Πανεπιστήμιο Θεσσαλίας

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων

Διπλωματική Διατριβή

Προστασία Γνώσης από Κατανεμημένους Αλγόριθμους Εξόρυξης Δεδομένων

Γουγούλας Δημήτριος

Επιτροπή:

Βερύκιος Βασίλειος, *Επίκουρος Καθηγητής Π.Θ.*
Μουστακίδης Γεώργιος, *Καθηγητής Π.Θ.*

Στην οικογένεια μου

Ευχαριστίες

Στην προσπάθεια μου αυτή συνέβαλλαν αρκετά άτομα τα οποία θα ήθελα να ευχαριστήσω. Πρώτα από όλα θα ήθελα να πω ένα μεγάλο ευχαριστώ στον καθηγητή κύριο Βερύκιο Βασίλειο ο οποίος στάθηκε η αιτία να ασχοληθώ με το αντικείμενο της Εξόρυξης Γνώσης. Ήταν αυτός που με βοήθησε όχι μόνο στην επιλογή του παρόντος θέματος, αλλά και με τις πολύτιμες συμβουλές του καθ' όλη τη διάρκεια της προσπάθειας μου, να φέρω σε πέρας τη διπλωματική μου διατριβή. Τον ευχαριστώ μέσα από τα βάθη της καρδιάς μου για την εμπιστοσύνη με την οποία με περιέβαλλε.

Επίσης, θα ήθελα να ευχαριστήσω τον καθηγητή κύριο Μουστακίδη Γεώργιο για την εξίσου σημαντική συμβολή του στην επιτυχή ολοκλήρωση της παρούσας μελέτης.

Ένα μεγάλο επίσης ευχαριστώ στα άτομα που με στηρίζουν όλα αυτά τα χρόνια, τους γονείς μου και την αδερφή μου οι οποίοι ήταν συνεχώς διπλά μου και χωρίς την πολύτιμη βοήθεια τους δεν θα τα είχα καταφέρει.

Τέλος θα ήθελα να ευχαριστήσω τους φίλους και συμφοιτητές μου για την υπομονή που έκαναν όλο αυτό το διάστημα στηρίζοντας με ο καθένας με τον τρόπο του.

Γουγούλας Δημήτριος

Βόλος, Σεπτέμβριος 2005

Περίληψη

Η πρόοδος της τεχνολογίας με τη συνεχή αύξηση των υπολογιστικών δυνατοτήτων και των χώρων αποθήκευσης, η διείσδυση του Internet στην καθημερινή μας ζωή, η αυτοματοποίηση πολλών εργασιών και αρκετοί άλλοι παράγοντες έχουν οδηγήσει σε μια θεαματική αύξηση της διακινούμενης πληροφορίας και δεδομένων. Τίθεται λοιπόν το ερώτημα: *πώς θα διαχωρίσουμε τη χρήσιμη από την άχρηστη πληροφορία;* Τη λύση έρχεται να δώσει η Εξόρυξη Δεδομένων, αντικείμενο της οποίας είναι ο προσδιορισμός προτύπων μέσα από μεγάλες ποσότητες δεδομένων. Επιπλέον, πολλές φορές οι πληροφορίες αυτές είναι φυσικά και γεωγραφικά, αποθηκευμένες σε πολλά διαφορετικά μέρη όπως για παράδειγμα τα δεδομένα ενός πολυκαταστήματος.

Ένα σημαντικό ζήτημα που προκύπτει είναι αυτό της ασφαλούς διακίνησης αλλά και αποθήκευσης των πληροφοριών. Για παράδειγμα, δεν μπορεί να έχει οποιοδήποτε άτομο πρόσβαση στη βάση δεδομένων μίας τράπεζας. Προκύπτει, λοιπόν, η ανάγκη δημιουργίας αποτελεσματικών αλγορίθμων που θα μπορούν να χειριστούν αποτελεσματικά τον τεράστιο όγκο των πληροφοριών χωρίς να υπάρχει κίνδυνος απώλειας ή υποκλοπής τους. Αυτό αποτελεί και το αντικείμενο της παρούσας μελέτης.

Συγκεκριμένα, θεωρώντας ότι οι πληροφορίες εξάγονται με τη μορφή κανόνων συσχέτισης, θα βασιστούμε σε ήδη υπάρχοντες αλγορίθμους απόκρυψης τους οποίους θα επιχειρήσουμε να τροποποιήσουμε και να συνδυάσουμε ώστε να λειτουργήσουν σε κατανεμημένο περιβάλλον. Σκοπός μας είναι να διαπιστώσουμε αν κάτι τέτοιο είναι αποτελεσματικό και να προτείνουμε τρόπους βελτίωσης τους. Αναλυτικότερα, θα εξεταστούν θέματα όπως οι συνθήκες υπό τις οποίες θα γίνουν τα πειράματα, ο τρόπος απόκρυψης των επιλεγμένων κανόνων συσχέτισης, τα ποσοστά απωλειών υπαρχόντων κανόνων συσχέτισης αλλά και δημιουργίας νέων και η αξιολόγηση των αλγορίθμων ώστε να διαπιστωθεί ποιος οδηγεί σε καλύτερα αποτελέσματα.

Συνεισφορές της Διπλωματικής Εργασίας

Η παρούσα μελέτη στηρίχθηκε στην ιδέα που περιγράφεται στο [41]. Η συγκεκριμένη δημοσίευση μελετά το πρόβλημα της αποτελεσματικής απόκρυψης σημαντικών δεδομένων που εκφράζονται με τη μορφή κανόνων συσχέτισης προτείνοντας τρεις αλγορίθμους οι οποίοι αποτελούν τη βάση του συστήματος που προτείνεται. Για την Εξόρυξη των κανόνων συσχέτισης χρησιμοποιήσαμε δύο γνωστούς αλγορίθμους της βιβλιογραφίας: τον Apriori [5] και τον Fast Distributed Mining[13] για εξόρυξη από μία κεντρική βάση δεδομένων και από κατανεμημένες βάσεις δεδομένων που επικοινωνούν μέσω δικτύου αντίστοιχα.

Οι σημαντικότερες συνεισφορές της παρούσας μελέτης συνοψίζονται στα εξής:

- Παρατηρήθηκε ότι είναι σχετικά εύκολο για κάποιον έμπειρο κακόβουλο χρήστη να καταφέρει, να επαναφέρει τα δεδομένα που προκύπτουν από την εφαρμογή ενός από τους τρεις αλγορίθμους στην αρχική τους κατάσταση και κατ' επέκταση να ανιχνεύσει τους «σημαντικούς» κανόνες συσχέτισης. Γι' αυτό το λόγο προτείνεται η εφαρμογή συνδυασμών των συγκεκριμένων αλγορίθμων.

- Στους προτεινόμενους αλγορίθμους υπάρχει μία παράμετρος που αποκαλείται Όριο Ασφάλειας, η οποία επηρεάζει το πλήθος των εγγραφών – συναλλαγών που θα υποστούν τροποποιήσεις γεγονός που επηρεάζει με τη σειρά του το ποσοστό των κανόνων συσχέτισης που χάνονται καθώς και της ύπαρξης νέων κανόνων που δεν προϋπήρχαν. Προτείνονται, λοιπόν δύο αλγόριθμοι οι οποίοι ψάχνουν και εντοπίζουν την ελάχιστη τιμή αυτής της παραμέτρου ώστε να διατηρούνται τα προαναφερόμενα ποσοστά στα χαμηλότερα δυνατά επίπεδα.

- Τέλος, προτείνεται ένας αλγόριθμος για την επανασύσταση των κανόνων συσχέτισης μετά την εφαρμογή κάποιας μεθόδου απόκρυψης μειώνοντας περαιτέρω τα ποσοστά των απωλειών υπαρχόντων αλλά και της δημιουργίας νέων κανόνων συσχέτισης.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1. Εισαγωγή

Η Εξόρυξη Γνώσης[15](*) εμφανίστηκε στο προσκήνιο ως ένα σημαντικό εργαλείο στενά συνδεδεμένο με την έννοια της ασφάλειας. Συνδεδεμένη με έννοιες όπως η ανίχνευση κάποιας απάτης, η αποτίμηση κινδύνων, η Εξόρυξη Γνώσης περιλαμβάνει τη χρήση διαφόρων εργαλείων ανάλυσης δεδομένων προκειμένου να ανακαλυφθούν άγνωστα, έγκυρα πρότυπα και σχέσεις μέσα από μεγάλα σύνολα δεδομένων όπως στατιστικά μοντέλα, μαθηματικούς αλγορίθμους και μεθόδους μηχανικής μάθησης[43] (αλγόριθμοι που μπορούν να βελτιώσουν την απόδοση τους αυτόματα μέσα από την εμπειρία όπως νευρωνικά δίκτυα[46] ή δέντρα απόφασης).

Οι διάφορες εφαρμογές Εξόρυξης Γνώσης μπορούν να χρησιμοποιήσουν μια μεγάλη ποικιλία παραμέτρων για να εξετάσουν τα διάφορα δεδομένα όπως *συσχετίσεις* (πρότυπα όπου κάποιο γεγονός συνδέεται αλληλένδητα με κάποιο άλλο όπως για παράδειγμα η αγορά μιας πένας συνεπάγεται την αγορά χαρτιού), *ακολουθιακή ανάλυση*[61] ή *ανάλυση μονοπατιού*[49] (πρότυπα όπου ένα γεγονός οδηγεί σε κάποιο άλλο όπως η γέννηση ενός παιδιού και η αγορά παιδικών τροφών), *κατηγοριοποίηση*[7] (προσδιορισμός νέων προτύπων), *συσταδοποίηση*[9] (εύρεση και ομαδοποίηση προηγουμένως άγνωστων μεταξύ τους γεγονότων όπως γεωγραφικές περιοχές και εμπορικές προτιμήσεις) και *πρόγνωση*[23] (ανακαλύπτοντας πρότυπα μέσω των οποίων κάποιος μπορεί να κάνει προβλέψεις για μελλοντικές δραστηριότητες όπως η πρόβλεψη ότι ένα άτομο το οποίο πηγαίνει σε έναν αθλητικό σύλλογο ενδεχομένως

(*) Οι αριθμοί που υπάρχουν ως δείκτες σε κάποιες λέξεις παραπέμπουν στο παράρτημα δίνοντας την αντίστοιχη ερμηνεία στα αγγλικά

να γυμνάζεται).

Η συνεχής πρόοδος της τεχνολογίας έχει συμβάλλει στην σταδιακή αύξηση του ενδιαφέροντος σχετικά με την Εξόρυξη Γνώσης. Μερικές από τις αλλαγές που έχουν επέλθει τα τελευταία χρόνια περιλαμβάνουν την αύξηση των δικτύων Ηλεκτρονικών Υπολογιστών που μπορεί να συνδέουν βάσεις δεδομένων, την ανάπτυξη ευριστικών τεχνικών[31] όπως νευρωνικά δίκτυα και εξειδικευμένοι αλγόριθμοι, τη διάδοση του μοντέλου πελάτη/εξυπηρετητή επιτρέποντας στους χρήστες την πρόσβαση σε κεντρικές βάσεις δεδομένων μέσα από έναν απλό σταθμό εργασίας και την αυξανόμενη ικανότητα συνδυασμού δεδομένων προερχόμενα από διαφορετικές πηγές. Επιπλέον, η συνεχώς αυξανόμενη ροή πληροφοριών και τα μειωμένα κόστη αποθήκευσης έχουν διαδραματίσει σημαντικό ρόλο. Από έρευνες έχει προκύψει ότι η ποσότητα της διακινούμενης πληροφορίας διπλασιάζεται κάθε χρόνο ενώ το κόστος αποθήκευσης δεδομένων έχει μειωθεί από μερικά ευρώ ανά Megabyte σε λίγα λεπτά.

1.1 Περιορισμοί της Εξόρυξης Γνώσης

Ενώ οι εφαρμογές της Εξόρυξης Γνώσης αποτελούν ένα πολύ χρήσιμο εργαλείο, δεν είναι αυτο-ελεγχόμενες. Απαιτούν άτομα πλήρως εξειδικευμένα που να μπορούν να παρακολουθούν την πορεία της ανάλυσης και να ερμηνεύουν τα τελικά αποτελέσματα.

Αν και η Εξόρυξη Γνώσης μπορεί να βοηθήσει αποκαλύπτοντας πρότυπα και σχέσεις, εντούτοις δεν παρέχει στο χρήστη επεξηγήσεις για τις τιμές και τη σημασία αυτών των προτύπων. Αντίθετα, πρέπει να γίνουν από τον ίδιο τον χρήστη. Επιπλέον, η αποτελεσματικότητα τους σχετίζεται με το κατά πόσο έχουν συγκριθεί με καταστάσεις του «πραγματικού κόσμου». Για παράδειγμα, για την αποτίμηση της λειτουργικότητας μιας εφαρμογής Εξόρυξης Γνώσης που σχεδιάστηκε για τον εντοπισμό ύποπτων ατόμων για τρομοκρατία, ο χρήστης θα πρέπει να ελέγξει το μοντέλο χρησιμοποιώντας δεδομένα που περιλαμβάνουν πληροφορίες για γνωστούς τρομοκράτες.

Έναν άλλο περιορισμό αποτελεί το γεγονός, ότι ενώ μπορεί να συνδέσει συμπεριφορές και χαρακτηριστικά, δεν σημαίνει απαραίτητα ότι είναι σε θέση να

προσδιορίσει την αιτία μιας κατάστασης. Για παράδειγμα, μια εφαρμογή μπορεί να προσδιορίσει ότι ένα πρότυπο συμπεριφοράς, όπως η κλίση για αγορά αεροπορικών εισιτηρίων λίγο πριν την αναχώρηση της πτήσης, αναφέρεται σε χαρακτηριστικά όπως το εισόδημα, το επίπεδο εκπαίδευσης και τη χρήση του παγκόσμιου ιστού. Αυτό δεν σημαίνει, ωστόσο, ότι η αγορά του εισιτηρίου συνδέεται απαραίτητα με κάποιο από αυτά τα χαρακτηριστικά. Στην πραγματικότητα, η συμπεριφορά ενός τέτοιου ατόμου μπορεί να επηρεάζεται από κάποια επιπλέον χαρακτηριστικά όπως η έλλειψη χρόνου εξαιτίας της φύσης της εργασίας του.

1.2 Χρήσεις της Εξόρυξης Γνώσης

Η Εξόρυξη Γνώσης χρησιμοποιείται σε μία ευρεία γκάμα τομέων και δραστηριοτήτων. Τράπεζες, ασφαλιστικές εταιρίες, εταιρίες παρασκευής φαρμάκων και λιανικών πωλήσεων χρησιμοποιούν την Εξόρυξη Γνώσης προκειμένου να μειώσουν το κόστος, να προάγουν την έρευνα και να αυξήσουν τις πωλήσεις. Για παράδειγμα, οι ασφαλιστικές εταιρίες και οι τράπεζες χρησιμοποιούν διάφορα εργαλεία Εξόρυξης Γνώσης για την ανίχνευση απειλών και την αποτίμηση κινδύνων (π.χ. αναλήψεις μεγάλων χρηματικών ποσών μέσω καρτών ανάληψης). Συλλέγοντας πληροφορίες σχετικές με τη συμπεριφορά των πελατών, οι εταιρίες μπορούν να αναπτύξουν μοντέλα που να μπορούν να προβλέπουν εάν ένας πελάτης διεκπεραιώνει εγκαίρως τις οικονομικές του υποχρεώσεις ή αν ο ισχυρισμός για κάποιο ατύχημα είναι ύποπτος και πρέπει να διερευνηθεί περαιτέρω. Η ιατρική κοινότητα χρησιμοποιεί την Εξόρυξη Γνώσης για την πρόβλεψη της αποτελεσματικότητας ενός νέου φαρμάκου ή κάποιας νέας ιατρικής μεθόδου όπως και οι φαρμακευτικές εταιρίες για την παρασκευή νέων χημικών μειγμάτων προκειμένου να ανακαλύψουν νέες θεραπείες για διάφορες ασθένειες. Οι μεγάλες αλυσίδες λιανικού εμπορίου χρησιμοποιούν τις πληροφορίες που συλλέγουν μέσω διαφόρων προγραμμάτων π.χ. μαγνητική κάρτα σούπερ μάρκετ, παράπονα και αγωγές πελατών, για την αποτίμηση της αποτελεσματικότητας της προώθησης ενός προϊόντος και της τοποθέτησης σε συγκεκριμένο ράφι, της προσφοράς εκπτωτικών κουπονιών και της παράλληλης αγοράς προϊόντων. Τέλος, εταιρίες όπως παροχείς τηλεπικοινωνιακών ή μουσικών υπηρεσιών μπορούν να χρησιμοποιήσουν την Εξόρυξη Γνώσης προκειμένου να αποτιμήσουν ποιοι πελάτες τους είναι πιθανό να παραμείνουν συνδρομητές τους και ποιοι θα απευθυνθούν σε ανταγωνιστές τους.

Αναφορικά με το δημόσιο τομέα, οι διάφορες τεχνικές Εξόρυξης Γνώσης ενώ χρησιμοποιήθηκαν αρχικά για την ανίχνευση απειλών, στις μέρες μας χρησιμοποιούνται και για άλλους σκοπούς όπως για τη παρακολούθηση και βελτίωση της απόδοσης προγραμμάτων. Η Εξόρυξη Γνώσης ήταν αυτή που βοήθησε το 2000 την ομοσπονδιακή κυβέρνηση των Η.Π.Α. να ανακτήσει εκατομμύρια δολάρια που προοριζόταν για τη μισθοδοσία του προσωπικού ιατρικής φροντίδας ηλικιωμένων [10]. Οι Αστυνομικές Υπηρεσίες χρησιμοποιούν την Εξόρυξη Γνώσης για να διαμορφώσουν πρότυπα εγκληματιών βοηθώντας στην πρόληψη και την εξιχνίαση εγκληματικών ενεργειών. Ένα άλλο παράδειγμα αποτελεί η υπηρεσία πολιτικής αεροπορίας που χρησιμοποιεί την Εξόρυξη Γνώσης προκειμένου να επανεξετάσει τα δεδομένα που σχετίζονται με τα αεροπορικά ατυχήματα ώστε να ανακαλύψει τυχόν ελλείψεις και τρόπους αντιμετώπισης τους με σκοπό πάντοτε τη βελτίωση της ασφάλειας των πτήσεων.

Πρόσφατα, η Εξόρυξη Γνώσης έχει συνδεθεί στενά με τον τομέα της εθνικής ασφάλειας. Πολλοί παρατηρητές έχουν προτείνει ότι η Εξόρυξη Γνώσης θα μπορούσε να χρησιμοποιηθεί για τον προσδιορισμό τρομοκρατικών ενεργειών όπως μεταφορές χρημάτων ή υποκλοπές επικοινωνιών καθώς και για τον εντοπισμό και την παρακολούθηση των τρομοκρατών μέσω ταξιδιωτικών ή μεταναστευτικών εγγραφών. Δύο παραδείγματα τέτοιων προγραμμάτων αποτελούν το Terrorism Information Awareness (TIA) [54] χρηματοδοτούμενο από το Defense Advanced Research Projects Agency (DARPA) [55] και το Computer-Assisted Passenger Prescreening System II (CAPPS II) [56] που υλοποιείται από το Transportation Security Administration (TSA) [57] και αντικαταστάθηκε από το Secure Flight.

Το υπόλοιπο της παρούσας μελέτης είναι οργανωμένο ως εξής: το κεφάλαιο 2 αναφέρεται στην Εξόρυξη Γνώσης από κατανεμημένες βάσεις δεδομένων. Συγκεκριμένα, παρατίθενται τα πλεονεκτήματα της κατανεμημένης Εξόρυξης Γνώσης και κάποια επιμέρους ζητήματα. Στο κεφάλαιο 3 αναφέρονται οι βασικοί ορισμοί των εννοιών της Εξόρυξης Γνώσης. Το κεφάλαιο 4 παραθέτει τη βιβλιογραφική έρευνα που έχει γίνει στο χώρο της Εξόρυξης Γνώσης σχετικά με το θέμα της ασφάλειας. Το κεφάλαιο 5 περιγράφει το πρόβλημα, που θα μελετηθεί στην παρούσα μελέτη καθώς και τους αλγορίθμους Εξόρυξης Γνώσης που έχουν προταθεί

τόσο για ένα κατανεμημένο περιβάλλον όσο και για μία απλή κεντρική βάση δεδομένων. Το κεφάλαιο 6 περιγράφει τις τεχνικές απόκρυψης των ευαίσθητων κανόνων συσχέτισης, τους αλγορίθμους υπολογισμού του Ορίου Ασφαλείας και τη μέθοδο ανάκτησης των κανόνων συσχέτισης. Το κεφάλαιο 7 περιλαμβάνει τα πειράματα για την αξιολόγηση του προτεινόμενου συστήματος. Το κεφάλαιο 8 παραθέτει τα βασικά συμπεράσματα που συνάγονται από τα πειράματα. Το κεφάλαιο 9 περιγράφει τους άξονες της μελλοντικής έρευνας. Τέλος, υπάρχει ένα παράρτημα όπου γίνεται αντιστοίχιση της ελληνικής ορολογίας σε αγγλική, ερμηνεία των συμβόλων που χρησιμοποιούνται, τα ονόματα των αλγορίθμων, οι πίνακες με τις μετρήσεις και η περιγραφή των περιεχομένων του συνοδευτικού CD.

2. Εξόρυξη Γνώσης από Κατανεμημένες Βάσεις Δεδομένων

Οι έννοιες της Κατανεμημένης και της Παράλληλης Εξόρυξης Γνώσης αναφέρονται μαζί σε πολλούς τίτλους της βιβλιογραφίας. Ενώ και οι δύο προσπαθούν να βελτιώσουν την απόδοση των παραδοσιακών συστημάτων Εξόρυξης Γνώσης, υποθέτουν διαφορετικές αρχιτεκτονικές και ακολουθούν διαφορετικές προσεγγίσεις. Στην πρώτη περίπτωση οι Ηλεκτρονικοί Υπολογιστές, που πιθανώς να βρίσκονται σε διαφορετικές γεωγραφικές περιοχές, επικοινωνούν μέσα από την ανταλλαγή μηνυμάτων μέσω δικτύου. Αντίθετα, στη δεύτερη περίπτωση υπάρχουν πολλοί επεξεργαστές οι οποίοι μοιράζονται την ίδια μνήμη ή και τον ίδιο δίσκο. Αυτή η διαφορά επιδρά στη σχεδίαση των αλγορίθμων, στο κόστος και στην απόδοση των προτεινόμενων μοντέλων.

2.1 Πλεονεκτήματα της Κατανεμημένης Εξόρυξης Γνώσης

Μέχρι πρόσφατα, η έρευνα στην Εξόρυξη Γνώσης θεωρούσε μόνο κεντρικές βάσεις δεδομένων όπου τα δεδομένα αποθηκεύονταν σε ένα μόνο μέρος[63]. Ωστόσο, αυτές οι τεχνικές δεν μπορούν να χρησιμοποιηθούν σε ένα κατανεμημένο περιβάλλον. Οι παρακάτω λόγοι έχουν καταστήσει την Κατανεμημένη Εξόρυξη Γνώσης αρκετά ενδιαφέρουσα:

- *Κατανεμημένα Δεδομένα*: Σε κάποιες εφαρμογές τα δεδομένα είναι κατανεμημένα αλλά πρέπει να αποκτήσουμε συνολική γνώση κάποιων

στοιχείων. Για παράδειγμα, κάθε υποκατάστημα μιας πολυεθνικής εταιρίας χειρίζεται τοπικά τα δικά του δεδομένα αλλά η κεντρική διεύθυνση θα πρέπει να αναλύσει τα δεδομένα όλων των υποκαταστημάτων προκειμένου να διαμορφώσει μια γενική στρατηγική.

Η άμεση λύση είναι να μεταφερθούν τα δεδομένα σε μία κεντρική βάση δεδομένων[5]. Ωστόσο, κάτι τέτοιο είναι αρκετά χρονοβόρο και δαπανηρό. Επιπλέον, μερικές φορές είναι αδύνατη η μεταφορά τους για λόγους ασφάλειας.

- *Απόδοση και κλιμάκωση*: Η Κατανεμημένη Εξόρυξη Γνώσης μπορεί να είναι απαραίτητη και στην περίπτωση που έχουμε μία μόνο βάση δεδομένων. Ένα τέτοιο σενάριο μπορεί να υπάρξει στην περίπτωση που η ποσότητα των δεδομένων είναι πολύ μεγάλη και η επεξεργασία τους δεν μπορεί να γίνει σε ένα μόνο μέρος. Σε αυτή την περίπτωση, θα πρέπει να σταλεί ένα τμήμα των δεδομένων σε άλλα μέρη.

  Ακόμη και αν ένα μέρος είναι σε θέση να επεξεργαστεί το σύνολο των δεδομένων[16] ενδεχομένως να αξίζει τον κόπο η μεταφορά ενός τμήματος ή ολόκληρης της ποσότητας των δεδομένων στα άλλα μέρη. Κάθε μέρος μπορεί να λαμβάνει και να επεξεργάζεται ένα τμήμα των δεδομένων (κατανομή δεδομένων[14]) ή να πραγματοποιεί ένα κομμάτι της διαδικασίας (κατανομή εργασιών[69]) σε πανομοιότυπα σύνολα δεδομένων. Με άλλα λόγια, τα αποτελέσματα από όλα τα μέρη συνδυάζονται μεταξύ τους καθιστώντας την όλη διαδικασία πιο γρήγορη, καθώς υπάρχει μείωση του χρόνου επεξεργασίας εξαιτίας της κατανομής των δεδομένων.

  Οι Κατανεμημένες τεχνικές Εξόρυξης Γνώσης είναι κλιμακωτές. Μία τεχνική Εξόρυξης Γνώσης είναι κλιμακωτή, όταν η απόδοση της δεν μειώνεται με την αύξηση του μεγέθους του συνόλου δεδομένων. Σε ένα κατανεμημένο περιβάλλον, ο αριθμός των μερών που περιλαμβάνονται σε μία δραστηριότητα μπορεί να αυξομειώνεται ανάλογα με το μέγεθος του συνόλου δεδομένων. Η συνολική απόδοση ενός τέτοιου συστήματος παραμένει σταθερή.

Φυσικά, η Κατανεμημένη Εξόρυξη Γνώσης έχει και κόστος το οποίο είναι πολύ μεγαλύτερο εν συγκρίσει με αυτό της Εξόρυξης Γνώσης από μια κεντρική βάση δεδομένων.

2.2 Επιμέρους ζητήματα

Η κατασκευή ενός συστήματος που θα έχει ως στόχο την Εξόρυξη Γνώσης από κατανεμημένα σύνολα δεδομένων δεν είναι απλή υπόθεση. Μερικά από τα ζητήματα που θα πρέπει να ληφθούν υπόψιν αναλύονται παρακάτω:

- *Ομοιογενή – Ετερογενή Δεδομένα*[32,30]: Οι περισσότερες μελέτες υποθέτουν ότι οι τοπικές βάσεις δεδομένων είναι ομοιογενείς δηλαδή βρίσκονται στην ίδια πλατφόρμα, ελέγχονται από ίδιο Σύστημα Διαχείρισης Βάσεων – DBMS[19] με το ίδιο σχήμα και τα αντίστοιχα χαρακτηριστικά έχουν το ίδιο domain. Αν αντίθετα είναι ετερογενείς, τα συστήματα Εξόρυξης Γνώσης θα πρέπει να προσαρμοστούν, ώστε να δουλεύουν σε τοπικές βάσεις δεδομένων. Επιπλέον, τα τοπικά σχήματα πρέπει να ενοποιηθούν σε ένα καθολικό σχήμα, διαφορετικά, είναι δύσκολο, αν όχι αδύνατο να ερμηνευτούν τα αποτελέσματα.

- *Κατάτμηση Δεδομένων*[24]: Ένας τοπικός πίνακας σε μία τοπική βάση δεδομένων μπορεί να αποτελεί κομμάτι ενός καθολικού πίνακα. Η κατάτμηση μπορεί να είναι κατακόρυφη ή οριζόντια. Στην οριζόντια κατάτμηση, κάθε τεμάχιο είναι υποσύνολο από πλειάδες[72] στον καθολικό πίνακα και τα τεμάχια διαμοιράζονται το ίδιο σχήμα. Στην κατακόρυφη κατάτμηση, κάθε τεμάχιο έχει ένα υποσύνολο από χαρακτηριστικά του καθολικού πίνακα. Ένας συνδυασμός των δύο μεθόδων είναι επίσης δυνατός σε μία βάση δεδομένων όπου κάποιοι πίνακες είναι κατακόρυφα κατατμημένοι και κάποιοι άλλοι οριζόντια.

