ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Τεχνικές Αναπαράστασης και Διαχείρισης Προτύπων – το Σύστημα Διαχείρισης Βάσεων Προτύπων

Διδακτορική Διατριβή

Ευάγγελος Η. Κοτσιφάκος

ΜΔΕ, Πληροφοριακά Συστήματα, Ο.Π.Α. (2003)

Πτυχίο, Πληροφορικής, Ο.Π.Α. (2001)

Αθήνα, Δεκέμβριος 2009

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Συμβουλευτική Επιτροπή:

Επιβλέπων:

Ιωάννης Θεοδωρίδης
Αν. Καθηγητής Πανεπιστημίου Πειραιώς

Μέλη:

Μιχάλης Βαζιργιάννης
Αν. Καθηγητής Οικονομικού Πανεπιστημίου Αθηνών

Μαρία Βίρβου
Καθηγήτρια Πανεπιστημίου Πειραιώς

Διατριβή
για την απόκτηση Διδακτορικού Διπλώματος του Τμήματος Πληροφορικής

ΕΥΑΓΓΕΛΟΥ Η. ΚΟΤΣΙΦΑΚΟΥ

“ Τεχνικές Αναπαράστασης και Διαχείρισης Προτύπων – το Σύστημα Διαχείρισης Βάσεων Προτύπων ”

Εξεταστική Επιτροπή:

Δημήτριος Δεσπότης
Καθηγητής Πανεπιστημίου Πειραιώς

Δημήτριος Αποστόλου
Λέκτορας Πανεπιστημίου Πειραιώς

Χαράλαμπος Κωνσταντόπουλος
Λέκτορας Πανεπιστημίου Πειραιώς

Άγγελος Πικράκης
Λέκτορας Πανεπιστημίου Πειραιώς

.....

ΕΥΑΓΓΕΛΟΣ Η. ΚΟΤΣΙΦΑΚΟΣ

Εγκεκριμένο από την Εξεταστική Επιτροπή, ...

.....	.....	.....
Ιωάννης Θεοδωρίδης	Μιχάλης Βαζιργιάννης	Μαρία Βίρβου
Αν. Καθηγητής, Πανεπιστημίου Πειραιώς	Αν. Καθηγητής, Οικον. Παν. Αθηνών	Καθηγήτρια, Παν. Πειραιώς
Επιβλέπων	Μέλος Συμβουλευτικής Επιτροπής	Μέλος Συμβουλευτικής Επιτροπής

.....	.....	.....
Δημήτριος Δεσπότης	Δημήτριος Αποστόλου	Χαράλαμπος Κωνσταντόπουλος
Καθηγητής, Παν, Πειραιώς	Λέκτορας, Παν. Πειραιώς	Λέκτορας, Παν. Πειραιώς
Μέλος Εξεταστικής Επιτροπής	Μέλος Εξεταστικής Επιτροπής	Μέλος Εξεταστικής Επιτροπής

.....

Άγγελος Πικράκης

Λέκτορας,
Παν. Πειραιώς

Μέλος Εξεταστικής Επιτροπής

Για τον Κωστή

Πρόλογος

Λόγω του μεγάλου όγκου των προτύπων που εξάγονται από βάσεις δεδομένων με τεχνικές εξόρυξης γνώσης, την πολυπλοκότητα και ετερογένειά τους, η ανάγκη για διαχείριση των προτύπων με ενιαίο τρόπο, είναι επιτακτική. Ένα Σύστημα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ - Pattern Base Management System (PBMS)) υποστηρίζει λειτουργίες προτύπων, όπως η αποθήκευση, η ανάκτηση και η σύγκριση και έχει ένα μεγάλο εύρος εφαρμογών σε κάθε επιστημονικό πεδίο.

Η παρούσα διατριβή αντιμετωπίζει τα θέματα του ορισμού ενός μοντέλου αναπαράστασης προτύπων για ένα ΣΔΒΠ και της σύγκρισης διακριτών και ασαφών προτύπων συστάδων. Προτείνουμε μία νέα μετρική σύγκρισης προτύπων συστάδων και παρουσιάζουμε ένα πρωτότυπο αλγόριθμο για διαισθητικά ασαφή συσταδοποίηση. Οι νέες μετρικές ενσωματώνονται στο πλαίσιο σύγκρισης PANDA και περιγράφουμε πραγματικές εφαρμογές και πειράματα. Επίσης, παρουσιάζουμε ένα πρωτότυπο ΣΔΒΠ, το PatternMiner, ένα ολοκληρωμένο και επεκτάσιμο περιβάλλον για τη διαχείριση των προτύπων. Επιπλέον, μελετάμε το πρόβλημα της αξιολόγησης των προτύπων και προτείνουμε τη χρήση οντολογιών για να παρέχεται στους ειδικούς πολύτιμη σημασιολογική πληροφορία για τα εξαγόμενα πρότυπα σε συγκεκριμένο πεδίο γνώσης.

Ευάγγελος Η. Κοτσιφάκος

Ευχαριστίες

Θα ήθελα να απευθύνω τις βαθειές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, Αν. Καθηγητή κ. Ιωάννη Θεοδωρίδη, για την καθοδήγησή του, την υποστήριξη και την έμπνευση που μου παρείχε όλα τα χρόνια της έρευνάς μου και της προσωπική μου εξέλιξης. Οι συμβουλές του με βοήθησαν να ξεπεράσω τις όποιες ερευνητικές δυσκολίες που παρουσιάζονταν ενώ με τη γνώση και την εμπειρία του μου παρείχε πάντα νέες ιδέες. Επιπλέον, αποτέλεσε παράδειγμα με τις αποτελεσματικές εκπαιδευτικές μεθόδους υποστηρίζοντας και προωθώντας πάντα τις ηθικές αξίες σε ακαδημαϊκό και προσωπικό επίπεδο.

Ιδιαίτερες ευχαριστίες για τους συν-ερευνητές μου, την Ειρήνη, το Νίκο, το Γεράσιμο, το Νίκο, τη Δέσποινα, τον Γιάννη και το Δημήτρη για τη συμβολή τους σε διάφορα πεδία της έρευνας μου. Η γνώση τους και η επιστημονική τους ωριμότητα μου έδωσαν την ευκαιρία να βελτιώσω την επιστημονική μου σκέψη και γραφή.

Θα ήθελα επίσης να ευχαριστήσω τον Αν. Καθηγητή κ. Μιχάλη Βαζιργιάννη και την Καθηγήτρια κα Μαρία Βίρβου για τα παραγωγικά σχόλια στην εργασία μου, τους Καθηγητές κ. Ευάγγελο Κοντιζά και την κα Μαίρη Κοντιζά για τη συνεργασία τους όπως επίσης και τους μεταπτυχιακούς φοιτητές Αντώνη, Βιβή, Γιάννη, Δήμητρα και Κωνσταντίνο για την βοήθειά τους σε διάφορα τεχνικά θέματα.

Ιδιαίτερες ευχαριστίες στα αδέρφια μου Αλέξη και Πρόδρομο για την πολύτιμη βοήθειά τους, στους γονείς μου που μου έδωσαν τα εφόδια να φτάσω στο σημείο αυτό της εκπαίδευσής μου, στην Κέλλυ που με υποστήριξε και με ενέπνευσε τα χρόνια αυτά, και στον Κωστή, που μου έδωσε έμπνευση για την έρευνά μου μέσα από τις ατελείωτες συζητήσεις μας στα πρώτα χρόνια του Μεταπτυχιακού και του Διδακτορικού μου, και του οποίου η μνήμη με ενθάρρυνε να ολοκληρώσω αυτή τη διατριβή.

Η παρούσα έρευνα υποστηρίχτηκε κατά κύριο λόγο από την υποτροφία ΠΕΝΕΔ 2003 της ΓΓΕΤ. Επιπρόσθετα, υποστηρίχθηκε από το έργο MetaOn, που χορηγήθηκε από το πρόγραμμα "Information Society" της Γενικής Γραμματείας Έρευνας και Τεχνολογίας (ΓΓΕΤ) του Υπ.Ανάπτυξης, συγχρηματοδοτούμενο και από την Ευρωπαϊκή Ένωση.

Η συλλογή των εικόνων που χρησιμοποιήθηκαν στην ενότητα 3.4 παρασχέθηκε από τον *Dr. T.M. Lehmann, Image Retrieval in Medical Application (IRMA) group, Dept. of Medical Informatics, RWTH Aachen, Germany, http://irma-project.org.*

Η συλλογή των εικόνων που χρησιμοποιήθηκαν στην ενότητα 3.5 παρασχέθηκε από το Ίδρυμα Μείζονος Ελληνισμού (ΙΜΕ), http://www.fhw.gr.

Τα δεδομένα στην ενότητα 6.3 συλλέχθηκαν από το Ελληνικό Γεωδυναμικό Ινστιτούτο του Εθνικού Αστεροσκοπείου Αθηνών (http://www.gein.noa.gr).

Τα δεδομένα που χρησιμοποιήθηκαν στη μελέτη της ενότητας 5.2 παρασχέθηκαν από το Sloan Digital Sky Survey (http://www.sdss.org/).

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1 Εισαγωγή

Η εξόρυξη γνώσης από δεδομένα (data mining) αποτελεί ένα από τα βήματα της διαδικασίας ανακάλυψης γνώσης και κυρίως διαθέτει τις μεθοδολογίες για την εξαγωγή προτύπων από μεγάλες βάσεις δεδομένων (Εικόνα 1-1). Οι κανόνες συσχέτισης (association rules), οι συστάδες (clusters), τα δέντρα απόφασης (decision trees), είναι κάποια γνωστά πρότυπα που προέρχονται από την περιοχή της εξόρυξης γνώσης. Τα πρότυπα όμως μπορεί επίσης να βρεθούν σε άλλες περιοχές όπως στα Μαθηματικά (πχ. πρότυπα σε ακολουθίες, σε αριθμούς, σε γράφους, σχήματα κλπ), Γεωμετρία, Επεξεργασία σήματος κλπ (Vazirgiannis et al., 2003). Λόγω της πολυμορφίας και της ποικιλίας των προτύπων, προκύπτει ένα σημαντικό ζήτημα: η διαχείριση όλων των προτύπων με έναν ενιαίο τρόπο, είτε αυτά έχουν αξιολογηθεί ή όχι.

Εικόνα 1-1 Η διαδικασία ανακάλυψης γνώσης (KDD)

Προς το παρόν, η πλειοψηφία των διαθέσιμων εργαλείων εξόρυξης γνώσης υποστηρίζει την οπτικοποίηση των προτύπων και, στην καλύτερη περίπτωση, την αποθήκευσή τους σε σχεσιακούς πίνακες. Σε συνδυασμό με το χαρακτηρισμό των προτύπων ως σύνθετη, συμπαγής και πλούσια σε σημασιολογία αναπαράσταση των δεδομένων (Rizzi et al., 2003), αυτό το ζήτημα είναι μία πρόκληση για την αποτελεσματική διαχείριση των προτύπων σε ένα Σύστημα. Σε αναλογία με το Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ - Database Management System (DBMS)), ένα Σύστημα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ - Pattern Base Management System (PBMS)) πρέπει

να είναι σε θέση να χειρίζεται τα πρότυπα όπως το ΣΔΒΔ χειρίζεται τα δεδομένα. Άρα, ένα ΣΔΒΠ μπορεί να χρησιμοποιηθεί για την αναπαράσταση, την αποθήκευση, την ανάκτηση, την ευρετηριοποίηση και την ενημέρωση των προτύπων (representing, storing, querying, indexing and updating patterns). Επιπλέον μπορούν να οριστούν προχωρημένες λειτουργίες για τα πρότυπα όπως είναι η σύγκριση και η παρακολούθηση της εξέλιξής τους στο χρόνο. Τα πρότυπα εξάγονται από τα αρχικά δεδομένα και για το λόγο αυτό υπάρχει μία σύνδεση μεταξύ των προτύπων και των δεδομένων από τα οποία εξήχθησαν, οπότε μία αλλαγή στα δεδομένα μπορεί να συνεπάγεται και αλλαγή στα σχετικά πρότυπα.

Γενικότερα, οι βασικές λειτουργίες για τα πρότυπα (ως αποτέλεσμα των αλγορίθμων εξόρυξης γνώσης) είναι:

- Αποθήκευση των προτύπων που εξάγονται από το ίδιο σύνολο δεδομένων με χρήση διαφορετικών αλγορίθμων και παραμέτρων.
- Ανάκτηση των προηγούμενα αποθηκευμένων προτύπων χρησιμοποιώντας διάφορα κριτήρια όπως είναι το σύνολο των αρχικών δεδομένων, η ώρα και ημερομηνία της εξαγωγής των προτύπων, οι παράμετροι που χρησιμοποιήθηκαν ή ακόμα και συγκεκριμένες ιδιότητες/ τιμές του αποτελέσματος (των προτύπων).
- Σύγκριση προτύπων που έχουν εξαχθεί από το ίδιο σύνολο δεδομένων, είτε με διαφορετικές παραμέτρους είτε σε διαφορετική χρονική στιγμή.
- Παρακολούθηση της εξέλιξης των προτύπων στο χρόνο.

Η σύγκριση των προτύπων είναι μία προχωρημένη και σημαντική λειτουργία σε πολλές πραγματικές εφαρμογές καθώς παρέχει μία υψηλού επιπέδου σύγκριση των δεδομένων. Η σύγκριση των (μεγάλου όγκου) δεδομένων είναι χρονοβόρα και απαιτεί μεγάλη επεξεργαστική ισχύ και λειτουργίες I/O. Με τη χρήση των προτύπων σαν μία μορφή αναπαράστασης των δεδομένων, η σύγκριση των προτύπων αντιστοιχεί σε σύγκριση των δεδομένων που τα πρότυπα αναπαριστούν αλλά απαιτεί πολύ λιγότερο χρόνο και γενικότερα λιγότερους πόρους.

Ο ορισμός των μετρικών ομοιότητας για τα πρότυπα κάνει εφικτές πολλές ενδιαφέρουσες εφαρμογές όπως παρουσιάζεται στο (Ntoutsi, 2008):

1. Απλές επερωτήσεις ομοιότητας των προτύπων

2. Παρακολούθηση των αλλαγών στη βάση δεδομένων
3. Σύγκριση συνόλων δεδομένων
4. Αξιολόγηση αλγορίθμων εξόρυξης γνώσης
5. Εξόρυξη γνώσης από κατανεμημένες πηγές δεδομένων
6. Ανακάλυψη ασυνήθιστων ή μη αναμενόμενων προτύπων (outliers)

Για την αποτελεσματική διαχείριση και υποστήριξη των παραπάνω λειτουργιών ένα ενιαίο μοντέλο αναπαράστασης προτύπων πρέπει να οριστεί, που να υποστηρίζει προχωρημένες λειτουργίες επερώτησης και σύγκρισης.

Στην παρούσα διατριβή υιοθετούμε την προσέγγιση του ερευνητικού έργου PANDA (PAtterns for Next-generation DAtabase systems) (PANDA, 2005) το οποίο αντιμετωπίζει θέματα αναπαράστασης και διαχείρισης των προτύπων, και προδιαγράφουμε ένα Σύστημα Διαχείρισης Βάσεων Προτύπων (PBMS). Ορίζουμε νέες συναρτήσεις για τη σύγκριση διακριτών και ασαφών προτύπων συστάδων και παρουσιάζουμε πειράματα σε πραγματικές εφαρμογές όπως η κατηγοριοποίηση και ανάκτηση εικόνων με βάση το περιεχόμενο. Μελετάμε την αναπαράσταση και τη διαχείριση των προτύπων συστάδων εστιάζοντας στη σύγκριση διακριτών και ασαφών προτύπων (εφαρμογές για συχνά στοιχειοσύνολα και δέντρα απόφασης έχουν ήδη συζητηθεί στο (Ntoutsi, 2008).

Παρουσιάζουμε ένα πρωτότυπο PBMS, το PatternMiner, ένα ολοκληρωμένο περιβάλλον βασισμένο στην XML για την εξαγωγή, αποθήκευση, ανάκτηση και σύγκριση των προτύπων. Ενσωματώνουμε τις νέες συναρτήσεις σύγκρισης συστάδων και παρουσιάζουμε εφαρμογές για ανάκτηση και κατηγοριοποίηση εικόνων. Ως ενδιαφέρουσα εφαρμογή, χρησιμοποιούμε το πλαίσιο του PBMS για την κατηγοριοποίηση ενός μεγάλου όγκου αστρονομικών δεδομένων, συγκεκριμένα, φασμάτων γαλαξιών.

Επιπλέον, αντιμετωπίζουμε το θέμα της αξιολόγησης των εξαγόμενων προτύπων με τη διαδικασία της εξόρυξης γνώσης προκειμένου να επεκτείνουμε το πλαίσιο του PBMS ώστε να περιέχει ένα ακόμα βήμα της διαδικασίας KDD (Εικόνα 1-1). Για το λόγο αυτό μελετάμε τη χρήση οντολογιών που περιγράφουν τη γνώση της περιοχής για την αξιολόγηση των προτύπων που εξάγονται από μεγάλα σύνολα δεδομένων.

Στην παρούσα διατριβή αντιμετωπίζουμε τη σύγκριση προτύπων συστάδων, χρησιμοποιούμε πρότυπα κανόνων συσχέτισης για τη μελέτη της αξιολόγησης

προτύπων με οντολογίες και δέντρα απόφασης για την κατηγοριοποίηση αστρονομικών δεδομένων για να γίνει κατανοητό το εύρος χρήσης του PBMS.

1.1 Οργάνωση Διατριβής

Η διατριβή αυτή οργανώνεται ως εξής:

Στο κεφάλαιο 2 αντιμετωπίζονται τα θέματα της αναπαράστασης και ανάκτησης των προτύπων. Ορίζουμε ένα μοντέλο για αναπαράσταση προτύπων βασισμένο στις σχετικές έννοιες του έργου PANDA και παρέχουμε ορισμούς και παραδείγματα όλων των εννοιών αυτών για τα πρότυπα που χρησιμοποιούνται στη διατριβή. Η κύρια συνεισφορά του κεφαλαίου 2 είναι να δειχθεί μέσω μίας ποιοτικής αξιολόγησης, ποιο είναι το πιο ταιριαστό μοντέλο αναπαράστασης για μία βάση προτύπων, μεταξύ του σχεσιακού (relational), του αντικειμενο-σχεσιακού (object relational) και του ημιδομημένου (semi-structured) XML μοντέλου. Στη μελέτη μας χρησιμοποιούμε ένα προσαρμοσμένο XML σχήμα που περιγράφει ένα πρότυπο σύμφωνα με την προσέγγιση του έργου PANDA αφού ο σκοπός δεν είναι να κατασκευαστεί μία βάση προτύπων αλλά να συμπεράνουμε για το αν το XML μοντέλο είναι πιο αποτελεσματικό για να χρησιμοποιηθεί σε μία βάση προτύπων.

