Shigang Chen. Privacy-Preserving Multivariate Statistical Analysis:Linear Regression and Classification. 2004.

[11] Wenliang Du, Mikhail J. Atallah. Secure MultiParty Computation Problems and Their Applications: A Review and Open Problems. 2001.

[12] Juggapong Natwichai, Xue Li, Maria E.Orlowska. A Reconstruction-based Algorithm for Classification Rules Hiding. 2006.

[13] Juggapong Natwichai, Xue Li, Maria E.Orlowska. Hiding Classification Rules for Data Sharing with Privacy Preservation. 2005.

[14] Juggapong Natwichai, Maria E.Orlowska and Xingzhi Sun. Hiding Sensitive Associative Classification Rule by Data Reduction. 2007.

[15] Md.Zahidul Islam, Ljiljana Brankovic. A Framework for Privacy Preserving Classification in Data Mining. 2004.