- *Ανταλλαγή Δεδομένων*[55]: Μερικά ή όλα τα δεδομένα στην τοπική βάση δεδομένων πρέπει να μεταδοθούν στα άλλα μέρη. Η ανταλλαγή βελτιώνει τη διαθεσιμότητα των δεδομένων αλλά κάνει δυσκολότερη τη διατήρηση της συνοχής τους. Κανονικά η ανταλλαγή των δεδομένων δεν γίνεται στα πλαίσια

της Εξόρυξης Γνώσης αλλά είναι μια απόφαση που σχετίζεται με τις υπολογιστικές δυνατότητες. Μπορούμε, ωστόσο, να ανταλλάξουμε δεδομένα στα πλαίσια της Εξόρυξης Γνώσης. Σε αυτή την περίπτωση το σύστημα πρέπει να αποφασίσει ποια δεδομένα ή μέρος αυτών θα αποστείλει.

- *Κόστος Μετάδοσης*: Στη διαδικασία της Εξόρυξη Γνώσης, από μία απλή βάση δεδομένων, κύριο μέτρο για την αξιολόγηση της αποτελεσματικότητας του αντίστοιχου αλγορίθμου αποτελεί ο χρόνος Εισόδου/Εξόδου[36] ή/και ο χρόνος επεξεργασίας της Κεντρικής Μονάδας Επεξεργασίας. Σε ένα κατανεμημένο περιβάλλον αυτό που πρέπει να ληφθεί υπόψιν είναι το κόστος μετάδοσης, το οποίο υπολογίζεται από το εύρος ζώνης[3] του δικτύου και τον αριθμό των μηνυμάτων που ανταλλάσσονται μέσω του δικτύου.

- *Ενοποίηση Αποτελεσμάτων*[37]: Η ενοποίηση δεν είναι τόσο εύκολη υπόθεση. Ένα μεγάλο στοιχειοσύνολο σε ένα τοπικό μέρος ενδεχομένως να μην είναι καθολικώς μεγάλο. Δεδομένου, ότι η επιτυχία της Κατανεμημένης Εξόρυξης Γνώσης συνίσταται στην εύρεση καθολικών προτύπων, τα πρότυπα και οι ιδιότητες τους θα πρέπει να επιλέγονται από όλα τα μέρη.

- *Ασύμμετρα Δεδομένα*[17]: Οι στατιστικές κατανομές δεδομένων, όπως οι τιμές των χαρακτηριστικών είναι συνήθως διαφορετικές μεταξύ των τοπικών βάσεων δεδομένων. Ένα τοπικό μοντέλο που αποκτάται, μέσω Εξόρυξης από μία τοπική βάση δεδομένων, αναπόφευκτα επηρεάζεται από μία τέτοια κατανομή. Μια τέτοια ασυμμετρία δεδομένων έχει ως αποτέλεσμα, τα τοπικά μοντέλα να είναι ανακριβή και μερικές φορές χωρίς καμία ουσία. Για παράδειγμα, ένας κατηγοριοποιητής που εκπαιδεύεται από μια τοπική βάση δεδομένων, η οποία έχει ελάχιστα ή καθόλου στιγμιότυπα της κλάσης, δεν θα είναι σε θέση να κατηγοριοποιήσει ένα μελλοντικό στιγμιότυπο αυτής της κλάσης.

Όλα τα επιμέρους ζητήματα που προαναφέρθηκαν δεν είναι ανεξάρτητα μεταξύ τους. Για παράδειγμα, η κατακόρυφη κατάτμηση ενός καθολικού πίνακα θα μετατρέψει τα δεδομένα των τοπικών βάσεων δεδομένων σε ετερογενή. Η οριζόντια κατάτμηση των

δεδομένων, ενδεχομένως να προκαλέσει ασυμμετρία στα δεδομένα, αν δεν γίνει προσεχτικά.

Υπάρχουν και αρκετά άλλα ζητήματα που σχετίζονται με τη διαδικασία της κατανεμημένης Εξόρυξης Γνώσης όπως η ασφάλεια, η αυτονομία των τοπικών βάσεων δεδομένων, η τοπολογία του δικτύου και το σχήμα μεταφοράς.

3. Ορισμοί – Υπόβαθρο

Έστω ότι έχουμε το σύνολο $I = \{i_1,...,i_n\}$ όπου $i_1,\ldots,i_n$ ονομάζονται στοιχεία[37] και μια βάση δεδομένων[18] ΒΔ από συναλλαγές[71], όπου κάθε συναλλαγή T είναι ένα στοιχειοσύνολο[38] τέτοιο ώστε $T \subseteq I$. Ένα κ-στοιχειοσύνολο αποτελείται από κ στοιχεία. Μία συναλλαγή T υποστηρίζει ένα σύνολο X από στοιχεία στο I αν $X \subseteq T$. Δηλαδή η υποστήριξη[66] του X είναι το ποσοστό των συναλλαγών που περιέχουν το X.

Ένας κανόνας συσχέτισης[1] έχει τη μορφή $X \Rightarrow Y$, όπου $X \subset I, Y \subset I$ και $X \cap Y = \text{Ø}$. Λέμε ότι ο κανόνας $X \Rightarrow Y$ έχει *εμπιστοσύνη*[11] c% αν $\frac{|X \cup Y| \times 100}{|X|} = c$, όπου |A| είναι ο αριθμός των εμφανίσεων του συνόλου των στοιχείων του στοιχειοσυνόλου A στις συναλλαγές της βάσης δεδομένων ΒΔ. Λέμε ότι ο κανόνας $X \Rightarrow Y$ έχει *υποστήριξη* s% αν $\frac{|X \cup Y| \times 100}{N} = s$, όπου N είναι ο αριθμός των συναλλαγών στη βάση δεδομένων ΒΔ. Θα πρέπει να παρατηρήσουμε ότι η υποστήριξη αποτελεί ένα μέτρο της συχνότητας εμφάνισης ενός κανόνα συσχέτισης, ενώ η εμπιστοσύνη αποτελεί ένα μέτρο, που μας δείχνει τη σχέση μεταξύ των συνόλων των στοιχείων.

Ένας κανόνας συσχέτισης r αντιστοιχεί σε ένα στοιχειοσύνολο, που αποτελεί την ένωση των στοιχείων που βρίσκονται στο αριστερό (l_r) και στο δεξί (r_r) άκρο του κανόνα. Ορίζουμε το στοιχειοσύνολο που αντιστοιχεί στον κανόνα r σαν Σ_r και αναφερόμαστε σε αυτό με τον όρο παραγόμενο στοιχειοσύνολο του r. Θα πρέπει να

σημειωθεί, ότι 2 διαφορετικοί κανόνες μπορούν, να έχουν το ίδιο παραγόμενο στοιχειοσύνολο.

Επειδή ο αριθμός των στοιχειοσυνόλων και των κανόνων συσχέτισης αυξάνεται εκθετικά σε σχέση με τον αριθμό των στοιχείων στη βάση δεδομένων, διατηρούμε μόνο τους κανόνες των οποίων οι τιμές της υποστήριξης και της εμπιστοσύνης είναι μεγαλύτερες από τις προσδιοριζόμενες από τον χρήστη κατώτατες τιμές: ΕΤΥ (Ελάχιστη Τιμή Υποστήριξης) και ΕΤΕ (Ελάχιστη Τιμή Εμπιστοσύνης).

Δοθέντος του συνόλου $I=\{i_1,i_2,\ldots,i_n\}$, μία συναλλαγή Τ μπορεί να αναπαρασταθεί ως ένα δυαδικό διάνυσμα $(t_1,t_2,...,t_n)$ όπου $t_j = 1$ αν και μόνο αν $t_j \in T$, διαφορετικά $t_j = 0$. Για παράδειγμα, έστω ότι διαθέτουμε το σύνολο I={A1,A2,A3,A4} από στοιχεία και τη συναλλαγή Τ={A1,A3} η οποία θα αναπαρασταθεί ως εξής: Τ=[1010]. Με βάση αυτόν τον τρόπο αναπαράστασης, μία συναλλαγή Τ υποστηρίζει ένα στοιχειοσύνολο Α $(A \subseteq T)$ αν ισχύει η σχέση $A \cap T = A$.

Μία συναλλαγή Τ *υποστηρίζει πλήρως*[26] ένα στοιχειοσύνολο Σ αν όλα τα στοιχεία του Σ έχουν την τιμή 1. Μία συναλλαγή Τ *υποστηρίζει μερικώς*[48] ένα στοιχειοσύνολο Σ αν υπάρχει τουλάχιστον ένα στοιχείο του Σ που δεν έχει την τιμή 1. Για παράδειγμα, έστω ότι διαθέτουμε όπως και προηγουμένως το σύνολο I={A1,A2,A3,A4}, το στοιχειοσύνολο Σ={A1,A2,A3}=[1110] και τις συναλλαγές Τ1={A1,A3}=[1010] και Τ2={A1,A2,A3}=[1110]. Λέμε λοιπόν ότι η συναλλαγή Τ1 υποστηρίζει μερικώς το στοιχειοσύνολο Σ και η συναλλαγή Τ2 υποστηρίζει πλήρως το στοιχειοσύνολο Σ.

Ένας κανόνας συσχέτισης ονομάζεται *ισχυρός*[64] αν η εμπιστοσύνη του είναι μεγαλύτερη από την ΕΤΕ. Ένα στοιχειοσύνολο ονομάζεται *μεγάλο*[41] αν η υποστήριξη του είναι μεγαλύτερη από την ΕΤΥ και συμβολίζουμε με L το σύνολο των μεγάλων στοιχειοσυνόλων και με $L_H \subseteq L$ το σύνολο των μεγάλων στοιχειοσυνόλων της βάσης δεδομένων που θέλουμε να κρύψουμε. Θα πρέπει να σημειώσουμε ότι το L_H είναι το σύνολο των παραγόμενων στοιχειοσυνόλων των κανόνων του συνόλου R_H. Ονομάζουμε T_Z το σύνολο των συναλλαγών που υποστηρίζουν το στοιχειοσύνολο Z. Χρησιμοποιούμε τον συμβολισμό T_r για να ορίσουμε το σύνολο των συναλλαγών που υποστηρίζουν πλήρως το παραγόμενο στοιχειοσύνολο του κανόνα r.

Εκτός των τιμών ‘0’ και ‘1’ εισάγουμε και την τιμή ‘?’ η οποία χρησιμοποιείται για την απόκρυψη της τιμής που περιέχεται στη θέση αυτή. Εξαιτίας της ύπαρξης του συμβόλου ‘?’ οι ορισμοί για την υποστήριξη και την εμπιστοσύνη διαφοροποιούνται. Δεν έχουμε πλέον μία μοναδική τιμή αλλά ένα διάστημα τιμών.

Το *διάστημα υποστήριξης*[67] για ένα στοιχειοσύνολο A αναπαρίσταται ως [Ελάχιστη_Υποστήριξη(A),Μέγιστη_Υποστήριξη(A)] όπου η πραγματική υποστήριξη του στοιχειοσυνόλου A βρίσκεται μεταξύ των τιμών Ελάχιστη_Υποστήριξη(A) και Μέγιστη_Υποστήριξη(A). Η Ελάχιστη_Υποστήριξη(A) αντιστοιχεί στην υποστήριξη που υπολογίζεται στη ΒΔ αν όλα τα ‘?’ αντικατασταθούν με ‘0’ και η Μέγιστη_Υποστήριξη(A) αντιστοιχεί στην υποστήριξη που υπολογίζεται αν όλα τα ‘?’ αντικατασταθούν με ‘1’ αντίστοιχα.

Το *διάστημα εμπιστοσύνης*[11] για έναν κανόνα συσχέτιση $A \Rightarrow B$ αναπαρίσταται ως [Ελάχιστη_Εμπιστοσύνη($A \Rightarrow B$),Μέγιστη_Εμπιστοσύνη($A \Rightarrow B$)] όπου η πραγματική εμπιστοσύνη του κανόνα $A \Rightarrow B$ βρίσκεται μεταξύ των τιμών Ελάχιστη_Εμπιστοσύνη($A \Rightarrow B$) και Μέγιστη_Εμπιστοσύνη($A \Rightarrow B$). Δοθέντων των ελάχιστων και μέγιστων τιμών υποστήριξης για τα στοιχειοσύνολα AB και A η ελάχιστη τιμή εμπιστοσύνης για τον κανόνα $A \Rightarrow B$ είναι:

$$Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{Ελάχιστη\_Υποστήριξη(AB) \times 100}{Μέγιστη\_Υποστήριξη(A)} \quad \textbf{(1)}$$

και η μέγιστη τιμή εμπιστοσύνης

$$Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{Μέγιστη\_Υποστήριξη(AB) \times 100}{Ελάχιστη\_Υποστήριξη(A)} \quad \textbf{(2)}$$

Θα πρέπει να σημειώσουμε πως όταν δεν υπάρχουν άγνωστες τιμές[73] (?) τότε Ελάχιστη_Εμπιστοσύνη($A \Rightarrow B$)=Μέγιστη_Εμπιστοσύνη($A \Rightarrow B$) και Ελάχιστη_Υποστήριξη(AB)=Μέγιστη_Υποστήριξη(AB).

Ας δούμε τα παραπάνω με ένα παράδειγμα. Έχουμε τη βάση δεδομένων του πίνακα 1 και τη βάση δεδομένων του πίνακα 2 που περιέχει και άγνωστες τιμές.

A1	A2	A3	A4
1	1	1	0
1	0	1	0
0	0	1	1
0	0	1	0
1	0	1	1

Πίνακας 1: βάση δεδομένων με 0 και 1

A1	A2	A3	A4
?	1	1	1
1	0	1	0
0	0	?	1
0	0	1	?
?	0	1	1

Πίνακας 2: βάση δεδομένων με 0, 1 και ?

Θα υπολογίσουμε την εμπιστοσύνη του κανόνα $A3 \Rightarrow A1$ από τη βάση δεδομένων του πίνακα 1 και του πίνακα 2 (με άγνωστες τιμές) αντίστοιχα. Καθώς ο πίνακας 1 δεν περιέχει άγνωστες τιμές η εμπιστοσύνη του κανόνα $A3 \Rightarrow A1$ υπολογίζεται από τη σχέση $Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{Υποστήριξη(A1A3)}{Υποστήριξη(A3)}$

όπου $Υποστήριξη(A1A3) = \frac{3}{5} = 60\%$ και $Υποστήριξη(A3) = \frac{5}{5} = 100\%$ οπότε $Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{60}{100} = 60\%$. Αντίθετα ο πίνακας 2 περιέχει άγνωστες τιμές, οπότε θα έχουμε το διάστημα εμπιστοσύνης [Ελάχιστη_Εμπιστοσύνη ($A3 \Rightarrow A1$), Μέγιστη_Εμπιστοσύνη ($A3 \Rightarrow A1$)] όπου

$Ελάχιστη\_Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{Ελάχιστη\_Υποστήριξη(A1A3)}{Μέγιστη\_Υποστήριξη(A3)}$ και

$Μέγιστη\_Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{Μέγιστη\_Υποστήριξη(A1A3)}{Ελάχιστη\_Υποστήριξη(A3)}$

Υπολογίζουμε τα διαστήματα υποστήριξης
[Ελάχιστη_Υποστήριξη(A3),Μέγιστη_Υποστήριξη(A3)] και
[Ελάχιστη_Υποστήριξη(A1A3),Μέγιστη_Υποστήριξη(A1A3)]. Θα έχουμε λοιπόν

$Ελάχιστη\_Υποστήριξη(A3) = \frac{4}{5} = 80\%$ και $Μέγιστη\_Υποστήριξη(A3) = \frac{5}{5} = 100\%$

$Ελάχιστη\_Υποστήριξη(A1A3) = \frac{1}{5} = 20\%$ και $Μέγιστη\_Υποστήριξη(A1A3) = \frac{3}{5} = 60\%$

Συνεπώς θα έχουμε $Ελάχιστη\_Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{1/5}{5/5} = 20\%$ και

$Μέγιστη\_Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{3/5}{4/5} = 75\%$.

Όπως αναφέρθηκε και παραπάνω η εμπιστοσύνη που προέκυψε για την πρώτη βάση δεδομένων (60%) είναι μεταξύ του Διαστήματος Εμπιστοσύνης [20%,70%] της δεύτερης βάσης δεδομένων.

Από το παραπάνω παράδειγμα προκύπτουν δύο σημαντικές διαπιστώσεις:

- $Ελάχιστη\_Υποστήριξη \le Πραγματική\,Υποστήριξη \le Μέγιστη\_Υποστήριξη$
- $Ελάχιστη\_Εμπιστοσύνη \le Πραγματική\,Εμπιστοσύνη \le Μέγιστη\_Εμπιστοσύνη$

Οι τιμές που υπολογίζονται με βάση τους ορισμούς της ελάχιστης και μέγιστης εμπιστοσύνης, που παρατέθηκαν παραπάνω, δεν οδηγούν σε σωστά αποτελέσματα στην περίπτωση που υπάρχει εξάρτηση μεταξύ του αριθμητή και του παρανομαστή. Έστω, ότι θέλουμε να υπολογίσουμε την ελάχιστη εμπιστοσύνη. Αν αντικαταστήσουμε το στοιχείο Α με '?' σε μια συγκεκριμένη συναλλαγή τότε δεν μπορούμε να τοποθετήσουμε '1' στο στοιχείο Α προκειμένου να υπολογίσουμε την $Μέγιστη\_Υποστήριξη(A)$ και ταυτόχρονα να τοποθετήσουμε '0' στο ίδιο στοιχείο, ώστε να υπολογίσουμε την $Ελάχιστη\_Υποστήριξη(AB)$. Συνεπώς, η εξάρτηση μεταξύ του αριθμητή και του παρανομαστή πρέπει να ληφθεί υπόψιν στον υπολογισμό της ελάχιστης εμπιστοσύνης. Η σχέση η οποία υπολογίζει ακριβώς την ελάχιστη εμπιστοσύνη του κανόνα $A \Rightarrow B$ είναι η εξής:

$$Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{\left|A \cup B\right|}{\left|A\right| + \left|A^{?} \cup B'\right| + \left|A^{?} \cup B^{?}\right|} \quad \textbf{(3)}$$

όπου $\left|\mathrm{A} \cup \mathrm{B}\right|$ αντιστοιχεί στο πλήθος των συναλλαγών όπου Α=1 και Β=1, $\left|\mathrm{A}^{?} \cup \mathrm{B}'\right|$ αντιστοιχεί στο πλήθος των συναλλαγών όπου Α=? Και Β=0 κ.ο.κ. Όταν ο κανόνας έχει περισσότερα από δύο στοιχεία είναι δηλαδή της μορφής $l_1 l_2 ... l_n \Rightarrow r_1 r_2 ... r_n$ ο προηγούμενος ορισμός επεκτείνεται ως εξής

$$Ελάχιστη\_Εμπιστοσύνη(l_1 l_2 ... l_n \Rightarrow r_1 r_2 ... r_n) = \frac{\left|l_1 l_2 ... l_n \cup r_1 r_2 ... r_n\right|}{\left|l_1 l_2 ... l_n\right| + \sum \left|(l_1 l_2 ... l_n)^{1?} \cup (r_1 r_2 ... r_n)'\right|} \quad \textbf{(4)}$$

όπου ο αριθμητής αντιστοιχεί στο πλήθος των συναλλαγών που υποστηρίζουν το παραγόμενο στοιχειοσύνολο $l_1 l_2 ... l_n \cup r_1 r_2 ... r_n$ του κανόνα και ο παρανομαστής αντιστοιχεί στο άθροισμα του πλήθους των συναλλαγών που υποστηρίζουν το στοιχειοσύνολο $l_1 l_2 ... l_n$ με το πλήθος των συναλλαγών που υποστηρίζουν το αριστερό

στοιχειοσύνολο του κανόνα και δεν υποστηρίζουν το δεξί στοιχειοσύνολο. Συνοπτικά, ο όρος $(l_1 l_2 ... l_n)^{1?}$ ορίζει ότι όλα τα στοιχεία $l_1 l_2 ... l_n$ πρέπει να είναι είτε ? είτε 1 και ο όρος $(r_1 r_2 ... r_n)'$ ορίζει ότι τουλάχιστον ένα στοιχείο του $r_1 r_2 ... r_n$ πρέπει να είναι 0.

Παρόμοια, προκύπτει, ότι η μέγιστη εμπιστοσύνη του κανόνα $A \Rightarrow B$ δίνεται από τον τύπο $Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B) \frac{|A \cup B| + |A \cup B^?| + |A^? \cup B| + |A^? \cup B^?|}{|A| + |A^? \cup B| + |A^? \cup B^?|}$ **(5)**

και γενικά η μέγιστη εμπιστοσύνη είναι

$$Μέγιστη\_Εμπιστοσύνη(l_1 l_2 ... l_n \Rightarrow r_1 r_2 ... r_m) = \frac{\sum |(l_1 l_2 ... l_n)^{1?} \cup (r_1 r_2 ... r_m)^{1?}|}{|l_1 l_2 ... l_n| + \sum |(l_1 l_2 ... l_n)^{1?} \cup (r_1 r_2 ... r_m)^{1?}|} \textbf{(6)}$$

Ας δούμε τους προαναφερόμενους ορισμούς με ένα παράδειγμα. Θεωρούμε και πάλι τη βάση δεδομένων του πίνακα 2 και τον κανόνα $A3 \Rightarrow A1$. Από τη σχέση (3) θα έχουμε $Ελάχιστη\_Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{1}{4+1+0} = \frac{1}{5} = 20\%$ και από τη σχέση (5) έχουμε $Μέγιστη\_Εμπιστοσύνη(A3 \Rightarrow A1) = \frac{1+2+0+0}{4+0+0} = \frac{3}{4} = 75\%$

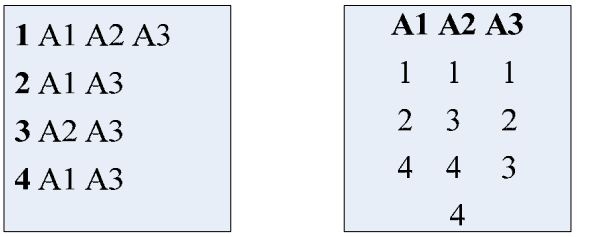
Παρατηρούμε ότι οι τιμές που προκύπτουν είναι ίδιες με αυτές που προέκυψαν από τις σχέσεις (1) και (2).

Οι προαναφερόμενοι ορισμοί ισχύουν για μία κεντρική βάση δεδομένων. Στην περίπτωση που διαθέτουμε περισσότερες βάσεις δεδομένων που επικοινωνούν μεταξύ τους θα πρέπει να εισάγουμε κάποιους επιπλέον όρους. Έστω, λοιπόν, ότι διαθέτουμε μία βάση δεδομένων ΒΔ και $ΒΔ_i$ οι διαμερίσεις αυτής, μεγέθους Μ και M_i αντίστοιχα. Ορίζουμε τα μεγέθη Χ.Υποστήριξη και Χ.Υποστήριξη$_{(i)}$ ως την υποστήριξη του στοιχειοσυνόλου Χ στις βάσεις ΒΔ και $ΒΔ_i$ αποκαλώντας τες *καθολική* και *τοπική υποστήριξη* αντίστοιχα. Το στοιχειοσύνολο Χ είναι *καθολικώς μεγάλο* αν $\text{Χ.Υποστήριξη} \geq s \cdot \text{Μ}$, όπου s η υποστήριξη του στοιχειοσυνόλου Χ. Το στοιχειοσύνολο Χ είναι *τοπικά μεγάλο* αν $\text{Χ.Υποστήριξη}_{(i)} \geq s \cdot \text{M}_i$. Συμβολίζουμε με ΚΜ το σύνολο των καθολικά μεγάλων στοιχειοσυνόλων στη βάση ΒΔ και με ΚΜ$_{(κ)}$ το σύνολο των καθολικά μεγάλων κ-στοιχειοσυνόλων. Αν ένα στοιχειοσύνολο είναι ταυτόχρονα καθολικά και τοπικά μεγάλο τότε αποκαλείται *κτ-μεγάλο*. Ορίζουμε

KTM_i το σύνολο των κτ-μεγάλων στοιχειοσυνόλων στη βάση $BΔ_i$, $KTM_{i(k)}$ το σύνολο των κτ-μεγάλων κ-στοιχειοσυνόλων στη βάση $BΔ_i$, $ΥΣ_{(k)}$ το σύνολο των υποψήφιων στοιχειοσυνόλων που προκύπτουν από το σύνολο $ΚΜ_{(k-1)}$ και $ΥΣ_{i(k)}$ το σύνολο των υποψήφιων στοιχειοσυνόλων που προκύπτουν από το σύνολο $KTM_{i(k-1)}$. Επίσης, με το συμβολισμό $TM_{i(k)}$ δηλώνουμε το σύνολο των τοπικά μεγάλων κ-στοιχειοσυνόλων στο σύνολο $ΥΣ_{i(k)}$.

Τέλος, θα πρέπει να αναφέρουμε ότι μια βάση βεδομένων μπορεί να είναι δομημένη με 2 τρόπους όπως φαίνεται στην εικόνα 1:

- *Οριζόντιο σχήμα*[33], όπου κάθε συναλλαγή σχετίζεται με έναν κωδικό (TID) και περιλαμβάνει όλα τα στοιχεία που περιέχονται σε αυτή
- *Κατακόρυφο σχήμα*[74], όπου κάθε στοιχείο σχετίζεται με μια λίστα όλων των κωδικών (TID) των συναλλαγών που περιλαμβάνουν αυτό το στοιχείο.



Εικόνα 1: Οριζόντιο – Κατακόρυφο σχήμα

4. Βιβλιογραφική Έρευνα

Το πρόβλημα της Προστασίας Γνώσης στον τομέα της Εξόρυξης Δεδομένων έχει απασχολήσει αρκετούς ερευνητές. Έχουν προταθεί, λοιπόν, αρκετές τεχνικές για την προστασία των δεδομένων από κακόβουλες επιθέσεις, οι οποίες μπορούν να κατηγοριοποιηθούν ως εξής:

- Ευριστικές τεχνικές
- Κρυπτογραφικές τεχνικές
- Τεχνικές ανακατασκευής των διαταραγμένων δεδομένων

4.1 Ευριστικές Τεχνικές

Βασίζονται στην υπόθεση ότι η τροποποίηση ή η απόκρυψη των επιλεγόμενων δεδομένων ανάγονται σε ένα NP-πρόβλημα. Διακρίνονται οι εξής κατηγορίες:

- Διατάραξη[50] κεντροποιημένων δεδομένων βασισμένη στη σύγχυση[13] των κανόνων συσχέτισης
- Μπλοκάρισμα[4] κεντροποιημένων δεδομένων βασισμένη στη σύγχυση των κανόνων συσχέτισης
- Μπλοκάρισμα κεντροποιημένων δεδομένων βασισμένη στη σύγχυση των κανόνων κατηγοριοποίησης.

4.1.1 Διατάραξη κεντροποιημένων δεδομένων βασισμένη στη σύγχυση των κανόνων συσχέτισης

Μία στοιχειοθετημένη απόδειξη ότι η βέλτιστη απόκρυψη ευαίσθητων μεγάλων στοιχειοσυνόλων ανάγεται σε NP-πρόβλημα παρουσιάζεται στο [6]. Τα στοιχεία αναπαρίστανται με '0' και '1' όπου όπως αναφέρθηκε και παραπάνω το '1' δηλώνει την παρουσία στοιχείου και το '0' την απουσία στοιχείου. Η μέθοδος προτείνει την αντικατάσταση των '1' με '0' έτσι ώστε να μειωθεί η υποστήριξη των ευαίσθητων κανόνων συσχέτισης κάτω από την ελάχιστη προσδιοριζόμενη από το χρήστη τιμή υποστήριξης ΕΤΥ.

Επέκταση της προαναφερθείσας μελέτης αποτελούν τα [16] και [48] όπου η απόκρυψη των ευαίσθητων μεγάλων στοιχειοσυνόλων επεκτείνεται στην απόκρυψη των ευαίσθητων κανόνων συσχέτισης. Παρουσιάζονται 2 διαφορετικές προσεγγίσεις:

- Μείωση της εμπιστοσύνης του κανόνα συσχέτισης, είτε αυξάνοντας την υποστήριξη του μέσω των συναλλαγών που υποστηρίζουν μερικώς το αριστερό μέρος του κανόνα, είτε μειώνοντας την υποστήριξη του μέσω των συναλλαγών που υποστηρίζουν και τα δύο μέρη του κανόνα.
- Μείωση της υποστήριξης του κανόνα μειώνοντας την υποστήριξη είτε του αριστερού είτε του δεξιού μέρους μέσω συναλλαγών που υποστηρίζουν πλήρως τον κανόνα.