Στα κεφάλαια 3 και 4 αντιμετωπίζουμε το πρόβλημα της σύγκρισης συστάδων. Στο κεφάλαιο 3 μελετάται η περίπτωση της διακριτής συσταδοποίησης, προτείνεται μία μετρική σύγκρισης για τον αλγόριθμο συσταδοποίησης Expectation-Maximization (Dempster et al., 1977) και παρουσιάζονται δύο διαφορετικές εφαρμογές του PBMS για ανάκτηση εικόνων με βάση το περιεχόμενο (content-based image retrieval- CBIR), που χρησιμοποιούν την προτεινόμενη μετρική. Το κεφάλαιο 4 αναφέρεται στη συσταδοποίηση της διαισθητικά ασαφούς συσταδοποίησης. Η θεωρία των διαισθητικά ασαφών συνόλων παρουσιάζεται αναλυτικά, ενώ προτείνεται μία παραλλαγή του αλγορίθμου Fuzzy C-Means (Bezdek, et al., 1984), που χρησιμοποιεί μία πρωτότυπη μετρική για διαισθητικά ασαφή δεδομένα. Το προτεινόμενο σχήμα αξιολογείται μέσω μίας εφαρμογής κατηγοριοποίησης εικόνων.

Στο Κεφάλαιο 5 παρουσιάζουμε μία εφαρμογή του PBMS για την κατηγοριοποίηση αστρονομικών δεδομένων, μία πραγματική εφαρμογή για το σε εξέλιξη πρόγραμμα GAIA της Ευρωπαϊκής Εταιρείας Διαστήματος (ESA). Στο Κεφάλαιο 6 παρουσιάζουμε ένα πρωτότυπο PBMS που αναπτύξαμε, το

PatternMiner, και μελετάμε την αξιολόγηση των προτύπων με χρήση οντολογιών.

Το PatternMiner χρησιμοποιεί XML έγγραφα συμβατά με το πρότυπο PMML (PMML, 2009), ειδικά τροποποιημένα για να περιλαμβάνουν όλη την απαραίτητη πληροφορία για την υποστήριξη των λειτουργιών του PBMS για τη σύγκριση και την παρακολούθηση των προτύπων.

Η μελέτη για την αξιολόγηση των προτύπων βασίζεται σε σεισμολογικά δεδομένα και την οντολογία SUMO (2009) για τον τομέα της οντολογίας.

Το Κεφάλαιο 7 συνοψίζει τις συνεισφορές αυτής της διατριβής και αναφέρει τα θέματα για μελλοντική έρευνα.

2 Αναπαράσταση και Ανάκτηση Προτύπων

Στο κεφάλαιο αυτό παρουσιάζονται οι βασικές έννοιες των *προτύπων*, ως αποτελέσματα της διαδικασίας εξόρυξης γνώσης και ως σημασιολογικά πλούσια αναπαράσταση των αρχικών δεδομένων γενικότερα. Περιγράφεται το μοντέλο ορισμού προτύπων PANDA και σχετικές εργασίες. Με την ανάδειξη της ανάγκης για ένα σύστημα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ), μελετώνται τρεις διαφορετικές προσεγγίσεις φυσικής αναπαράστασης των προτύπων σε μια βάση προτύπων, η σχεσιακή, η αντικειμενο-σχεσιακή και η ημι-δομημένη (XML). Ο σκοπός της μελέτης είναι να αναδειχθεί το καλύτερο μοντέλο αναπαράστασης (και δομής) για ένα ΣΔΒΠ. Μέσω μιας ποιοτικής αξιολόγησης καταλήγουμε στην ημι-δομημένη (XML) προσέγγιση ως την καταλληλότερη. Ενώ τα πειράματα που διεξάγονται για την αξιολόγηση αυτή γίνονται με χρήση του Oracle DBMS, στη συνέχεια (κεφάλαιο 5) θα χρησιμοποιηθεί το επιλεγμένο μοντέλο αναπαράστασης (ημι-δομημένο) για την κατασκευή ενός ΣΔΒΠ που δεν θα είναι βασισμένο σε εμπορικά προϊόντα.

2.1 Εισαγωγή

Λόγω της πληθώρας εφαρμογών πληροφορικής, οι βάσεις δεδομένων που τις υποστηρίζουν είναι ιδιαίτερα μεγάλες, δυναμικές και με δεδομένα που προέρχονται από πολλές και διαφορετικές πηγές, ενώ πολλά και πολύπλοκα πρότυπα μπορούν να εξαχθούν από αυτές. Προκειμένου να είναι δυνατό για κάποιον να μπορεί να εκμεταλλευτεί την πληροφορία που τα πρότυπα αυτά αναπαριστούν, ένα γενικής χρήσης ΣΔΒΠ για τη διαχείριση (αποθήκευση/ επεξεργασία/ ανάκτηση) των προτύπων είναι απαραίτητο για πολλές επιστημονικές περιοχές και εκτός της εξόρυξης γνώσης (Rizzi et al., 2003). Επιστήμονες από κάθε τομέα έχουν τις δικές τους ιδιαίτερες ανάγκες για τη δημιουργία και τη διαχείριση των προτύπων και η προσέγγιση του ΣΔΒΠ θα

ήταν η λύση στην ανά εφαρμογή εξατομικευμένη αντιμετώπιση των αναγκών αυτών.

```
kMeans

======

Number of iterations: 4

Within cluster sum of squared errors: 80.4216834631833

Cluster centroids:

Cluster 0

       Mean/Mode:      68.0345    10.5172      6.5172      6.8621    10.4483
18.069      19.0345    27.8621   101          296.5862   254.2069   249.2759
389.0345   388.3793   337.3103   302.1034   282.8621   246.9655   223.9655
223.3448   194.4483   161.8621    97.8276    71.7931    57.4138    35.8276    12
2          0.3103     0.1379     0          0

       Std  Devs:      260.2523    28.8551    16.0326    15.3034    22.5556
38.0863     33.7104     49.6529   185.6823   459.7586   288.1144   206.0688
381.6582   346.0339   194.6038   168.1244   171.1642   157.144    169.7554
210.7751   204.1347   184.5747   127.573     104.3912    87.4894    57.857
25.9986    6.2393     1.0387     0.5158     0          0

Cluster 1

       Mean/Mode:   1958.4        161.15        86.1         77.7         91.8
104.95       88.15       98.85       84.8        78.35       79.7        99.25
125.85      114.4        87           68.4        65.55       67.95       79.2
84.95      87.85      81.05      55.85      42.1       34         35.7       29.7
21.3        5.75       0.2        0          0

       Std  Devs:     1507.5583     99.3109     54.0369     47.7715     59.9733
83.4856     75.985     109.9293   101.0714     94.0863   105.6305   128.4617
185.1507   169.1525   100.3615    74.6151    71.832      71.6075    89.7661
118.888    125.2615   122.393      94.5985    76.8313    63.8699    67.1488
56.7396    40.8232    11.7109     0.6156     0          0

=== Clustering stats for training data ===

Clustered Instances

0       29 ( 59%)

1       20 ( 41%)
```

Εικόνα 2-1 το αποτέλεσμα του αλγορίθμου K-means algorithm στο εργαλείο εξόρυξης γνώσης WEKA

Για να αναδειχθεί καλύτερα το πρόβλημα της διαχείρισης των προτύπων και η ανάγκη για ένα ΣΔΒΠ, υποθέτουμε ότι έχουμε ένα μεγάλο σύνολο δεδομένων στα οποία πρέπει να εφαρμοστεί ο δημοφιλής αλγόριθμος συσταδοποίησης K-Means (MacQueen, 1967). Με τη χρήση ενός συνόλου από παραμέτρους ορισμένους από τους χρήστες, κάποιο εργαλείο εξόρυξης γνώσης θα καταλήξει σε *k* το πλήθος συστάδες. Το αποτέλεσμα παρουσιάζεται σε μία μορφή κειμένου που περιγράφει γενικά το κέντρο κάθε συστάδας και την κατανομή των

δεδομένων σε κάθε μία από αυτές. Ανάλογα το εργαλείο που χρησιμοποιείται για την εκτέλεση του αλγορίθμου, η μορφή των αποτελεσμάτων θα είναι διαφορετική. Ακόμα και αν υπάρχει η επιλογή να αποθηκευτεί το αποτέλεσμα, ο χρήστης δεν μπορεί να αναζητήσει προηγούμενες συσταδοποιήσεις κάνοντας χρήση κριτηρίων όπως οι παράμετροι του αλγορίθμου ή το σύνολο των δεδομένων που χρησιμοποιήθηκαν. Επιπλέον, ο χρήστης δεν μπορεί να συνδυάσει ή να συγκρίνει διαφορετικές συσταδοποιήσεις που προέρχονται από το ίδιο σύνολο αρχικών δεδομένων (ως αποτέλεσμα πχ. διαφορετικών ρυθμίσεων του αλγορίθμου).

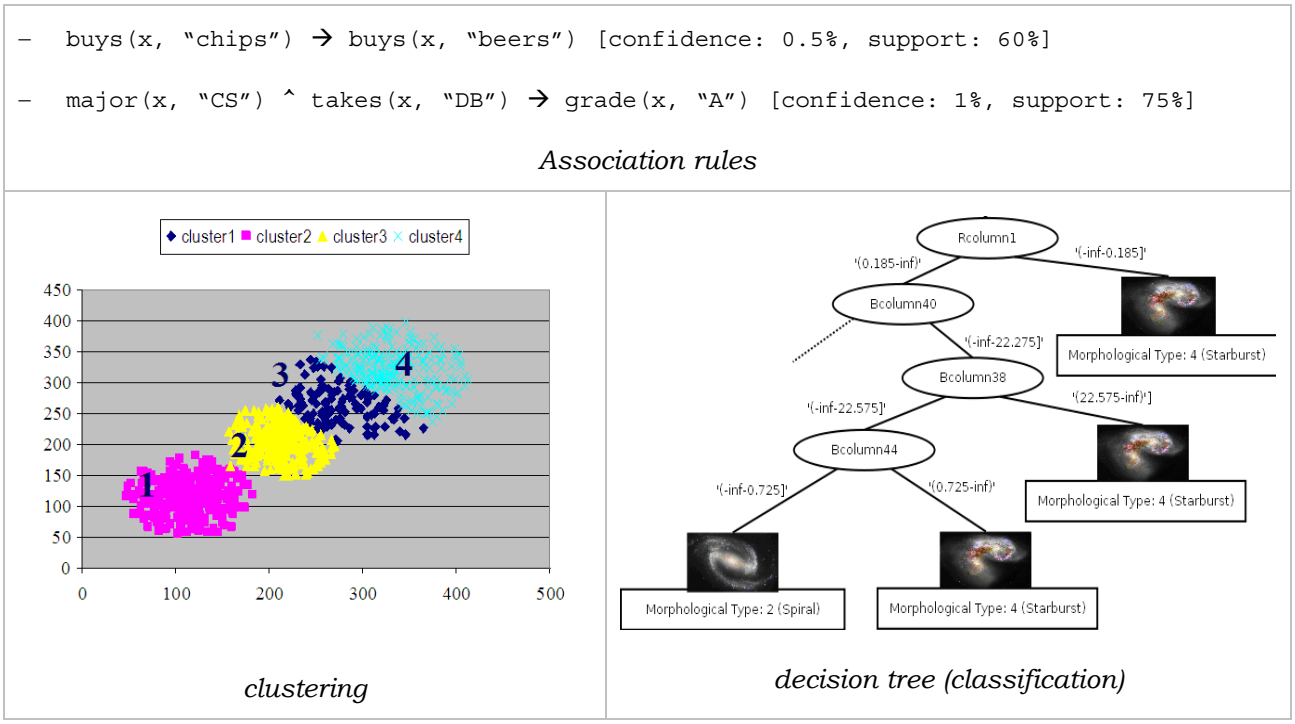
```
=== Classifier model (full training set) ===

J48 pruned tree

------------------
Bcolumn 39 = '(-inf-23.995]'
|  Rcolumn 10 = '(-inf-9.675]'
|   |   Rcolumn 1 = '(-inf-0.185]': 4 (12583.0)
|   |   Rcolumn 1 = '(0.185-inf)'
|   |   |   Bcolumn 40 = '(-inf-22.275]'
|   |   |   |   Bcolumn 38 = '(-inf-22.575]'
|   |   |   |   |   Bcolumn 44 = '(-inf-0.725]': 2 (2.0)
|   |   |   |   |   Bcolumn 44 = '(0.725-inf)': 4 (6.0)
|   |   |   |   Bcolumn 38 = '(22.575-inf)': 4 (439.0/1.0)


[...]


=== Summary ===

Correctly Classified Instances  26625    92.1759 %
Incorrectly Classified Instances  2260    7.8241 %
Kappa statistic       0.8712
Mean absolute error      0.0586
Root mean squared error     0.172
Relative absolute error     18.9389 %
Root relative squared error    43.7147 %
Total Number of Instances   28885

=== Confusion Matrix ===

   a          b         c          d     <-- classified as
 1231   1585      0          0   |   a = 1
 125    10377   51         16   |   b = 2
   0      194     1304        2   |   c = 3
   0       65       222   13713 |   d = 4
```

Εικόνα 2-2 Το αποτέλεσμα του αλγορίθμου κατηγοριοποίησης J48 στο εργαλείο εξόρυξης γνώσης WEKA

Η Εικόνα 2-1 και η Εικόνα 2-2 παρουσιάζουν το αποτέλεσμα του αλγορίθμου συσταδοποίησης K-means και του αλγορίθμου κατηγοριοποίησης J48

αντίστοιχα, στο γνωστό και ευρέως χρησιμοποιούμενο εργαλείο εξόρυξης γνώσης WEKA (Witten and Frank, 2005). Παρατηρείται η πολύ συγκεκριμένη μορφή που παρουσιάζεται το αποτέλεσμα.

Επιπλέον, οι διαδικασίες εξόρυξης γνώσης που μπορούν να εφαρμοστούν σε ένα σύνολο δεδομένων, περιλαμβάνουν όχι μόνο συσταδοποίηση αλλά, στις περισσότερες περιπτώσεις, κατηγοριοποίηση και ανακάλυψη κανόνων συσχέτισης των δεδομένων.



Εικόνα 2-3 Παραδείγματα των αποτελεσμάτων τριών περισσότερο κοινών διαδικασιών εξόρυξης γνώσης

Παραδείγματα των αποτελεσμάτων αυτών των διαδικασιών παρουσιάζονται στην Εικόνα 2-3. Συγκεκριμένα, παρουσιάζονται πρότυπα των τριών πιο κοινών διαδικασιών εξόρυξης γνώσης· κανόνες συσχέτισης, συστάδες και δέντρα απόφασης. Στο παράδειγμα των κανόνων συσχέτισης, γίνεται φανερή η διμερής δομή τους (head και body) καθώς φαίνονται και τα μέτρα ποιότητας των κανόνων (confidence και support). Παρουσιάζεται επίσης το αποτέλεσμα μιας συσταδοποίησης, με τέσσερις συστάδες (ομαδοποίηση των δεδομένων βάση πυκνότητας/ εγγύτητας) και ένα δέντρο απόφασης που κατηγοριοποιεί αστρονομικά δεδομένα επιλέγοντας τις κατάλληλες ιδιότητες/γνωρίσματά τους. Κάθε μία προσέγγιση παρέχει σημαντική πληροφόρηση για τα δεδομένα που χρησιμοποιήθηκαν, με διαφορετικό όμως τρόπο και με χρήση διαφορετικών

αλγορίθμων και παραμέτρων. Το διαφορετικό αποτέλεσμα και αναπαράσταση των διαδικασιών αυτών κάνουν περισσότερο πολύπλοκο το πρόβλημα της κοινής διαχείρισης των προτύπων.

2.2 Σχετικές Προσεγγίσεις

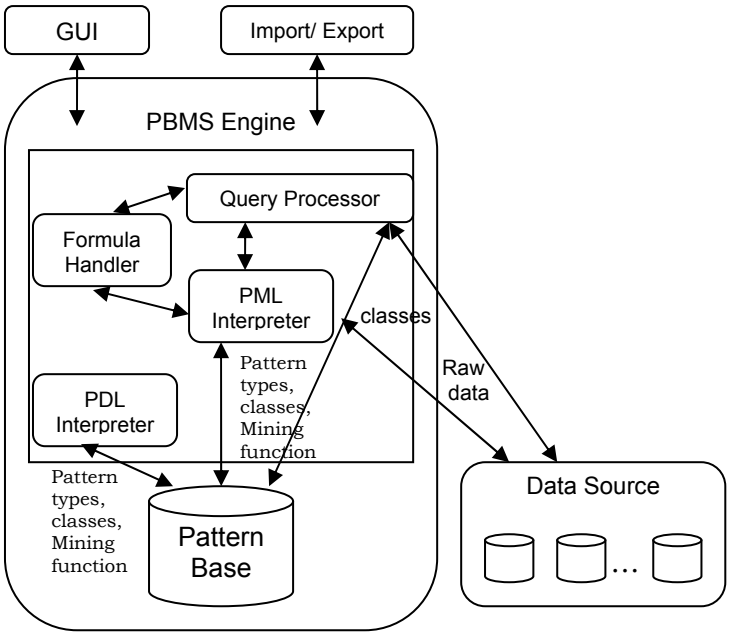
Τα σύγχρονα συστήματα βάσεων δεδομένων δεν υποστηρίζουν αποτελεσματικά τη διαχείριση των προτύπων (παρά μόνο την αποθήκευση των βασικότερων προτύπων, με απλοϊκό όμως τρόπο που δεν επιτρέπει ιδιαίτερες διαδικασίες επεξεργασίας) που εξάγονται από δεδομένα με τη χρήση εργαλείων εξόρυξης γνώσης. Η περιοχή της αναπαράστασης και διαχείρισης προτύπων είναι σχετικά πρόσφατη και μόνο λίγες σχετικές εργασίες έχουν γίνει. Η PMML (PMML, 2009), η SQL/MM (SQL/MM, 2001), το CWM (CWM, 2007), το JDMAPI (JDMAPI, 2007) και η PQL (PQL, 2007) είναι πρότυπα και συστήματα που έχουν αναπτυχθεί για την αποθήκευση προτύπων εξόρυξης γνώσης και στατιστικής.