Βάση για τη μελέτη στο [32] αποτέλεσε το [16] στοχεύοντας στη διατήρηση μιας ισορροπίας ανάμεσα στην ασφάλεια και στην αποκάλυψη πληροφορίας προσπαθώντας να ελαχιστοποιήσει τις επιδράσεις από την απόκρυψη στοιχείων στις συναλλαγές και τις έμμεσες απώλειες[62] (εμφάνιση νέων κανόνων και απόκρυψη υπαρχόντων επιπλέον των ευαίσθητων κανόνων) σε κανόνες συσχέτισης.

Οι αλγόριθμοι που υλοποιούνται στο [36] χρησιμοποιούν ως είσοδο μία δυαδική βάση δεδομένων και εντοπίζουν όλους τους κανόνες συσχέτισης. Στη συνέχεια, κρύβουν ένα υποσύνολο αυτών των κανόνων που θεωρούνται ευαίσθητοι από τον χρήστη, αντιστρέφοντας κάποιες από τις δυαδικές τιμές (π.χ. αντικαθιστούν τα '1' με '0'). Η διαδικασία απόκρυψης μπορεί, να επηρεάσει το σύνολο των κανόνων, που μπορούν να εξαχθούν από τη βάση δεδομένων, είτε κρύβοντας κανόνες που δεν είναι

ευαίσθητοι («απολεσθέντες κανόνες»), είτε εισάγοντας κανόνες που δεν υποστηρίζονται από την αρχική βάση δεδομένων («κανόνες φαντάσματα»).

4.1.2 Μπλοκάρισμα κεντροποιημένων δεδομένων βασισμένη στη σύγχυση των κανόνων συσχέτισης

Μία από τις πιο σημαντικές τεχνικές τροποποίησης δεδομένων που έχουν χρησιμοποιηθεί για τη σύγχυση των κανόνων συσχέτισης είναι το μπλοκάρισμα των δεδομένων [11] όπου έχουμε αντικατάσταση κάποιων στοιχείων με '?'.

Η προαναφερθείσα προσέγγιση ακολουθείται στα [40] και [41] . Η εισαγωγή του '?' στη βάση δεδομένων επιφέρει κάποιες αλλαγές στους ορισμούς της υποστήριξης και της εμπιστοσύνης ενός κανόνα συσχέτισης. Εισάγονται οι έννοιες του Διαστήματος Υποστήριξης και Διαστήματος Εμπιστοσύνης. Καθώς, η υποστήριξη ή/και η εμπιστοσύνη του ευαίσθητου κανόνα γίνεται μικρότερη από το μέσο αυτών των διαστημάτων, εξασφαλίζεται η μη παραβίαση της εμπιστευτικότητας των δεδομένων. Για να έχουμε τα προσδοκώμενα αποτελέσματα θα πρέπει να γίνεται εισαγωγή των '?' τόσο στη θέση των '0' όσο και στη θέση των '1'.

Στο [37] παρουσιάζεται ένας νέος αλγόριθμος ο οποίος

- Μειώνει την ελάχιστη εμπιστοσύνη των ευαίσθητων κανόνων κάτω από την προσδιορισμένη από τον χρήστη ελάχιστη τιμή εμπιστοσύνης ΕΤΕ.
- Δεν μειώνει την ελάχιστη εμπιστοσύνη των μη-ευαίσθητων κανόνων
- Αυξάνει την μέγιστη εμπιστοσύνη των κανόνων αρνητικού ορίου[45] με σκοπό τη δημιουργία όσο το δυνατόν περισσότερων κανόνων που δεν προϋπήρχαν.

4.1.3 Μπλοκάρισμα κεντροποιημένων δεδομένων βασισμένη στη σύγχυση των κανόνων κατηγοριοποίησης

Στο [12] παρουσιάζεται ένα νέο περιβάλλον εργασίας[25] που συνδυάζει ανάλυση των κανόνων κατηγοριοποίησης και φειδωλή υποβάθμιση[47]. Ο όρος φειδωλή υποβάθμιση συνδέεται με την τυποποίηση του φαινομένου της μεταφοράς πληροφορίας από ένα ασφαλές ιδιωτικό περιβάλλον (αναφέρεται ως 'High') σε ένα κοινόχρηστο περιβάλλον (αναφέρεται ως 'Low') μέσω των καναλιών εξαγωγής συμπερασμάτων[35].

Στόχος της μελέτης είναι να διαπιστωθεί, αν η απώλεια της πληροφορίας εξαιτίας της υποβάθμισης, αντισταθμίζεται από επιπλέον αύξηση της εμπιστευτικότητας.

Στο [26] οι συγγραφείς παρουσιάζουν ένα εργαλείο λογισμικού (Rational Down Grader) που βασίζεται στην ιδέα της φειδωλής υποβάθμισης που αναφέρθηκε στην προηγούμενη παράγραφο. Ο αλγόριθμος εντοπίζει ποιοι από τους κανόνες που προκύπτουν από το δέντρο απόφασης,[20] είναι απαραίτητοι για την κατηγοριοποίηση των ιδιωτικών δεδομένων. Δεδομένα που δεν υποστηρίζουν τους κανόνες που βρίσκονται κατ' αυτόν τον τρόπο αποκλείονται από την υποβάθμιση. Από τα δεδομένα που παραμένουν ο αλγόριθμος θα πρέπει να αποφασίσει ποιες τιμές θα μετατραπούν σε ελλιπείς.

4.2 Κρυπτογραφικές τεχνικές

Έχουν υλοποιηθεί αρκετές κρυπτογραφικές τεχνικές προκειμένου να υφίστανται όλες εκείνες οι απαραίτητες προδιαγραφές που απαιτούνται για την προστασία των δεδομένων. Το πρόβλημα αυτό, που αποκαλείται Ασφαλής Πολυμερής Υπολογισμός[57] (ΑΠΥ), παρουσιάστηκε για πρώτη φορά στο [51] και αναλύθηκε περαιτέρω στο [21] και ορίζεται ως εξής: δύο η περισσότερα μέρη επιθυμούν να μεταφέρουν πληροφορίες, έτσι ώστε στο τέλος κάθε τμήμα να μην γνωρίζει τίποτα περισσότερο παρά μόνο τις δικές του πληροφορίες.

Στο [17] προτείνεται ένα περιβάλλον εργασίας που επιτρέπει τον συστηματικό μετασχηματισμό υπολογισμών μεταξύ διαφόρων μερών σε Ασφαλείς Πολυμερής Υπολογισμούς. Περιγράφονται διάφορες τεχνικές που περιλαμβάνουν κατηγοριοποίηση δεδομένων, συσταδοποίηση δεδομένων, γενίκευση δεδομένων[27], εξόρυξη κανόνων συσχέτισης[2], παρουσίαση συνόψεων[65] δεδομένων και χαρακτηρισμό[6] δεδομένων.

Στο [15] προτείνονται διάφορες τεχνικές, που μπορούν να εξασφαλίσουν την προστασία των δεδομένων κατά τη διαδικασία εξόρυξης δεδομένων όπως το Ασφαλές Άθροισμα (Secure Sum), η Ασφαλής Ένωση Συνόλων (Secure Set Union), το Ασφαλές Μέγεθος της Τομής των Συνόλων (Secure Size of Set Intersection) και το Εσωτερικό Γινόμενο Διανυσμάτων (Scalar Product).

4.2.1 Ασφαλής Εξόρυξη Κανόνων Συσχέτισης από κατακόρυφα κατανεμημένα δεδομένα

Η εξαγωγή ασφαλών κανόνων συσχέτισης από κατακόρυφα κατανεμημένα δεδομένα μπορεί να επιτευχθεί μέσω της εύρεσης της υποστήριξης των στοιχειοσυνόλων, σημαντικό στοιχείο για τον ασφαλή υπολογισμό των οποίων αποτελεί το εσωτερικό γινόμενο των διανυσμάτων που αναπαριστούν τα υπο-στοιχειοσύνολα στις διάφορες βάσεις δεδομένων. Αν το εσωτερικό γινόμενο μπορεί να υπολογιστεί, τότε μπορεί να υπολογιστεί με ασφάλεια και η υποστήριξη των στοιχειοσυνόλων. Ο αλγόριθμος που υπολογίζει το εσωτερικό γινόμενο περιγράφεται στο [47]. Η ασφάλεια στον υπολογισμό του εσωτερικού γινομένου εξασφαλίζεται από το γεγονός ότι κανένα από τα συμβαλλόμενα μέρη δεν μπορεί να λύσει περισσότερες από k εξισώσεις με περισσότερους από k αγνώστους [24]. Ένας άλλος τρόπος για τον υπολογισμό των υποστηρίξεων είναι μέσω του Ασφαλούς Μεγέθους της Τομής των Συνόλων [15].

Μια διαφορετική προσέγγιση ακολουθείται στο [19] όπου χρησιμοποιείται η έννοια της δεσμευμένης πιθανότητας $P[\beta|\alpha]$ όπου α,β είναι 2 οποιαδήποτε χαρακτηριστικά και ο κανόνας του Baye's σε συνδυασμό με τους νόμους de Morgan, προκειμένου να προσδιοριστούν τα στατιστικά για οποιαδήποτε λογική συνάρτηση των τιμών των χαρακτηριστικών.

4.2.2 Ασφαλής Εξόρυξη Κανόνων Συσχέτισης από οριζόντια κατανεμημένα δεδομένα

Στο [25] οι συγγραφείς προτείνουν 2 διαφορετικά πρωτόκολλα: ασφαλή ένωση των τοπικά μεγάλων[42] στοιχειοσυνόλων και έλεγχο της ελάχιστης τιμής υποστήριξης χωρίς την αποκάλυψη της τιμής της υποστήριξης. Το πρώτο πρωτόκολλο χρησιμοποιεί κρυπτογραφία για την κρυπτογράφηση των τιμών των τοπικών υποστηρίξεων εξασφαλίζοντας, ότι δεν αποκαλύπτεται η πηγή προέλευσης κάθε στοιχειοσυνόλου. Αποκαλύπτει ωστόσο τον αριθμό των στοιχειοσυνόλων που έχουν κοινή υποστήριξη. Το δεύτερο πρωτόκολλο προσθέτει έναν τυχαίο αριθμό σε κάθε υποστήριξη και οι τιμές που προκύπτουν, στέλνονται στη δεύτερη βάση δεδομένων, η οποία δεν γνωρίζει απολύτως τίποτα σχετικά με το μέγεθος της πρώτης βάσης ή τις υποστηρίξεις. Η δεύτερη βάση προσθέτει ξανά έναν τυχαίο αριθμό και τα

αποτελέσματα στέλνονται στην τρίτη βάση δεδομένων. Η διαδικασία αυτή επαναλαμβάνεται μέχρι την τελευταία βάση δεδομένων. Επιπλέον, το πρωτόκολλο ανακαλύπτει μόνο τα στοιχειοσύνολα που είναι καθολικά μεγάλα[29] και όχι την πραγματική υποστήριξη κάθε στοιχειοσυνόλου.

Μία άλλη μελέτη έχει γίνει στο [7] όπου εντοπίζεται η ακριβής τιμή της υποστήριξης κάθε καθολικά μεγάλου στοιχειοσυνόλου χωρίς την αποκάλυψη των τιμών των υποστηρίξεων των υποψήφιων στοιχειοσυνόλων που προέρχονται από διαφορετικά μέρη. Η προτεινόμενη τεχνική έχει το πλεονέκτημα, ότι ελαχιστοποιεί το πρόβλημα της συνωμοσίας χωρίς την αύξηση του συνολικού κόστους μετάδοσης. Επιπλέον, εξαλείφει το πρόβλημα της ανακατασκευής το οποίο αυξάνεται όταν μεταβάλλουμε τις συναλλαγές χρησιμοποιώντας διάφορες τεχνικές τυχαιοποίησης[54]. Ο αλγόριθμος αποτελείται από 3 φάσεις: την εύρεση του συνολικού αριθμού όλων των συναλλαγών, την εύρεση των καθολικών υποστηρίξεων για τα κ-στοιχειοσύνολα και την παραγωγή των καθολικών κανόνων συσχέτισης.

Στο [49] παρουσιάζεται μια αποτελεσματική προσέγγιση για την εξόρυξη μεγάλων στοιχειοσυνόλων από οριζόντια κατανεμημένες βάσεις δεδομένων. Αντί για το διαμοιρασμό πιθανώς ασφαλών ευαίσθητων δεδομένων, επιλέγουν να διαμοιράσουν ένα μικρό μόνο τμήμα κάθε τοπικού μοντέλου και τα οποία τμήματα χρησιμοποιούνται για να κατασκευάσουν το καθολικό μοντέλο των μεγάλων στοιχειοσυνόλων. Αυτή η επιλογή καθιστά τη συγκεκριμένη προσέγγιση πολύ αποτελεσματική αναφορικά με το κόστος μετάδοσης[10] καθιστώντας την ιδανική επιλογή για εξόρυξη από γεωγραφικά κατανεμημένες βάσεις δεδομένων. Ανέπτυξαν επίσης έναν μηχανισμό μετάδοσης για να εξασφαλίσουν ασφάλεια μεταξύ των μερών που συμμετέχουν στη όλη διαδικασία. Απέδειξαν τέλος ότι το καθολικό μοντέλο που παράγεται από αυτή την προσέγγιση είναι απόλυτα ακριβές μέσω μιας σειράς πειραμάτων κάτω από διάφορες συνθήκες.

4.2.3 Ασφαλής επαγωγή[34] κανόνων μέσω δέντρου απόφασης από κατακόρυφα κατανεμημένα δεδομένα

Η μελέτη που έγινε στο [18] ερευνά τη διαδικασία κατασκευής του δέντρου απόφασης ενός κατηγοριοποιητή[8] για μια βάση δεδομένων κατακόρυφα κατανεμημένη. Το πρωτόκολλο που παρουσιάζεται βασίζεται στο εσωτερικό γινόμενο διανυσμάτων χρησιμοποιώντας έναν εξυπηρετητή δικτύου μιας τρίτης ανεξάρτητης πηγής.

4.2.4 Ασφαλής επαγωγή κανόνων μέσω δέντρου απόφασης από οριζόντια κατανεμημένα δεδομένα

Η μελέτη στο [29] προτείνει μία λύση στο πρόβλημα της προστασίας γνώσης κατά την κατηγοριοποίηση χρησιμοποιώντας μια ασφαλή μέθοδο πολυμερούς υπολογισμού. Συγκεκριμένα οι συγγραφείς βασίζονται στον ID3 αλγόριθμο που χρησιμοποιείται για την επαγωγή κανόνων συσχέτισης από ένα δέντρο απόφασης. Ο ID3 επιλέγει το «καλύτερο» χαρακτηριστικό συγκρίνοντας τις εντροπίες. Όταν οι τιμές των εντροπιών διαφορετικών χαρακτηριστικών συγκλίνουν μεταξύ τους, προσδοκάται, ότι τα αποτελέσματα στο δέντρο που προκύπτει, επιλέγοντας κάποιο από αυτά τα χαρακτηριστικά, θα έχει σχεδόν την ίδια ικανότητα πρόβλεψης. Ένα ζεύγος χαρακτηριστικών έχει δ-ισοδύναμο κέρδος αν η διαφορά μεταξύ των δύο κερδών είναι μικρότερη από δ. Ορίζοντας ως *ID3* το σύνολο όλων των πιθανών δέντρων που παράγονται εφαρμόζοντας τον ID3 και επιλέγοντας κάποιο από τα χαρακτηριστικά στην περίπτωση που είναι δ-ισοδύναμα, προτείνεται από τους συγγραφείς ένα πρωτόκολλο για ασφαλούς υπολογισμούς που ονομάζεται $ID3_{\delta}$.

4.3 Τεχνικές ανακατασκευής των διαταραγμένων δεδομένων

Έχουν υλοποιηθεί αρκετές τεχνικές σχετικά με την προστασία γνώσης διαταράσσοντας τα δεδομένα και ανακατασκευάζοντας τα με τέτοιο τρόπο που να είναι ασφαλής η μεταφορά-κοινοποίηση τους ενώ την ίδια στιγμή τα αποτελέσματα των αλγόριθμων Εξόρυξης Γνώσης στα διαταραγμένα δεδομένα να είναι παρόμοια με

αυτά από την εφαρμογή στα αρχικά δεδομένα. Κάποιες από αυτές τις τεχνικές παρουσιάζονται παρακάτω.

4.3.1 Τεχνικές ανακατασκευής για αριθμητικά δεδομένα

Η μελέτη που παρουσιάζεται στο [2] παρουσιάζει το πρόβλημα της κατασκευής ενός δέντρου απόφασης από δεδομένα εκπαίδευσης[70] στα οποία οι τιμές των εγγραφών έχουν διαταραχθεί. Καθώς δεν είναι δυνατός ο ακριβής υπολογισμός των αρχικών τιμών, προτείνεται από τους συγγραφείς μια διαδικασία ανακατασκευής που θα μπορεί, αποτελεσματικά, να εκτιμήσει τις αρχικές τιμές. Είναι σε θέση να κατασκευάσουν κατηγοριοποιητές των οποίων η ακρίβεια είναι συγκρίσιμη με αυτή των κατηγοριοποιητών, που δημιουργούνται από τα πραγματικά δεδομένα. Για την ανακατασκευή των αρχικών τιμών ακολουθείται Bayesian προσέγγιση και προτείνονται τρεις αλγόριθμοι για την κατασκευή αποτελεσματικών ως προς την ακρίβεια δέντρων απόφασης.

Στο [1] προτείνεται μια βελτίωση της Bayesian διαδικασίας που αναφέρθηκε στην προηγούμενη παράγραφο χρησιμοποιώντας τον Expectation Maximization (EM) αλγόριθμο για την ανακατασκευή. Συγκεκριμένα οι συγγραφείς απέδειξαν ότι, όταν μια μεγάλη ποσότητα από δεδομένα είναι διαθέσιμη, ο EM αλγόριθμος οδηγεί σε καλύτερα αποτελέσματα ως προς την ανακατασκευή των αρχικών δεδομένων και ότι οι υπολογισμοί στο [2] δεν δίνουν καλά αποτελέσματα, όταν η επιπλέον γνώση που αποκτά το σύστημα από τα συνολικά ανακατασκευασμένα δεδομένα περιλαμβάνεται στη διατύπωση του προβλήματος.

4.3.2 Τεχνικές ανακατασκευής για δυαδικά και κατηγοριοποιημένα δεδομένα

Στο [38] παρουσιάζεται ένα σχήμα[58], βασιζόμενο σε μια πιθανοτική παραμόρφωση[21] των δεδομένων εισόδου χρησιμοποιώντας τυχαίους αριθμούς που παράγονται από μία προκαθορισμένη συνάρτηση κατανομής, που εξασφαλίζει συνεχώς έναν υψηλό βαθμό ασφάλειας και ακρίβειας όσον αφορά τα αποτελέσματα της εξόρυξης. Η αξιολόγηση του μοντέλου γίνεται κάνοντας πειράματα με πραγματικές και τεχνητές[68] βάσεις δεδομένων.

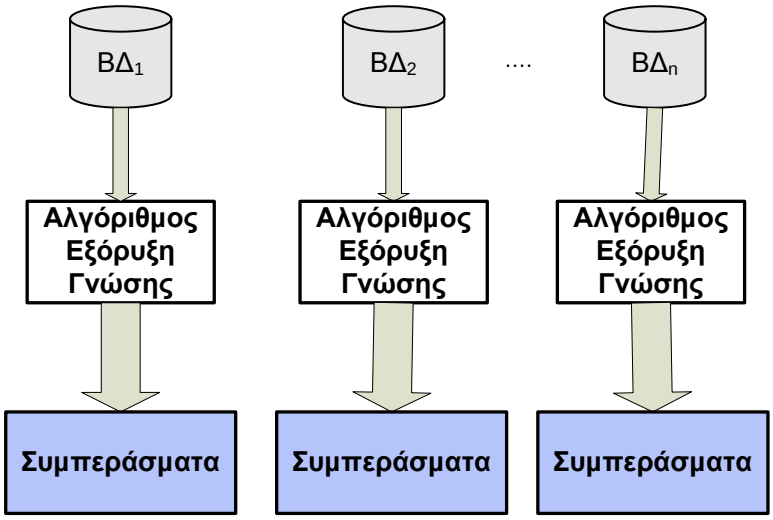
Στο [20] γίνεται αναφορά σε ένα περιβάλλον εργασίας εξαγωγής κανόνων συσχέτισης από κατηγοριοποιημένα στοιχεία όπου τα δεδομένα είναι τυχαία διατεταγμένα. Ενώ είναι εφικτή η ανάκτηση κανόνων συσχέτισης και η επαρκής προστασία τους χρησιμοποιώντας «απλές» τεχνικές τυχαιοποίησης, οι κανόνες που «γεννιούνται» ενδεχομένως να εμπεριέχουν κενά όσον αφορά την ασφάλεια. Οι συγγραφείς αναλύουν αυτά τα κενά ασφάλειας[51] και προτείνουν μια σειρά τεχνικών τυχαιοποίησης, που είναι πιο αποτελεσματικές σε σχέση με τις «απλές» τεχνικές. Παρέχουν μία μέθοδο για τον αμερόληπτο υπολογισμό της υποστήριξης που μας επιτρέπει να ανακτήσουμε τις υποστηρίξεις των στοιχειοσυνόλων από τις κατανεμημένες βάσεις δεδομένων και περιγράφουν τον τρόπο για την ενσωμάτωση αυτής της μεθόδου στους αλγόριθμους εξόρυξης.

5. Προστασία Γνώσης από Κατανεμημένους Αλγόριθμους Εξόρυξης Δεδομένων

Όπως αναφέρθηκε, η Εξόρυξη Γνώσης είναι μία διαδικασία που αναλύει ογκώδη ψηφιακά δεδομένα με σκοπό την ανακάλυψη κρυμμένων αλλά χρήσιμων προτύπων. Ωστόσο, η ανακάλυψη τέτοιου είδους προτύπων ενδεχομένως να έχει στατιστική σημασία και να αποκαλύπτει κάποια ευαίσθητη πληροφορία. Φυσικό είναι, λοιπόν, η ασφάλεια να κατέχει κυρίαρχη θέση στις προσπάθειες της ερευνητικής κοινότητας. Στην παρούσα μελέτη παρουσιάζεται ένα σύστημα, το οποίο παράγει κανόνες συσχέτισης, χωρίς να αποκαλύπτει σημαντικές πληροφορίες επιτυγχάνοντας ένα ικανοποιητικό βαθμό ακρίβειας των αποτελεσμάτων.

5.1 Περιγραφή προβλήματος

Έστω, ότι έχουμε ένα σύνολο βάσεων δεδομένων που επικοινωνούν μεταξύ τους μέσω δικτύου και ενδεχομένως να βρίσκονται στον ίδιο ή και σε διαφορετικούς γεωγραφικούς χώρους. Υποθέτουμε, ότι στις βάσεις αυτές υπάρχουν κάποια ευαίσθητα δεδομένα, μερικά εκ’ των οποίων για διάφορους λόγους, θα πρέπει να είναι γνωστά μόνο στη βάση στην οποία ανήκουν. Η διαχείριση των δεδομένων και η εξαγωγή συμπερασμάτων γίνονται μέσω εργαλείων Εξόρυξης Γνώσης όπως φαίνεται στην εικόνα 2.

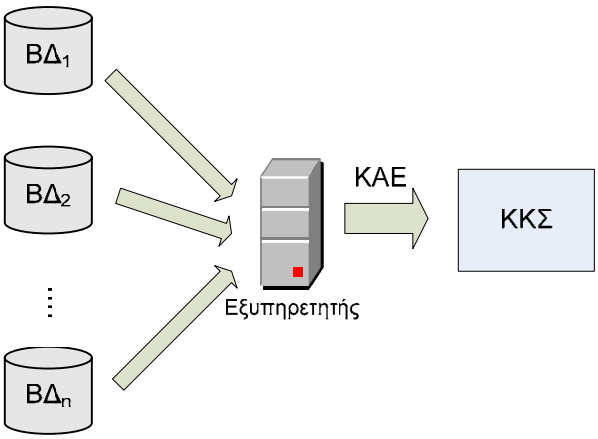


Εικόνα 2: Σύστημα βάσεων δεδομένων

Τα αποτελέσματα της παραπάνω διαδικασίας μπορούν να εκφράζονται με διάφορους τρόπους, ένας εκ των οποίων είναι οι κανόνες συσχέτισης. Όπως αναφέρθηκε και στην εισαγωγή του κεφαλαίου κάποια από τα συμπεράσματα που συνάγονται μπορεί να είναι πολύ σημαντικά. Για παράδειγμα, είναι ευνόητο ότι σε διαφορετικές πληροφορίες επιτρέπεται να έχουν πρόσβαση ο διευθυντής μιας εταιρίας και ένας απλός υπάλληλος. Ας υποθέσουμε, λοιπόν, ότι έχουμε τον κανόνα συσχέτισης με βάση τον οποίο προκύπτει ότι το 70% των πελατών της επιχείρησης αγοράζοντας το προϊόν Α θα αγοράσει και το προϊόν Β. Όπως γίνεται εύκολα αντιληπτό η πληροφορία αυτή είναι ιδιαίτερα σημαντική για την εταιρία και θα πρέπει να είναι γνωστή μόνο στα άτομα που πρέπει.

Οι κανόνες συσχέτισης που επιθυμούμε να αποκρύψουμε ονομάζονται *ευαίσθητοι κανόνες*[59] και αναπαριστούμε το σύνολο τους ως R_H όπου $R_H \subset R$ του συνόλου των κανόνων που εξάγονται από τη βάση δεδομένων.

Έστω, λοιπόν ότι διαθέτουμε n βάσεις δεδομένων $B\Delta_1$, $B\Delta_2$,...,$B\Delta_n$ όπως φαίνεται στην εικόνα 3, που έχουν δομηθεί με βάση το οριζόντιο σχήμα. Θα πρέπει να εφαρμόσουμε έναν κατανεμημένο αλγόριθμο Εξόρυξης Γνώσης επιλέγοντας τις κατάλληλες τιμές κατωφλίου για την υποστήριξη και την εμπιστοσύνη. Μετά την εφαρμογή του συγκεκριμένου αλγορίθμου προκύπτει το *σύνολο ΚΚΣ* με τους καθολικούς κανόνες συσχέτισης που ικανοποιούν τις προαναφερθείσες ελάχιστες τιμές υποστήριξης και εμπιστοσύνης.

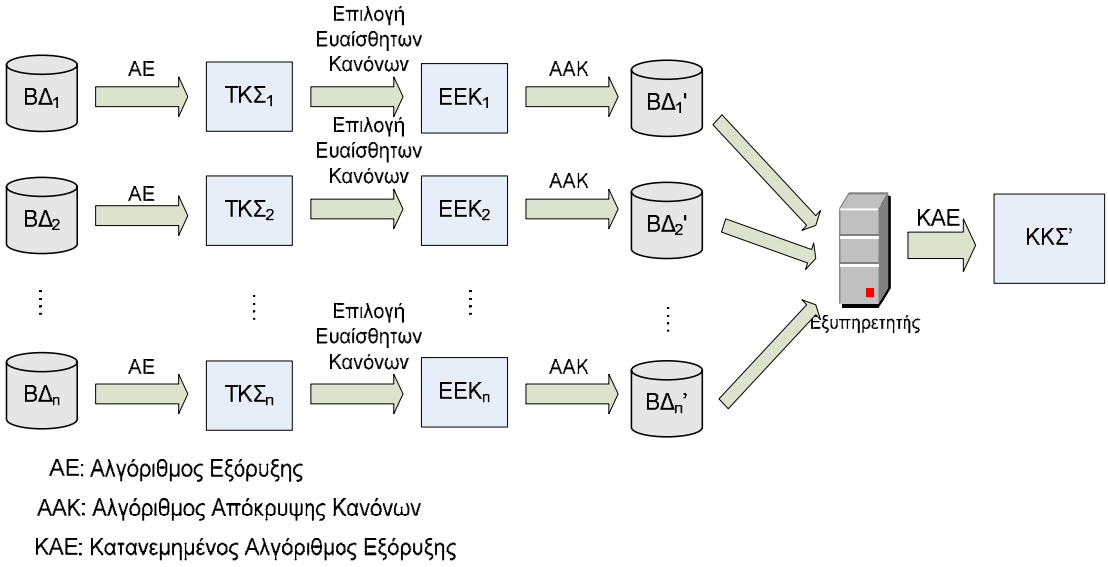


Εικόνα 3: Εφαρμογή Κατανεμημένου Αλγόριθμου Εξόρυξης Γνώσης

Σκοπός μας είναι, μετά την απόκρυψη των επιλεγόμενων κανόνων συσχέτισης στις επιμέρους βάσεις δεδομένων, το *σύνολο ΚΚΣ'*, που θα προκύψει εφαρμόζοντας και πάλι τον ίδιο κατανεμημένο αλγόριθμο, να περιέχει όλους τους κανόνες που είχε το σύνολο ΚΚΣ εκτός από εκείνους που περιλαμβάνονται στην ένωση των συνόλων $EEK = EEK_1 \cup EEK_2 \cup ... \cup EEK_n$ που περιέχουν τους προς απόκρυψη κανόνες συσχέτισης. Είναι πολύ πιθανό να μην ισχύει κάτι τέτοιο αλλά να προκύπτει ένα νέο σύνολο κανόνων που αποκαλούμε «έμμεσες απώλειες» και στο οποίο εκτός από τους κανόνες που υπήρχαν πριν από τη διαδικασία απόκρυψης συμπεριλαμβάνονται κάποιοι νέοι κανόνες («κανόνες φαντάσματα») ενώ κάποιοι άλλοι χάνονται («απολεσθέντες κανόνες»).

Συνοπτικά, η διαδικασία περιλαμβάνει τα εξής: Επιλέγουμε την ή τις βάσεις δεδομένων στην οποία/ες ενδιαφερόμαστε να αποκρύψουμε κάποιους ευαίσθητους κανόνες συσχέτισης και εφαρμόζουμε τοπικά έναν απλό αλγόριθμο Εξόρυξης Γνώσης προκειμένου να αποκτήσουμε όλους τους δυνατούς κανόνες που ικανοποιούν τις προδιαγραφές που έχουμε ορίσει. Από τα προκύπτοντα σύνολα $ΤΚΣ_i$ επιλέγουμε τους προς απόκρυψη κανόνες συσχέτισης οπότε προκύπτουν τα σύνολα $ΕΕΚ_i$. Στη συνέχεια πρέπει να γίνει επιλογή κάποιου αλγορίθμου απόκρυψης ώστε να επιτευχθεί ο στόχος μας. Τέλος στις μετασχηματισμένες βάσεις δεδομένων $ΒΔ_i$ που προκύπτουν, εφαρμόζουμε και πάλι τον κατανεμημένο αλγόριθμο Εξόρυξης, προκειμένου να πάρουμε το σύνολο *ΚΚΣ'* που όπως προαναφέρθηκε δεν θα πρέπει να περιέχει τους

κανόνες που περιέχονται στα σύνολα $ΕΕΚ_i$, δηλαδή $ΚΚΣ' = ΚΚΣ - ΕΕΚ$. Η περιγραφείσα διαδικασία αναπαρίσταται στην εικόνα 4.



Εικόνα 4: Περιγραφή του συστήματος

5.2 Βασικοί Αλγόριθμοι Εξόρυξης Γνώσης

Στο εδάφιο αυτό θα αναφέρουμε τους βασικούς αλγόριθμους Εξόρυξης Γνώσης για μία κεντρική βάση δεδομένων αλλά και για ένα κατανεμημένο σύστημα.

5.2.1 Εξόρυξη Κανόνων Συσχέτισης για μία Κεντρική Βάση Δεδομένων

Έχουν υλοποιηθεί αρκετοί αλγόριθμοι Εξόρυξης Κανόνων Συσχέτισης. Οι σημαντικότεροι από αυτούς είναι οι AIS [4], SETM [23], Apriori, Apriori-TID και Apriori-Hybrid [5], Off-line Candidate Determination (OCD) [30], Partitioning [39], Sampling [46], Dynamic Itemset Counting (DIC) [9], Continuous Association Rule Mining Algorithm (CARMA) [22], Dynamic Hasing and Pruning (DHP) [35], Sequential Efficient Association Rule (SEAR), Sequential Partitioning with TID (SPTID), SPEAR και SPINC [31], Eclat, MaxEclat, Clique και MaxClique [52].

5.2.1.1 Apriori

Στην παρούσα υλοποίηση επιλέχτηκε ο Apriori. Ο αλγόριθμος μπορεί να συνοψιστεί σε 2 βήματα:

- Εύρεση των μεγάλων στοιχειοσυνόλων
- Παραγωγή των ισχυρών κανόνων συσχέτισης από τα μεγάλα στοιχειοσύνολα

Αρχικά παράγεται το σύνολο των 1-στοιχειοσυνόλων C_1 από το οποίο παράγεται το σύνολο L_1 που περιέχει τα στοιχειοσύνολα που έχουν υποστήριξη μεγαλύτερη από την ΕΤΥ. Τα στοιχεία του L_1 συνδυάζονται για να μας δώσουν το σύνολο C_2. Η διαδικασία επαναλαμβάνεται επαναληπτικά μέχρι το L_{k-1} != Ø δηλαδή κάθε φορά από το σύνολο L_{k-1} κατασκευάζουμε το σύνολο C_k. Στο τέλος κατασκευάζουμε το σύνολο L όπου $L=\cup L_i$. Στη συνέχεια από τα στοιχειοσύνολα του συνόλου L βρίσκουμε τους κανόνες συσχέτισης ως εξής: Για κάθε μεγάλο στοιχειοσύνολο Σ προσδιορίζουμε όλα τα δυνατά υποσύνολα του. Για κάθε υποσύνολο s του Σ δημιουργούμε έναν κανόνα $s \Rightarrow l - s$ και υπολογίζουμε την εμπιστοσύνη του, $Εμπιστοσύνη(r) = \frac{Υποστήριξη(\Sigma)}{Υποστήριξη(s)}$.

Αν είναι μεγαλύτερη από την ΕΤΕ τότε ο κανόνας είναι ισχυρός. Η περιγραφή του αλγορίθμου συνοψίζεται στην εικόνα 5α.

```
1. L1 = {μεγάλα 1-στοιχειοσύνολα}
2. Για ( k = 2; Lk-1 ≠ Ø; k++) κάνε επαναληπτικά
3.       Ck = Apriori-gen (Lk-1); //νέα υποψήφια στοιχειοσύνολα
4.       Για όλες τις συναλλαγές t ∈ ΒΔ κάνε επαναληπτικά
5.              Ct = υποσύνολο (Ck, t); //υποψήφια στοιχειοσύνολα που
                                          περιλαμβάνονται στην t
6.              Για όλα τα υποψήφια στοιχειοσύνολα c ∈ Ct κάνε
7.                      c.count++;
8.       Τέλος
9.       Lk = { c ∈ Ck | c.count >= ETY}
10. Τέλος
11. L = ∪ Lk;
```

Εικόνα 5α: Περιγραφή Apriori

Η συνάρτηση Apriori-gen αποτελείται από 2 βήματα: το *βήμα συνένωσης*[40] και το *βήμα αποκοπής*[53]. Κατά τη διάρκεια του βήματος της συνένωσης παράγεται το σύνολο C_k των υποψήφιων k-στοιχειοσυνόλων συνενώνοντας το L_{k-1} με τον εαυτό

του ως εξής: Δύο στοιχειοσύνολα Σ_i και Σ_j του συνόλου L_{k-1} συνενώνονται αν τα πρώτα (k-2) στοιχεία τους είναι ίδια π.χ. έστω ότι έχουμε το σύνολο L_2 = { {A1,A2}, {A1,A3} } από το οποίο παίρνουμε το υποψήφιο 3-στοιχειοσύνολο {A1,A2,A3}. Το βήμα συνένωσης περιγράφεται στην εικόνα 5β.

1. **Για κάθε ζεύγος** στοιχειοσυνόλων p και q, $p \in L_{k-1}$, $q \in L_{k-1}$ **κάνε επαναληπτικά**
2. **Αν** (p.στοιχείο1 = q.στοιχείο1 και p.στοιχείο2 = q.στοιχείο2 και...και p.στοιχείο$_{κ-2}$=q.στοιχείο$_{κ-2}$ και p.στοιχείο$_{κ-1}$ < q.στοιχείο$_{κ-1}$) **κάνε**
3. temp = $p \cup q$;
4. **Εισήγαγε** το temp **στο** σύνολο C_k;
5. **Τέλος**
6. **Τέλος**

Εικόνα 5β: Βήμα Συνένωσης της Apriori-gen

Το βήμα αποκοπής χρησιμεύει στη μείωση των υποψήφιων στοιχειοσυνόλων. Στηρίζεται στην *Apriori ιδιότητα* σύμφωνα με την οποία, αν ένα στοιχειοσύνολο δεν είναι μεγάλο κανένα από τα υπερσύνολά του δεν θα είναι μεγάλο. Η διαδικασία αποκοπής περιγράφεται στην εικόνα 5γ.

1.**Για όλα** τα στοιχειοσύνολα $c \in C_k$ **κάνε**
2. **Για όλα** τα (k-1) υποσύνολα s του c **κάνε**
3. **Αν** ($s \notin L_{k-1}$)
4. **Σβήσε** το c από το C_k

Εικόνα 5γ: Βήμα Αποκοπής της Apriori-gen

Ας δούμε τον αλγόριθμο με ένα παράδειγμα. Έστω ότι έχουμε τη βάση δεδομένων που φαίνεται στον πίνακα 3α με ΕΤΥ = 40% και ΕΤΕ = 60%

T1	A1A2
T2	A1A2A3
T3	A1
T4	A2A3
T5	A1A3
T6	A1A2A3

Πίνακας 3α: Παράδειγμα Apriori

Στον πίνακα 3β φαίνονται τα μεγάλα στοιχειοσύνολα και στον πίνακα 3γ οι ισχυροί κανόνες συσχέτισης.

Μεγάλα Στοιχειοσύνολο	Υποστήριξη
A1	83,3%
A2	66,6%
A3	66,6%
A1A2	50%
A1A3	50%
A2A3	50%

Πίνακας 3β: Μεγάλα Στοιχειοσύνολα

Ισχυροί Κανόνες	Υποστήριξη	Εμπιστοσύνη
A2 ⇒ A1	50%	75%
A3 ⇒ A1	50%	75%
A2 ⇒ A3	50%	75%
A3 ⇒ A2	50%	75%

Πίνακας 3γ: Ισχυροί κανόνες Συσχέτισης

5.2.2 Αλγόριθμοι Εξόρυξης Κανόνων Συσχέτισης για ένα Κατανεμημένο Σύστημα Βάσεων Δεδομένων

Έχουν υλοποιηθεί αρκετοί κατανεμημένοι αλγόριθμοι Εξόρυξης κανόνων συσχέτισης. Οι σημαντικότεροι από αυτούς είναι οι εξής: Fast Distributed Algorithm (FDM) [13], Distributed Data Mining (DMA) [14], Mining of N common Association rules (MNA) [27], Count Distribution (CD) και Data Distribution (DD) [3] , Distributed Decision Miner (DDM), Preemptive Distributed Decision Miner, Distributed Dual Decision Miner και Distributed Decision Confidence Miner [42], D-Sampling, A Modified Distributed Decision Miner (MDDM) και M-Max [43], ParDCI [33], Majority Rule [50], Private Majority Rule [44], PARECLAT και PARMAXECLAT [53], ZIGZAG [34], Optimized Distributed Association Rule Mining Algorithm (ODAM) [8], Approximate Partition (AP) [45] και Distributed Higher-Order Association Rule Miner (DiHO) [28].

5.2.2.1 Fast Distributed Algorithm (FDM)

Στο σύστημα που υλοποιήθηκε, από τους προαναφερθέντες αλγορίθμους επιλέχθηκε ο FDM ο οποίος βασίζεται στον Count Distribution και προτείνει νέες τεχνικές προκειμένου να μειώσει το πλήθος των υποψήφιων στοιχειοσυνόλων γεγονός που οδηγεί στη μείωση του κόστους μετάδοσης. Επίσης, χρησιμοποιείται για οριζόντια κατανεμημένες βάσεις δεδομένων. Ο αλγόριθμος συντίθεται από 4 επαναληπτικά βήματα που είναι τα εξής:

1. **Παραγωγή των υποψήφιων συνόλων:** Παραγωγή των υποψήφιων συνόλων $ΥΣ\_ΚΤΜ_{i(k)}$ με βάση τα κτ-μεγάλα στοιχειοσύνολα ($ΚΤΜ_{i(k-1)}$) που προέκυψαν στη βάση $ΒΔ_i$ κατά την (k-1) επανάληψη.

2. **Τοπικό ψαλίδισμα:** Για κάθε στοιχειοσύνολο $X \in \text{ΥΣ\_ΚΤΜ}_{i(k)}$ ελέγχει τη βάση ΒΔ_i για να υπολογίσει τα Χ.Υποστήριξη$_{(i)}$. Αν το Χ δεν είναι τοπικά μεγάλο τότε αποκλείεται από το σύνολο $\text{ΤΜ}_{i(k)}$.

3. **Ανταλλαγή των υποστηρίξεων:** Αποστολή των στοιχειοσυνόλων που βρίσκονται στο σύνολο $\text{ΤΜ}_{i(k)}$ στις άλλες βάσεις, υπολογισμός των καθολικών υποστηρίξεων και εύρεση των κτ-μεγάλων κ-στοιχειοσυνόλων στη βάση ΒΔ_i.

4. **Αποστολή των αποτελεσμάτων:** Αποστολή των κτ-μεγάλων κ-στοιχειοσυνόλων στις υπόλοιπες βάσεις.

Ας δούμε τον αλγόριθμο με ένα παράδειγμα. Έστω ότι έχουμε 3 βάσεις δεδομένων ΒΔ_1, ΒΔ_2 και ΒΔ_3 με 30 συναλλαγές έκαστη και ΕΤΥ = 20%. Υποθέτουμε ότι κατά την πρώτη επανάληψη έχει βρεθεί το σύνολο $\text{ΚΜ}_{(1)}$={Α1,Α2,Α3,Α4,Α5} όπου τα στοιχεία Α1,Α2,Α3 είναι τοπικά μεγάλα στη βάση ΒΔ_1, τα στοιχεία Α1,Α4,Α5 στη βάση ΒΔ_2 και τα στοιχεία Α2,Α4,Α5 στη βάση ΒΔ_3 αντίστοιχα δηλαδή

- $\text{ΤΜ}_{1(1)}$={Α1,Α2,Α3}
- $\text{ΤΜ}_{2(1)}$={Α1,Α4,Α5}
- $\text{ΤΜ}_{3(1)}$={Α2,Α4,Α5}

Δεδομένου πως όλα τα παραπάνω στοιχεία είναι τοπικά και καθολικά μεγάλα θα είναι και κτ-μεγάλα, οπότε

- $\text{ΚΤΜ}_{1(1)}$={Α1,Α2,Α3}
- $\text{ΚΤΜ}_{2(1)}$={Α1,Α4,Α5}
- $\text{ΚΤΜ}_{3(1)}$={Α2,Α4,Α5}

Από την εφαρμογή της συνάρτησης Apriori-gen στα σύνολα $\text{ΚΤΜ}_{i(1)}$ προέκυψαν τα εξής σύνολα $\text{ΥΣ\_ΚΤΜ}_{i(2)}$:

- $\text{ΥΣ\_ΚΤΜ}_{1(2)}$={Α1Α2,Α1Α3,Α2Α3}
- $\text{ΥΣ\_ΚΤΜ}_{2(2)}$={Α1Α4,Α1Α5,Α4Α5}
- $\text{ΥΣ\_ΚΤΜ}_{3(2)}$={Α2Α4,Α2Α5,Α4Α5}

Στη συνέχεια περνάμε στο βήμα 2 όπου υπολογίζεται η τοπική υποστήριξη κάθε στοιχειοσυνόλου και οι οποίες απεικονίζονται στον ακόλουθο πίνακα 4

X.Υποστήριξη₁		***X.Υποστήριξη₂***		***X.Υποστήριξη₃***	
A1A2	5	**A1A4**	2	**A2A4**	13
A1A3	11	**A1A5**	4	**A2A5**	6
A2A3	10	**A4A5**	12	**A4A5**	10

Πίνακας 4: Παράδειγμα του Fast Distributed Mining Algorithm

Για να ανήκει ένα στοιχειοσύνολο στο σύνολο των τοπικά μεγάλων στοιχειοσυνόλων θα πρέπει $X.Υποστήριξη_i \geq ΕΤΥ*ΒΔ_i \geq 0.2*50 \geq 10$. Συνεπώς προκύπτουν τα εξής $TM_{i(2)}$

- $TM_{1(2)}$={A1A3,A2A3}
- $TM_{2(2)}$={A4A5 }
- $TM_{3(2)}$={A2A4,A4A5}

Ακολουθεί το βήμα 3 κατά το οποίο τα στοιχειοσύνολα των συνόλων $TM_{i(2)}$ αποστέλλονται σε όλες τις βάσεις δεδομένων και υπολογίζονται τα κτ-μεγάλα στοιχειοσύνολα. Στον πίνακα 5 παρουσιάζονται τα τοπικά μεγάλα στοιχειοσύνολα σε συνδυασμό με την τοπική υποστήριξη κάθε βάσης

Τοπικά Μεγάλα Στοιχειοσύνολα	***Αποστολή Από***	***X.Υποστήριξη₁***	***X.Υποστήριξη₂***	***X.Υποστήριξη₃***	***X.Υποστήριξη***
A1A3	ΒΔ1	11	15	10	36
A2A3	ΒΔ1	10	6	6	22
A4A5	ΒΔ2,ΒΔ3	4	8	13	25
A2A4	ΒΔ3	12	12	10	34

Πίνακας 5: Παράδειγμα του Fast Distributed Mining Algorithm (συνέχεια)

Για να ανήκει ένα στοιχειοσύνολο στο σύνολο των καθολικά μεγάλων στοιχειοσυνόλων θα πρέπει $X.Υποστήριξη \geq ΕΤΥ*ΒΔ \geq 0.2*150 \geq 30$ οπότε συνάγονται τα παρακάτω $ΚΤM_{i(2)}$

- $ΚΤM_{1(2)}$={A1A3}
- $ΚΤM_{2(2)}$={Ø}

- $KTM_{3(2)}=\{A2A4\}$

Μετά την αποστολή των κτ-μεγάλων στοιχειοσυνόλων όλες οι βάσεις επιστρέφουν τα καθολικά μεγάλα 2-στοιχειοσύνολα, δηλαδή $KT_{(2)}=\{A1A3,A2A4\}$

Η παραπάνω διαδικασία επαναλαμβάνεται επαναληπτικά έως ότου $KT_{(k)}=\{Ø\}$

6. Τεχνικές Απόκρυψης Κανόνων Συσχέτισης

Στο κεφάλαιο αυτό θα παραθέσουμε τους βασικούς αλγορίθμους για την απόκρυψη των ευαίσθητων κανόνων συσχέτισης, θα μελετήσουμε την αποτελεσματικότητα τους καθώς και μεθόδους για τη βελτίωση της. Τέλος, θα μελετήσουμε πως μπορούμε να ανακτήσουμε τους αρχικούς κανόνες συσχέτισης με τις λιγότερες δυνατές απώλειες.

6.1 Στρατηγικές Απόκρυψης Κανόνων Συσχέτισης

Με βάση τη μελέτη που έγινε στα [40] και [41] μπορούμε να αποκρύψουμε έναν κανόνα συσχέτισης με δύο τρόπους:

- Μειώνοντας την υποστήριξη του
- Μειώνοντας την εμπιστοσύνη του

μέχρι οι τιμές της υποστήριξης και της εμπιστοσύνης να γίνουν μικρότερες από τις προσδιοριζόμενες από τον χρήση τιμές ETY και ETE αντίστοιχα. Αυτό μπορεί να επιτευχθεί τοποθετώντας '?'στη θέση των πραγματικών τιμών.

Έστω ότι θέλουμε να κρύψουμε τον κανόνα $A \Rightarrow B$. Μπορούμε να μειώσουμε την υποστήριξη του αν αντικαταστήσουμε το '1' με '?' για όλα τα στοιχεία στο στοιχειοσύνολο AB. Το διάστημα εμπιστοσύνης του κανόνα $A \Rightarrow B$ είναι [Ελάχιστη_Εμπιστοσύνη $(A \Rightarrow B)$,Μέγιστη_Εμπιστοσύνη $(A \Rightarrow B)$]. Σκοπός μας

είναι η τιμή της Ελάχιστης_Εμπιστοσύνης $(A \Rightarrow B)$ να γίνει μικρότερη από την ΕΤΕ.

Όπως αναφέρθηκε $Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{Ελάχιστη\_Υποστήριξη(AB) \times 100}{Μέγιστη\_Υποστήριξη(A)}$

Η μείωση της εμπιστοσύνης μπορεί να επιτευχθεί με 2 τρόπους:

- Μειώνοντας την Ελάχιστη_Υποστήριξη(AB)
- Αυξάνοντας την Μέγιστη_Υποστήριξη(A)

Η Ελάχιστη_Υποστήριξη(AB) μπορεί να μειωθεί αντικαθιστώντας τους '1' με '?' για όλα τα στοιχεία στο στοιχειοσύνολο B ενώ η Μέγιστη_Υποστήριξη(A) μπορεί να αυξηθεί αντικαθιστώντας τα '0' με '?' για όλα τα στοιχεία στο στοιχειοσύνολο A.

6.1.1 Αλγόριθμοι Απόκρυψης Κανόνων Συσχέτισης

Στηριζόμενοι στα διαστήματα υποστήριξης και εμπιστοσύνης που αναφέρθηκαν προηγουμένως έχουν αναπτυχθεί αλγόριθμοι μείωσης της Ελάχιστης_Υποστήριξης και της Ελάχιστης_Εμπιστοσύνης κάτω από τις προσδιοριζόμενες από το χρήστη ΕΤΥ και ΕΤΕ καθώς δίνεται ένα όριο ασφάλειας[56] (ΟΑ). Δοθέντος του κανόνα $(A \Rightarrow B)$ στο τέλος της διαδικασίας θα ισχύει ένα από τα εξής:

- $Ελάχιστη\_Υποστήριξη(A \Rightarrow B) \leq ETY - OA$ ή
- $Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) \leq ETE - OA$

6.1.1.1 Αλγόριθμος Απόκρυψης Κανόνων Συσχέτισης Μειώνοντας την Υποστήριξη (Αλγόριθμος ΜΥ)

Ο αλγόριθμος κρύβει τους ευαίσθητους κανόνες μειώνοντας την ελάχιστη υποστήριξη των παραγόμενων στοιχειοσυνόλων μέχρι η ελάχιστη υποστήριξη του κανόνα να γίνει μικρότερη από την παράσταση (ΕΤΥ-ΟΑ). Τα παραγόμενα στοιχειοσύνολα[28] των κανόνων του R_h (σύνολο των ευαίσθητων κανόνων) αποθηκεύονται στο L_h (σύνολο των μεγάλων στοιχειοσυνόλων) και κρύβονται ένα προς ένα μειώνοντας την ελάχιστη υποστήριξη τους. Τα στοιχειοσύνολα στο L_h ταξινομούνται σε φθίνουσα σειρά με βάση το μέγεθος τους και την ελάχιστη υποστήριξη τους. Αν υπάρχουν περισσότερα από ένα στοιχειοσύνολα με το ίδιο μέγεθος τότε επιλέγεται για κρύψιμο εκείνο με τη μεγαλύτερη ελάχιστη υποστήριξη.

Ο αλγόριθμος δουλεύει ως εξής: ας θεωρήσουμε ως Z το επόμενο προς απόκρυψη στοιχειοσύνολο. Ο αλγόριθμος κρύβει το Z μειώνοντας την υποστήριξη του. Πρώτα ταξινομεί τα στοιχεία του Z σε φθίνουσα σειρά με βάση την ελάχιστη υποστήριξη τους και ταξινομεί τις συναλλαγές στο T_z (συναλλαγές που υποστηρίζουν το στοιχειοσύνολο Z) σε αύξουσα σειρά με βάση το μέγεθός τους. Το μέγεθος μιας συναλλαγής προσδιορίζεται από τον αριθμό των στοιχείων που περιέχει. Σε κάθε βήμα επιλέγεται το στοιχείο $j \in Z$ με τη μεγαλύτερη ελάχιστη υποστήριξη και ένα '?' τοποθετείται για το στοιχείο j στη συναλλαγή με το μικρότερο μέγεθος. Η περιγραφή του αλγορίθμου συνοψίζεται στην εικόνα 6.

ΕΙΣΟΔΟΣ: το σύνολο L των μεγάλων στοιχειοσυνόλων, το σύνολο L_h των μεγάλων στοιχειοσυνόλων που θα κρυφτούν, η βάση δεδομένων ΒΔ, ΕΤΥ και ΟΑ

ΕΞΟΔΟΣ: Η βάση δεδομένων ΒΔ τροποποιημένη με τη διαγραφή των μεγάλων στοιχειοσυνόλων στο L_h

Αρχή

1. Ταξινόμησε το L σε φθίνουσα σειρά μεγέθους και ελάχιστης υποστήριξης των μεγάλων στοιχειοσυνόλων

Για κάθε στοιχειοσύνολο Z **στο** L_h {

2. Ταξινόμησε τις συναλλαγές στο T_Z σε αύξουσα σειρά με βάση το μέγεθος των συναλλαγών
3. Αριθμός_Επαναλήψεων= $|T_z| - (ETY - OA) \times |B\Delta|$

Για κ=1 **μέχρι** Αριθμός_Επαναλήψεων **κάνε** {

4. Τοποθέτησε ? στο στοιχείο με τη μεγαλύτερη ελάχιστη υποστήριξη στο στοιχειοσύνολο Z στην επόμενη συναλλαγή στο σύνολο T_Z
5. Ενημέρωση της υποστήριξης των επηρεαζόμενων στοιχειοσυνόλων
6. Ενημέρωση της βάσης δεδομένων ΒΔ

}

}

Τέλος

Εικόνα 6: Περιγραφή Αλγορίθμου ΜΥ

Ας δούμε τον αλγόριθμο με ένα παράδειγμα. Θεωρούμε την βάση δεδομένων του πίνακα 6 με ΕΤΥ = 15%, ΕΤΕ = 40% και ΟΑ = 5%

	A1	**A2**	**A3**	**A4**	**Μέγεθος**
T1	1	1	1	0	3
T2	1	0	0	1	2
T3	0	0	1	1	2
T4	1	1	0	0	2
T5	0	0	0	1	1

Πίνακας 6: Παράδειγμα βάσης δεδομένων

και ας υποθέσουμε ότι θέλουμε να κρύψουμε τον κανόνα $A1A2 \Rightarrow A3$ ο οποίος έχει

$$Υποστήριξη(A1A2 \Rightarrow A3) = \frac{1}{5} = 20\%$$

και $Εμπιστοσύνη(A1A2 \Rightarrow A3) = \frac{Υποστήριξη(A1A2A3)}{Υποστήριξη(A1A2)} = \frac{1/5}{2/5} = 50\%$.

Τότε L_h = { {A1A2}, {A3} }

<u>1ο πέρασμα:</u> Z = { {A1}, {A2} }

Υπολογίζουμε την υποστήριξη κάθε στοιχείου

$$Υποστήριξη(A1) = \frac{3}{5} = 60\% \text{ και } Υποστήριξη(A2) = \frac{2}{5} = 40\%$$

Τότε Z = { {A1}, {A2} }

Επιλέγουμε το πρώτο στοιχείο του Z δηλαδή το {A1} και ταξινομούμε τις συναλλαγές που το περιέχουν σε αύξουσα σειρά

T_Z = { T1, T2, T4 } ➔ T_Z = { T2, T4, T1 }

Στη θέση του A1 κάθε μιας από τις παραπάνω συναλλαγές τοποθετούμε '?'

Η παραπάνω διαδικασία επαναλαμβάνεται για όλα τα στοιχειοσύνολα του L_h. Μετά τον τερματισμό του αλγορίθμου προκύπτει η βάση δεδομένων που φαίνεται στον πίνακα 7.