Η PMML (Predictive Model Markup Language) που έχει προταθεί από το data mining Group (DMG) είναι η πιο δημοφιλής προσέγγιση. Με τη χρήση XML εγγράφων παρέχει ένα γρήγορο και εύκολο τρόπο στις διάφορες εφαρμογές για να ορίζονται μοντέλα και να διαμοιράζονται μεταξύ συμβατών με την PMML εφαρμογών. Στην PMML ορίζονται κάποια συγκεκριμένα πρότυπα εξόρυξης γνώσης (όπως είναι τα δέντρα απόφασης, οι κανόνες συσχέτισης, τα νευρωνικά δίκτυα κλπ) αλλά δεν υποστηρίζονται τύποι προτύπων ορισμένοι από τους χρήστες. Στην έκδοση 3.2 της PMML παρέχονται ορισμοί για περισσότερα πρότυπα όπως και κάποιες συναρτήσεις για την επεξεργασία των δεδομένων (PMML, 2009). Μία επισκόπηση των προσεγγίσεων αυτών σε σχέση με τη διαχείριση προτύπων γίνεται στο (Catania & Maddalena, 2006). Οι παραπάνω προσεγγίσεις αντιμετωπίζουν τα πιο κοινά πρότυπα εξόρυξης γνώσης και δεν έχουν δυνατότητες διαχείρισής τους. Επικεντρώνονται περισσότερο στον ορισμό στατιστικών μοντέλων και μοντέλων εξόρυξης γνώσης και την ανταλλαγή των προτύπων με συγκεκριμένα χαρακτηριστικά μεταξύ εφαρμογών, παρά στη δημιουργία ενός γενικού συστήματος για την αναπαράσταση και τη διαχείριση προτύπων διαφορετικών τύπων. Δυνατότητες όπως αποθήκευση και ανάκτηση ή ακόμα και αντιστοίχιση προτύπων με τα αρχικά δεδομένα, δεν παρέχονται από τις προσεγγίσεις αυτές.

Κατά τα τελευταία χρόνια δύο ερευνητικά έργα, το CINQ (CINQ, 2005) και το PANDA (PANDA, 2005), όρισαν το πρόβλημα της αποθήκευσης και διαχείρισης των προτύπων και προτάθηκαν κάποιες λύσεις. Το CINQ στοχεύει στη μελέτη και ανάπτυξη τεχνικών ανάκτησης για τις επαγωγικές βάσεις δεδομένων (inductive databases), δηλαδή τις βάσεις δεδομένων που αποθηκεύουν τα αρχικά δεδομένα μαζί με τα πρότυπα που εξάγονται από αυτά. Από την άλλη μεριά το PANDA στοχεύει στον ορισμό και σχεδίαση ενός ΣΔΒΠ για την αποτελεσματική αναπαράσταση και διαχείριση διαφόρων τύπων προτύπων που παράγονται από διαφορετικά πεδία εφαρμογών (και όχι αποκλειστικά από την εξόρυξη γνώσης). Τα πρότυπα θα διαχειρίζονται σε ένα ΣΔΒΠ όπως τα αρχικά δεδομένα αποθηκεύονται και διαχειρίζονται από ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ – DBMS). Στο ΣΔΒΠ, διαφορετικοί τύποι προτύπων θα διαχειρίζονται αποτελεσματικά (γενικότητα – generality) και νέοι τύποι προτύπων θα μπορούν εύκολα να ενσωματωθούν (επεκτασιμότητα – extensibility). Σημαντική απόφαση για ένα ΣΔΒΠ είναι αν αυτό δημιουργηθεί εξ' αρχής ή δημιουργηθεί ως επέκταση ενός ήδη υπάρχοντος ΣΔΒΔ σαν ένα επιπλέον επίπεδο.

Ένα πρωτότυπο PBMS βασισμένο επίσης στο μοντέλο του PANDA, που ονομάζεται PSYCHO (Catania, Maddalena & Mazza, 2005) παρουσιάστηκε πρόσφατα. Το PSYCHO διαχειρίζεται διάφορους τύπους προτύπων με ενιαίο και ομοιόμορφο τρόπο και έχει αναπτυχθεί με συγκεκριμένα εργαλεία πάνω από το object-relational Oracle DBMS.

Η αρχιτεκτονική του PSYCHO παρουσιάζεται στην Εικόνα 2-4. Το σύστημα αποτελείται από τρία διαφορετικά και διακριτά επίπεδα. Το φυσικό επίπεδο που περιλαμβάνει τη βάση προτύπων (*Pattern Base)* που αποθηκεύονται τα πρότυπα και τις πηγές δεδομένων (*Data Source*) που αποθηκεύονται τα αρχικά δεδομένα από τα οποία προέρχονται τα εξαγόμενα πρότυπα. Το μεσαίο επίπεδο, το *PBMS Engine*, υποστηρίζει λειτουργίες για την αποθήκευση και ανάκτηση των προτύπων. Το εξωτερικό επίπεδο αντιστοιχεί σε ένα σύνολο από διεπαφών με το χρήστη (ένα shell και ένα γραφικό – GUI) από τα οποία ο χρήστης μπορεί να στείλει εντολές στην engine και να εισάγει/ εξάγει δεδομένα σε άλλα formats.



Εικόνα 2-4 Η αρχιτεκτονική του PSYCHO

Στο παρόν κεφάλαιο ακολουθούμε χρησιμοποιούμε ένα υπάρχον ΣΔΒΔ για να μελετηθεί το πιο είναι το καλύτερο μοντέλο αναπαράστασης για τα πρότυπα σε ένα ΣΔΒΠ. Για το σκοπό αυτό, εξετάζουμε τρεις γνωστές προσεγγίσεις που ακολουθούνται στα ΣΔΒΔ, τη σχεσιακή, την αντικειμενο-σχεσιακή και την ημι-δομημένη (XML).

2.3 Αναπαράσταση Προτύπων σε ένα Σύστημα Διαχείρισης Βάσεων Προτύπων

Η έννοια του προτύπου είναι βασική για το ΣΔΒΠ. Υιοθετώντας την ορολογία που προτάθηκε στο ερευνητικό έργο PANDA (Theodoridis et al, 2003), ένα *πρότυπο* είναι μία συμπαγής και πλούσια σε σημασιολογία αναπαράσταση των αρχικών δεδομένων. Μία *βάση προτύπων* είναι μία συλλογή από αποθηκευμένα πρότυπα. Ένα ΣΔΒΠ είναι ένα σύστημα για τη διαχείριση των προτύπων που προέρχονται από τα αρχικά δεδομένα (raw data) και είναι οργανωμένα σε βάσεις προτύπων προκειμένου να υποστηρίζεται αποτελεσματικά το ταίριασμα μεταξύ των προτύπων και να εκμεταλλεύονται οι σχετικές με τα πρότυπα λειτουργίες ώστε να παράγεται σημαντική πληροφορία. Σε ένα ΣΔΒΠ τα πρότυπα έχουν τον ρόλο που έχουν τα αρχικά δεδομένα σε ένα ΣΔΒΔ.

Προκειμένου να γίνεται αποτελεσματική διαχείριση των προτύπων, ένα ΣΔΒΠ πρέπει να πληροί τις παρακάτω προδιαγραφές (Theodoridis et al, 2003):

- *Πολυπλοκότητα Υλοποίησης*: Η βάση προτύπων θα πρέπει να υλοποιείται εύκολα χωρίς να απαιτεί τη χρήση πολύπλοκων τύπων.
- *Υλοποίηση Περιορισμών:* Στο ΣΔΒΠ θα πρέπει να υλοποιούνται οι περιορισμοί που ορίζονται από το λογικό μοντέλο του προτύπου, καθώς επίσης να πραγματοποιείται έλεγχος εγκυρότητας των προτύπων.
- *Αξιοποίηση των ιδιαίτερων χαρακτηριστικών των προτύπων:* Στο ΣΔΒΠ θα πρέπει να λαμβάνονται υπόψη τα ιδιαίτερα γνωρίσματα των προτύπων ώστε να βελτιώνονται λειτουργίες όπως η ευρετηριοποίηση και η ανάκτηση.
- *Αποτελεσματικότητα Ανάκτησης*: Στο ΣΔΒΠ θα πρέπει να επιτρέπεται η απλή αλλά και αποτελεσματική κατασκευή ερωτημάτων. Οι χρήστες θα πρέπει να μπορούν να δημιουργήσουν μικρές και εύκολες επερωτήσεις για κάθε στοιχείο των προτύπων.
- *Εγκυρότητα Προτύπων*: Στο ΣΔΒΠ θα πρέπει να πραγματοποιείται έλεγχος εγκυρότητας των προτύπων σύμφωνα με τον ορισμό του τύπου τους και να απορρίπτονται τα πρότυπα με λανθασμένη δομή.
- *Επεκτασιμότητα*: Το ΣΔΒΠ θα πρέπει να μπορεί να επεκταθεί εύκολα και να συμπεριλάβει νέους τύπους προτύπων που προέρχονται από νέες και πρωτοποριακές εφαρμογές.
- *Γενικότητα*: Το ΣΔΒΠ θα πρέπει να μπορεί να διαχειριστεί διαφορετικούς τύπους προτύπων που προέρχονται από διαφορετικές εφαρμογές.
- *Επαναχρησιμοποίηση:* Οι διάφορες συνιστώσες του ΣΔΒΠ θα πρέπει να μπορούν να επαναχρησιμοποιηθούν μελλοντικά ώστε να αποφεύγεται ο ορισμός ήδη υπαρχόντων στοιχείων.

Ένα λογικό μοντέλο για το ΣΔΒΠ αποτελείται από τρεις βασικές οντότητες: τύπος προτύπου, πρότυπο και κλάση, που ορίζονται ως εξής(Rizzi et al. 2003):

Ορισμός 2-1. (Τύπος Προτύπου): Ένας τύπος προτύπου (pattern type) αποτελείται από πέντε στοιχεία $pt = (n, ss, ds, ms, f)$ όπου n είναι το όνομα του τύπου, ss η *δομή* (structure), ds η *πηγή* (source), ms το *μέτρο ποιότητας* (measure) και f η *έκφραση* (expression). Η δομή χαρακτηρίζει το πρότυπο στο χώρο των προτύπων, περιγράφει απλά τη δομή του. Το μέτρο ποιότητας

χαρακτηρίζει το πρότυπο μετρώντας το κατά πόσο η απεικόνιση των δεδομένων αντιπροσωπεύει την πραγματική φύση των αρχικών δεδομένων. Η πηγή περιγράφει τα αρχικά δεδομένα με τα οποία σχετίζεται το πρότυπο ενώ η έκφραση περιγράφει προσεγγιστικά την αντιστοίχιση (mapping) μεταξύ του προτύπου και των αρχικών δεδομένων.

■

Ένα παράδειγμα τύπου προτύπου κανόνα συσχέτισης είναι:

```
n: AssociationRule
ss: TUPLE(head: SET(STRING), body: SET(STRING))
ds: BAG(transaction: SET(STRING))
ms: TUPLE(confidence: REAL, support: REAL)
f: head U body ⊆ transaction
```

Ορισμός 2-2. (Πρότυπο): Ένα πρότυπο (pattern) *p*, αποτελείται από πέντε στοιχεία *p* = (*pid*, *s*, *d*, *m*, *e)* είναι ένα στιγμιότυπο ενός τύπου προτύπου *pt*, και έχει τις αντίστοιχες τιμές για κάθε στοιχείο του.

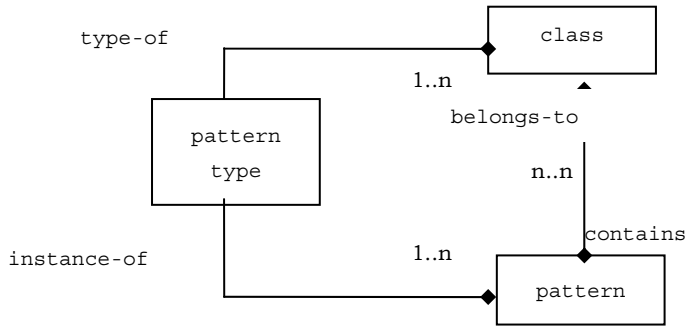
Ένα παράδειγμα ενός προτύπου κανόνα συσχέτισης, στιγμιότυπο ενός τύπου προτύπου κανόνα συσχέτισης όπως ορίστηκε παραπάνω, είναι το παρακάτω:

```
pid: 413
s: (head={'Boots'}, body={'Socks', 'Hat'})
d: 'SELECT SETOF(article) AS transaction FROM sales GROUP BY transactionId'
m: (confidence=0.75, support=0.55)
e: {transaction: {'Boots', 'Socks', 'Hat'} ⊆ transaction}
```

Ορισμός 2-3. (Κλάση): Μία κλάση (class) *c*, ορισμένη σε ένα τύπο προτύπου *pt*, ορίζεται σαν μία τριπλέτα *c* = (*cid*, *pt*, *pc*) όπου *cid* είναι το μοναδικό αναγνωριστικό της κλάσης, *pt* είναι ο τύπος προτύπου και *pc* είναι ένα σύνολο από πρότυπα του τύπου *pt*.

■

Ένα παράδειγμα κλάσης ορίζεται για ένα συγκεκριμένο τύπο προτύπου και περιέχει μόνο πρότυπα του τύπου αυτού. Κάθε πρότυπο πρέπει να ανήκει σε τουλάχιστον μία κλάση. Η σχέσεις μεταξύ των τριών οντοτήτων ενός ΣΔΒΠ, δηλαδή τύπος προτύπου, πρότυπο και κλάση, παρουσιάζονται στην Εικόνα 2-5:



Εικόνα 2-5 Σχέση μεταξύ τύπου προτύπου, προτύπου και κλάσης

Τα πρότυπα στο πλαίσιο του PANDA μπορεί να είναι *απλά* ή *σύνθετα*. Τα απλά πρότυπα προέρχονται από τα αρχικά δεδομένα με τη χρήση τεχνικών εξόρυξης γνώσης (πχ. συστάδες από δεδομένα), ενώ τα σύνθετα πρότυπα προέρχονται από απλά πρότυπα (πχ. συσταδοποίηση σε ένα σύνολο συστάδων – συστάδες συστάδων). Σε ένα σύνθετο πρότυπο η δομή του περιγράφει απλά πρότυπα και το μέτρο ποιότητας είναι είτε κενό είτε είναι συναθροιστικό μέτρο που εξαρτάται από τα μέτρα των απλών προτύπων. Στην επόμενη ενότητα περιγράφονται παραδείγματα από απλά και σύνθετα πρότυπα.

Έχοντας ορίσει τις βασικές έννοιες ενός ΣΔΒΠ και τη δομή και ιδιότητες του προτύπου, αμέσως μετά τα παραδείγματα προτύπων θα εξετάσουμε τις διαφορετικές επιλογές για τη φυσική αναπαράσταση των προτύπων σε μία βάση προτύπων.

2.3.1 Παραδείγματα συνηθισμένων τύπων προτύπων

Στην ενότητα αυτή παρουσιάζονται τρεις βασικοί τύποι προτύπων που θα μας απασχολήσουν στην παρούσα διατριβή, η βασική αναπαράστασή τους σύμφωνα με το μοντέλο του PANDA και παραδείγματά τους.

Συχνά στοιχειοσύνολα – Κανόνες Συσχέτισης (Itemsets – Association Rules)

Οι Κανόνες Συσχέτισης (Association Rules) παρουσιάζουν σχέσεις μεταξύ δεδομένων και βασίζονται στην εξόρυξη συχνών στοιχειοσυνόλων (ΕΣΣ - Frequent Itemset Mining, FIM) (Agrawal et al, 1993). Η ΕΣΣ έχει μεγάλη εφαρμογή σε εφαρμογές καταστημάτων όπου κανόνες συσχέτισης βασισμένοι σε συχνά στοιχειοσύνοολα εξάγονται για να υποστηρίξουν τη διαχείριση του καταστήματος, τη διαφήμιση κτλ.

Τα στοιχειοσύνολα (itemsets) μπορούν να περιγραφούν σαν πρότυπα με τη χρήση της αναπαράστασης του μοντέλου PANDA ως εξής:

Itemset =
(*SS*: {String},
MS: sup: (Real))

Ένα σύνολο στοιχεισυνόλων μπορεί να εκφραστεί σαν ένα σύνθετο πρότυπο ως εξής:

SetOfItemsets =
(*SS*: {*Itemset*},
MS: ⊥)

Τα πρότυπα Κανόνων Συσχέτισης έχουν δύο μέρη, την κεφαλή (head) και το σώμα (body) του κανόνα, τα οποία είναι σύνολα αντικειμένων δηλαδή στοιχειοσύνολα, ενώ χαρακτηρίζονται από τα μέτρα ποιότητας υποστήριξη (support) και εμπιστοσύνη (confidence). Η αναπαράστασή τους σύμφωνα με το μοντέλο PANDA, είναι η ακόλουθη:

AssociationRule =
(*SS*: (head: Itemset, body: Itemset)
MS: (sup: (Real), conf: (Real)))

Ας σημειωθεί ότι ορίζουμε μόνο τα στοιχεία της δομής και του μέτρου ποιότητας καθώς τα άλλα τρία στοιχεία (το όνομα, τα δεδομένα και η συνάρτηση αντιστοίχησης) είναι δευτερεύουσας σημασίας και εξαρτώνται από την εφαρμογή.

Συστάδες (Clusters)

Οι συστάδες είναι πολύ συχνοί τύποι προτύπων, καθώς οι αλγόριθμοι συσταδοποίησης χρησιμοποιούνται σε πολλές και διάφορες εφαρμογές. Συνήθως, οι συστάδες είναι είτε σφαιρικές είτε βασίζονται στην πυκνότητα, ανάλογα με τον αλγόριθμο συσταδοποίησης.