	A1	A2	A3	A4	Μέγεθος
T1	1	1	1	0	3
T2	1	0	0	1	2
T3	0	0	1	1	2
T4	1	1	0	0	2
T5	0	0	0	1	1

Αρχική βάση δεδομένων

	A1	A2	A3	A4	Μέγεθος
T1	?	?	?	0	3
T2	?	0	0	1	2
T3	0	0	?	1	2
T4	?	?	0	0	2
T5	0	0	0	1	1

Μετασχηματισμένη βάση δεδομένων

Πίνακας 7: Παράδειγμα εφαρμογής του αλγορίθμου MY

6.1.1.2 Αλγόριθμος Απόκρυψης Κανόνων Συσχέτισης Μειώνοντας την Ελάχιστη Υποστήριξη (Αλγόριθμος ΜΕ)

Ο αλγόριθμος κρύβει τους ευαίσθητους κανόνες μειώνοντας την υποστήριξη των παραγόμενων στοιχειοσυνόλων. Η διαφορά με τον αλγόριθμο που παρουσιάστηκε στην ενότητα 6.1.1.1 είναι ότι επιλέγονται για απόκρυψη μόνο τα στοιχεία που βρίσκονται στο δεξί μέρος των κανόνων. Πρώτα παράγεται το σύνολο T_r των συναλλαγών που υποστηρίζουν πλήρως τον κανόνα r και υπολογίζεται ο αριθμός των στοιχείων που υποστηρίζονται από κάθε συναλλαγή. Στη συνέχεια το σύνολο T_r ταξινομείται σε αύξουσα σειρά με βάση το μέγεθος των συναλλαγών. Επιλέγεται η πρώτη συναλλαγή και τοποθετείται '?' στο στοιχείο με τη μεγαλύτερη υποστήριξη που βρίσκεται στο δεξί μέρος (r_r) της συναλλαγής. Η διαδικασία επαναλαμβάνεται μέχρι η ελάχιστη εμπιστοσύνη του κανόνα r να γίνει μικρότερη από (ΕΤΕ – ΟΑ). Η περιγραφή του αλγορίθμου συνοψίζεται στην εικόνα 7.

ΕΙΣΟΔΟΣ: το σύνολο R_h των κανόνων που θα κρυφτούν, η βάση δεδομένων ΒΔ, ΕΤΥ, ΕΤΕ και ΟΑ

ΕΞΟΔΟΣ: Η βάση δεδομένων ΒΔ τροποποιημένη ώστε οι κανόνες του συνόλου R_h να μην μπορούν να εξαχθούν

Αρχή
Για κάθε κανόνα r στο R_h **κάνε** {
1. T_r = {t στην ΒΔ | Τ υποστηρίζει πλήρων τον r}
2. Για κάθε t στο T_r μέτρησε τον αριθμό των στοιχείων στην t
3. Ταξινόμησε τις συναλλαγές στο T_r σε αύξουσα σειρά με βάση τον αριθμό των υποστηριζόμενων στοιχείων

Επανέλαβε μέχρι *Ελάχιστη_Εμπιστοσύνη* (r)< $ETE - OA$ {
4. Επέλεξε την πρώτη συναλλαγή $t \in T_r$
5. Επέλεξε το στοιχείο j στο r_r με τη μεγαλύτερη ελάχιστη υποστήριξη
6. Τοποθέτησε ? στη θέση του j στην t
7. Υπολόγισε της Ελάχιστης_Υποστήριξης(r)
8. Υπολόγισε της Ελάχιστης_Εμπιστοσύνης(r)
9. Υπολόγισε της Ελάχιστης Εμπιστοσύνης των επηρεαζόμενων κανόνων
10. Απομάκρυνε την t από το T_r

}
11. Απομάκρυνε τον κανόνα r από το R_h

}
Τέλος

Εικόνα 7: Περιγραφή Αλγορίθμου ΜΕ

Ας δούμε τον αλγόριθμο με ένα παράδειγμα. Θεωρούμε και πάλι τη βάση δεδομένων του πίνακα 6 με ΕΤΥ = 15%, ΕΤΕ = 40% και ΟΑ = 5% και ας υποθέσουμε ότι θέλουμε να κρύψουμε τον κανόνα $A1A2 \Rightarrow A3$ ο οποίος έχει $Υποστήριξη(A1A2 \Rightarrow A3) = \frac{1}{5} = 20\%$

και $Εμπιστοσύνη(A1A2 \Rightarrow A3) = \frac{Υποστήριξη(A1A2A3)}{Υποστήριξη(A1A2)} = \frac{1/5}{2/5} = 50\%$.

Τότε θα έχουμε T_r = { Τ1 } που διαθέτει μία μόνο συναλλαγή ενώ και το δεξί μέρος του παραπάνω κανόνα έχει μόνο ένα στοιχείο το Α3. Συνεπώς θα τοποθετηθεί '?' μόνο στο στοιχείο Α3 στη συναλλαγή Τ1. Μετά τον τερματισμό του αλγορίθμου προκύπτει η βάση δεδομένων που φαίνεται στον πίνακα 8.

	A1	A2	A3	A4	Μέγεθος
T1	1	1	1	0	3
T2	1	0	0	1	2
T3	0	0	1	1	2
T4	1	1	0	0	2
T5	0	0	0	1	1

Αρχική βάση δεδομένων

	A1	A2	A3	A4	Μέγεθος
T1	1	1	?	0	3
T2	1	0	0	1	2
T3	0	0	1	1	2
T4	1	1	0	0	2
T5	0	0	0	1	1

Μετασχηματισμένη βάση δεδομένων

Πίνακας 8: Παράδειγμα εφαρμογής του αλγορίθμου ΜΕ

6.1.1.3 Αλγόριθμος Απόκρυψης Κανόνων Συσχέτισης Αυξάνοντας την Μέγιστη Υποστήριξη (Αλγόριθμος ΜΕ2)

Ο αλγόριθμος κρύβει τους ευαίσθητους κανόνες αυξάνοντας την μέγιστη υποστήριξη του αριστερού μέρους του κανόνα (l_r) τοποθετώντας '?' στη θέση των '0' για τα στοιχεία στο l_r . Πρώτα παράγεται το σύνολο T'_{lr} που περιέχει τις συναλλαγές που υποστηρίζουν μερικώς το αριστερό μέρος του κανόνα (l_r) αλλά δεν υποστηρίζουν το δεξί μέρος του κανόνα (r_r). Για κάθε συναλλαγή του T'_{lr} υπολογίζεται ο αριθμός των στοιχείων που περιέχεται στο l_r και γίνεται ταξινόμηση των συναλλαγών σε φθίνουσα σειρά με βάση τις προαναφερθείσες μετρήσεις. Επιλέγεται η συναλλαγή T που περιέχει τον μεγαλύτερο αριθμό στοιχείων στο l_r και τοποθετείται '?' για τα στοιχεία στο l_r που δεν υποστηρίζονται από τη συναλλαγή T. Η διαδικασία επαναλαμβάνεται μέχρι η ελάχιστη εμπιστοσύνη του κανόνα r να γίνει μικρότερη από (ΕΤΕ – ΟΑ). Η περιγραφή του αλγορίθμου συνοψίζεται στην εικόνα 8.

ΕΙΣΟΔΟΣ: το σύνολο R_h των κανόνων που θα κρυφτούν, η βάση δεδομένων ΒΔ, ΕΤΥ, ΕΤΕ και ΟΑ

ΕΞΟΔΟΣ: Η βάση δεδομένων ΒΔ τροποποιημένη ώστε οι κανόνες του συνόλου R_h να μην μπορούν να εξαχθούν

Αρχή

Για κάθε κανόνα r στο R_h **κάνε** {

1. T'_{lr} = {Τ στη ΒΔ | Τ υποστηρίζει μερικώς το αριστερό μέρος l_r του κανόνα και δεν υποστηρίζει πλήρως το δεξί μέρος r_r του κανόνα}
2. Για κάθε συναλλαγή του T'_{lr} μέτρησε τον αριθμό των στοιχείων στο αριστερό μέρος l_r του κανόνα
3. Ταξινόμησε τις συναλλαγές T'_{lr} σε φθίνουσα σειρά με βάση τις υπολογιζόμενες μετρήσεις

Επανέλαβε μέχρι *Ελάχιστη_Εμπιστοσύνη* (r) < ΕΤΕ − ΟΑ {

4. Επέλεξε την πρώτη συναλλαγή $t \in T'_{lr}$
5. Τοποθέτησε ? στην t για όλα τα στοιχεία στο αριστερό μέρος l_r του κανόνα που δεν υποστηρίζονται από την t
6. Υπολόγισε την Μέγιστη_Υποστήριξη(l_r)
7. Υπολόγισε την Ελάχιστη_Εμπιστοσύνη(r)
8. Υπολόγισε την Ελάχιστη Εμπιστοσύνη των επηρεαζόμενων κανόνων
9. Απομάκρυνε την t από το T'_{lr}

}

10. Απομάκρυνε την t από το R_h

}

Τέλος

Εικόνα 8: Περιγραφή Αλγορίθμου ΜΕ2

Ας δούμε τον αλγόριθμο με ένα παράδειγμα. Θεωρούμε και πάλι τη βάση δεδομένων του πίνακα 6 με ΕΤΥ = 15%, ΕΤΕ = 40% και ΟΑ = 5% και ας υποθέσουμε ότι θέλουμε να κρύψουμε τον κανόνα $A1A2 \Rightarrow A3$ ο οποίος έχει

$$Υποστήριξη(A1A2 \Rightarrow A3) = \frac{1}{5} = 20\%$$

και $Εμπιστοσύνη(A1A2 \Rightarrow A3) = \dfrac{Υποστήριξη(A1A2A3)}{Υποστήριξη(A1A2)} = \dfrac{1/5}{2/5} = 50\%$.

Τότε θα έχουμε T'_{lr} = { Τ2 } που διαθέτει μία μόνο συναλλαγή. Θα τοποθετήσουμε '?' στο στοιχείο Α2. Μετά τον τερματισμό του αλγορίθμου προκύπτει η βάση δεδομένων που φαίνεται στον πίνακα 9.

	A1	A2	A3	A4	Μέγεθος
T1	1	1	1	0	3
T2	1	0	0	1	2
T3	0	0	1	1	2
T4	1	1	0	0	2
T5	0	0	0	1	1

Αρχική βάση δεδομένων

	A1	A2	A3	A4	Μέγεθος
T1	1	1	1	0	3
T2	1	?	0	1	2
T3	0	0	1	1	2
T4	1	1	0	0	2
T5	0	0	0	1	1

Μετασχηματισμένη βάση δεδομένων

Πίνακας 9: Παράδειγμα εφαρμογής του Αλγορίθμου ΜΕ2

6.2 Αποτελεσματικότητα των τεχνικών Απόκρυψης

Στην ενότητα αυτή θα μελετηθεί η αποτελεσματικότητα των αλγορίθμων απόκρυψης κανόνων συσχέτισης που περιγράφτηκαν στις ενότητες 6.1.1.1-6.1.1.3. Ας ξεκινήσουμε με κάποιες υποθέσεις σχετικά με το τι γνωρίζει ο κακόβουλος χρήστης:

- Τη μετασχηματισμένη βάση δεδομένων
- Ότι οι αλγόριθμοι αντικαθιστούν είτε τα ‘0’ είτε τα ‘1’ με ‘?’
- Η αρχική βάση δεδομένων δεν περιείχε άλλες άγνωστες τιμές.

Υποθέτουμε επίσης ότι έχει επιλεγεί για απόκρυψη μόνο ένας κανόνας ($L \Rightarrow R$).

Γίνεται εύκολα αντιληπτό ότι προκύπτουν 2 οικογένειες αλγορίθμων:

- Αυτοί που αντικαθιστούν τα ‘1’ με ‘?’ (ΜΥ και ΜΕ)
- Αυτοί που αντικαθιστούν τα ‘0’ με ‘?’ (ΜΕ2)

Ο επιτιθέμενος έχει δύο επιλογές

1. Να αντικαταστήσει όλα τα ‘?’ με ‘1’ και να χρησιμοποιήσει έναν αλγόριθμο Εξόρυξης.
2. Να αντικαταστήσει όλα τα ‘?’ με ‘0’ και να χρησιμοποιήσει έναν αλγόριθμο Εξόρυξης.

Αν προσεγγίσουμε το πρόβλημα από τη σκοπιά της υποστήριξης των ευαίσθητων κανόνων:

- Στην πρώτη περίπτωση, αν ο ΜΥ είναι γνωστός ο κακόβουλος χρήστης θα αποκτήσει ένα υπερσύνολο των μεγάλων στοιχειοσυνόλων μιας και όλα τα ‘1’ που είχαν μετατραπεί σε ‘0’ από τον αλγόριθμο επαναφέρονται σε ‘1’.

Παράλληλα όλα τα ‘0’ που είχαν μετατραπεί σε ‘?’ γίνονται και αυτά ‘1’ οδηγώντας στην παραγωγή επιπλέον μεγάλων στοιχειοσυνόλων γεγονός που οδηγεί τον επιτιθέμενο στο να διαπιστώσει ποια μεγάλα στοιχειοσύνολα μπορούν να παράγουν ευαίσθητους κανόνες.

- Στην δεύτερη περίπτωση, ο επιτιθέμενος δεν είναι σε θέση να ανακτήσει τα μεγάλα στοιχειοσύνολα που παράγουν τους ευαίσθητους κανόνες αν έχει εφαρμοστεί ο MY.

Αν προσεγγίσουμε το πρόβλημα από τη σκοπιά της εμπιστοσύνης των ευαίσθητων κανόνων:

- Στην πρώτη περίπτωση, τα ‘?’ μετατρέπονται σε ‘1’ από τον επιτιθέμενο χωρίς να γνωρίζει αν αρχικά ήταν ‘0’ ή ‘1’.

  Αν ήταν ‘0’, αυτό σημαίνει ότι χρησιμοποιήθηκε ο ΜΕ2. Σε αυτή την περίπτωση, η αντικατάσταση των ‘?’ με ‘1’ οδηγεί σε αύξηση της Ελάχιστης_Υποστήριξης(L) και μείωση της
  $Μέγιστης\_Εμπιστοσύνης(L \Rightarrow R)$ καθώς
  $Μέγιστη\_Εμπιστοσύνη(L \Rightarrow R) = \frac{Μέγιστη\_Υποστήριξη(L \Rightarrow R)}{Ελάχιστη\_Υποστήριξη(L)}$. Θα πρέπει να υπενθυμίσουμε ότι ο ΜΕ3 αντικαθιστά τα στοιχεία στο αριστερό μέρος του κανόνα σε συναλλαγές που περιέχουν το L αλλά όχι $L \cup R$. Η $Ελάχιστη\_Εμπιστοσύνη(L \Rightarrow R)$ δεν μεταβάλλεται. Δεδομένου λοιπόν ότι η μέγιστη εμπιστοσύνη του ευαίσθητου κανόνα θα μειωθεί και η ελάχιστη εμπιστοσύνη θα παραμείνει σταθερή ο επιτιθέμενος δεν μπορεί να τον εξάγει.

  Αν ήταν ‘1’, αυτό σημαίνει ότι αντικαταστάθηκε από ‘?’ από οποιονδήποτε από τους τρεις αλγορίθμους. Συνεπώς αν ο επιτιθέμενος αντικαταστήσει τα ‘?’ με ‘1’ η ελάχιστη υποστήριξη και/ή μέγιστη εμπιστοσύνη του κανόνα $L \Rightarrow R$ θα αυξηθεί κάτι που δεν είναι επιθυμητό.

- Στην δεύτερη περίπτωση τα ‘?’ μετατρέπονται σε ‘0’.

Αν τα ‘?’ ήταν ‘0’, αυτό σημαίνει ότι χρησιμοποιήθηκε ο ΜΕ2 μειώνοντας την υποστήριξη. Αντικαθιστώντας τα ‘?’ με ‘0’ προκαλείται αύξηση της εμπιστοσύνης του κανόνα $L \Rightarrow R$ κάνοντας τον φανερό.

Αν τα ‘?’ ήταν ‘1’, αυτό σημαίνει ότι χρησιμοποιήθηκε είτε ο ΜΕ είτε ο ΜΥ και η αντικατάσταση των ‘?’ με ‘0’ θα οδηγήσει σε μείωση της Μέγιστης_Υποστήριξης(LR) και συνεπώς μείωση της $Μέγιστης\_Εμπιστοσύνης(L \Rightarrow R)$ επιτρέποντας στον επιτιθέμενο να εντοπίσει τον ευαίσθητο κανόνα.

Από όλα τα παραπάνω γίνεται φανερό ότι ο επιτιθέμενος μπορεί να συμπεράνει τους ευαίσθητους κανόνες. Θα πρέπει λοιπόν να συνδυάσουμε τους αλγορίθμους έτσι ώστε να κρύβονται ταυτόχρονα ‘0’ και ‘1’. Προκύπτουν λοιπόν δύο συνδυασμοί:

- ΜΕ – ΜΕ2
- ΜΥ – ΜΕ2

6.3 Υπολογισμός του Ορίου Ασφαλείας

Το Όριο Ασφαλείας (Ο.Α.) επηρεάζει το πλήθος των στοιχείων που θα κρυφτούν, δηλαδή θα τοποθετηθεί ‘?’, στη βάση δεδομένων που διαθέτουμε. Όσο μεγαλύτερο είναι το Ο.Α. τόσο μεγαλύτερο είναι το πλήθος των επικείμενων αλλαγών. Κάτι τέτοιο όμως οδηγεί σε μεγαλύτερα ποσοστά έμμεσων απωλειών.

Στις ενότητες που ακολουθούν παρατείθονται οι τρόποι με τους οποίους το σύστημα μπορεί να υπολογίζει μόνο του το ελάχιστο Ο.Α. για τους δύο συνδυασμούς αλγορίθμων που προαναφέραμε:

- ΜΕ – ΜΕ2
- ΜΥ – ΜΕ2

6.3.1 Αλγόριθμος Υπολογισμού Ορίου Ασφαλείας 1 (ΟΑ1)

Αρχικά υπολογίζουμε μία παράμετρο την οποία αποκαλούμε *όριο* σαν το γινόμενο της ΕΤΥ με το μέγεθος ολόκληρης της βάσης δεδομένων. Σκοπός μας είναι να μειώσουμε την καθολική υποστήριξη του κανόνα $L \Rightarrow R$ τόσο όσο η εμπιστοσύνη του να γίνει μικρότερη από την ΕΤΕ ώστε να μην ανιχνεύεται από τον FDM. Θα πρέπει όμως σε κάθε βήμα να ελέγχουμε αν η καθολική υποστήριξη του στοιχειοσυνόλου LR παραμένει μεγαλύτερη από το όριο καθώς σε διαφορετική περίπτωση το στοιχειοσύνολο LR θα έχει κρυφτεί. Στη συνέχεια υπολογίζουμε την τοπική υποστήριξη του στοιχειοσυνόλου LR αφαιρώντας την τιμή του μετρητή, που καταμετρά το πλήθος των επαναλήψεων του αλγορίθμου, από την αρχική τιμή της τοπικής υποστήριξης του στοιχειοσυνόλου LR. Ακολουθεί ο υπολογισμός της εμπιστοσύνης του κανόνα τοπικά και η προκύπτουσα τιμή αφαιρείται από την ΕΤΕ για να πάρουμε το Ο.Α.. Σε περίπτωση που έχουμε περισσότερους από έναν κανόνες αποθηκεύουμε τις τιμές των Ο.Α. σε ένα πίνακα και το τελικό Ο.Α. είναι η μέγιστη τιμή τους. Η περιγραφή του αλγορίθμου συνοψίζεται στην εικόνα 9.

ΕΙΣΟΔΟΣ: το σύνολο R_h των κανόνων που θα κρυφτούν, η συνολική βάση δεδομένων ΒΔ, οι επιμέρους βάσεις δεδομένων $ΒΔ_i$, ΕΤΥ, ΕΤΕ και Μέγεθος Β.Δ. (D)

ΕΞΟΔΟΣ: Το Όριο Ασφάλειας ΟΑ

Αρχή

Για κάθε κανόνα $L \Rightarrow R$ στο R_h **κάνε** {

1. Μετρητής = -1
2. Όριο = D * ΕΤΥ
3. Εμπιστοσύνη ($L \Rightarrow R$) = $\frac{Υποστήριξη(LR)}{Υποστήριξη(L)}$

Επανέλαβε όσο (Καθολική_Υποστήριξη(LR) $\geq$ Όριο) **και** (Εμπιστοσύνη($L \Rightarrow R$) $\geq$ ΕΤΕ)) {

4. Αύξηση Μετρητή κατά 1
5. Μείωση Καθολικής_Υποστήριξης(LR) κατά 1
6. Υπολογισμός Εμπιστοσύνης($L \Rightarrow R$)

}

Αν (Μετρητής = 0) {

7. Ο.Α. = 0

}

Διαφορετικά {

8. Τοπική_Υποστήριξη$(LR)_i$ = Τοπική_Υποστήριξη$(LR)_i$ – Μετρητής
9. Υπολογισμός Τοπικής_Εμπιστοσύνης($L \Rightarrow R$)$_i$

```
            10. Ο.Α. = (Ε.Τ.Ε. – Τοπική_Εμπιστοσύνη( L ⇒ R ))
        }
        11.  Αποθήκευση στον πίνακα ΟΑ
    }

    Για όλα τα στοιχεία του πίνακα ΟΑ εντόπισε το μεγαλύτερο
            12.  Ο.Α. = μεγαλύτερο στοιχείο του πίνακα ΟΑ
Τέλος
```

Εικόνα 9: Περιγραφή αλγορίθμου ΟΑ1

Ας δούμε τον παραπάνω αλγόριθμο με ένα απλό παράδειγμα. Έστω ότι έχουμε δύο βάσεις δεδομένων ΒΔ1 και ΒΔ2 με 22 συναλλαγές η κάθε μία. Άρα Όριο = 44 * 0.2 = 8.8 Θέλουμε να κρύψουμε τον κανόνα $A1A2 \Rightarrow A3$ από τη βάση ΒΔ1 με ΕΤΥ = 20% και ΕΤΕ = 60%. Έστω ότι τα στοιχειοσύνολα που απαρτίζουν τον παραπάνω κανόνα εμφανίζονται όπως στον πίνακα 10.

	ΒΔ1	**ΒΔ2**	**ΒΔ**
A1A2	7	6	13
A1A2A3	6	6	12

Πίνακας 10: Πλήθος εμφανίσεων του κανόνα $A1A2 \Rightarrow A3$

Στον πίνακα 11 που ακολουθεί φαίνονται τα βήματα εφαρμογής του αλγορίθμου για τον παραπάνω κανόνα.

#επανάληψης	**μετρητής**	**Υποστήριξη(A1A2A3)**	**Εμπιστοσύνη(** $A1A2 \Rightarrow A3$**)**
1	-1	12	12/13 = 0.923
2	0	11	11/13 = 0.846
3	1	10	10/13 = 0.769
4	2	9	9/13 = 0.692
5	3	8	8/13 = 0.615

Πίνακας 11: Πλήθος επαναλήψεων του αλγορίθμου ΟΑ1

Παρατηρούμε ότι Καθολική_Υποστήριξη(A1A2A3) = 8 < 8.8.

Άρα Τοπική_Υποστήριξη$(A1A2A3)_1$ = 6 – 3 = 3 και

Εμπιστοσύνη($A1A2 \Rightarrow A3$**)** = 3/7 = 0.428

Συνεπώς, Ο.Α. = 0.6 – 0.428 = 0.172 ή 17.2%

6.3.2 Αλγόριθμος Υπολογισμού Ορίου Ασφαλείας 2 (ΟΑ2)

Όπως και στον προηγούμενο αλγόριθμο υπολογίζουμε την παράμετρο *όριο* από τη σχέση Μέγεθος Β.Δ. * ΕΤΥ. Σκοπός του αλγορίθμου είναι και πάλι να μειώσει την καθολική υποστήριξη(LR) ώστε η εμπιστοσύνη του κανόνα $L \Rightarrow R$ να γίνει μικρότερη από την τιμή της ΕΤΕ. Στην περίπτωση που το μέγεθος του αριστερού μέρους (υπενθυμίζεται ότι μέγεθος του κανόνα αποκαλείται το πλήθος των στοιχείων που τον αποτελούν) του κανόνα είναι μικρότερο ή μεγαλύτερο από το μέγεθος του δεξιού μέρους μειώνουμε, όπως και στον ΟΑ1, την καθολική υποστήριξη του στοιχειοσυνόλου LR κατά ένα μέχρις ότου είτε αυτή γίνει μικρότερη από το όριο ή η εμπιστοσύνη του κανόνα γίνει μικρότερη από την ΕΤΕ. Σε διαφορετική περίπτωση, δηλαδή, αν το μέγεθος του αριστερού μέρους είναι ίσο με το μέγεθος του δεξιού θα πρέπει να εξετάσουμε αν η τοπική υποστήριξη του αριστερού τμήματος είναι μικρότερη από αυτή του δεξιού. Αν αυτό ισχύει τότε κάνουμε ότι και προηγουμένως. Διαφορετικά θα πρέπει να μειώνουμε κατά ένα τόσο την καθολική υποστήριξη του LR όσο και του L. Το Ο.Α. υπολογίζεται από το πηλίκο της τιμής του μετρητή μείον την τοπική υποστήριξη του LR συν το γινόμενο του μεγέθους της τοπικής βάσης δεδομένων με την ΕΤΥ προς το μέγεθος της τοπικής βάσης δεδομένων. Όπως και στον ΟΑ1 σε περίπτωση περισσότερων του ενός κανόνων αποθηκεύουμε τις τιμές των Ο.Α. που προκύπτουν σε ένα πίνακα και σαν τελικό Ο.Α. θεωρούμε τη μέγιστη τιμή του. Η περιγραφή του αλγορίθμου συνοψίζεται στην εικόνα 10.

ΕΙΣΟΔΟΣ: το σύνολο R_h των κανόνων που θα κρυφτούν, η συνολική βάση δεδομένων ΒΔ, οι επιμέρους βάσεις δεδομένων $ΒΔ_i$, ΕΤΥ, ΕΤΕ και Μέγεθος Β.Δ. (D)

ΕΞΟΔΟΣ: Το Όριο Ασφάλειας ΟΑ

Αρχή

Για κάθε κανόνα $L \Rightarrow R$ στο R_h **κάνε** {

1. Μετρητής = 0
2. Όριο = D * ΕΤΥ
3. Εμπιστοσύνη ($L \Rightarrow R$) = $\frac{Υποστήριξη(LR)}{Υποστήριξη(L)}$

Αν (Μέγεθος (L) > Μέγεθος (R)) **ή** (Μέγεθος (L) < Μέγεθος (R)) {

Επανέλαβε όσο (Καθολική_Υποστήριξη(LR) ≥ Όριο) **και** (Εμπιστοσύνη($L \Rightarrow R$) ≥ ΕΤΕ) {

4. Αύξηση Μετρητή κατά 1

5. Μείωση Καθολικής_Υποστήριξης(LR) κατά 1
6. Υπολογισμός Εμπιστοσύνης($L \Rightarrow R$)
}
}
Διαφορετικά {
Αν (Τοπική_Υποστήριξη(L)$_i$ < Τοπική_Υποστήριξη(R)$_i$ {
Επανέλαβε όσο ((Καθολική_Υποστήριξη(LR) ≥ Όριο) **και** (Εμπιστοσύνη($L \Rightarrow R$) ≥ ΕΤΕ) {
7. Αύξηση Μετρητή κατά 1
8. Μείωση Καθολικής_Υποστήριξης(LR) κατά 1
9. Υπολογισμός Εμπιστοσύνης($L \Rightarrow R$)
}
}
Διαφορετικά {
Επανέλαβε όσο ((Καθολική_Υποστήριξη(LR) ≥ Όριο) **και** (Εμπιστοσύνη($L \Rightarrow R$) ≥ ΕΤΕ) {
10. Αύξηση Μετρητή κατά 1
11. Μείωση Καθολικής_Υποστήριξης(LR) κατά 1
12. Μείωση Καθολικής_Υποστήριξης(L) κατά 1
13. Υπολογισμός Εμπιστοσύνης($L \Rightarrow R$)
}
}
}
14. Ο.Α. = $\frac{Μετρητής - Τοπική\_Υποστήριξη(LR)_i + ETY \times D_i}{0.01 \times D_i}$
Αν (Ο.Α. < 0) {
15. Ο.Α. = 0
}
16. Αποθήκευση στον πίνακα ΟΑ
}

Για όλα τα στοιχεία του πίνακα ΟΑ εντόπισε το μεγαλύτερο
17. Ο.Α. = μεγαλύτερο στοιχείο του πίνακα ΟΑ

Τέλος

Εικόνα 10: Περιγραφή αλγορίθμου ΟΑ2

Ας θεωρήσουμε ότι θέλουμε να κρύψουμε και πάλι τον κανόνα $A1A2 \Rightarrow A3$. Από τον πίνακα 10 παρατηρούμε ότι: Μέγεθος(Α1Α2) > Μέγεθος(Α3). Συνεπώς θα μειώνουμε την καθολική υποστήριξη του κανόνα έως ότου είτε η εμπιστοσύνη του να γίνει μικρότερη από την ΕΤΕ, δηλαδή 60%, είτε η τιμή της καθολικής υποστήριξης να γίνει μικρότερη από την τιμή του ορίου, δηλαδή 8.8.