Μία σφαιρική (πχ. βάσει Ευκλείδειας απόστασης) συστάδα, όπως αυτές που παράγονται από τον αλγόριθμο k-means, μπορεί να μοντελοποιηθεί χρησιμοποιώντας το κέντρο και την ακτίνα τους, τα οποία αποτελούν και το στοιχείο της δομής (structure) της συστάδας. Για το στοιχείο του μέτρου ποιότητας, μπορεί να χρησιμοποιηθεί η υποστήριξη (support), δηλαδή το ποσοστό των αντικειμένων που ανήκουν στη συστάδα, η μέση απόσταση των

στοιχείων της συστάδας (intra-cluster distance) ή η μέση απόσταση μεταξύ αυτής και των άλλων συστάδων (inter-cluster distance):

EuclideanCluster =
(*SS* : (center: (Real), radius: (Real)),
MS: sup: (Real))

Ένα παράδειγμα στιγμιοτύπου αυτού του προτύπου είναι το παρακάτω:

Cluster1 =
(*SS* : (center = 0.1, radius = 0.77),
MS: sup = 0.15)

Οι συστάδες που βασίζονται στην πυκνότητα (density-based clusters) παράγονται από αλγόριθμους όπως ο αλγόριθμος Expectation-Maximization (Dempster et al., 1977), που χρησιμοποιεί κατανομές για να ομαδοποιήσει τα δεδομένα σε συστάδες. Ένα πρότυπο βασισμένο στην πυκνότητα μπορεί να μοντελοποιηθεί με τη χρήση του μέσου και της τυπικής απόκλισης της κατανομής σαν στοιχείο δομής του προτύπου και με την υποστήριξη (το ποσοστό των δεδομένων που ανήκουν στη συστάδα) σαν το στοιχείο μέτρου ποιότητας:

DensityBasedCluster =
(*SS* : (mean: (Real), stdDev: (Real)),
MS: sup: (Real))

Ένα παράδειγμα στιγμιοτύπου αυτού του προτύπου είναι το εξής:

DensCluster =
(*SS* : (mean = 15.5, stdDev = 3.6),
MS: sup = 0.33)

Σημειώνεται ότι στις περισσότερες περιπτώσεις τα δεδομένα είναι πολλών διαστάσεων οπότε και τα παραπάνω στοιχεία αναπαρίστανται σαν πολυδιάστατα διανύσματα (vectors).

Δέντρα Απόφασης (Decision Trees)

Τα δέντρα απόφασης είναι ιδιαίτερα δημοφιλή σαν μέθοδο κατηγοριοποίησης δεδομένων (classification) και παρέχουν μία εύκολα αντιληπτή αναπαράσταση της κατηγοριοποίησης.

Οι κόμβοι-φύλλα του δέντρου είναι οι κλάσεις στις οποίες κατηγοριοποιούνται τα δεδομένα, ενώ τα μονοπάτια του δέντρου είναι περιορισμοί που «σπρώχνουν» τα δεδομένα προς τα φύλλα. Τα δέντρα απόφασης μπορούν να περιγραφούν με την αναπαράσταση του μοντέλου PANDA με την περιγραφή των περιορισμών αυτών για όλα τα γνωρίσματα (attributes):

Path =
(SS : [(ValueFrom: Real, ValueTo: Real)]$_1^N$,
MS: sup: Real)

DecisionTree =
(SS : {Path},
MS: $\perp$)

Ένα παράδειγμα προτύπου δέντρου απόφασης (σε ένα σύνολο δεδομένων με τρία γνωρίσματα) είναι το παρακάτω:

aPath =
(SS : [(0, 8),(4, 6),(1, 2)],
MS: sup: 0.17)

aDecisionTree =
(SS : {Path},
MS: $\perp$)

Στην παρούσα διατριβή θα ασχοληθούμε με πρότυπα κανόνων συσχέτισης, συστάδων βασισμένων στην πυκνότητα και δέντρων απόφασης σε διάφορες εφαρμογές.

2.4 Φυσική Αναπαράσταση σε μία Βάση Προτύπων

Για την αναπαράσταση και αποθήκευση των προτύπων σε μία βάση προτύπων, εξετάζουμε τρεις κλασικές προσεγγίσεις από τη θεωρία Βάσεων Δεδομένων: το σχεσιακό μοντέλο, το αντικειμενο-σχεσιακό και το ημι-δομημένο (XML) με τη χρήση οντοτήτων όπως παρουσιάστηκαν στην προηγούμενη ενότητα.

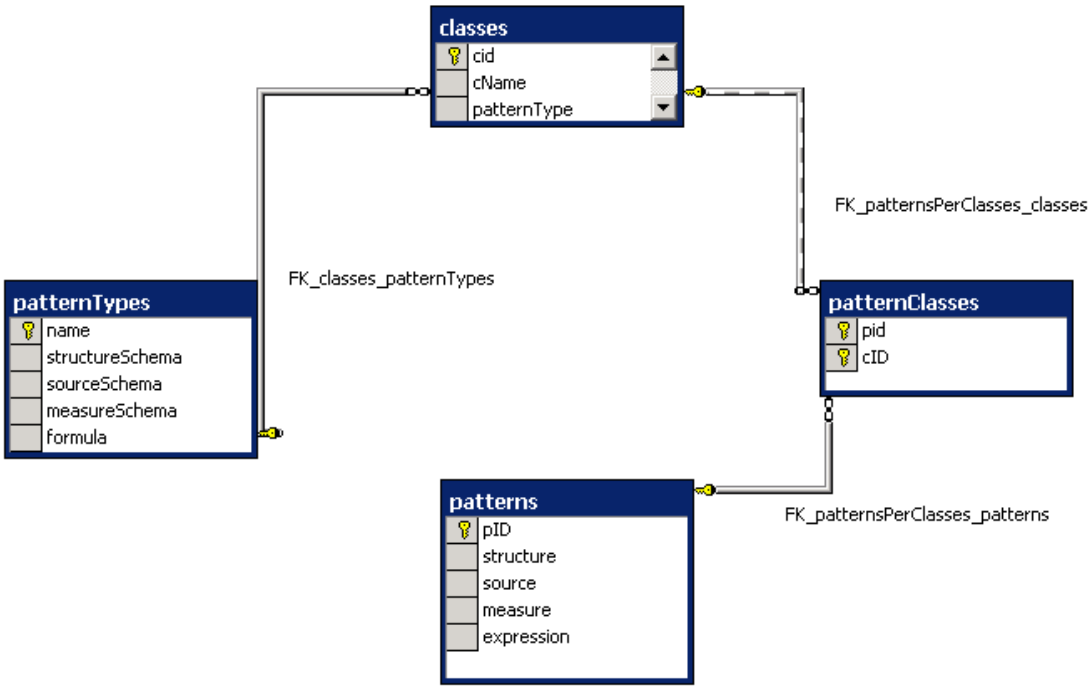
Στη συνέχεια παρουσιάζουμε κάθε μία προσέγγιση (μοντέλο) και δίνουμε κάποιες αντιπροσωπευτικές επερωτήσεις που επισημαίνουν τα πλεονεκτήματα και μειονεκτήματα κάθε μίας. Αυτή η σύγκριση στοχεύει στην εξέταση της πιθανής χρήσης του λογικού μοντέλου των προτύπων σε τρέχουσες τεχνολογίες Συστημάτων Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) και βασίζεται σε ποιοτικά

και όχι ποσοτικά κριτήρια. Βασικός στόχος είναι να εξεταστεί αν ένα ΣΔΒΠ μπορεί να δημιουργηθεί βασισμένο σε κάποιο από αυτά τα τρία μοντέλα και ποιο από αυτά είναι το καταλληλότερο στην υποστήριξη των ιδιαίτερων χαρακτηριστικών των προτύπων.

2.4.1 Σχεσιακή Προσέγγιση

Βασικός μας στόχος κατά τη σχεδίαση και υλοποίηση της βάσης προτύπων ήταν η ικανοποίηση των τριών βασικών προϋποθέσεων του λογικού μοντέλου: γενικότητα, επεκτασιμότητα και αξιοποίηση των ιδιαίτερων χαρακτηριστικών των προτύπων (Theodoridis et al, 2003). Το σχεσιακό σχήμα απεικονίζεται στην Εικόνα 2-6.

Διάφοροι τύποι προτύπων αποθηκεύονται στον πίνακα *patternTypes*, τα πρότυπα αποθηκεύονται στον πίνακα *patterns* και οι κλάσεις των προτύπων αποθηκεύονται στον πίνακα *classes*. Ο πίνακας *patternClasses* σχετίζει τα πρότυπα με τις κλάσεις (μία κλάση περιέχει ένα ή περισσότερα πρότυπα του ίδιου τύπου και κάθε πρότυπο ανήκει τουλάχιστον σε μία κλάση).



Εικόνα 2-6 Το σχεσιακό σχήμα μίας βάσης προτύπων

Παρακάτω παρουσιάζονται κάποιες αντιπροσωπευτικές επερωτήσεις. Οι επερωτήσεις περιγράφονται πρώτα σε φυσική γλώσσα και κατόπιν σε σύνταξη τύπου SQL.

RQ1) *Ανάκτηση της δομής των κανόνων συσχέτισης που ανήκουν στην κλάση Association_Rule_1.*

```
select patterns.structure from classes
inner join patternclasses on classes.cid = patternclasses.cid
inner join patterns on patternclasses.pid = patterns.pid
where (classes.cname='Association_Rule_1');
```

RQ2) *Ανάκτηση των μερών «head» και «body» της δομής των προτύπων που ανήκουν στην κλάση Association_Rule_1.*

```
Select Substr(structure,1,instr(structure,'body')-2) as head,
Substr(structure,instr(structure,'body')) as body from classes
inner join patternclasses on classes.cid = patternclasses.cid
inner join patterns on patternclasses.pid = patterns.pid
where (classesr.cname='Association_Rule_1');
```

RQ3) *Ανάκτηση του μέτρου ποιότητας εμπιστοσύνης (confidence) από όλους τους κανόνες συσχέτισης.*

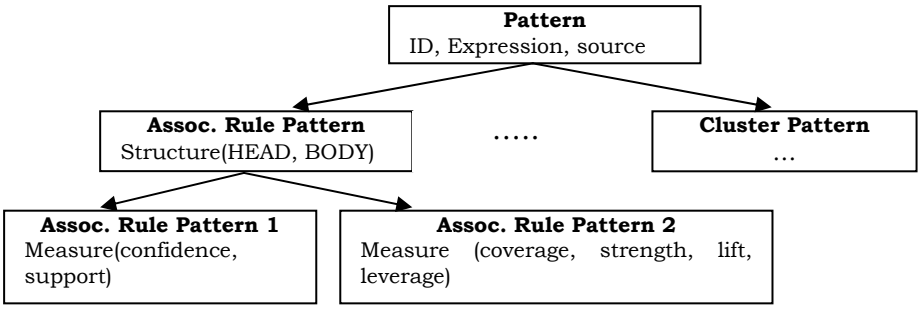
```
Select Substr(measure,1,instr(measure,'confidence')-2) as confidence, from
patterns;
```

Η Σχεσιακή προσέγγιση χαρακτηρίζεται από απλότητα και ευκολία υλοποίησης. Εντούτοις, είναι φανερό ότι έχει πολλά μειονεκτήματα που προέρχονται από το γεγονός ότι δεν λαμβάνεται υπόψη η δομή των στοιχείων του προτύπου (δομή, μέτρο ποιότητας κλπ) αλλά τα πρότυπα αντιμετωπίζονται σαν απλό κείμενο (texts/ strings). Αυτό κάνει την σύνταξη των επερωτήσεων μία ιδιαίτερα πολύπλοκη, χρονοβόρα και αναποτελεσματική διαδικασία.

2.4.2 Αντικειμενο-σχεσιακή Προσέγγιση

Το αντικειμενο-σχεσιακό μοντέλο (Stonebraker, 1997; Stonebraker et al., 1999) αντιμετωπίζει τα βασικά μειονεκτήματα του σχεσιακού, ορίζοντας διαφορετικά αντικείμενα και ιδιότητες για κάθε στοιχείο των προτύπων και χρησιμοποιώντας την κληρονομικότητα. Με τον τρόπο αυτό μειώνεται η πολυπλοκότητα αφού και η ανάκτηση είναι πιο απλή.

Η βασική ιδέα του αντικειμενο-σχεσιακού μοντέλου (ενός μέρους αυτού) αναπαρίσταται στην Εικόνα 2-7. Στη ρίζα του μοντέλου βρίσκεται η οντότητα «*Πρότυπο*» που περιέχει βασική πληροφορία για το πρότυπο, όπως το αναγνωριστικό, την εξίσωση αντιστοίχισης (formula) και την πηγή των δεδομένων. Στο επόμενο επίπεδο της δενδρικής δομής το *Πρότυπο* εξειδικεύεται με βάση τον τύπο προτύπου στον οποίο ανήκει, για παράδειγμα σε πρότυπα Κανόνων Συσχέτισης, Συστάδων κλπ. Αυτές οι οντότητες διαφέρουν στα μέρη της δομής και μέτρου ποιότητας αλλά επιπλέον έχουν κάποια κοινά γνωρίσματα, αυτά που κληρονομούνται από την οντότητα *Πρότυπο*. Για παράδειγμα, το αντικείμενο Κανόνας Συσχέτισης περιέχει κάθε γνώρισμα από το αντικείμενο *Πρότυπο* και επιπλέον περιέχει το γνώρισμα *Δομή* που αποτελείται από την *head* και το *body* του κανόνα. Αυτό το αντικείμενο μπορεί να εξειδικευτεί περαιτέρω με βάση το μέρος του μέτρου ποιότητας. Όπως φαίνεται στην Εικόνα 2-7, στο αντικείμενο *Association Rule Pattern 1* το μέρος του μέτρου ποιότητας αποτελείται από τα μέτρα *confidence* και *support*, ενώ στο αντικείμενο *Association Rule Pattern 2* το μέτρο ποιότητας αποτελείται από τα μέτρα *coverage*, *strength*, *lift* και *leverage*.



Εικόνα 2-7 Η βασική ιδέα της αντικειμενοσχεσιακής προσέγγισης

Παρακάτω παρουσιάζονται κάποιες αντιπροσωπευτικές επερωτήσεις για το αντικειμενο-σχεσιακό μοντέλο:

OQ1) *Ανάκτηση της δομής των προτύπων κανόνων συσχέτισης.*

```
select p.id, treat(value(p) as hr.assrule_pattern).structureschema from
hr.tbl_patterns p;
```

OQ2) *Ανάκτηση του μέρους body της δομής των προτύπων κανόνων συσχέτισης.*

```
select p.id, value(e), value(f)  from hr.tbl_patterns p,
table(treat(value(p) as hr.assrule_pattern).structureschema.head) e,
```

```
table(treat(value(p) as hr.assrule_pattern).structureschema.body) f;
```

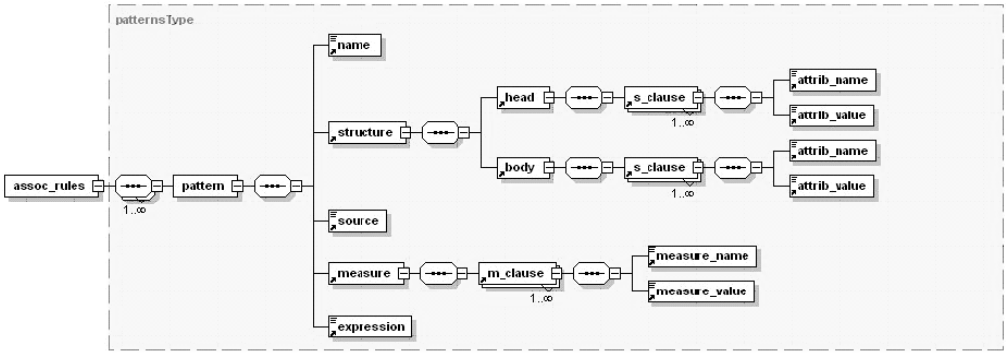
OQ3) *Ανάκτηση των μέτρων ποιότητας confidence των προτύπων κανόνων συσχέτισης.*

```
select p.id, treat(value(p) as
hr.assrule_pattern_1).measureschema.confidence as
confidence from hr.tbl_patterns p;
```

Η αντικειμενο-σχεσιακή προσέγγιση δεν παρουσιάζει κάποιους από τους περιορισμούς που έχει η σχεσιακή προσέγγιση λόγω της ικανότητάς της να μοντελοποιεί σύνθετες οντότητες ως αντικείμενα. Επιπλέον εκμεταλλεύεται τις ομοιότητες των αντικειμένων μέσω της κληρονομικότητας. Το αντικειμενο-σχεσιακό μοντέλο είναι περισσότερο ευέλικτο και αποτελεσματικό από το σχεσιακό αλλά, από την άλλη πλευρά απαιτεί ακριβείς ορισμούς όλων των νέων αντικειμένων και των μερών που τα αποτελούν.

2.4.3 Ημι-δομημένη (XML) προσέγγιση

Αντίθετα με τις παραδοσιακές βάσεις δεδομένων, σε μία XML βάση η μορφή/δομή των δεδομένων δεν είναι τόσο αυστηρά ορισμένη. Αυτή η ιδιότητα είναι ιδιαίτερα χρήσιμη στην περίπτωση των προτύπων καθώς αυτά μπορεί να εξάγονται από διαφορετικά πεδία εφαρμογών έχοντας διαφορετικά χαρακτηριστικά. Για την υλοποίηση της XML βάσης πρέπει να δημιουργηθεί ένα XML schema για κάθε τύπο προτύπου. Πρότυπα ενός συγκεκριμένου τύπου θα είναι τα έγγραφα – στιγμιότυπα του XML σχήματος του τύπου αυτού.



Εικόνα 2-8 Το σχήμα association_rule.xsd

```
<assoc_rules ptype="association_rule">
        <pattern id="1"> <name>rule 1</name>
                <structure>
                        <head>
                                <s_clause>
                                        <attrib_name>buys</attrib_name>
                                        <attrib_value>scarf</attrib_value>
                                </s_clause>
                        </head>
                        <body>
                                <s_clause>
                                        <attrib_name>buys</attrib_name>
                                        <attrib_value>gloves</attrib_value>
                                </s_clause>
                        </body>
                </structure>
                <source>SELECT * FROM orders</source>
                <measure>
                        <m_clause>
                                <measure_name>support</measure_name>
                                <measure_value>0.35</measure_value>
                        </m_clause>
                        <m_clause>
                                <measure_name>confidence</measure_name>
                                <measure_value>0.75</measure_value>
                        </m_clause>
                </measure>
<expression>
{buys="hat",buys="cap",buys="gloves"}
</expression>
        </pattern>
```

Εικόνα 2-9 association_rule.xml

Για παράδειγμα, το πρότυπο τύπου κανόνα συσχέτισης περιγράφεται μέσω του σχήματος “association_rule.xsd” (Εικόνα 2-8), ενώ το XML έγγραφο “pattern-association_rules.xml” (Εικόνα 2-9) περιέχει πρότυπα του τύπου προτύπου κανόνα συσχέτισης του αντίστοιχου σχήματος.