Στον πίνακα 12 που ακολουθεί φαίνονται τα βήματα εφαρμογής του αλγορίθμου για τον παραπάνω κανόνα.

#επανάληψης	**μετρητής**	**Υποστήριξη(A1A2A3)**	**Εμπιστοσύνη(** $A1A2 \Rightarrow A3$**)**
1	0	12	12/13 = 0.923
2	1	11	11/13 = 0.846
3	2	10	10/13 = 0.769
4	3	9	9/13 = 0.692
5	4	8	8/13 = 0.615

Πίνακας 12: Πλήθος επαναλήψεων του αλγορίθμου ΟΑ2

Παρατηρούμε ότι Καθολική_Υποστήριξη(A1A2A3) = 8 < 8.8

Συνεπώς $\text{O.A.} = \frac{4 - 6 + 4.4}{22 * 0.01} = 10.90$

Από τα 2 παραδείγματα που περιγράφτηκαν παρατηρούμε ότι το Ο.Α. που προέκυψε από τον ΟΑ2 είναι μικρότερο από αυτό που προέκυψε στην περίπτωση του ΟΑ1.

6.4 Τεχνική Ανάκτησης Κανόνων Συσχέτισης

Τελευταίο στάδιο αποτελεί η επανασύσταση των κανόνων συσχέτισης προκειμένου να διαπιστωθεί αν όντως οι ευαίσθητοι κανόνες έχουν κρυφτεί. Η βάση δεδομένων μας μετά και την εφαρμογή ενός από τους δύο συνδυασμούς αλγορίθμων που αναφέρονται στην ενότητα 6.5 αποτελείται από ‘1’, ‘0’ και ‘?’. Όπως έχει αναφερθεί στην ενότητα 3 όταν στη βάση δεδομένων εκτός των ‘1’ και ‘0’ εισέρχεται το ‘?’ έχουμε τα διαστήματα υποστήριξης και εμπιστοσύνης. Υπενθυμίζεται ότι

$$Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{Ελάχιστη\_Υποστήριξη(AB) \times 100}{Μέγιστη\_Υποστήριξη(A)} \quad \textbf{(7)}$$ και

$$Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{Μέγιστη\_Υποστήριξη(AB) \times 100}{Ελάχιστη\_Υποστήριξη(A)} \quad \textbf{(8)}$$

Σε πρώτη φάση αντικαθιστούμε τα ‘?’ με ‘0’ και υπολογίζουμε την Ελάχιστη Υποστήριξη όλων των στοιχειοσυνόλων, στη συνέχεια αντικαθιστούμε τα ‘?’ με ‘1’ και υπολογίζουμε την Μέγιστη Υποστήριξη όλων των στοιχειοσυνόλων. Δηλαδή για κάθε στοιχειοσύνολο προκύπτει ένα διάστημα υποστήριξης [α,β] όπου η τιμή α αντιστοιχεί στην Ελάχιστη Υποστήριξη και η τιμή β στη Μέγιστη Υποστήριξη. Από τις σχέσεις (7) και (8) υπολογίζονται οι τιμές της Ελάχιστης και Μέγιστης

Εμπιστοσύνης και τέλος από τη σχέση που ακολουθεί προκύπτει η εμπιστοσύνη του εκάστοτε κανόνα συσχέτισης

$$Εμπιστοσύνη(A \Rightarrow B) = \frac{Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) + Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B)}{2}$$

η οποία συγκρίνεται με την ΕΤΕ προκειμένου να διαπιστωθεί αν ο κανόνας ικανοποιεί τις προϋποθέσεις που έθεσε ο χρήστης ή όχι. Θα πρέπει να σημειωθεί ότι στις περιπτώσεις όπου Μέγιστη_Υποστήριξη(Α) = 0 ή Ελάχιστη_Υποστήριξη(Α) = 0 θεωρούμε $Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) = 0$

και $Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B) = 0$ αντίστοιχα. Συνοπτικά η περιγραφείσα διαδικασία αναπαρίσταται από τον αλγόριθμο της εικόνας 11.

ΕΙΣΟΔΟΣ: η βάση δεδομένων ΒΔ, το σύνολο L των στοιχειοσυνόλων, ΕΤΥ, ΕΤΕ και ΟΑ

ΕΞΟΔΟΣ: αποδεκτοί κανόνες συσχέτισης

Αρχή
Για κάθε στοιχειοσύνολο **κάνε** {
Για όλους τους δυνατούς συνδυασμούς **κάνε** {
Αν Μέγιστη_Υποστήριξη(Α) = 0 {
1. $Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) = 0$
}
Διαφορετικά {
2. $Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{Ελάχιστη\_Υποστήριξη(AB) \times 100}{Μέγιστη\_Υποστήριξη(A)}$
3. **Αν** Ελάχιστη_Υποστήριξη(Α) = 0 {
4. $Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B) = 0$
}
Διαφορετικά {
5. $Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B) = \frac{Μέγιστη\_Υποστήριξη(AB) \times 100}{Ελάχιστη\_Υποστήριξη(A)}$
}
6. $Εμπιστοσύνη(A \Rightarrow B) = \frac{Ελάχιστη\_Εμπιστοσύνη(A \Rightarrow B) + Μέγιστη\_Εμπιστοσύνη(A \Rightarrow B)}{2}$
}
Αν $Εμπιστοσύνη(A \Rightarrow B) \geq ΕΤΕ$ {
7. αποδεκτός κανόνας
}
}
}
Τέλος

Εικόνα 11: Περιγραφή αλγορίθμου ανάκτησης κανόνων συσχέτισης (ΑΚ)

Ας δούμε όλα όσα φαίνονται στη εικόνα 11 με ένα μικρό παράδειγμα. Έστω ότι έχουμε τα στοιχειοσύνολα A1, A2 και A1A2 με τα παρακάτω διαστήματα υποστήριξης:

A1 [30, 50]

A2 [20, 40]

A1A2 [5, 20]

και θέλουμε να ελέγξουμε αν υφίσταται ο κανόνας $A1 \Rightarrow A2$ για ΕΤΕ = 20%.

Θα έχουμε λοιπόν

$$Ελάχιστη\_Εμπιστοσύνη(A1 \Rightarrow A2) = \frac{Ελάχιστη\_Υποστήριξη(AB) \times 100}{Μέγιστη\_Υποστήριξη(A)} = \frac{5}{50} = 0.1$$ ή

10%

$$Μέγιστη\_Εμπιστοσύνη(A1 \Rightarrow A2) = \frac{Μέγιστη\_Υποστήριξη(AB) \times 100}{Ελάχιστη\_Υποστήριξη(A)} = \frac{20}{30} = 0.666$$ ή

66.6%

Συνεπώς $Εμπιστοσύνη(A1 \Rightarrow A2) = \frac{10 + 66.6}{2} = \frac{76.6}{2} = 38.3\%$

Παρατηρούμε ότι $Εμπιστοσύνη(A1 \Rightarrow A2) \geq$ ΕΤΕ οπότε ο κανόνας $A1 \Rightarrow A2$ είναι αποδεκτός.

7. Αξιολόγηση – Πειράματα

Έγιναν αρκετά πειράματα με διαφορετικές παραμέτρους ώστε να αξιολογηθεί, όσο το δυνατόν καλύτερα, η αποτελεσματικότητα και η ακρίβεια του συστήματος που περιγράφτηκε στις προηγούμενες ενότητες. Τρέξαμε τους προτεινόμενους αλγορίθμους σε Windows XP ενός centrino mobile Intel Pentium στα 1.7 GHz με μνήμη RAM στα 512 MB. Οι βάσεις δεδομένων οι οποίες θα αποτελέσουν τη βάση για την αξιολόγηση του συστήματος μας παράχθηκαν μέσω του εργαλείου IBM Synthetic Data Generator [58]. Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε έτοιμα σύνολα δεδομένων από το UCI Machine Learning Repository [59] και το Frequent Itemset Mining Implementations Repository (FIMI) [60]. Όλοι οι αλγόριθμοι υλοποιήθηκαν χρησιμοποιώντας τη γλώσσα προγραμματισμού Perl.

Υποθέσαμε ότι το σύστημα μας αποτελείται από 2 βάσεις δεδομένων που έχουν ακριβώς το ίδιο πλήθος εγγραφών και πως οι κανόνες που επιλέγουμε για απόκρυψη είναι καθολικά (κανόνες που προκύπτουν από την εφαρμογή του κατανεμημένου αλγόριθμου Εξόρυξης Γνώσης) και όχι τοπικά ισχυροί. Αποφασίσαμε να επαναλάβουμε τα πειράματα επιλέγοντας για απόκρυψη διαφορετικό κάθε φορά πλήθος (4, 6 και 8) κανόνων συσχέτισης για 500, 1000, 2000, 4000 και 8000 συναλλαγές. Επίσης, τα πειράματα επαναλήφθηκαν για διαφορετικό αριθμό στοιχείων, την πρώτη φορά οι βάσεις δεδομένων αποτελούνταν από 10 στοιχεία με μέσο μήκος συναλλαγής 5 στοιχεία και τη δεύτερη από 15 στοιχεία με μέσω μήκος συναλλαγής 8 στοιχεία.

Υπάρχει η δυνατότητα πραγματοποίησης αρκετών πειραμάτων. Θα μπορούσαμε, για παράδειγμα, να εφαρμόζαμε τους αλγορίθμους απόκρυψης σε ολόκληρη τη βάση δεδομένων και στη συνέχεια να επιλέγαμε για απόκρυψη κάποιους κανόνες από την $ΒΔ_1$ και κάποιους από την $ΒΔ_2$ ώστε να συγκρίναμε τα αποτελέσματα. Μια άλλη σειρά πειραμάτων θα μπορούσε, να αποτελέσει η απόκρυψη κανόνων από μία εκ' των δύο βάσεων δεδομένων από την οποία θα επιλέγουμε κάποιους κανόνες για απόκρυψη και στη συνέχεια εφαρμόζοντας τον κατανεμημένο αλγόριθμο Εξόρυξης Γνώσης να διαπιστώνουμε ποιοι κανόνες παρέμειναν, ποιοι χάθηκαν και ποιοι δημιουργήθηκαν. Στα πειράματα, που ακολουθούν στις επόμενες ενότητες, ακολουθείται η δεύτερη από τις παραπάνω αναφορές.

Αρχικά τρέχουμε τον αλγόριθμο FDM με ETY = 1% και ETE = 50% ώστε να αποκτήσουμε τους καθολικούς κανόνες συσχέτισης. Στη συνέχεια το σύστημα μας ζητάει να επιλέξουμε τη βάση δεδομένων από την οποία θέλουμε να αποκρύψουμε κάποιον ή κάποιους κανόνες που δεν επιθυμούμε να γνωρίζει η άλλη βάση δεδομένων και τρέχουμε τον Apriori με τις ίδιες τιμές ETY και ETE. Αφού επιλέξουμε τον αλγόριθμο απόκρυψης δηλώνουμε τους κωδικούς των ευαίσθητων κανόνων και το σύστημα προχωράει στον υπολογισμό της ελάχιστης τιμής του Ορίου Ασφαλείας και ξεκινάει τη διαδικασία απόκρυψης.

Σε κάθε πείραμα ενδιαφερόμαστε για την καταγραφή του ποσοστού των έμμεσων απωλειών και του χρόνου επεξεργασίας της CPU που θα πρέπει να σημειώσουμε ότι αρχίζουμε να τον μετράμε από τη στιγμή που επιλέξουμε τον αλγόριθμο απόκρυψης.

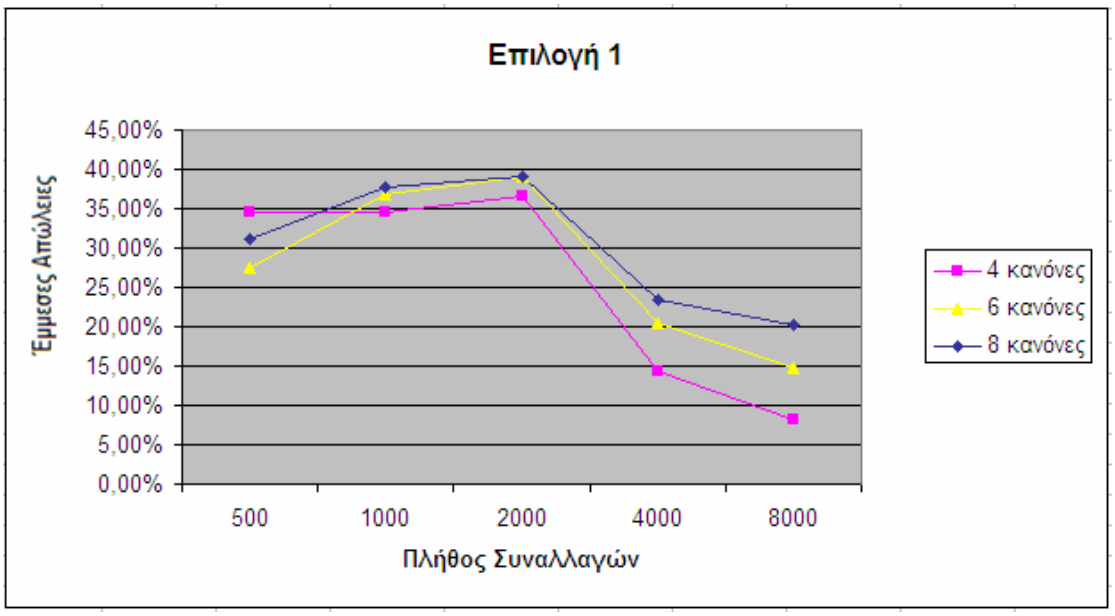
Για χάριν συντομίας στο εξής θα αναφερόμαστε στους 2 συνδυασμούς αλγορίθμων MY – ME2 και ME – ME2 με τους όρους Επιλογή 1 και Επιλογή 2 αντίστοιχα.

7.1 Πρώτη Σειρά Πειραμάτων με Βάσεις Δεδομένων με δέκα στοιχεία

Στα πειράματα που ακολουθούν χρησιμοποιήσαμε βάσεις δεδομένων με 10 στοιχεία και μέσο μήκος συναλλαγής 5 στοιχεία. Τα πειράματα επαναλήφθηκαν επιλέγοντας 4, 6 και 8 κανόνες για απόκρυψη χρησιμοποιώντας την πρώτη φορά το συνδυασμό MY – ME2 (Επιλογή 1) και τη δεύτερη φορά το συνδυασμό ME – ME2 (Επιλογή 2)

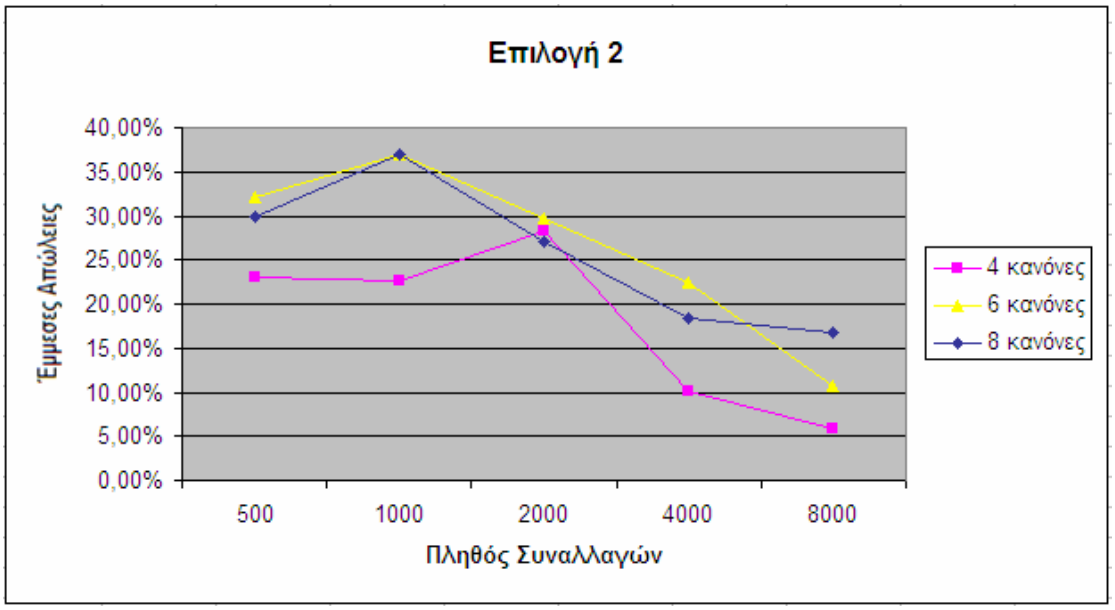
καταγράφοντας τα ποσοστά των έμμεσων απωλειών και τους χρόνους επεξεργασίας της Κεντρικής Μονάδας Επεξεργασίας.

Στην εικόνα 12 που ακολουθεί παρουσιάζονται τα ποσοστά των έμμεσων απωλειών με βάση την Επιλογή 1. Ενώ αρχικά παρατηρείται μία μικρή αύξηση των έμμεσων απωλειών στη συνέχεια καθώς αυξάνεται το πλήθος των συναλλαγών της βάσης δεδομένων το ποσοστό αυτό μειώνεται σχεδόν στο μισό. Παρατηρούμε επίσης ότι καθώς αυξάνεται το πλήθος των προς απόκρυψη κανόνων αυξάνεται και το ποσοστό των απωλειών.



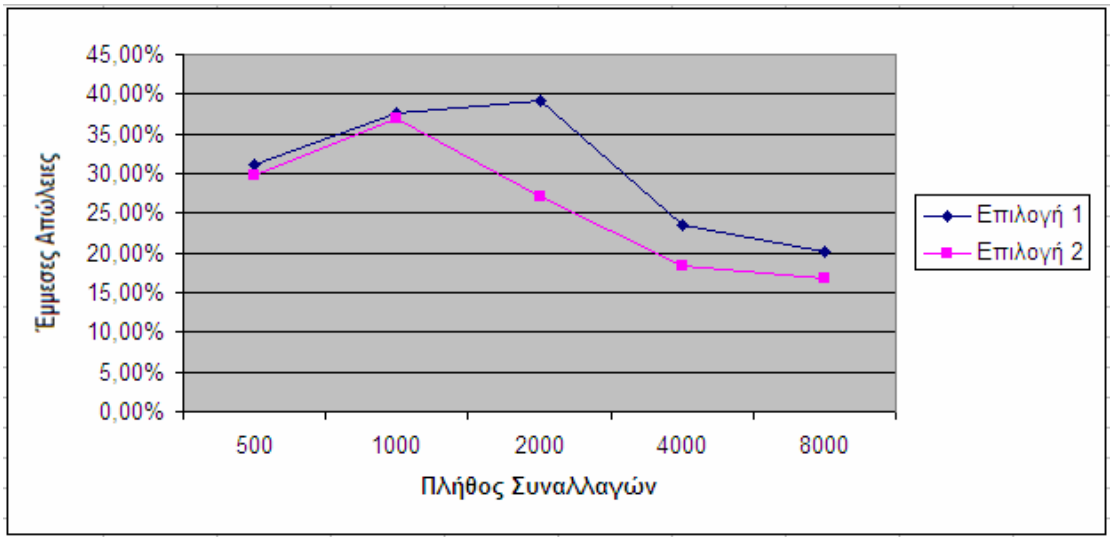
Εικόνα 12: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 1 για 10 στοιχεία

Στην εικόνα 13 παρουσιάζονται και πάλι οι έμμεσες απώλειες χρησιμοποιώντας την Επιλογή 2. Όπως και στο προηγούμενο πείραμα παρατηρείται και εδώ η τάση της μείωσης των απωλειών καθώς αυξάνεται το πλήθος των συναλλαγών όπως επίσης αύξηση αυτού του ποσοστού αυξάνοντας το πλήθος των κανόνων που επιθυμούμε να κρύψουμε.



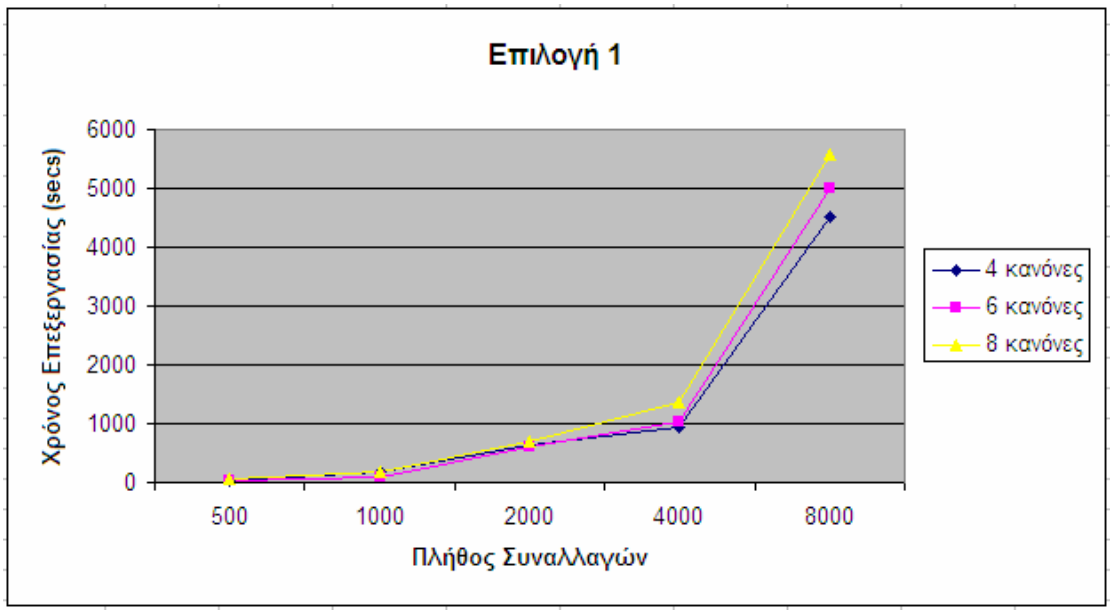
Εικόνα 13: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 2 για 10 στοιχεία

Στην εικόνα 14 γίνεται σύγκριση των ποσοστών των έμμεσων απωλειών για την Επιλογή 1 και Επιλογή 2 αποκρύπτοντας 8 κανόνες συσχέτισης. Παρατηρούμε ότι η Επιλογή 2 οδηγεί σε λιγότερες απώλειες κανόνων σε σχέση με την Επιλογή 1.



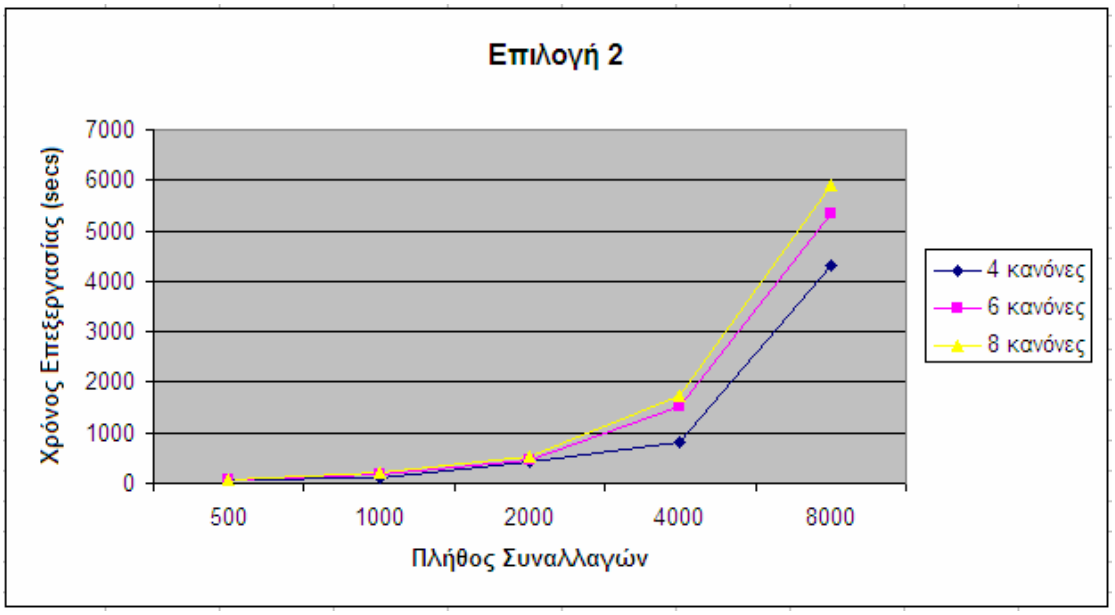
Εικόνα 14: Σύγκριση ποσοστών απωλειών Επιλογής 1 και Επιλογής 2 για απόκρυψη 8 κανόνων συσχέτισης

Στην εικόνα 15 παρατίθενται οι χρόνοι επεξεργασίας της Κ.Μ.Ε. για την Επιλογή 1 επιλέγοντας και για απόκρυψη 4, 6 και 8 κανόνων. Παρατηρείται μια εκθετική αύξηση του χρόνου επεξεργασίας σε σχέση με το πλήθος των συναλλαγών της βάσης δεδομένων. Επίσης, προκύπτει ότι η αύξηση του αριθμού των προς απόκρυψη κανόνων έχει ως αποτέλεσμα μια μικρή αύξηση στον χρόνο επεξεργασίας.



Εικόνα 15: Χρόνος Επεξεργασίας για την Επιλογή 1 για 10 στοιχεία

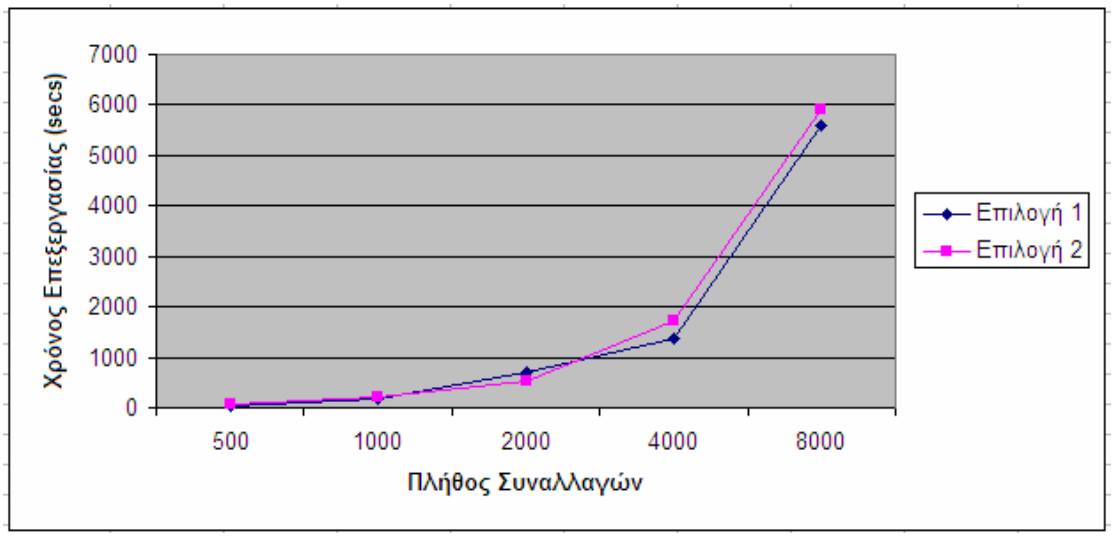
Στην εικόνα 16 παρατίθενται οι χρόνοι επεξεργασίας της Κ.Μ.Ε. για την Επιλογή 2. Παρατηρείται μια αντίστοιχη τάση με αυτή της εικόνας 14, δηλαδή η αύξηση του πλήθους των συναλλαγών οδηγεί σε αύξηση του χρόνου επεξεργασίας όπως επίσης και η αύξηση του αριθμού των προς απόκρυψη κανόνων.



Εικόνα 16: Χρόνος Επεξεργασίας για την Επιλογή 2 για 10 στοιχεία

Στην εικόνα 17 γίνεται σύγκριση των χρόνων επεξεργασίας για την Επιλογή 1 και την Επιλογή 2 επιλέγοντας για απόκρυψη 8 κανόνες συσχέτισης. Όπως παρατηρείται οι

χρόνοι είναι παραπλήσιοι. Ωστόσο στην περίπτωση της Επιλογής 2 παρατηρείται μία ελαφρά υπεροχή.

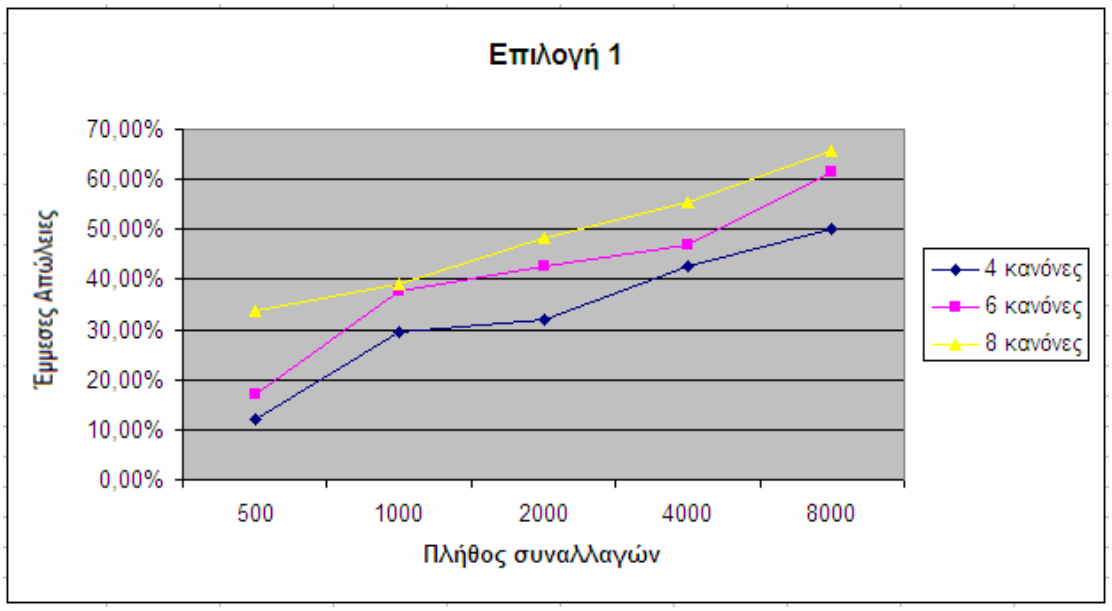


Εικόνα 17: Σύγκριση Χρόνων Επεξεργασίας Επιλογής 1 και Επιλογής 2 για απόκρυψη 8 κανόνων συσχέτισης

7.2 Δεύτερη Σειρά Πειραμάτων με Βάσεις Δεδομένων με δεκαπέντε στοιχεία

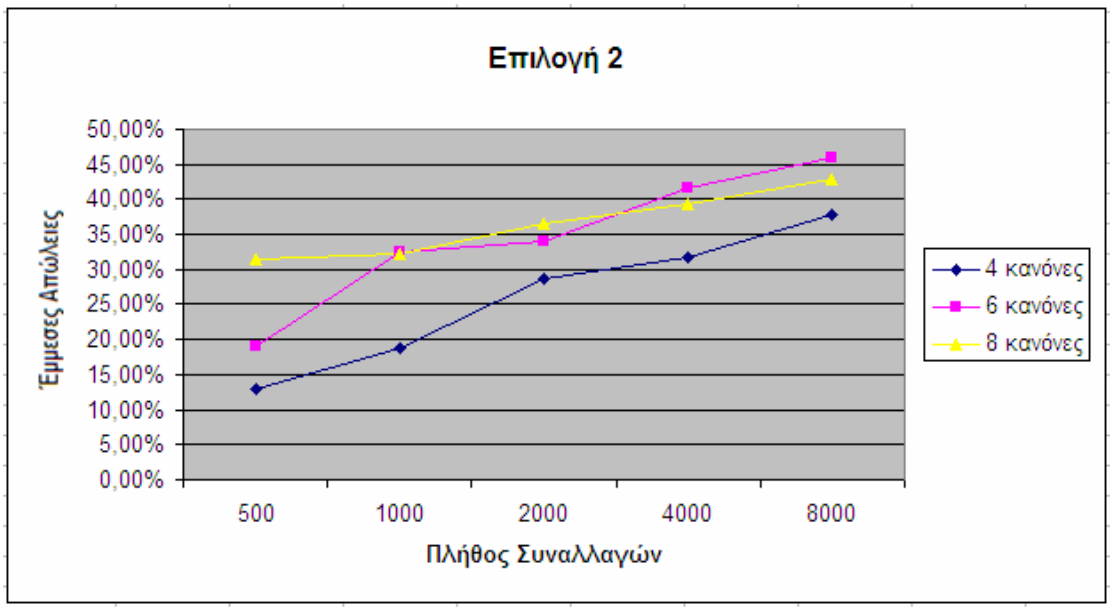
Στα πειράματα που ακολουθούν χρησιμοποιήσαμε βάσεις δεδομένων με 15 στοιχεία και μέσο μήκος συναλλαγής 8 στοιχεία. Τα πειράματα επαναλήφθηκαν όπως και στην προηγούμενη ενότητα επιλέγοντας 4, 6 και 8 κανόνες για απόκρυψη χρησιμοποιώντας την πρώτη φορά την Επιλογή 1 και τη δεύτερη φορά την Επιλογή 2 καταγράφοντας τα ποσοστά των έμμεσων απωλειών και τους χρόνους επεξεργασίας της Κεντρικής Μονάδας Επεξεργασίας.

Στην εικόνα 18 που ακολουθεί παρουσιάζονται τα ποσοστά των έμμεσων απωλειών με βάση την Επιλογή 1. Παρατηρείται, λοιπόν, μία σταδιακή αύξηση του ποσοστού των έμμεσων απωλειών σε σχέση με το πλήθος των συναλλαγών της βάσης δεδομένων καθώς και ότι το ποσοστό αυτό αυξάνεται σε σχέση με το πλήθος των επιλεγόμενων κανόνων.



Εικόνα 18: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 1 για 15 στοιχεία

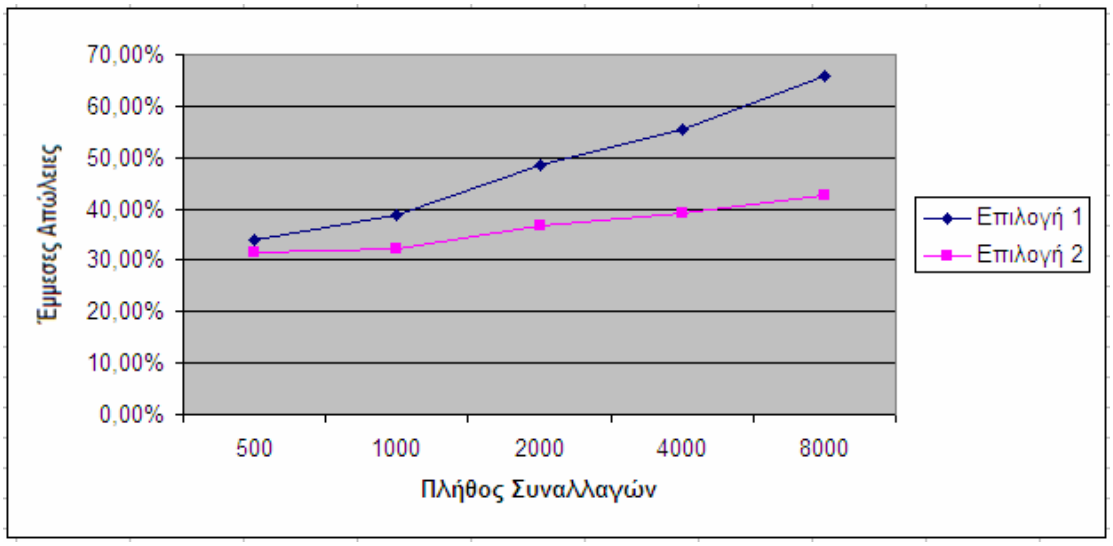
Στην εικόνα 19 έγιναν παρόμοιες μετρήσεις με βάση την Επιλογή 2. Όπως και στην περίπτωση της Επιλογής 1 παρατηρείται μια σταδιακή αύξηση των έμμεσων απωλειών καθώς αυξάνεται το μέγεθος της βάσης δεδομένων ενώ και η αύξηση του αριθμού των κανόνων επιδρά αρνητικά.



Εικόνα 19: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 2 για 15 στοιχεία

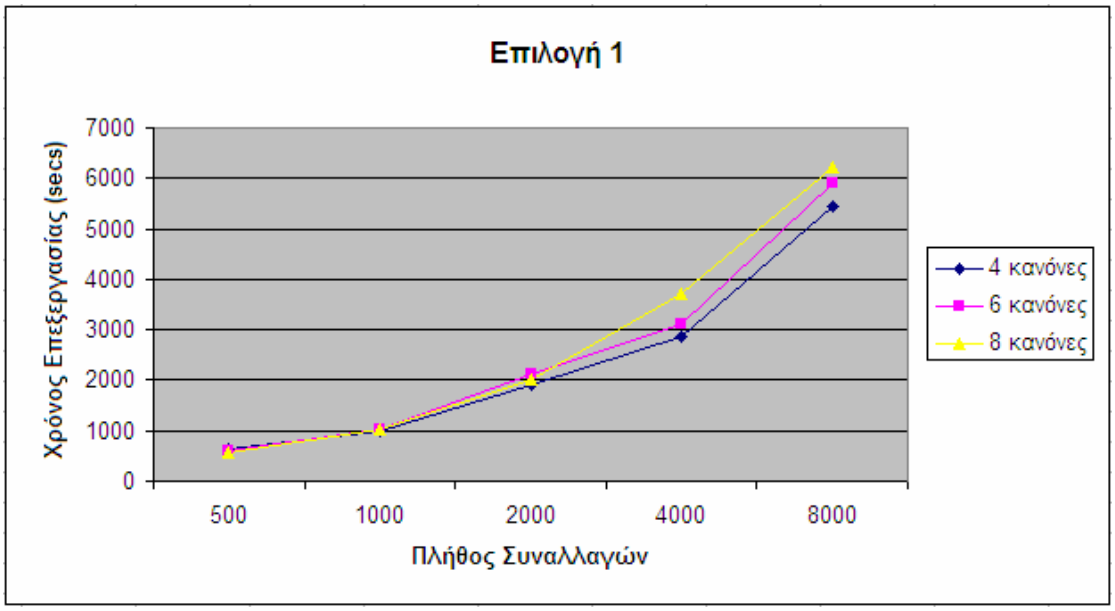
Στην εικόνα 20 γίνεται σύγκριση των ποσοστών των έμμεσων απωλειών για τις Επιλογές 1 και 2 αποκρύπτοντας 8 κανόνες συσχέτισης. Γίνεται φανερό, ότι η

Επιλογή 1 οδηγεί σε περισσότερες απώλειες σε σχέση με την Επιλογή 2 οι οποίες συγκρινόμενες μεταξύ τους αυξάνονται διαρκώς.

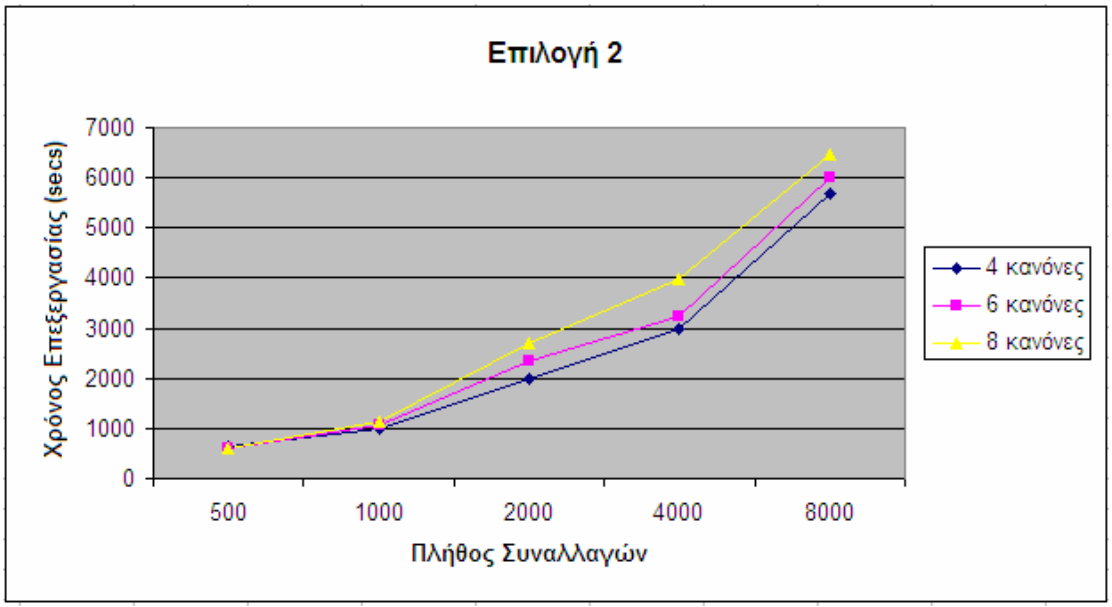


Εικόνα 20: Σύγκριση ποσοστών απωλειών Επιλογής 1 και Επιλογής 2 για απόκρυψη 8 κανόνων συσχέτισης

Στις εικόνες 21 και 22 παρουσιάζονται οι χρόνοι επεξεργασίας για τις Επιλογές 1 και 2 αντίστοιχα Όπως και στις προηγούμενες περιπτώσεις η αύξηση που παρατηρείται είναι εκθετική και είναι μεγαλύτερη στην περίπτωση της επιλογής 8 κανόνων συσχέτισης.

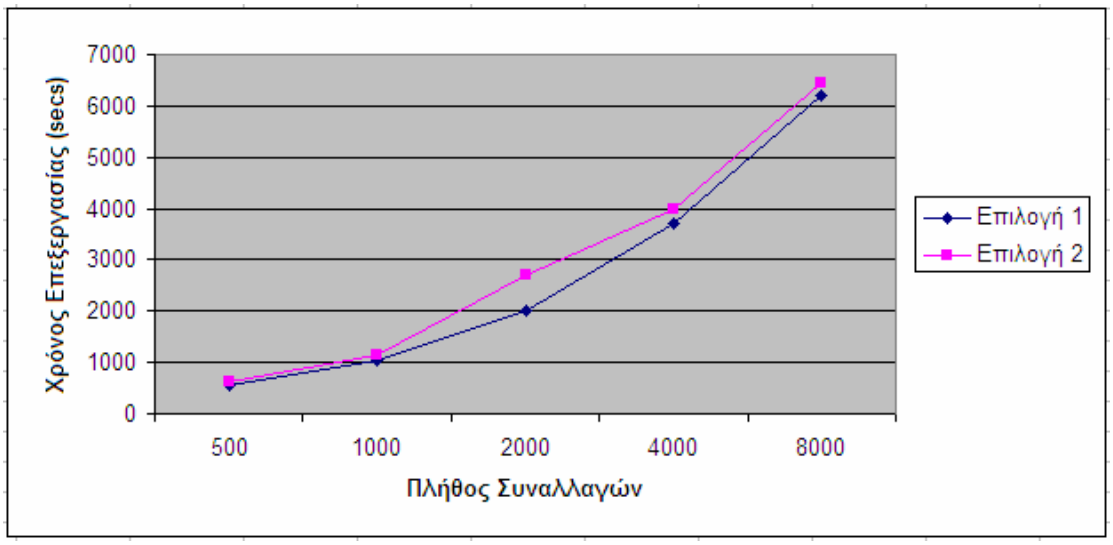


Εικόνα 21: Χρόνος Επεξεργασίας για την Επιλογή 1 για 15 στοιχεία



Εικόνα 22: Χρόνος Επεξεργασίας για την Επιλογή 2 για 15 στοιχεία

Στην εικόνα 23 γίνεται σύγκριση των χρόνων επεξεργασίας των Επιλογών 1 και 2 στην περίπτωση της επιλογής για απόκρυψη 8 κανόνων συσχέτισης από την οποία προκύπτει ότι η Επιλογή 2 είναι πιο αργή σε σχέση με την Επιλογή 1.

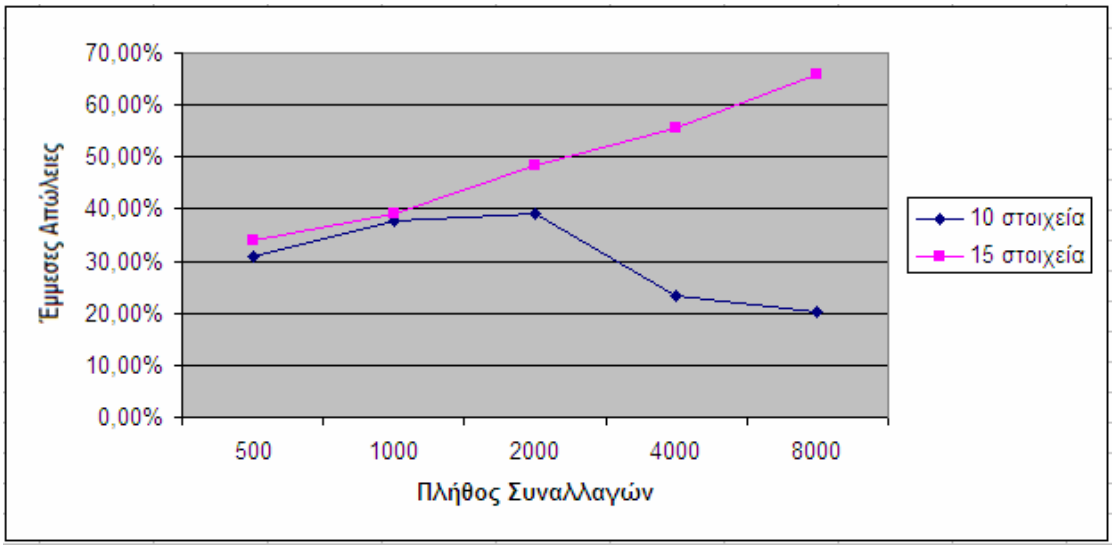


Εικόνα 23: Σύγκριση Χρόνων Επεξεργασίας Επιλογής 1 και Επιλογής 2 για απόκρυψη 8 κανόνων συσχέτισης

7.3 Σύγκριση Αποτελεσμάτων

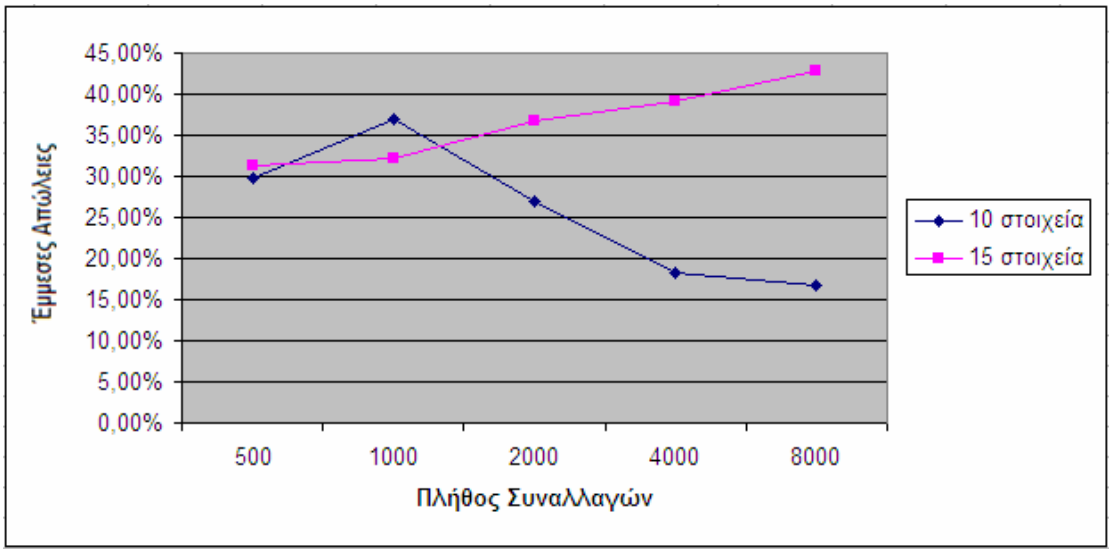
Στην εικόνα 24 γίνεται σύγκριση των ποσοστών των έμμεσων απωλειών για 2 βάσεις δεδομένων με 10 και 15 στοιχεία στην περίπτωση της Επιλογής 1 και για απόκρυψη 8 κανόνων συσχέτισης από την οποία προκύπτει ότι τα ποσοστά είναι μεγαλύτερα για

τη βάση δεδομένων που αποτελείται από 15 στοιχεία. Παρατηρείται, επίσης, ότι ενώ στη βάση δεδομένων με 10 στοιχεία το ποσοστό από τις 2000 συναλλαγές και έπειτα αρχίζει να φθίνει ενώ αντίθετα στη βάση δεδομένων με 15 στοιχεία αυξάνεται συνεχώς.



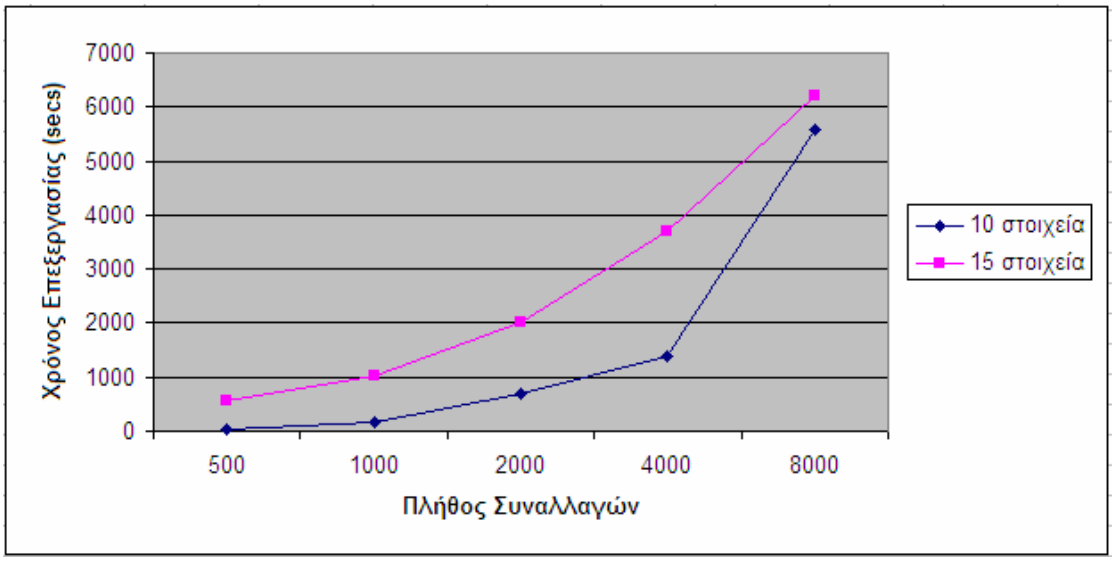
Εικόνα 24: Σύγκριση Έμμεσων Απωλειών για την Επιλογή 1 για 10 και 15 στοιχεία για απόκρυψη 8 κανόνων συσχέτισης

Στην εικόνα 25 γίνεται σύγκριση των ποσοστών των έμμεσων απωλειών για 2 βάσεις δεδομένων με 10 και 15 στοιχεία στην περίπτωση της Επιλογής 2 και για απόκρυψη 8 κανόνων συσχέτισης απ' όπου προκύπτει ότι είναι συνεχώς υψηλότερα για τη βάση δεδομένων με 15 στοιχεία με εξαίρεση ένα μικρό διάστημα (περίπου από 700 μέχρι 1300 συναλλαγές) όπου είναι υψηλότερα για τη βάση δεδομένων με 10 στοιχεία. Όπως και στην περίπτωση της Επιλογής 1, τα ποσοστά αυξάνονται συνεχώς στην περίπτωση της βάσης δεδομένων με 15 στοιχεία ενώ στην περίπτωση της βάσης δεδομένων με 10 στοιχεία μετά τις 1000 συναλλαγές μειώνονται διαρκώς.



Εικόνα 25: Σύγκριση Έμμεσων Απωλειών για την Επιλογή 2 για 10 και 15 στοιχεία για απόκρυψη 8 κανόνων συσχέτισης

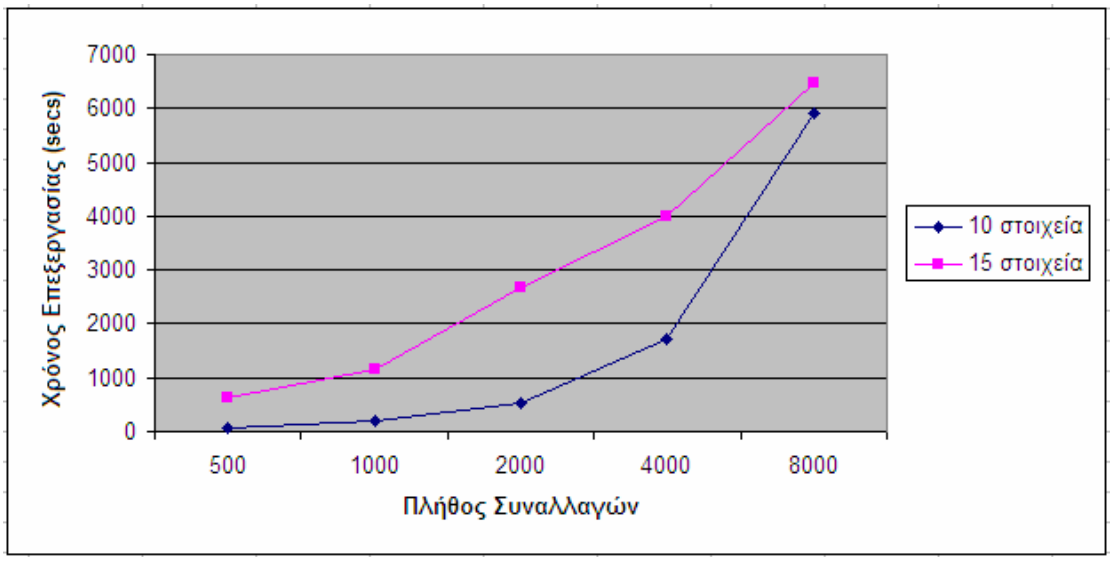
Στην εικόνα 26 γίνεται σύγκριση των χρόνων επεξεργασίας για 2 βάσεις δεδομένων με 10 και 15 στοιχεία στην περίπτωση της Επιλογής 1 και για απόκρυψη 8 κανόνων συσχέτισης. Παρατηρούμε ότι για τη βάση δεδομένων με 15 στοιχεία έχουμε συνεχώς υψηλότερους χρόνους σε σχέση με τη βάση δεδομένων με 10 στοιχεία.



Εικόνα 26: Σύγκριση Χρόνων Επεξεργασίας για την Επιλογή 1 για 10 και 15 στοιχεία για απόκρυψη 8 κανόνων συσχέτισης

Στην εικόνα 27 γίνεται σύγκριση των χρόνων επεξεργασίας για 2 βάσεις δεδομένων με 10 και 15 στοιχεία στην περίπτωση της Επιλογής 2 και για απόκρυψη 8 κανόνων συσχέτισης. Παρατηρούμε ότι τα αποτελέσματα είναι αντίστοιχα με αυτά στην

περίπτωση της Επιλογής 1, δηλαδή, μεγαλύτεροι χρόνοι επεξεργασίας για τη βάση δεδομένων με 15 στοιχεία.