Παρακάτω παρουσιάζονται αντιπροσωπεύτηκες επερωτήσεις για το μοντέλο XML υλοποιημένα σε ORACLE XML-SQL:

XQ1) *Ανάκτηση της δομής των προτύπων κανόνων συσχέτισης που ανήκουν στην κλάση “class1”.*

```
Select
extract(value(y),'//pattern[@id="'||extract(value(e),
'pid/text()')||'"]/structure') as structures from assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x),
'class[@name="class1"]//pids/pid'))) e where
existsNode(value(y),'//pattern[@id="'||extract(value(e),'pid/text()')||'"]/s
tructure') = 1
```

XQ2) *Ανάκτηση του μέρους 'head' και του μέρους 'body' της δομής των προτύπων που ανήκουν στην κλάση Association_Rule_1.*

```
select
extract(value(y),'//pattern[@id="'||extract(value(e),'pid/text()')||'"]//s_c
lause') as pattern_name from assoc_rules y, classes x,
TABLE(XMLsequence(extract(value(x),'class[@name="class1"]//pids/pid'))) e
where
existsNode(value(y),'//pattern[@id="'||extract(value(e),'pid/text()')||'"]//
s_clause') = 1;
```

XQ3) *Ανάκτηση των τιμών του μέτρου ποιότητας confidence των προτύπων κανόνων συσχέτισης.*

```
select (extractvalue(value(val),'//text()')) as confidence from assoc_rules
a,
TABLE(xmlsequence(extract(value(a),'//m_clause[measure_name="confidence"]/me
asure_value'))) val
```

XQ4) *Ανάκτηση όλων των διαφορετικών μέτρων ποιότητας των προτύπων κανόνων συσχέτισης.*

```
select distinct extractValue(value(r), '//m_clause/measure_name/text()') as
measures from assoc_rules y, classes
x,TABLE(XMLsequence(extract(value(x),'//pids/pid'))) e,
TABLE(XMLsequence(extract(value(y),'//pattern[@id="'||
extract(value(e),'pid/text()')||'"]//m_clause'))) r;
```

Στην XML βάση προτύπων ο ορισμός ενός νέου τύπου προτύπου είναι απλή διαδικασία (επεκτασιμότητα). Επιπλέον, είναι πιθανό να δημιουργηθεί το XML σχήμα για ένα τύπο προτύπου, αρκετά γενικό για να περιλαμβάνει κάθε εναλλακτικό πρότυπο του τύπου αυτού (γενικότητα). Το XML σχήμα επηρεάζει επιπλέον την αποτελεσματικότητα των επερωτήσεων. Επερωτήσεις όπως η *XQ4* "ανάκτηση όλων των διαφορετικών μέτρων ποιότητας των προτύπων κανόνων συσχέτισης", μπορούν εύκολα να υλοποιηθούν, σε αντίθεση με τη σχεσιακή και αντικειμενο-σχεσιακή προσέγγιση.

2.4.4 Ποιοτική Σύγκριση

Στην ενότητα αυτή παρουσιάζουμε τα κριτήρια για την σύγκριση των τριών εναλλακτικών αναπαραστάσεων και τα συμπεράσματα της σύγκρισης

 # 1. Πολυπλοκότητα Υλοποίησης Βάσης Προτύπων

Και τα τρία μοντέλα που παρουσιάστηκαν μπορούν εύκολα να υλοποιηθούν. Το πιο απλό μοντέλο είναι το σχεσιακό, όπου και η κατασκευή της βάσης προτύπων και οι λειτουργίες εισαγωγής μπορούν να πραγματοποιηθούν εύκολα και γρήγορα. Το αντικειμενο-σχεσιακό μοντέλο παρουσιάζει περισσότερη δυσκολία αφού απαιτεί τον ορισμό διαφορετικών αντικειμένων για κάθε τύπο προτύπου (και τις παραλλαγές τους). Οι λειτουργίες εισαγωγής είναι επίσης δυσκολότερες αφού πρέπει να είναι διαφορετικές για κάθε τύπο προτύπου και τις παραλλαγές του. Τέλος, η δυσκολία του XML μοντέλου έγκειται στο γεγονός ότι η επιτυχία του εξαρτάται στην ποιότητα (γενικότητα) του XML σχήματος κάθε τύπου προτύπου. Εντούτοις, μετά την κατασκευή του κατάλληλου σχήματος, οι λειτουργίες εισαγωγής μπορούν πολύ εύκολα να πραγματοποιηθούν. Επιπλέον, εάν το σχήμα είναι αρκετά γενικό, οι παραλλαγές των προτύπων που ανήκουν σε ένα συγκεκριμένο τύπο προτύπου μπορούν εύκολα να υποστηριχθούν από το ίδιο σχήμα.

 # 2. Υλοποίηση Περιορισμών

Οι βασικοί περιορισμοί που τίθενται από το λογικό μοντέλο (Rizzi et al., 2003) είναι οι εξής: (α) κάθε πρότυπο είναι ένα στιγμιότυπο ενός τύπου προτύπου, (β) κάθε πρότυπο ανήκει σε τουλάχιστον μία κλάση, (γ) μία κλάση προτύπων πρέπει να περιέχει πρότυπα του ίδιου τύπου.

Αυτοί οι περιορισμοί μπορούν εύκολα να υλοποιηθούν στο σχεσιακό μοντέλο μέσω των περιορισμών ξένων κλειδιών. Στο αντικειμενο-σχεσιακό μοντέλο, αυτοί οι περιορισμοί υποστηρίζονται άμεσα από τον ορισμό των τύπων προτύπων, για παράδειγμα είναι αδύνατο να ανατεθεί ένα πρότυπο συστάδας σε ένα τύπο προτύπου κανόνα συσχέτισης. Τέλος, στο XML μοντέλο η υλοποίηση των περιορισμών υποστηρίζεται από το ΣΔΒΔ με μηχανισμούς συσχέτισης των XML εγγράφων.

 # 3. Εκμετάλλευση των Ιδιαίτερων Χαρακτηριστικών των Προτύπων

Σύμφωνα με το λογικό μοντέλο (Rizzi et al., 2003), κάθε πρότυπο αποτελείται από πέντε βασικά μέρη: όνομα, δομή, πηγή δεδομένων, μέτρο ποιότητας και συνάρτηση αντιστοίχισης. Εντούτοις, διαφορετικοί τύποι προτύπου διαφοροποιούνται σε κάποια από αυτά τα μέρη, για παράδειγμα στη δομή και στο μέτρο ποιότητας. Αν εκμεταλλευτούμε τα ειδικά χαρακτηριστικά κάθε τύπου προτύπου μπορούμε να βελτιώσουμε λειτουργίες όπως ευρετηριοποίηση

και ανάκτηση. Το σχεσιακό μοντέλο δεν εκμεταλλεύεται τη δομή των προτύπων αφού θεωρεί κάθε μέρος του προτύπου σαν συμβολοσειρά, ενώ τόσο το αντικειμενο-σχεσιακό όσο και το XML μοντέλο λαμβάνουν υπόψη τα ιδιαίτερα χαρακτηριστικά όλων των μερών των προτύπων ανάλογα τον τύπο προτύπου.

 # 4. Αποτελεσματικότητα Επερωτήσεων

Η βάση προτύπων δε στοχεύει μόνο στην αποθήκευση των προτύπων αλλά κυρίως στην εύκολη διαχείρισή τους και άρα η αποτελεσματικότητα των επερωτήσεων έχει ιδιαίτερη σημασία. Από τις αντιπροσωπευτικές επερωτήσεις που παρουσιάσαμε για κάθε υλοποίηση, είναι φανερό ότι η κατασκευή των επερωτήσεων στο σχεσιακό μοντέλο είναι πολύπλοκη και χρονοβόρα (αφού έχει αποκλειστικά να κάνει με διαχείριση συμβολοσειρών). Τα υπόλοιπα δύο μοντέλα εκμεταλλεύονται τη δομή των προτύπων και έτσι οι επερωτήσεις εκφράζονται ευκολότερα και είναι περισσότερο αποτελεσματικές.

 # 5. Επεκτασιμότητα

Η επεκτασιμότητα αφορά στη δυνατότητα ενσωμάτωσης νέων τύπων προτύπων στη βάση προτύπων. Όσο περισσότερο εύκολη είναι η διαδικασία αυτή τόσο περισσότερο επεκτάσιμο είναι το σύστημα. Το σχεσιακό μοντέλο είναι ιδιαίτερα επεκτάσιμο. Ένας νέος τύπος προτύπου είναι μία επιπλέον εγγραφή στον πίνακα των τύπων προτύπων. Στο αντικειμενο-σχεσιακό απαιτείται η δημιουργία νέων αντικειμένων για κάθε νέο τύπο προτύπου και τα μέρη που αποτελείται (το ίδιο ισχύει για κάθε παραλλαγή του τύπου προτύπου). Αυτό έχει ως αποτέλεσμα να απαιτούνται πιθανώς περισσότερα από ένα σχήματα κανόνων συσχέτισης για την ενσωμάτωση των διαφορών στη δομή κάθε κανόνα συσχέτισης. Στο XML μοντέλο ένα νέο σχήμα απαιτείται για κάθε νέο τύπο προτύπου, αλλά εάν έχει οριστεί ένα τέτοιο σχήμα, όλες οι παραλλαγές των προτύπων αυτού του τύπου μπορούν να ενσωματωθούν χωρίς τροποποιήσεις.

 # 6. Εγκυρότητα προτύπων

Ο έλεγχος εγκυρότητας στις λειτουργίες εισαγωγή/ ενημέρωση στην βάση προτύπων είναι ιδιαίτερης σημασίας. Με τον όρο εγκυρότητα εννοείται ότι κάθε πρότυπο στη βάση προτύπων πρέπει να είναι σύμφωνο με τον ορισμό του τύπου προτύπου του. Το παραπάνω δεν ισχύει στο σχεσιακό μοντέλο, ενώ ισχύει για το αντικειμενο-σχεσιακό και για το XML λόγω του ορισμού των αντικειμένων και των XML σχημάτων αντίστοιχα.

 # 7. Επαναχρησιμοποίηση

Το κριτήριο της επαναχρησιμοποίησης ικανοποιείται από τη σχεσιακή και την XML βάση αφού η σχεσιακή προσέγγιση δεν υποστηρίζει κληρονομικότητα ή τον ορισμό ημι-δομημένων εγγράφων όπως αντίστοιχα υποστηρίζουν οι άλλες δύο προσεγγίσεις.

 # 8. Γενικότητα

Όλες οι προσεγγίσεις ικανοποιούν αυτό το κριτήριο αφού σε κάθε μία είναι δυνατόν να οριστεί κάθε είδος τύπου προτύπου, αν και ειδικά στη σχεσιακή προσέγγιση έχει περισσότερη πολυπλοκότητα.

Τα συμπεράσματα της σύγκρισης είναι συγκεντρωμένα στον Πίνακα 2-1.

Πίνακα 2-1 Συγκριτικός πίνακας των τριών προσεγγίσεων για την αναπαράσταση της βάσης προτύπων

	Σχεσιακή βάση προτύπων	**Αντικειμενο-Σχεσιακή βάση προτύπων**	**XML βάση προτύπων**
Πολυπλοκότητα υλοποίησης	Υψηλή	Μεσαία	Υψηλή
Υλοποίηση περιορισμών	Ναι	Ναι	Ναι
Εκμετάλλευση χαρακτηριστικών των προτύπων	Όχι	Ναι	Ναι
Αποτελεσματικότητα Επερωτήσεων	Χαμηλή	Μεσαία	Μεσαία
Εγκυρότητα προτύπων	Όχι	Ναι	Ναι
Επεκτασιμότητα	Υψηλή	Μεσαία	Υψηλή
Επαναχρησιμοποίηση	Όχι	Ναι	Ναι
Γενικότητα	Ναι	Ναι	Ναι

Από τον παραπάνω πίνακα είναι φανερό ότι η XML βάση προτύπων είναι η καλύτερη από τις τρεις επιλογές.

Στο σημείο αυτό πρέπει να γίνει αποσαφήνιση της έννοιας της «εγκυρότητας» που αναφέρεται στην εγκυρότητα των XML εγγράφων, την ορθή δομή τους

δηλαδή σε σχέση με το αντίστοιχο σχήμα που έχει οριστεί για αυτά, σύμφωνα με το πλαίσιο του PANDA.

2.5 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάζουμε έννοιες που σχετίζονται με τα πρότυπα και τα μοντέλα αναπαραστάσεών τους για μία βάση προτύπων προκειμένου να ενσωματωθούν σε ένα Σύστημα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ). Λόγω του ότι τα πρότυπα αποτελούν συμπαγή και πλούσια σε σημασιολογία αναπαράσταση των αρχικών δεδομένων (Theodoridis et al., 2003), μοιράζονται κάποια κοινά χαρακτηριστικά, αλλά διαφοροποιούνται επίσης σύμφωνα με τον τύπο με τον οποίο ανήκουν. Επιπλέον, υπάρχουν παραλλαγές μεταξύ των προτύπων του ίδιου τύπου. Η ανάγκη για τα Συστήματα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ) είναι όλο και περισσότερο επιτακτική λόγω της μεγάλης σημασίας που έχουν τα πρότυπα σε πολλές εφαρμογές.

Το λογικό μοντέλο για ένα ΣΔΒΠ ορίζεται στο (Rizzi et al. 2003), και περιλαμβάνει τρεις βασικές έννοιες: τον *τύπο προτύπου*, το *πρότυπο* και την *κλάση*. Μία βάση προτύπων πρέπει να υποστηρίζει αποτελεσματικά αυτές τις έννοιες, οπότε πρέπει να οριστεί και το κατάλληλο μοντέλο αναπαράστασης. Για το πρόβλημα της αναπαράστασης των προτύπων, το ημι-δομημένο μοντέλο είναι περισσότερο κατάλληλο από το σχεσιακό ή το αντικειμενο-σχεσιακό μοντέλο. Με τη χρήση της XML για την υλοποίηση της βάσης προτύπων, μπορεί να επιτευχθεί ένα περισσότερο γενικό και πλήρες ΣΔΒΠ.

Άλλες προσεγγίσεις έχουν προταθεί, αλλά ο στόχος τους δεν είναι ένα ολοκληρωμένο και γενικό σύστημα διαχείρισης βάσης προτύπων. Ανάμεσα στις πιθανές αναπαραστάσεις, η PMML είναι η προτιμότερη. Αν και η PMML είναι βασισμένη στην XML και τείνει να υποστηρίζει όλο και περισσότερους τύπους προτύπων, μία περισσότερο γενική προσέγγιση πρέπει να υιοθετηθεί για ένα ΣΔΒΠ. Τα πρότυπα ορίζονται ανά εφαρμογή ή επιστημονική περιοχή, ώστε το σύστημα να είναι ανοικτό σε επεκτάσεις του χρήστη. Οι επερωτήσεις για τα πρότυπα και η αντιστοίχιση των δεδομένων και των προτύπων είναι θέματα που η PMML δεν λαμβάνει υπόψη, αν και είναι βασικά θέματα για τη δημιουργία ενός περισσότερο πλήρους ΣΔΒΠ.

Η PMML μπορεί ωστόσο να χρησιμοποιηθεί για την αναπαράσταση των προτύπων εξόρυξης γνώσης και μπορεί να βελτιωθεί με μεταδεδομένα για να υποστηριχθούν χαρακτηριστικά απαραίτητα για ένα ΣΔΒΠ. Στο Κεφάλαιο 5,

παρουσιάζεται μία περισσότερο λεπτομερής περιγραφή του PMML σχήματος και των απαιτούμενων μεταδεδομένων.

Έχοντας ορίσει το κατάλληλο μοντέλο αναπαράστασης (XML σχήματα και έγγραφα), μπορούμε να ασχοληθούμε με περισσότερο σύνθετες λειτουργίες προτύπων, όπως η σύγκριση και η εγκυρότητα των προτύπων. Αυτές οι προχωρημένες λειτουργίες είναι ιδιαίτερης σημασίας σε πολλές πραγματικές εφαρμογές.

Στα επόμενα κεφάλαια παρουσιάζουμε τη σύγκριση διακριτών και ασαφών συστάδων για την επέκταση του πλαισίου σύγκρισης PANDA με αλγορίθμους σύγκρισης συστάδων. Έως τώρα στη βιβλιογραφία δεν είχε οριστεί συνάρτηση σύγκρισης για τα πρότυπα του αλγορίθμου EM ή του αλγορίθμου Fuzzy-C-means για διακριτές και ασαφείς συστάδες, αντίστοιχα.

3 Σύγκριση Προτύπων – Η Περίπτωση των Διακριτών Προτύπων Συστάδων

Στο κεφάλαιο αυτό παρουσιάζονται μέθοδοι και αλγόριθμοι με την χρήση των οποίων είναι δυνατή η σύγκριση προτύπων, ώστε να επιτευχθεί μία σύγκριση των αρχικών δεδομένων σε ένα υψηλότερο επίπεδο. Στην παρούσα διατριβή επικεντρωνόμαστε στα πρότυπα συστάδων που εξάγονται από τον αλγόριθμο EM (Dempster et al., 1977). Οι συστάδες αυτές αναπαρίστανται ως κατανομές και έτσι παρουσιάζεται μία κατάλληλη συνάρτηση σύγκρισης κατανομών. Σαν μελέτη περίπτωσης, παρουσιάζουμε δύο πραγματικά σενάρια σύγκρισης εικόνων μέσω της σύγκρισης των προτύπων συστάδων που τις αποτελούν, για να γίνει κατανοητή η πιθανή χρήση του συστήματος σύγκρισης προτύπων.

3.1 Εισαγωγή

Ο λόγος για τον οποίο εστιάζουμε στη σύγκριση πρότυπα συστάδων και ειδικά σε συστάδες που έχουν εξαχθεί από τον αλγόριθμο EM είναι επειδή δεν έχει οριστεί ήδη συνάρτηση σύγκρισης τέτοιων προτύπων στο σύστημα σύγκρισης PANDA (το οποίο θα παρουσιαστεί στην Ενότητα 3.2).