Εικόνα 27: Σύγκριση Χρόνων Επεξεργασίας για την Επιλογή 2 για 10 και 15 στοιχεία για απόκρυψη 8 κανόνων συσχέτισης

8. Συμπεράσματα

Σκοπός των πειραμάτων που έγιναν στο κεφάλαιο 7 ήταν να οδηγηθούμε σε κάποια συμπεράσματα σχετικά με την απόδοση των αλγορίθμων που απαρτίζουν το προτεινόμενο σύστημα ως προς τους χρόνους επεξεργασίας και τα ποσοστά των έμμεσων απωλειών. Χρησιμοποιώντας, λοιπόν, τα δεδομένα βάσεων δεδομένων που παρήχθησαν με τον Synthetic Dataset Generator της IBM [58] προέκυψαν τα παρακάτω συμπεράσματα:

- Όταν η βάση δεδομένων αποτελείται από μικρό αριθμό στοιχείων, η αύξηση του αριθμού των συναλλαγών από τις οποίες αποτελείται οδηγεί σε σταδιακή μείωση του ποσοστού των έμμεσων απωλειών. Αντίθετα, σε βάσεις δεδομένων που απαρτίζονται από περισσότερα στοιχεία η αύξηση των συναλλαγών έχει ως αποτέλεσμα τη συνεχή αύξηση των έμμεσων απωλειών. Συνεπώς, το σύστημα μας είναι πιο αποδοτικό για βάσεις δεδομένων με σχετικά μικρό αριθμό στοιχείων.
- Η αύξηση του αριθμού των προς απόκρυψη κανόνων συσχέτισης έχει ως συνέπεια την αύξηση του ποσοστού των έμμεσων απωλειών.
- Η Επιλογή 2 οδηγεί σε μικρότερα ποσοστά έμμεσων απωλειών σε σχέση με την Επιλογή 1.
- Όσο περισσότερες συναλλαγές διαθέτει η βάση δεδομένων τόσο περισσότερο καθυστερεί η ολοκλήρωση της επεξεργασίας των δεδομένων από το σύστημα μας.
- Η αύξηση του αριθμού των προς απόκρυψη κανόνων συσχέτισης αυξάνει τη διάρκεια επεξεργασίας των δεδομένων.

- Η επεξεργασία των δεδομένων ολοκληρώνεται σε μικρότερο χρονικό διάστημα όταν επιλέξουμε την Επιλογή 1.

Συνοψίζοντας τα παραπάνω συμπεράσματα, γίνεται αντιληπτό ότι, το προτεινόμενο σύστημα έχει καλύτερα αποτελέσματα για βάσεις δεδομένων που απαρτίζονται από ένα σχετικά μικρό αριθμό στοιχείων (μέχρι 10). Όσον αφορά την επιλογή της Επιλογής 1 ή 2, αυτή εξαρτάται από το τι μας ενδιαφέρει περισσότερο. Αν ενδιαφερόμαστε να έχουμε τις λιγότερες δυνατές απώλειες θα προτιμήσουμε την Επιλογή 2 ενώ αν επιθυμούμε την ταχύτερη δυνατή επεξεργασία θα προτιμήσουμε την Επιλογή 1. Η επιλογή, λοιπόν, επαφίεται στον χρήστη του συστήματος και στους σκοπούς για τους οποίους ενδιαφέρεται να το χρησιμοποιήσει.

9. Μελλοντική Έρευνα

Η παρούσα μελέτη αναφέρεται σε ένα κατανεμημένο περιβάλλον με την έννοια ότι υπάρχουν αρκετές βάσεις δεδομένων που βρίσκονται αποθηκευμένες σε διαφορετικούς σταθμούς εργασίας. Εντούτοις, στην προσέγγιση που ακολουθήσαμε, χρησιμοποιήθηκε μία βάση δεδομένων η οποία διασπάται σε επιμέρους βάσεις οι οποίες βρίσκονται στον ίδιο σκληρό δίσκο. Αντικείμενο περαιτέρω μελέτης θα αποτελέσει η επέκταση του συστήματος ώστε να λειτουργήσει για βάσεις δεδομένων που βρίσκονται διασκορπισμένες σε διαφορετικούς Η/Υ και ενδεχομένως σε διαφορετικές γεωγραφικές περιοχές λαμβάνοντας υπόψιν και το κόστος μετάδοσης που δεν λήφθηκε υπόψιν στη συγκεκριμένη ανάλυση και αποτελεί ένα πολύ σημαντικό παράγοντα για την απόδοση ενός τέτοιου συστήματος.

Ένα άλλο θέμα που θα μελετηθεί αποτελούν οι έμμεσες απώλειες. Όπως, αναφέρθηκε για βάσεις δεδομένων με μεγάλο πλήθος στοιχείων και συναλλαγών, τα ποσοστά είναι αρκετά μεγάλα. Η μείωση, λοιπόν, του ποσοστού των έμμεσων απωλειών θα πρέπει να τύχει περαιτέρω έρευνας και μελέτης.

Τέλος, παρατηρήθηκε ότι η αύξηση του πλήθους των συναλλαγών της βάσης δεδομένων αυξάνει και μάλιστα εκθετικά τους χρόνους επεξεργασίας. Κάτι τέτοιο σημαίνει ότι για βάσεις δεδομένων με πολλές χιλιάδες ή ακόμη και εκατομμύρια συναλλαγές, το σύστημα θα χρειαστεί αρκετά μεγάλο χρονικό διάστημα (της τάξης μερικών ωρών) μέχρι να ολοκληρώσει την επεξεργασία όλων των δεδομένων.

Συνεπώς, θα πρέπει να μελετηθούν τρόποι που θα επιτύχουν τα καλύτερα δυνατά αποτελέσματα στο συντομότερο χρονικό διάστημα.

10. Βιβλιογραφία

[1] Dakshi Agrawal and Charu C. Aggarwal, *On the design and quantification of privacy preserving data mining algorithms*, In Proceedings of the 20th ACM Symposium on Principles of Database Systems (2001), 247-255

[2] Rakesh Agrawal and Ramakrishnan Srikant, *Privacy-preserving data mining*, In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), 439–450

[3] R. Agrawal and J. Shafer, *Parallel mining of association rules*, IEEE Transactions on Knowledge and Data Engineering, 8(6): 962-969, 1996

[4] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, *Mining Association Rules Between Sets of Items in Large Databases*, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, D.C., May 1993

[5] Rakesh Agrawal and Ramakrishnan Srikant, *Fast Algorithms for Mining Association Rules in Large Databases*, In Proceedings of the Twentieth International Conference on Very Large Databases, pp. 487-499, Santiago, Chile, 1994

[6] Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim and Vassilios S. Verykios, *Disclosure Limitation of Sensitive Rules*, In Proceedings of the IEEE Knowledge and Data Engineering Workshop (1999), 45-52

[7] Ashrafi, M. Z., Taniar, D. and Smith K.A., *PPDAM: Privacy-preserving distributed association rule mining algorithm*, International Journal of Intelligent Information Technologies, vol. 1, no. 1, pp. 49-69, 2005

[8] Mafruz Zaman Ashrafi, David Taniar, Kate Smith, *ODAM: An Optimized Distributed Association Rule Mining Algorithm*, In IEEE Distributed Systems Online, March 2004

[9] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur, *Dynamic Itemset Counting and Implication Rules for Market Basket Data*, In Proceedings of the ACM SIGMOD Conference, pp. 255-264, 1997

[10] George Cahlink, *Data Mining Taps the Trends*, Government Executive Magazine, October 1,2000, [http://www.govexec.com/tech/articles/1000managetech.htm]

[11] Liwu Chang and Ira S. Moskowitz, *An integrated framework for database inference and privacy protection*, Data and Applications Security (2000), 161-172, Kluwer, IFIP WG 11.3, The Netherlands

[12] Liwu Chang and Ira S. Moskowitz, *Parsimonious downgrading and decision trees applied to the inference problem*, In Proceedings of the 1998 New Security Paradigms Workshop (1998), 82-89

[13] D. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu. A *fast distributed algorithm for mining association rules*. In Proc. Of 1996 Int'l. Conf. on Parallel and Distributed Information Systems, pages 31 . 44, Miami Beach, Florida, December 1996.

[14] D. W. Cheung, V. T. Ng, A. W. Fu, and Y. Fu, *Efficient Mining of Association Rules in Distributed Databases*, IEEE Transactions On Knowledge And Data Engineering, 8:911-922, 1996

[15] Chris Clifton, Murat Kantarcioglou, Xiadong Lin, and Michael Y. Zhu, *Tools for privacy preserving distributed data mining*, SIGKDD Explorations 4 (2002), no. 2

[16] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid and Elisa Bertino, *Hiding Association Rules by using Confidence and Support*, In Proceedings of the 4th Information Hiding Workshop (2001), 369-383

[17] Wenliang Du and Mikhail J. Atallah, *Secure Multi-Party Computation Problems and their Applications: A Review and Open Problems* , In New Security Paradigms Workshop 2001. September 11th - 13th, 2001, Cloudcroft, New Mexico, USA. Pages 11-20

[18] Wenliang Du and Zhijun Zhan, *Building decision tree classifier on private data*, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002)

[19] Cynthia Dwork and Kobbi Nissim. *Privacy-preserving Data mining on Vertically Partitioned Databases*, In Proceedings of the 24rd Annual International Cryptology Conference (CRYPTO 2004), Santa Barbara, CA, August 2004

[20] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, *Privacy preserving mining of association rules*, In Proceedings of the 8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining (2002)

[21] O. Goldreich, S. Micaly and A. Wigderson, *How to play any mental game*, In Proceedings of the 19th annual ACM symposium on Theory of computing, pages 218-229, 1987

[22] Christian Hidber, *Online Association Rule Mining*, SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadephia, Pennsylvania, pp.145-156.

[23] M. Houtsma and A. Swami, *Set-Oriented Mining for Association Rules in Relational Databases*, In Proceedings of the 11th IEEE International Conference on Data Engineering, pp. 25-34, Taipei, Taiwan, March 1995

[24] Ioannis Ioannidis, Ananth Grama and Mikhail Atallah, *A secure protocol for computing dot products in clustered and distributed environments*, In Proceedings of the International Conference on Parallel Processing (2002)

[25] Murat Kantarcioglou and Chris Clifton, *Privacy-preserving distributed mining of association rules on horizontally partitioned data*, In Proceedings of the ACM SIGMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery (2002), 24–31

[26] Ira S. Moskowitz and Liwu Chang, *A decision theoretical based system for information downgrading*, In Proceedings of the 5th Joint Conference on Information Sciences (2000)

[27] Y. Li, X. Lin, C. Tsang, *An Efficient Distributed Algorithm for Computing Association Rules*, In Proceedings of 1st International Conference on Web Age Information Management, LNCS 1846, Springer-Verlag, 2000

[28] Li, Shenzhi and Pottenger, William M., *DiHO: A Distributed Higher-Order Association Rule Miner*, In the Proceedings of the 24th ACM SIGMOD International Conference on Management of Data. Baltimore, MD, June 2004

[29] Yehuda Lindell and Benny Pinkas, *Privacy preserving data mining*, In Advances in Cryptology - CRYPTO 2000 (2000), 36–54

[30] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo, *Efficient Algorithms for Discovering Association Rules*, In Proceedings of the AAAI Workshop on Knowledge Discovery in Databases (KDD-94), pp. 181-192, July 1994

[31] A. Mueller. *Fast sequential and parallel algorithms for association rule mining: A comparison*. Technical Report CS-TR-3515, University of Maryland, College Park, August 1995.

[32] Stanley R. M. Oliveira and Osmar R. Zaiane, *Privacy preserving frequent itemset mining*, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002), 43– 54

[33] Salvatore Orlando, Paolo Palmerini, Raffaele Perego, Fabrizio Silvestri: *An Efficient Parallel and Distributed Algorithm for Counting Frequent Sets*. VECPAR 2002: 421-435

[34] M. Otey, A. Veloso, C. Wang, S. Parthasarathy and W. Meira, *Mining Frequent Itemsets in Distributed and Dynamic Databases,* IEEE International Conference on Data Mining, 2003

[35] J. S. Park, M. Chen, and P. S. Yu. *An effective hash based algorithm for mining association rules*. In ACM SIGMOD Intl. Conf. Management of Data, May 1995.

[36] E. Pontikakis, A. Tsitonis, V. Verykios, *An Experimental Study of Distortion-based Techniques in Association Rule Hiding,* DBSec 2004: 325-339

[37] Emmanuel D. Pontikakis, Yannis Theodoridis, Achilleas A. Tsitsonis, LiWu Chang, Vassilios S. Verykios, *A quantitative and qualitative analysis of blocking in association rule hiding,* WPES 2004: 29-30

[38] Shariq J. Rizvi and Jayant R. Haritsa, *Maintaining data privacy in association rule mining,* In Proceedings of the 28th International Conference on Very Large Databases, Hong Kong, China, 2002

[39] Ashoka Savasere, Edward Omiecinski, and Shamkant B. Navathe, *An Efficient Algorithm for Mining Association Rules in Large Databases,* In Proceedings of the 21nd International Conference on Very Large Databases, pp. 432-444, Zurich, Swizerland, 1995

[40] Yucel Saygin, Vassilios Verykios and Ahmed K. Elmagarmid, *Privacy preserving association rule mining,* In Proceedings of the 12th International Workshop on Research Issues in Data Engineering (2002), 151-158

[41] Yucel Saygin, Vassilios Verykios and Chris Clifton, *Using unknows to prevent discovery of association rules,* SIGMOD Record 30 (2001), no 4, 45-54

[42] Assaf Schuster and Ran Wolff, *Communication Efficient Distributed Mining of Association Rules*, In Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, volume 30, pages 473-484, California, USA, June 2001

[43] Assaf Schuster, Ran Wolff, and Dan Trock, *A High-Performance Distributed Algorithm for Mining Association Rules*, In Third IEEE International Conference on Data Mining, Florida , USA, November 2003

[44] Assaf Schuster, Ran Wolff, and Bobi Gilburd, *Privacy-Preserving Association Rule Mining in Large-Scale Distributed Systems*, In 4th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid'04), Illinois , USA, April 2004

[45] C. Silvestri and S. Orlando, *Distributed association mining: an approximate method*, Proceedings of the 7th International Workshop on High Performance and Distributed Data Mining, in conjunction with Third International SIAM Conference on Data Mining, April 22-24, 2004

[46] Hannu Toivonen, *Sampling Large Databases for Association Rules*, In Proceedings of the 22nd International Conference on Very Large Databases, pp. 134-145, Mumbai, India, 1996

[47] Jaideep Vaidya and Chris Clifton, *Privacy preserving association rule mining in vertically partitioned data*, In the 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (2002), 639–644

[48] Vassilios S. Verykios, Ahmed K. Elmagarmid, Bertino Elisa, Yucel Saygin and Dasseni Elena, *Association Rule Hiding*, IEEE Transactions on Knowledge and Data Enigineering (2003)

[49] A. Veloso, W. Meira, S. Parthasarathy and M. B. de Carvalho, *Efficient, Accurate and Privacy-Preserving Data Mining for Frequent Itemsets in Distributed Databases***,** Brazilian Symposium on Databases 2003

[50] R. Wolf and A. Schuster, *Association rule mining in peer-to-peer systems*, In Proc. ICDM'03, November 2003

[51] A. C. Yao, *Protocols for Secure Computations*, In Proceedings of the 23th Annual IEEE Symposium on Foundations of Computer Science, 1982

[52] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New *algorithms for fast discovery of association rules*. In 3rd Intl. Conf. on Knowledge Discovery and Data Mining, August 1997.

[53] M .Zaki, S. Parthasarathy, M. Ogihara and W. Li, *New Parallel algorithms for fast discovery of association rules*, Data Mining and Knowledge Discovery: An international Journal, vol. 4, no. 1, pp. 343-373, December 1997

[54] http://www.epic.org/privacy/profiling/tia/

[55] http://www.darpa.mil

[56] http://www.gao.gov/cgi-bin/getrpt?GAO-04-385

[57] http://www.tsa.gov/public/index.jsp

[58] http://www.almaden.ibm.com/software/quest/Resources/datasets/syndata.html

[59] http://www.ics.uci.edu/~mlearn/MLRepository.html

[60] http://fimi.cs.helsinki.fi/

Παράρτημα

Α. Αντιστοίχιση Όρων

Α/Α	Αγγλικός όρος	Ελληνικός όρος
1	Association Rule	Κανόνας Συσχέτισης
2	Association Rule Mining (ARM)	Εξόρυξη Κανόνων Συσχέτισης
3	Bandwidth	Εύρος Ζώνης
4	Blocking	Μπλοκάρισμα
5	Centralized Database	Κεντρική βάση Δεδομένων
6	Characterization	Χαρακτηρισμός
7	Classification	Κατηγοριοποίηση
8	Classifier	Κατηγοριοποιητής
9	Clustering	Συσταδοποίηση
10	Communication Cost	Κόστος μετάδοσης
11	Confidence	Εμπιστοσύνη
12	Confidence Interval	Διάστημα Εμπιστοσύνης
13	Confusion	Σύγχυση
14	Data Distribution	Κατανομή Δεδομένων
15	Data Mining	Εξόρυξη Γνώσης
16	Data Set	Σύνολο Δεδομένων
17	Data Skewness	Ασυμμετρία Δεδομένων
18	Database	Βάση Δεδομένων
19	Database Management System (DBMS)	Σύστημα Διαχείρισης Βάσεων
20	Decision Tree	Δέντρο Απόφασης
21	Distorting	Παραμόρφωση
22	Distributed Database	Κατανεμημένη Βάση Δεδομένων
23	Forecasting	Πρόγνωση
24	Fragmentation	Κατάτμηση
25	Framework	Περιβάλλον Εργασίας
26	Fully Supports	Υποστηρίζει Πλήρως
27	Generalization	Γενίκευση
28	Generating Itemsets	Παραγόμενα Στοιχειοσύνολα
29	Globally Large	Καθολικά Μεγάλο
30	Heterogeneous data	Ετερογενή Δεδομένα
31	Heuristic techniques	Ευριστικές τεχνικές
32	Homogeneous Data	Ομοιογενή Δεδομένα
33	Horizontal Layout	Οριζόντιο Σχήμα
34	Induction	Επαγωγή
35	Inference Channels	Κανάλια Συμπερασμάτων
36	Input / Output Time	Χρόνος Εισόδου / Εξόδου

37	Integration of Results	Ενοποίηση Αποτελεσμάτων
38	Item	Στοιχείο
39	Itemset	Στοιχειοσύνολο
40	Join Step	Βήμα Συνένωσης
41	Large Itemsets	Μεγάλα Στοιχειοσύνολα
42	Locally Large	Τοπικά μεγάλο
43	Machine Learning	Μηχανική Μάθηση
44	Missing Data	Ελλιπή Δεδομένα
45	Negative Border	Αρνητικό Περιθώριο
46	Neural Networks	Νευρωνικά Δίκτυα
47	Parsimonious Downgrading	Φειδωλή Υποβάθμιση
48	Partially Supports	Υποστηρίζει Μερικώς
49	Path Analysis	Ανάλυση Μονοπατιού
50	Perturbation	Διατάραξη
51	Privacy Breaches	Κενά ασφάλειας
52	Privacy Preserving	Προστασία γνώσης
53	Prune Step	Βήμα Αποκοπής
54	Randomization Techniques	Τεχνικές Τυχαιοποίησης
55	Replication of Data	Ανταλλαγή Δεδομένων
56	Safety Margin	Όριο Ασφάλειας
57	Secure Multiparty Computation	Ασφαλής Πολυμερής Υπολογισμός
58	Schema	Σχήμα
59	Sensitive Rules	Ευαίσθητοι Κανόνες
60	Sensitivity Level	Επίπεδο Εμπιστοσύνης
61	Sequence Analysis	Ακολουθιακή Ανάλυση
62	Side Effects	Έμμεσες απώλειες
63	Site	Μέρος
64	Strong Rule	Ισχυρός Κανόνας
65	Summarization	Παρουσίαση Συνόψεων
66	Support	Υποστήριξη
67	Support Interval	Διάστημα Υποστήριξης
68	Synthetic	Τεχνητός
69	Task Distribution	Κατανομή Εργασιών
70	Training Data	Δεδομένα Εκπαίδευσης
71	Transaction	Συναλλαγή
72	Tuple	Πλειάδα
73	Unknown Attributes	Άγνωστες Τιμές
74	Vertical Layout	Κατακόρυφο Σχήμα

Β. Ερμηνεία Συμβόλων

Σύμβολο	Ερμηνεία
I	σύνολο στοιχείων της βάσης δεδομένων
i_n	στοιχείο της βάση δεδομένων
ΒΔ	κεντρική βάση δεδομένων
$ΒΔ_i$	τοπική βάση δεδομένων
T	συναλλαγή
κ-στοιχειοσύνολο	στοιχειοσύνολο με κ στοιχεία
r	κανόνας συσχέτισης
l_r	σύνολο στοιχείων που βρίσκονται στο αριστερό μέρος του κανόνα r
r_r	σύνολο στοιχείων που βρίσκονται στο δεξί μέρος του κανόνα r
$Σ_r$	στοιχειοσύνολο που αντιστοιχεί στον κανόνα r
ΕΤΥ	Ελάχιστη Τιμή Υποστήριξης
ΕΤΕ	Ελάχιστη Τιμή Εμπιστοσύνης
L	σύνολο μεγάλων στοιχειοσυνόλων
L_H	σύνολο μεγάλων στοιχειοσυνόλων για απόκρυψη
R	σύνολο κανόνων συσχέτισης
R_H	σύνολο κανόνων συσχέτισης για απόκρυψη
T_Z	σύνολο συναλλαγών που υποστηρίζουν το στοιχειοσύνολο Z
T_r	σύνολο συναλλαγών που υποστηρίζουν πλήρως το παραγόμενο στοιχειοσύνολο του κανόνα συσχέτισης r
M	μέγεθος κεντρικής βάσης δεδομένων
M_i	μέγεθος τοπικών βάσεων δεδομένων
Χ.Υποστήριξη	υποστήριξη του στοιχειοσυνόλου Χ στην κεντρική βάση δεδομένων
$Χ.Υποστήριξη_{(i)}$	υποστήριξη του στοιχειοσυνόλου Χ στην τοπική βάση δεδομένων
ΚΜ	σύνολο καθολικά μεγάλων στοιχειοσυνόλων στην κεντρική βάση ΒΔ
$ΚΜ_{(k)}$	σύνολο καθολικά μεγάλων k-στοιχειοσυνόλων στην κεντρική βάση ΒΔ
$ΚΤΜ_i$	σύνολο κτ-μεγάλων στοιχειοσυνόλων στη τοπική βάση $ΒΔ_i$
$ΚΤΜ_{i(k)}$	σύνολο κτ-μεγάλων κ-στοιχειοσυνόλων στη τοπική βάση $ΒΔ_i$
$ΥΣ_{(k)}$	σύνολο υποψήφιων στοιχειοσυνόλων που προκύπτουν από το σύνολο $ΚΜ_{(k-1)}$
$ΥΣ_{i(k)}$	σύνολο υποψήφιων στοιχειοσυνόλων που προκύπτουν από το σύνολο $ΚΜ_{i(k-1)}$
$ΤΜ_{i(k)}$	σύνολο τοπικά μεγάλων κ-στοιχειοσυνόλων που προκύπτουν από το σύνολο $ΥΣ_{i(k)}$
ΑΠΥ	Ασφαλής Πολυμερής Υπολογισμός
ΚΚΣ	σύνολο καθολικών κανόνων συσχέτισης
ΤΚΣ	σύνολο τοπικών κανόνων συσχέτισης
ΕΕΚ	σύνολο ευαίσθητων κανόνων συσχέτισης
ΚΑΕ	Κατανεμημένος Αλγόριθμος Εξόρυξης
ΑΕ	Αλγόριθμος Εξόρυξης

ΑΑΚ	Αλγόριθμος Απόκρυψης Κανόνων
Κ.Μ.Ε.	Κεντρική Μονάδα Επεξεργασίας

Γ. Ονοματολογία Αλγορίθμων

Όνομα Αλγορίθμου	Περιγραφή
Apriori	Αλγόριθμος Εξόρυξης κανόνων συσχέτισης από μία κεντρική βάση δεδομένων
FDM	Κατανεμημένος Αλγόριθμος Εξόρυξης κανόνων συσχέτισης
MY	Αλγόριθμος απόκρυψης κανόνων συσχέτισης μειώνοντας την υποστήριξη
ME	Αλγόριθμος απόκρυψης κανόνων συσχέτισης μειώνοντας την ελάχιστη υποστήριξη
ME2	Αλγόριθμος απόκρυψης κανόνων συσχέτισης αυξάνοντας την μέγιστη υποστήριξη
OA1	Αλγόριθμος υπολογισμού Ορίου Ασφαλείας για τον συνδυασμό αλγορίθμων ME-ME2
OA2	Αλγόριθμος υπολογισμού Ορίου Ασφαλείας για τον συνδυασμό αλγορίθμων MY-ME2
AK	Αλγόριθμος ανάκτησης κανόνων συσχέτισης

Δ. Πίνακες Μετρήσεων

Εδώ παρατίθενται οι πίνακες με τα στοιχεία που προέκυψαν από τα πειράματα που πραγματοποιήθηκαν στο κεφάλαιο 7.

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	34,48%	27,58%	31,03%
1000	34,52%	36,90%	37,71%
2000	36,48%	39,18%	39,18%
4000	14,28%	20,40%	23,45%
8000	8,14%	14,67%	20,12%

Πίνακας 13: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 1 για 10 στοιχεία

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	22,98%	32,18%	29,88%
1000	22,61%	36,90%	36,90%
2000	28,37%	29,72%	27,02%
4000	10,12%	22,44%	18,36%
8000	5,90%	10,63%	16,74%

Πίνακας 14: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 2 για 10 στοιχεία

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	12,07%	17,20%	33,85%
1000	29,50%	37,61%	38,96%
2000	32,12%	42,56%	48,44%
4000	42,79%	46,78%	55,56%
8000	50,22%	61,40%	65,79%

Πίνακας 15: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 1 για 15 στοιχεία

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	12,90%	19,02%	31,38%
1000	18,69%	32,43%	32,20%
2000	28,57%	34,12%	36,65%
4000	31,62%	41,67%	39,23%
8000	37,89%	45,89%	42,78%

Πίνακας 16: Ποσοστά Απωλειών κανόνων συσχέτισης για την Επιλογή 2 για 15 στοιχεία

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	36,578	39,813	46,605
1000	176,468	102,016	177,25
2000	631,594	616,313	688,312
4000	934,906	1024,515	1375,422
8000	4512,289	5013,346	5578,178

Πίνακας 17: Χρόνοι Επεξεργασίας για την Επιλογή 1 για 10 στοιχεία

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	56,187	57,047	66,626
1000	110,782	169,688	198,234
2000	407,219	450,907	544,438
4000	828,512	1536,969	1733,031
8000	4310,179	5341,223	5896,265

Πίνακας 18: Χρόνοι Επεξεργασίας για την Επιλογή 2 για 10 στοιχεία

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	621,484	607,968	570,187
1000	1000,093	1016,25	1038,281
2000	1923,678	2112,698	2002,427
4000	2879,127	3109,371	3696,12
8000	5458,127	5890,645	6210,22

Πίνακας 19: Χρόνοι Επεξεργασίας για την Επιλογή 1 για 15 στοιχεία

Πλήθος συναλλαγών	4 κανόνες	6 κανόνες	8 κανόνες
500	623,734	621,816	611,187
1000	1002,61	1054,843	1150,25
2000	2001,102	2341,997	2689,145
4000	2997,139	3221,21	3994,521
8000	5698,793	5996,326	6456,197

Πίνακας 20: Χρόνοι Επεξεργασίας για την Επιλογή 2 για 15 στοιχεία

Ε. Περιεχόμενα CD

Η παρούσα μελέτη συνοδεύεται από ένα CD το οποίο έχει τα εξής περιεχόμενα:

- Το αρχείο thesis.pdf που περιέχει το κείμενο της διπλωματικής
- Το αρχείο thesis.ppt που περιέχει την παρουσίαση της διπλωματικής
- Τον φάκελο code με τα αρχεία devide.pl και thesis_code.pl που περιέχουν τον κώδικα με βάση τον οποίο πραγματοποιήθηκαν τα πειράματα του κεφαλαίου 7
- Το αρχείο readme.txt που περιέχει οδηγίες τα βήματα εκτέλεσης του κώδικα των αρχείων του φακέλου code
- Το αρχείο arrays.xls που περιέχει τους πίνακες με τις μετρήσεις που προέκυψαν από τα πειράματα
- Το αρχείο results.pdf που περιέχει τις γραφικές παραστάσεις των μετρήσεων
- Τον φάκελο IBM_Generator που περιέχει IBM Quest Market-Basket Synthetic Data Generator για περιβάλλον Windows. Συστήνεται η εκτέλεση των αρχείων σε περιβάλλον Microsoft Visual Studio