Ο αλγόριθμος συσταδοποίησης EM (Expectation-Maximization), ο οποίος είναι ιδιαίτερα γνωστός και ευρέως χρησιμοποιούμενος, αναζητά συστάδες με τη χρήση μοντέλων μίξης (mixture models) και κατανομών. Μπορεί να αναγνωρίσει ομάδες δεδομένων που επικαλύπτονται ή που διαφέρουν σε μέγεθος και σχήματα. Ο EM χρησιμοποιεί κατανομές για να δείξει ποια δεδομένα ανήκουν σε κάθε συστάδα. Ο EM είναι αρκετά γενικότερος από έναν αλγόριθμο συσταδοποίησης, αφού μπορεί να ανακαλύψει έναν αριθμό κατανομών στα δεδομένα και να χτίσει “mixture models”.

Προκειμένου να συγκρίνουμε συστάδες που αναπαρίστανται ως κατανομές, πρέπει να συγκρίνουμε τις ίδιες τις κατανομές. Για το σκοπό αυτό,

χρησιμοποιούμε την συνάρτηση απόστασης *Cohen's d* (Cohen, 1988). Η γενική διαδικασία σύγκρισης (μεθοδολογία), υποστηρίζεται από τις συναρτήσεις του πλαισίου σύγκρισης PANDA (Ntoutsi et al., 2007). Στο κεφάλαιο αυτό ορίζουμε νέες συναρτήσεις σύγκρισης στο πλαίσιο PANDA, επεκτείνοντάς το για να υποστηρίζει πρότυπα συστάδων του αλγορίθμου EM και παρουσιάζουμε πραγματικές εφαρμογές, ενώ στο Ntoutsi (2008) παρουσιάζονται εφαρμογές που κυρίως έχουν να κάνουν με τη σύγκριση προτύπων συχνών στοιχειοσυνόλων και δέντρων απόφασης.

3.2 Ορισμός Ομοιότητας Προτύπων

Το πλαίσιο PANDA (Ntoutsi et al., 2007) παρέχει συναρτήσεις σύγκρισης για απλά και σύνθετα πρότυπα. Τα απλά πρότυπα εξάγονται από τα αρχικά δεδομένα με διαδικασίες εξόρυξης γνώσης (πχ. συστάδες από δεδομένα), ενώ τα σύνθετα πρότυπα αποτελούνται από απλά πρότυπα (πχ. συστάδες από συστάδες). Με τη χρήση αλγορίθμων του πλαισίου PANDA μπορούμε να συγκρίνουμε πρότυπα του ίδιου τύπου (δηλαδή συστάδες με συστάδες, ή κανόνες συσχέτισης με κανόνες συσχέτισης κλπ).

Λόγω της συμπαγούς και πλούσιας σε σημασιολογία αναπαράστασης των προτύπων, το πλαίσιο PANDA μπορεί να χρησιμοποιηθεί για να συγκριθούν πρότυπα με μεγάλο βαθμό πολυπλοκότητας. Το πλαίσιο PANDA χρησιμοποιεί τη διμερή ιδιότητα για τη σύγκριση δύο προτύπων. Η βασική ιδέα της ιδιότητας αυτής είναι ότι η πλειοψηφία των προτύπων μπορούν να περιγραφούν επαρκώς από τα δύο μέρη, τη δομή και το μέτρο ποιότητας (structure και measure).

Η σύγκριση γίνεται με την έκφραση της απόστασης *dis*, όπου η ελάχιστη απόσταση δείχνει την περισσότερη ομοιότητα δύο προτύπων p_1, p_2 του ίδιου τύπου. Η απόσταση αυτή μπορεί να υπολογιστεί συνδυάζοντας, με μία συνάρτηση συνάθροισης f_{aggr}, την απόσταση μεταξύ τόσο της δομής (structure) s όσο και του μέτρου (measure) m (Ntoutsi et al., 2007):

$$dis(p_1, p_2) = f_{\text{aggr}}(dis_{\text{struct}}(p_1.s, p_2.s), dis_{\text{meas}}(p_1.m, p_2.m)) \quad (3\text{-}1)$$

όπου τα $p_i.s$ και $p_i.m$ δηλώνουν τη δομή και το μέτρο, αντίστοιχα, του προτύπου p_i.

Η σημειολογία με την τελεία δηλώνει ότι η μεταβλητή στα δεξιά της, είναι μέλος του προτύπου αριστερά της, σύμφωνα με τη σημειολογία που χρησιμοποιείται στο αντικειμενοστραφές μοντέλο.

Εάν και τα δύο πρότυπα προς σύγκριση έχουν ίδια ακριβώς δομή, τότε η συνάρτηση σύγκρισης λαμβάνει υπόψη μόνο την απόσταση μεταξύ των μέτρων ποιότητας (measure components).

Ο αποτελεσματικός ορισμός της δομής και του μέτρου των προτύπων που έχουν εξαχθεί από τα αρχικά δεδομένα, όπως επίσης η κατάλληλη επιλογή της λογικής συνάθροισης και της συνάρτησης απόστασης για την εύρεση των αντίστοιχων αποστάσεων, είναι ιδιαίτερης σημασίας για κάθε μία και διαφορετική εφαρμογή.

Στη συνέχεια θα περιγράψουμε τη μεθοδολογία της σύγκρισης των προτύπων συστάδων, βασιζόμενοι στις έννοιες που περιγράφηκαν παραπάνω.

Στην ενότητα 2.3 ορίσαμε την έννοια του τύπου προτύπου σαν μία πεντάδα στοιχείων pt = (n, ss, ds, ms, f). Προκειμένου να ορίσουμε συναρτήσεις σύγκρισης για τις συστάδες και άλλα πρότυπα, μόνο τα μέρη *ss* και *ms* είναι απαραίτητα. Τα τρία άλλα μέρη δεν χρησιμοποιούνται για τη σύγκριση και γι' αυτό το λόγο δεν θα μας απασχολήσουν στην πορεία. Βάσει των παραπάνω, επαναορίζουμε την έννοια του τύπου προτύπου σαν ζευγάρι *PT* = <*SS*, *MS*>, όπου το *SS* (structure schema) αναφέρεται στη δομή του προτύπου, ενώ το *MS* (measure schema) ποσοτικοποιεί την ποιότητα της αναπαράστασης των αρχικών δεδομένων από το πρότυπο. Στη συνέχεια θα αναφερόμαστε σε αυτά τα δύο μόνο μέρη του τύπου προτύπου εκτός αν απαιτείται μία πιο λεπτομερής περιγραφή.

Σαν παράδειγμα, θεωρούμε ένα τύπο προτύπου που αναπαριστά σφαιρικές συστάδες σε ένα D-διάστατο χώρο, με χρήση Ευκλείδειας απόστασης – Ευκλείδειες συστάδες. Η δομή ενός τέτοιου προτύπου μπορεί να περιγραφεί ορίζοντας το κέντρο της συστάδας (ένα D-διάστατο διάνυσμα – vector) και μία ακτίνα. Το μέτρο ποιότητας της συστάδας μπορεί να είναι για παράδειγμα, η υποστήριξη (support), ο λόγος δηλαδή των στοιχείων/δεδομένων που αναπαρίστανται από τη συστάδα. Έτσι:

$$EuclideanCluster = \begin{pmatrix} SS : \left(center : [\text{Real}]_1^{\text{D}}, radius : \text{Real}\right) \\ MS : \left(\sup p : \text{Real}\right) \end{pmatrix}$$

Όπως ήδη αναφέρθηκε στο προηγούμενο κεφάλαιο, ένας τύπος προτύπου *PT* λέγεται *σύνθετος*, αν το μέρος της δομής του (SS) περιλαμβάνει έναν άλλο τύπο προτύπου, αλλιώς αυτός λέγεται *απλός*.

Έτσι, ένα *EuclideanCluster* είναι ένας απλός τύπος προτύπου, ενώ μία συσταδοποίηση που εξήχθη από π.χ. έναν διαμεριστικό αλγόριθμο συσταδοποίησης θεωρείται σύνθετος τύπος προτύπου αφού μπορεί να περιγραφεί σαν ένα σύνολο από συστάδες χωρίς κάποιο μέτρο ποιότητας:

$$PartitioningClustering = \begin{pmatrix} SS : \{EuclideanCluster\} \\ MS : \perp \end{pmatrix}$$

Εάν *PT* είναι ένας τύπος προτύπου, τότε το $p = < s, m>$ είναι το πρότυπο-στιγμιότυπο του *PT*, όπου *s, m* είναι οι αντίστοιχες τιμές για τη δομή και το μέτρο ποιότητας του προτύπου. Σύμφωνα με το προηγούμενο παράδειγμα, ένα πιθανό στιγμιότυπο ενός 3-διάστατου *EuclideanCluster* θα είναι:

$$Cluster1 = \begin{pmatrix} s : (center : [0.1, 0.3, 0.45], radius : 0.77) \\ m : (\sup\ p : 0.15) \end{pmatrix}$$

Σύμφωνα με το πλαίσιο PANDA, όπως προαναφέρθηκε, η απόσταση *dis* μεταξύ δύο απλών προτύπων p_1, p_2 υπολογίζεται από τη συνάρτηση (3-1).

Από την άλλη μεριά η απόσταση μεταξύ δύο σύνθετων προτύπων ορίζεται ως η συναθροιστική απόσταση μεταξύ των προτύπων από τα οποία αποτελούνται τα αρχικά σύνθετα πρότυπα (αυτός είναι ένας αναδρομικός ορισμός, αφού τα σύνθετα πρότυπα μπορεί να αποτελούνται από άλλα σύνθετα πρότυπα κ.ο.κ.).

Το πλαίσιο PANDA παρέχει έναν αριθμό από συναρτήσεις απόστασης, συνάθροισης και ταιριάσματος. Περισσότερα για το πλαίσιο PANDA και τις συναρτήσεις που υποστηρίζει μπορεί να βρεθούν στα (Ntoutsi et al., 2007; Ntoutsi, 2008). Ακολουθώντας την μεθοδολογία σύγκρισης που ορίζεται από το πλαίσιο PANDA, ορίζουμε τις συναρτήσεις σύγκρισης και συνάθροισης που θα χρησιμοποιηθούν για τη σύγκριση προτύπων συστάδων από τον αλγόριθμο EM.

3.3 Σύγκριση Προτύπων Συστάδων

Προκειμένου να συγκρίνουμε πρότυπα συστάδων, θα πρέπει να περιγραφεί η όλη διαδικασία, από τη δημιουργία των προτύπων έως το ταίριασμα και τη σύγκριση. Η ομοιότητα των προτύπων εξαρτάται από τα μέρη της δομής και του μέτρου ποιότητας και έτσι, οι λεπτομέρειες των προτύπων συστάδων και της αναπαράστασής τους θα πρέπει να αναλυθούν.

Για την ανάλυση, τα παραδείγματα αλλά και τις εφαρμογές που θα παρουσιαστούν, θα χρησιμοποιηθούν δεδομένα εικόνων ενώ ο αλγόριθμος συσταδοποίησης θα είναι ο ΕΜ όπως προαναφέρθηκε.

Η μεθοδολογία της εξαγωγής και της σύγκρισης προτύπων συστάδων περιλαμβάνει τα παρακάτω βήματα:

1. Εξαγωγή χαρακτηριστικών από τα αρχικά δεδομένα
2. Εφαρμογή του κατάλληλου αλγορίθμου εξόρυξης γνώσης/ συσταδοποίησης
3. Αναπαράσταση και δημιουργία των προτύπων
4. Υπολογισμός της ομοιότητας των προτύπων

Λεπτομερώς,

1. Εξαγωγή χαρακτηριστικών από τα αρχικά δεδομένα

Το πρώτο βήμα είναι να γίνει η εξαγωγή εκείνων των χαρακτηριστικών των αρχικών δεδομένων που θα χρησιμοποιηθούν στη συσταδοποίηση. Τα αρχικά δεδομένα μπορεί να είναι για παράδειγμα κείμενο ή εικόνες.

Στη δεύτερη περίπτωση, η εικόνα σαρώνεται με την τεχνική sliding window με μέγεθος ορισμένο από τον χρήστη, χωρίζοντας έτσι την εικόνα σε blocks με συγκεκριμένη απόσταση/βήμα. Το βήμα αυτό μπορεί να επιτρέπει δύο συνεχόμενα blocks να υπερκαλύπτονται. Για κάθε μπλοκ, ένα σύνολο από N χαρακτηριστικά f_i, i = 1, …, N, υπολογίζεται ώστε να οριστεί ένα μόνο διάνυσμα χαρακτηριστικών F (feature vector). Ο αριθμός των διανυσμάτων χαρακτηριστικών που παράγεται για κάθε εικόνα εξαρτάται από το μέγεθος και τις διαστάσεις του παραθύρου και του βήματος. Για δεδομένα εικόνας, το χρώμα, η υφή και το σχήμα είναι τρεις βασικές κλάσεις χαρακτηριστικών που χρησιμοποιούνται συχνά.

Το αποτέλεσμα του βήματος της εξαγωγής των χαρακτηριστικών είναι ένα σύνολο διανυσμάτων (vectors) με N χαρακτηριστικά.

2. Εφαρμογή του κατάλληλου αλγορίθμου εξόρυξης γνώσης/ συσταδοποίησης

Στο επόμενο βήμα, τα διανύσματα των χαρακτηριστικών συσταδοποιούνται με τη χρήση mixture models που ομαδοποιούν τα δεδομένα με έναν αριθμό από Γκαουσιανές κατανομές. Μία συστάδα αντιστοιχεί σε ένα σύνολο κατανομών,

μία για κάθε διάσταση των δεδομένων. Κάθε κατανομή περιγράφεται ως μία μέση τιμή και μία απόκλιση. Για την ανάθεση των διανυσμάτων δεδομένων στις συστάδες, χρησιμοποιείται μία πιθανοτική (probabilistic) μέθοδος.

Για μονοδιάστατα σύνολα δεδομένων, ένα mixture είναι ένα σύνολο από *c* Γκαουσιανές πιθανοτικές κατανομές, που αναπαριστούν *c* συστάδες. Οι παράμετροι ενός mixture model ορίζονται από τον αλγόριθμο *Expectation Maximization* (EM) (Dempster et al., 1977). Με *c* Γκαουσιανές, η συνάρτηση πυκνότητας πιθανότητας μιας μεταβλητής *X* είναι

$$f(X|\theta)=\sum_{i=1}^{c} pp_i \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} e^{-\frac{1}{2}(X-\mu_i)^T \sum_i^{-1}(X-\mu_i)} \quad (3\text{-}2)$$

όπου $pp_i > 0$, $\sum_{i=1}^{c} pp_i = 1$, και d είναι η διάσταση του διανύσματος χαρακτηριστικών. Το σύνολο των παραμέτρων του μοντέλου $\theta\{pp_i, \mu_i, \Sigma_i\}$, $i = 1, \ldots, c$, αποτελείται από τις εκ-των-προτέρων πιθανοτήτων pp_i της Γκαουσιανής i, το μέσο διάνυσμα μ_i και τη μήτρα διακύμανσης Σ_i για την Γκαουσιανή i, αντίστοιχα. Ο αλγόριθμος EM χρησιμοποιείται για να υπολογίσει το maximum likelihood L του θ δοθέντος ενός συνόλου χαρακτηριστικών $\{x_1,\ldots,x_N\}$:

$$L(\theta|X)=\log \prod_{j=1}^{N} f(x_j|\theta) \quad (3\text{-}3)$$

Οι παράμετροι του μοντέλου αρχικοποιούνται με τυχαίες τιμές. Ο αλγόριθμος ξεκινά με τον υπολογισμό των πιθανοτήτων που ένα διάνυσμα θα ανήκει σε κάθε συστάδα. Αυτές οι πιθανότητες χρησιμοποιούνται για να υπολογίσουν μία νέα εκτίμηση για τις παραμέτρους. Η όλη διαδικασία επαναλαμβάνεται μέχρι οι παράμετροι να συγκλίνουν σε μία σταθερή ή σχεδόν σταθερή εκτίμηση. Ο αλγόριθμος καταλήγει σε ένα σύνολο από κατανομές, ένα διάνυσμα ζευγών από μέσους μ και τυπικές αποκλίσεις σ, κάθε ένα από τα οποία αντιστοιχούν σε ένα χαρακτηριστικό, και στο μέγεθος της συστάδας (ο αριθμός των διανυσμάτων που ανήκουν στη συστάδα). Το διάνυσμα των μέσων μ των κατανομών για κάθε χαρακτηριστικό αναπαριστά το κέντρο της συστάδας.

Ο αλγόριθμος EM παρουσιάζει πολλά πλεονεκτήματα σε σχέση με άλλους αλγορίθμους συσταδοποίησης. Συνδυάζοντας τον EM με την τεχνική υ-fold cross-validation (Stone, 1974), ο αριθμός των συστάδων μπορεί να οριστεί

αυτόματα. Η τεχνική αυτή λειτουργεί με το να διαιρεί τα δεδομένα σε υ ίσα τμήματα. Ξεκινώντας με μία συστάδα, ο ΕΜ εκτελείται υ φορές αφήνοντας κάθε φορά ένα τμήμα εκτός για λόγους δοκιμής και η τιμή likelihood υπολογίζεται σαν μέσος όρος από όλα τα δεδομένα. Στη συνέχεια ο ΕΜ εκτελείται για δύο συστάδες και εάν η τιμή likelihood αυξηθεί, ο αριθμός των συστάδων ορίζεται σαν δύο και η διαδικασία επαναλαμβάνεται μέχρι η τιμή Likelihood αρχίζει να μειώνεται (Witten and Frank, 2005).

Επιπλέον, ο αλγόριθμος ΕΜ είναι περισσότερο γενικός από τον *K*-means (Hartigan, 1975), για παράδειγμα, αφού μπορεί να ανακαλύψει συστάδες διαφόρων μεγεθών και ελλειψοειδών σχημάτων. Ιδιαίτερα σημαντικό είναι και το γεγονός ότι το αποτέλεσμα του αλγορίθμου, οι κατανομές δηλαδή που αναπαριστούν τις συστάδες, μπορούν εύκολα να χρησιμοποιηθούν σαν πρότυπα στο πλαίσιο PANDA.

3. Αναπαράσταση και δημιουργία των προτύπων

Τα πρότυπα που παράγονται από τον αλγόριθμο ΕΜ αναπαρίστανται και διαχειρίζονται σύμφωνα με φορμαλισμό του PANDA όπως περιγράφηκε στην ενότητα 3.2. Έτσι, δεδομένου ενός αντικειμένου που έχει συσταδοποιηθεί (πχ. μία εικόνα) που αποτελείται από M απλά πρότυπα P_i, i = 1,…,M, και λαμβάνοντας υπόψη το αποτέλεσμα του ΕΜ, ένα πρότυπο P_i αναπαριστά ένα πρότυπο στα δεδομένα ως εξής:

$$P_i = \begin{pmatrix} SS : (D : [[\mu : [\text{Real}], \sigma : [\text{Real}]]_1^N), \\ MS : (pp : \text{Real}), (\text{SV} : \text{Real}) \end{pmatrix}$$

Πιο συγκεκριμένα, το μέρος της δομής *SS* ενός προτύπου αναπαρίσταται από ένα ζευγάρι (μ, σ) της κατανομής D_j για κάθε ένα από τα N χαρακτηριστικά (j=1, …, N) στο πρότυπο P_i, αντίστοιχα. Σε αντιστοιχία, το μέρος του μέτρου ποιότητας *MS* ενός προτύπου αναπαρίσταται από δύο τιμές, την εκ των προτέρων πιθανότητα (*prior probability - pp))* και την τιμή διασποράς (*Scatter Value - SV*) του P_i.

Η εκ των προτέρων πιθανότητα *pp* ορίζεται ως το μέρος των διανυσμάτων χαρακτηριστικών του *Object* που ανήκουν στο πρότυπο P_i. Διαισθητικά, η *pp* είναι αντίστοιχη με το μέτρο *support* που ευρέως χρησιμοποιείται στα μοντέλα εξόρυξης γνώσης. Στην περίπτωση αυτή, είναι μία ένδειξη του μεγέθους του *Object*. Από την άλλη μεριά, η τιμή *SV* είναι ένα μέτρο της συνεκτικότητας των δεδομένων σε μία συστάδα με βάση το κέντρο (centroid) της συστάδας, και

είναι ένα μέτρο που συχνά χρησιμοποιείται για τη μέτρηση της ποιότητας της συστάδας (Littau, 2003). Η τιμή της τιμής διασποράς scatter value SV ενός αντικειμένου ορίζεται ως εξής:

$$SV = \sum_{k \in P_i} \left(x_k - c_{P_i} \right)^2 \tag{3-4}$$

όπου x_k είναι τα διανύσματα χαρακτηριστικών που ανήκουν στο πρότυπο P_i και c_{P_i} είναι το αντίστοιχο κέντρο (centroid), που επίσης είναι ένα διάνυσμα με την ίδια διάσταση όπως το x_k, και η τιμή σε κάθε διάσταση υπολογίζεται σαν ο μέσος όρος από τις αντίστοιχες τιμές των χαρακτηριστικών που ανήκουν στο πρότυπο P_i. Μία χαμηλή τιμή scatter value δείχνει καλύτερη ποιότητα διασποράς, αλλά πρέπει να σημειωθεί ότι αυτό είναι ένα μέτρο ποιότητας σχετικό, αφού εξαρτάται από το νούμερο των αντικειμένων σε μία συστάδα.

Στο πλαίσιο αυτό, ένα Object θεωρείται ένα σύνθετο πρότυπο:

$$Object = \begin{pmatrix} SS : \{ P \}, \\ MS : \perp \end{pmatrix}$$

Αποτελούμενο από ένα σύνολο από απλά πρότυπα.

4. Υπολογισμός της ομοιότητας των προτύπων

Στοχεύοντας στην εκτίμηση της ομοιότητας μεταξύ δύο αντικειμένων - Objects (που ορίζονται ως σύνθετα πρότυπα), πρώτα πρέπει να οριστεί η απόσταση μεταξύ των μερών της δομής και του μέτρου ποιότητας των δύο απλών προτύπων P_1 και P_2. Αφού τα σύνθετα πρότυπα αποτελούνται από έναν αριθμό απλών προτύπων, για τη σύγκριση δύο αντικειμένων O_1 και O_2, χρειάζεται ένας τρόπος για να συσχετιστούν τα πρότυπα που αποτελούν το O_1 με αυτά του O_2. Προς αυτή την κατεύθυνση, ο τύπος σύζευξης (*coupling type)* περιορίζει τον τρόπο που τα πρότυπα (που αποτελούν το αντικείμενο) μπορούν να συσχετιστούν (δηλαδή να ταιριαστούν). Παρακάτω, προτείνουμε πρώτα έναν αποτελεσματικό τρόπο για τη μέτρηση της απόστασης μεταξύ δύο απλών προτύπων, και κατόπιν παρουσιάζουμε (βλέπε Eq. (3-12)) αυτό που επιλέγουμε ως τρόπο για τη σύζευξή τους.

Η απόσταση μεταξύ των μέτρων δύο προτύπων προτείνεται να οριστεί ως η απόλυτη διαφορά των τιμών διασποράς, βάζοντας βάρος στην κάθε μία την

αντίστοιχη τιμή της prior probability των προτύπων, και κανονικοποιώντας με το άθροισμα των δύο τιμών διασποράς. Πιο τυπικά:

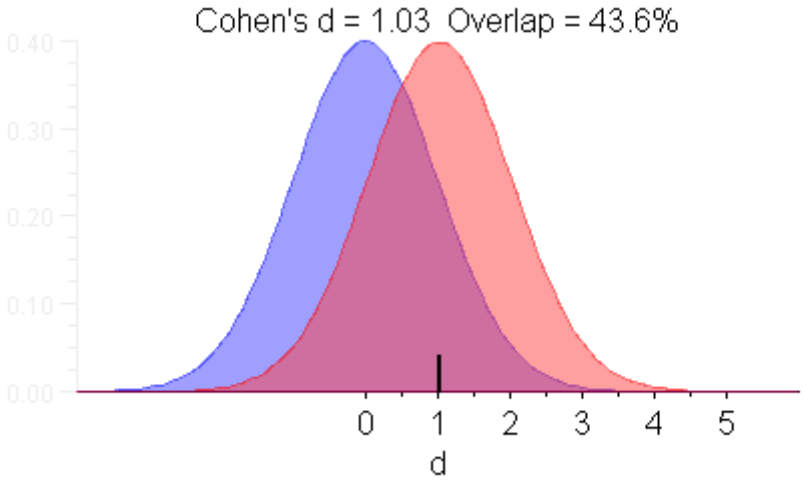
$$dis_{meas}(P_1, P_2) = \frac{\left|P_1.pp \cdot P_1.\mathrm{SV} - P_2.pp \cdot P_2.\mathrm{SV}\right|}{P_1.\mathrm{SV} + P_2.\mathrm{SV}} \tag{3-5}$$

Διαισθητικά, η εξίσωση (3-5) ποσοτικοποιεί την απόκλιση των προτύπων μεταξύ της συνεκτικότητας δύο συστάδων. Πρέπει να σημειωθεί ότι αυτός ο ορισμός απορροφά την αναποτελεσματικότητα της σχετικότητας της τιμής διασποράς με βάση τον αριθμό των αντικειμένων σε μία συστάδα, καθώς κάθε τιμή διασποράς δέχεται το βάρος του λόγου των διανυσμάτων χαρακτηριστικών της εικόνας, που ανήκουν στο πρότυπο P_i.

Σχετικά με την ομοιότητα της δομής μεταξύ των P_1 και P_2, αναζητούμε ένα μέτρο για την αποτίμηση της «εγγύτητας» δύο συνόλων κατανομών, όπως είναι οι P_1 και P_2. Περαιτέρω αποσυνθέτοντας το πρόβλημα, πρέπει πρώτα να οριστεί μία μέθοδος για τον υπολογισμό της ομοιότητας μεταξύ δύο κατανομών D_1 και D_2. Για να επιτευχθεί αυτό, χρησιμοποιούμε την τυποποιημένη διαφορά *d* μεταξύ δύο κατανομών, όπως ορίζεται από τον Cohen (1988). Η απόσταση *Cohen's d* (*dis*) ορίζεται ως η απόλυτη διαφορά μεταξύ των μέσων των κατανομών, διαιρεμένη με τη ρίζα του μέσου των τετραγώνων των δύο τυπικών αποκλίσεων.

$$dis(D_1, D_2) = \begin{cases} d = \dfrac{\left|D_1.mean - D_2.mean\right|}{\sqrt{\dfrac{D_1.stdDev^2 + D_2.stdDev^2}{2}}}, \\ if\ D_1.stdDev \neq 0\ and\ D_2.stdDev \neq 0 \\ \\ \left|D_1.mean - D_2.mean\right|, otherwise \end{cases} \tag{3-6}$$

Η απόσταση *Cohen's d* είναι ένας μη-αρνητικός πραγματικός αριθμός που ερμηνεύει την επικάλυψη δύο κατανομών. Εάν το *d* είναι μηδέν, οι κατανομές είναι πανομοιότυπες. Χαμηλές τιμές του *d* δείχνουν πολύ όμοιες κατανομές ενώ υψηλές τιμές του *d* δείχνουν ανόμοιες κατανομές. Εάν και οι δύο τυπικές αποκλίσεις είναι μηδέν, η απόλυτη διαφορά των μέσων χρησιμοποιείται ως η απόσταση μεταξύ των κατανομών.



Εικόνα 3-1 Γραφική αναπαράσταση της ομοιότητας δύο κατανομών με τη χρήση του μέτρου απόστασης Cohen's d.

Η απόσταση του *Cohen* είναι ένα μέσο για την αυτοματοποίηση και την υλοποίηση της διαισθητικής επικάλυψης μεταξύ δύο κατανομών. Με βάση τα παραπάνω, ορίζουμε ότι η δομική ομοιότητα μεταξύ δύο συνόλων κατανομών (δηλ. δύο προτύπων P_1 και P_2) πρέπει να είναι το αποτέλεσμα μιας συναθροιστικής συνάρτησης g_{aggr} (Εξίσωση (3-7)), που αλληλο-συσχετίζει τις διαφορετικές τιμές των διαφόρων αποστάσεων που έχει το κάθε ζευγάρι κατανομών:

$$dis_{struct}(P_1, P_2) = g_{aggr}\left(\frac{d(D_j^1, D_j^2)}{\delta}\right),\ \forall\ j = 1, 2, \ldots, N \qquad (3\text{-}7)$$

όπου το d είναι η απόσταση του Cohen και δ είναι ένας παράγοντας κανονικοποίησης του πεδίου ορισμού της συνάρτησης g_{aggr} (g_{aggr}: $[0,1] \rightarrow [0,1]$), η οποία διαισθητικά αντιστοιχεί στην τιμή της *Cohen's d* για την οποία δύο κατανομές θεωρούνται απόλυτα ανόμοιες (δηλ. δεν επικαλύπτονται καθόλου). Σύμφωνα με αυτά, η συνάρτηση g_{aggr} μπορεί να είναι οποιαδήποτε αντιστοίχιση που αρχικά πραγματοποιεί μία επιλογή χαρακτηριστικών και κατόπιν εφαρμόζει την συναθροιστική συνάρτηση πάνω σε αυτά. Παραδείγματα τέτοιων συναρτήσεων είναι: (α) η συνάρτηση ελαχίστου (*minimum)* $g_{\min}$ (δηλ. επιλογή των πιο όμοιων κατανομών), (β) η συνάρτηση μέσου (*average*) g_{avg} (δηλ. επιλογή του μέσου όρου των αποστάσεων που υπολογίζονται για κάθε ζευγάρι των N χαρακτηριστικών) και (γ) η συνάρτηση του μέσου όρου των k πιο κοντινών κατανομών (*average of the k Nearest Distributions*) $g_{\text{avg\_kND}}$ (δηλ. επιλογή των

$k \leq N$ πιο όμοιων ζευγών κατανομών). Στην τελευταία περίπτωση, η παράμετρος k μπορεί να μην ορίζεται αποκλειστικά, αλλά μπορεί να ορίζεται «χαλαρώνοντας» την παράμετρο δ. Τυπικά:

$$g_{\min} = \min_{j=1}^{N} \{d(D_j^1, D_j^2)\} \tag{3-8}$$

$$g_{avg} = \frac{1}{N} \sum_{j=1}^{N} d(D_j^1, D_j^2) \tag{3-9}$$

$$g_{avg\_kND} = \frac{1}{k} \sum_{j=1}^{k} kND(d(D_j^1, D_j^2)) \tag{3-10}$$

όπου η συνάρτηση *kND* επιστρέφει τις k πιο όμοιες κατανομές.

Μέχρι το σημείο αυτό, έχουμε ορίσει την dis_{meas} και την dis_{struct} (Εξισώσεις (3-5) και (3-7), αντίστοιχα) μεταξύ δύο προτύπων. Στη συνέχεια, συναθροίζουμε αυτές τις αποστάσεις με τη χρήση μιας συνάρτησης αθροίσματος με βάρη. Τυπικά, η απόσταση $dis(P_1, P_2)$ μεταξύ δύο προτύπων P_1 και P_2 ορίζεται ως:

$$dis(P_1, P_2) = dis_{struct}(P_1, P_2) + (1 - dis_{struct}(P_1, P_2)) \cdot dis_{meas}(P_1, P_2)^2 \tag{3-11}$$

Το νόημα πίσω από την επιλογή αυτή είναι ότι όσο περισσότερο όμοια είναι η δομή τόσο η απόσταση των μέτρων ποιότητας πρέπει να συμμετέχει με μεγαλύτερο βάρος στην τιμή της συνολικής απόστασης.

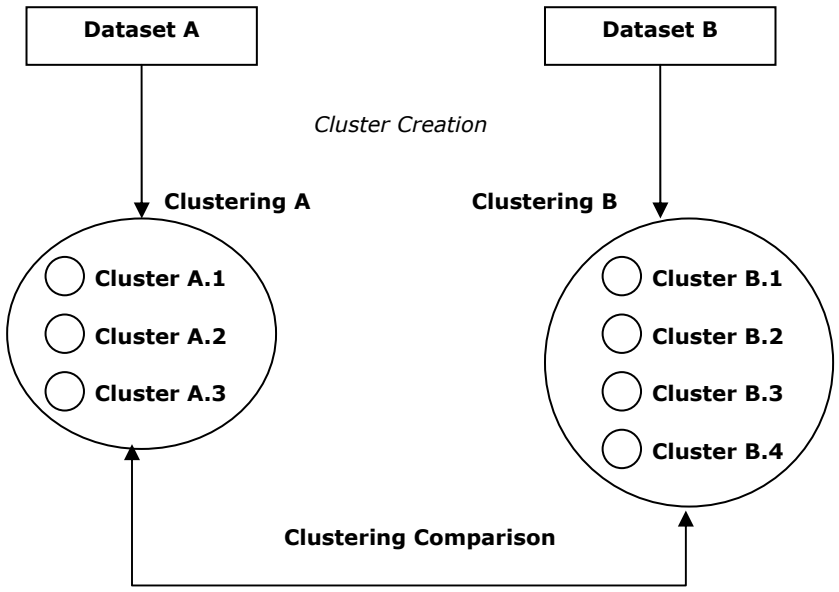
Αυτό υπονοεί ότι εάν οι δομές είναι τελείως διαφορετικές, η απόσταση πρέπει να είναι 1, ανεξάρτητα με το μέρος του μέτρου ποιότητας. Αυτή η επιλογή υπονοεί επιπλέον ότι δίνουμε έμφαση στην ομοιότητα της δομής. Αυτό ενισχύεται επιπλέον με τον πολλαπλασιασμό του παράγοντα $1 - dis_{\text{struct}}$ (δηλαδή την ομοιότητα μεταξύ της δομής των δύο προτύπων) με μία μικρότερη τιμή από την πραγματική τιμή της απόστασης του μέτρου ποιότητας dis_{meas}. Υπενθυμίζεται ότι το πεδίο τιμών της dis_{meas} είναι το $[0,1]$, έτσι με το τετράγωνό της γίνεται η «χαλάρωση» της συμμετοχής της dis_{meas}.

Έχοντας ορίσει την απόσταση μεταξύ απλών κατανομών, για να συγκρίνουμε δύο αντικείμενα (Objects) O_1 και O_2 (δηλαδή δύο σύνθετα πρότυπα) υιοθετούμε την συζευκτική μεθοδολογία μεταξύ των διαφορετικών προτύπων κάθε αντικειμένου ως εξής:

$$dis(O_1, O_2) = \frac{1}{M \cdot K} \cdot \sum_{i=1}^{M} \sum_{j=1}^{K} dis\left(P_i^{O_1}, P_j^{O_2}\right) \tag{3-12}$$

όπου M και K είναι ο αριθμός των απλών προτύπων από τα οποία αποτελείται κάθε αντικειμενο με βάση το αποτέλεσμα του αλγορίθμου EM. Διάφοροι τύποι σύζευξης μπορούν να εφαρμοστούν στο πλαίσιο του PANDA (Bartolini et al., 2004), αλλά ο όλα-προς-όλα τύπος όπως φαίνεται στην Eq. (3-12) αποφεύγει την τάση προς συγκεκριμένα πρότυπα. Το τελικό αποτέλεσμα είναι ο μέσος όρος όλων των πιθανών ταιριασμάτων. Ο καλύτερος τύπος ταιριάσματος για τη σύγκριση δύο προτύπων είναι υπό συζήτηση και εξαρτάται από την εφαρμογή. Οι διαφορετικές περιπτώσεις περιγράφονται παρακάτω.

Στην Εικόνα 3-2, δύο συσταδοποιήσεις παρουσιάζονται. Η Clustering A αποτελείται από 3 συστάδες ενώ η clustering B αποτελείται από 4 συστάδες.



Εικόνα 3-2 Σύγκριση δύο συσταδοποιήσεων *Clustering A* και *Clustering B*

Προκειμένου να γίνει η σύγκριση των δύο συσταδοποιήσεων, τα παρακάτω ταιριάσματα μπορεί να γίνουν:

Περίπτωση 1. Σύγκριση κάθε συστάδας της Clustering A με κάθε μία της Clustering B.

Σε αυτή την περίπτωση γίνονται όλα τα ζευγάρια αντιστοίχησης και για κάθε συστάδα διατηρείται η καλύτερη αντιστοίχιση. Όλες οι τιμές αντιστοίχισης συναθροίζονται στο τέλος της διαδικασίας (με χρήση πιθανώς του μέσου).

Περίπτωση 2. Σύγκριση κάθε συστάδας της Clustering A με κάθε συστάδα της Clustering B αλλά χωρίς να επιτρέπονται διπλο-αντιστοιχίσεις.

Στην περίπτωση αυτή εάν η συστάδα Α.1 για παράδειγμα έχει την καλύτερη (υψηλότερη τιμή) αντιστοίχισης με την Β.1, αυτές οι συστάδες δε θα ξαναελεγχθούν. Στη συνέχεια, η συστάδα Α.2 θα ελεγχθεί μόνο με τις Β.2, Β.3 και Β.4.

Περίπτωση 2α. Αυτή η διαδικασία μπορεί να ακολουθείται από την αντίθετη φορά, που σημαίνει, να ελεγχθεί κάθε συστάδα της Clustering B για την καλύτερη αντιστοίχιση με τις συστάδες της Clustering A. Ένας πίνακας μπορεί να κατασκευαστεί και παραμένουν οι καλύτερες αντιστοιχίσεις.

Περίπτωση 2b. Είναι φανερό ότι σε περίπτωση που υπάρχουν δύο συσταδοποιήσεις με διαφορετικό αριθμό συστάδων (προτύπων), όπως στο παράδειγμά μας, δε θα αντιστοιχηθούν όλες οι συστάδες με κάποια άλλη συστάδα. Μία πιθανή λύση στην περίπτωση αυτή είναι να ταξινομηθούν οι δύο συσταδοποιήσεις από τη μεγαλύτερη προς το μικρότερη συστάδα, δεδομένου ότι οι μεγαλύτερες συστάδες θα αντιστοιχηθούν καλύτερα μεταξύ τους αφήνοντας έτσι τις μικρότερες συστάδες χωρίς αντιστοίχιση.

Περίπτωση 2c. Ένας άλλος τρόπος για να ξεπεραστεί αυτό το πρόβλημα, είναι να επιτραπούν σε ανάγκη οι διπλο-αντιστοιχίσεις, χωρίς όμως να υπολογίζονται όλοι οι πιθανοί συνδυασμοί – κάτι που θα αντιστοιχούσε στην *Περίπτωση1*.

Το ποιος είναι ο πιο κατάλληλος τύπος αντιστοίχισης εξαρτάται από την εφαρμογή και το χρήστη.

Οι έννοιες και η μεθοδολογία που παρουσιάστηκαν σε αυτή την ενότητα μπορούν να χρησιμοποιηθούν σε πολλές πραγματικές εφαρμογές για την κατηγοριοποίηση και τη σύγκριση προτύπων που εξάγονται από μία ποικιλία δεδομένων. Στις επόμενες ενότητες παρουσιάζονται δύο εφαρμογές για τη σύγκριση συσταδοποιήσεων που εξάγονται από εικόνες. Και στις δύο περιπτώσεις χρησιμοποιείται μία βάση με εικόνες και τεχνικές συσταδοποίησης

για την κατηγοριοποίηση των χαρακτηριστικών που εξάγονται από τις εικόνες αυτές. Ο σκοπός των δύο περιπτώσεων είναι να δημιουργηθεί μία μεθοδολογία ανάκτησης εικόνων με βάση το περιεχόμενο.

Η μεθοδολογία περιλαμβάνει τέσσερα βήματα: α) εξαγωγή χαρακτηριστικών από κάθε αποθηκευμένη εικόνα και της εικόνας επερώτησης, β) συσταδοποίηση των εξαγόμενων διανυσμάτων χαρακτηριστικών ανά εικόνα, γ) αναπαράσταση των προτύπων συστάδων, και δ) υπολογισμός των ομοιοτήτων μεταξύ των προτύπων. Η καταχώρηση μίας νέας εικόνας στη βάση εικόνων περιλαμβάνει τα πρώτα τρία βήματα της ανάκτησης (α, β, και γ).

Η πρώτη μελέτη (Iakovidis et al., 2007) χρησιμοποιεί εικόνες πολιτιστικής κληρονομιάς που προέρχονται από τη βάση δεδομένων του Ιδρύματος Μείζονος Ελληνισμού (FHW, 2009) ενώ η δεύτερη μελέτη (Iakovidis et al., 2006) χρησιμοποιεί ακτινολογικές εικόνες από τη βάση IRMA (Image Retrieval in Medical Applications) (Lehmann, 2003).

3.4 Εφαρμογή Ι: Σύγκριση συστάδων από ιατρικές εικόνες

Ένα από τα βασικότερα εργαλεία που χρησιμοποιούνται από ειδικούς ιατρούς είναι η σύγκριση προηγουμένων και νέων ιατρικών εικόνων που σχετίζονται με παθολογικές καταστάσεις. Λόγω της συνεχούς αύξησης της πληροφορίας σε μορφή εικόνων που αποθηκεύονται τόσο σε τοπικές όσο και δημοσίως διαθέσιμες βάσεις ιατρικών δεδομένων, η αποτελεσματική ευρετηριοποίηση και ανάκτησή της είναι απαραίτητη.

Την τελευταία δεκαετία οι εξελίξεις στην πληροφορική επέτρεψαν την ανάπτυξη συστημάτων Content-Based Image Retrieval (CBIR), ικανά να ανακτήσουν εικόνες βασισμένα στην ομοιότητα των χαρακτηριστικών τους με μία ή περισσότερες εικόνες επερώτησης. Μερικά από τα συστήματα αυτά είναι τα QBIC (Faloutsos et al., 1994), VisualSEEK (Smith & Chang, 1996), Virage (Hampapur et al., 1997), Netra (Ma & Manjunath, 1999), PicSOM (Laaksonen et al., 2000), SIMPLicity Wang et al., 2001), CIRES (Iqbal & Aggarwal, 2002), και FIRE (Deselaers et al., 2004). Περισσότερα από πενήντα CBIR συστήματα αναλύονται στο (Veltcamp & Tanase, 2000).

Τα πλεονεκτήματα της εφαρμογής προσεγγίσεων ανάκτησης ιατρικών εικόνων με βάση το περιεχόμενο ποικίλουν από την υποστήριξη κλινικών αποφάσεων έως την ιατρική εκπαίδευση και έρευνα (Müller et al., 2004). Τα

πλεονεκτήματα αυτά έχουν κινητοποιήσει τους ερευνητές στην εφαρμογή γενικής χρήσης συστημάτων CBIR ή στην ανάπτυξη εξειδικευμένων συστημάτων αποκλειστικά προσανατολισμένων σε συγκεκριμένους ιατρικούς τομείς.

Τα εξειδικευμένα CBIR συστήματα έχουν αναπτυχθεί για να υποστηρίζουν την ανάκτηση διαφόρων ειδών ιατρικών δεδομένων, συμπεριλαμβανομένων τομογραφικών εικόνων υψηλής ανάλυσης (High Resolution Computed Tomographic (HRCT) images) (Shyu et al., 1999), ακτινογραφίες βιοψίας καρκίνου του μαστού (Schnorrenberg et al., 2000), λειτουργικές εικόνες τομογραφιών εκπομπής ποζιτρονίων (Positron Emission Tomographic (PET) functional images) (Cai et al., 2000), εικόνες υπερήχων (Kwak, 2002), εικόνες παθολογίας (Zheng et al., 2003) και ακτινογραφικών εικόνων (El-Naqa et al., 2004). Κοινό στοιχείο για τα περισσότερα από τα συστήματα που αναφέρθηκαν είναι ότι η ανάκτηση των εικόνων βασίζεται σε μέτρα ομοιότητας που υπολογίζονται κατευθείαν από τα χαμηλού επιπέδου χαρακτηριστικά της εικόνας. Αυτή η προσέγγιση μπορεί να οδηγήσει σε ανάκτηση εικόνων με αρκετές διαφορές από την εικόνα-επερώτησης, αφού τα χαμηλού επιπέδου χαρακτηριστικά συνήθως δεν έχουν σημασιολογία. Αυτό έχει δώσει το κίνητρο στους ερευνητές να επικεντρωθούν στη χρήση υψηλότερου επιπέδου σημασιολογικών αναπαραστάσεων του περιεχομένου της εικόνας για την ανάκτηση εικόνας με βάση το περιεχόμενο.

Πρόσφατες προσεγγίσεις περιλαμβάνουν τη σημασιολογική αντιστοίχιση μέσω υβριδικών Bayesian δικτύων, Semantic Error-Correcting output Codes (SECC) που βασίζονται σε συνδυασμούς ανεξάρτητων κατηγοριοποιητών (Yaoa et al., 2006), και ένα πλαίσιο που χρησιμοποιεί μηχανική μάθηση και στατιστικές τεχνικές ταιριάσματος με σχετική ανάδραση (Rahman et al., 2007). Εντούτοις, αυτές οι προσεγγίσεις περιλαμβάνουν εποπτευόμενες (supervised) μεθοδολογίες που απαιτούν προηγούμενη γνώση σχετικά με τα δεδομένα και εισάγουν περιορισμούς για τη σημασιολογία που απαιτείται για την διαδικασία της ανάκτησης της εικόνας.

Μία state-of-the-art CBIR προσέγγιση έχει παρουσιαστεί στο (Greenspan & Pinhas, 2007). Χρησιμοποιεί ένα συνεχές και πιθανοτικό (probabilistic) σχήμα αναπαράστασης εικόνας που σχετίζεται με Gaussian mixture modelling (GMM) και με ταίριασμα εικόνας με βάση τη θεωρία της πληροφορίας μέσω του μέτρου Kullback–Leibler (KL). Τα αποτελέσματα που αναφέρονται στο (Greenspan &

Pinhas., 2007) δείχνουν ότι αυτή η προσέγγιση είναι πολύ αποτελεσματική στην ανάκτηση ραδιογραφικών εικόνων (ακτινογραφιών). Εντούτοις, η αποτελεσματικότητα για μεγάλης κλίμακας διαδικασίες ανάκτησης εικόνων παραμένει μία πρόκληση.

Στην παρούσα μελέτη, προτείνουμε μία μη-εποπτευόμενη προσέγγιση για την αποτελεσματική ανάκτηση ιατρικών εικόνων που βασίζεται σε μέτρα ομοιότητας που ορίζονται πάνω στα υψηλότερου επιπέδου πρότυπα που σχετίζονται με συστάδες πάνω στα χαμηλού επιπέδου χαρακτηριστικά των εικόνων. Η προτεινόμενη προσέγγιση συνδυάζει τα πλεονεκτήματα των CBIR μεθοδολογιών που βασίζονται στη συσταδοποίηση (Stehling et al., 2001; Carson et al., 2002; Yixin Chen et al., 2005) με μία σημασιολογικά πλούσια αναπαράσταση των ιατρικών εικόνων. Επιπλέον, αντίθετα με παρόμοιες CBIR προσεγγίσεις που εκμεταλλεύονται πολυδιάστατες τεχνικές ευρετηριοποίησης, όπως είναι τα *R*-trees (Faloutsos et al., 1994), (Petrakis and Faloutsos, 1997), iconic index trees (Wu & Narasimhalu, 1994), και τα meshes of trees (Jeng & Hsiao, 2005), η αποτελεσματικότητα της προτεινόμενης προσέγγισης επηρεάζεται δύσκολα από την αύξηση της διάστασης των χαμηλού επιπέδου χαρακτηριστικά.

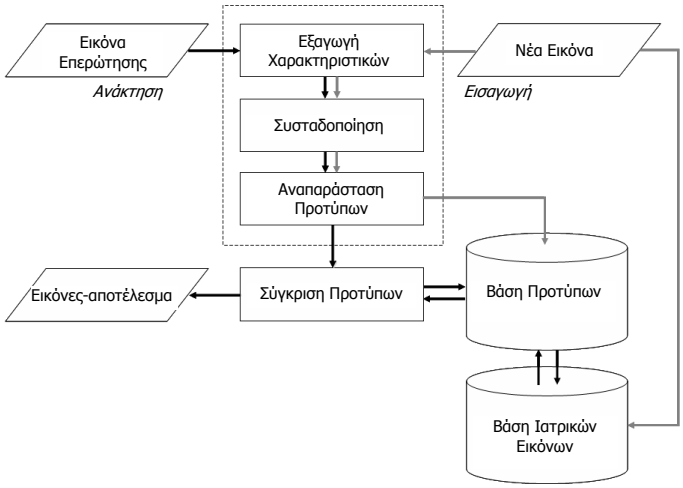
Στη συνέχεια παρουσιάζουμε την πρώτη εφαρμογή σύγκρισης προτύπων συστάδων (πρότυπα που προέρχονται από ιατρικές εικόνες), οι κύριες συνεισφορές της οποίας είναι οι ακόλουθες:

- Ορίζουμε μία πρωτότυπη αναπαράσταση ιατρικών εικόνων, ως πλούσια σημασιολογικά *σύνθετα πρότυπα*. Κάθε σύνθετο πρότυπο αποτελείται από ένα σύνολο απλών προτύπων που αναπαριστούν συστάδες περιοχών της εικόνας που σχετίζονται με ανατομικά δείγματα, με μη-εποπτευόμενο τρόπο. Η αναπαράσταση των προτύπων βασίζεται σε στοιχεία της δομής τους αλλά και σε μέτρα ποιότητας.
- Προτείνουμε ένα πρωτότυπο σχήμα για την αποτίμηση της ομοιότητας μεταξύ σύνθετων προτύπων (των ιατρικών εικόνων δηλαδή) για σκοπούς ανάκτησης των εικόνων με βάση το περιεχόμενό τους.
- Παρουσιάζουμε ένα περιεκτικό σύνολο πειραμάτων σε μία δημοσίως διαθέσιμη βιβλιοθήκη ραδιογραφικών εικόνων (ακτινογραφιών), προκειμένου να αποτιμήσουμε την προσέγγισή μας και να δείξουμε την

αποτελεσματικότητα και την αποδοτικότητά της σε σύγκριση με τεχνικές state-of-the-art.

3.4.1 Η προτεινόμενη μεθοδολογία

Το προτεινόμενο σχήμα ανάκτησης ιατρικών εικόνων με βάση το περιεχόμενο παρουσιάζεται στην Εικόνα 3-3. Περιλαμβάνει τέσσερα βήματα: α) εξαγωγή των χαμηλού επιπέδου χαρακτηριστικών από κάθε εικόνα στη βάση και από την εικόνα-επερώτηση, β) συσταδοποίηση των διανυσμάτων χαρακτηριστικών ανά εικόνα, γ) αναπαράσταση των προτύπων από τις εξαγόμενες συστάδες, και δ) υπολογισμός των ομοιοτήτων μεταξύ των προτύπων. Η εισαγωγή και καταχώρηση μίας νέας εικόνας στη βάση δεδομένων περιλαμβάνει τα βήματα α, β, και γ, ενώ το βήμα δ πραγματοποιείται κατά τη διαδικασία της ανάκτησης.



Εικόνα 3-3 Περίγραμμα της προτεινόμενης μεθοδολογίας ανάκτησης εικόνας με βάση το περιεχόμενο. Τα σκούρα μαύρα βέλη δείχνουν την ροή των δεδομένων για την ανάκτηση εικόνων, ενώ τα γκρί βέλη δείχνουν τη ροή δεδομένων για την καταχώρηση μίας νέας εικόνας.

Στην περίπτωση της ανάκτησης ακτινογραφιών, η ένταση του τοπικού επιπέδου του γκρι και τα χαρακτηριστικά υφής έχει αποδειχθεί να διακρίνουν καλύτερα τα απεικονιζόμενα αντικείμενα. Τέτοια χαρακτηριστικά είναι οι τιμές των pixel που χρησιμοποιούνται με ένα μοντέλο ομοιότητας διαστρέβλωσης της εικόνας (image distortion similarity model), τα ιστογράμματα τοπικών χαρακτηριστικών (local feature histograms) και τα τοπικά σχετικά χαρακτηριστικά (local relational features) (Deselaers et al., 2004b), (Setia et al., 2006). Πρόσφατα στο (Greenspan and Pinhas, 2007) παρουσιάστηκε ότι η μεγαλύτερη ακρίβεια ανάκτησης μπορεί να επιτευχθεί με το συνδυασμό της έντασης και της

αντίθεσης της υφής μαζί με τις αντίστοιχες χωρικές συντεταγμένες. Εντούτοις, η εισαγωγή χωρικής πληροφορίας στα διανύσματα χαρακτηριστικών τα καθιστά εξαρτώμενα από τη θέση των ασθενών που έκαναν τις ακτινογραφίες. Αν και συνήθως η στάση/θέση των ασθενών είναι καθορισμένη για μία ακτινογραφία, υπάρχουν πολλές περιπτώσεις όπου αυτό δεν είναι πρακτικά εφικτό. Για παράδειγμα όταν η ακτινογραφία γίνεται σε ιδιαίτερα άρρωστα άτομα με φορητές συσκευές (Bongard and Sue, 2002) και με στις ακτινογραφίες άνω και κάτω άκρων (Karkanis et al., 2003).

Στην παρούσα μελέτη, υιοθετούμε μία γνωστή-Multiscale στατιστική προσέγγιση για την αναπαράσταση των περιοχών των ακτινογραφιών που διατηρούν τα τοπικά χαρακτηριστικά και δεν εξαρτώνται από τις χωρικές συντεταγμένες. Βασίζεται σε ένα διδιάστατο Discrete Wavelet Transform (2D-DWT), έναν αποτελεσματικό και αποδοτικό μετασχηματισμό που έχει αποδειχθεί χρήσιμος σε μία ποικιλία εφαρμογών επεξεργασίας και ανάλυσης ιατρικών εικόνων, συμπεριλαμβανομένων των CBIR (Müller et al., 2004), (Mallat, 1999, Wang et al., 1998, Karkanis et al., 2003, Wang, 2001). Καθιστά δυνατή την κωδικοποίηση της υφής της εικόνας σε λεπτομερείς (υψηλότερης συχνότητας - higher frequency) συντελεστές, ενώ η πληροφορία για την ένταση της εικόνας μπορεί να εξαχθεί από τους συντελεστές προσέγγισης (lower frequency) (Mallat, 1999). Μία συμπαγής αναπαράσταση των κατανομών των προσεγγιστικών και των λεπτομερών συντελεστών μπορεί να παρασχεθεί από την πρωτοβάθμια στατιστική προσέγγιση.

Παρόλα αυτά πρέπει να σημειωθεί ότι η παρούσα μελέτη εστιάζει στη χρησιμότητα του προτεινόμενου σχήματος ομοιότητας προτύπων αντί στην επιλογή των καλύτερων χαρακτηριστικών για μία συγκεκριμένη διαδικασία ανάκτησης εικόνας.

Σύμφωνα με το σχήμα αναπαράστασης προτύπων που παρουσιάστηκε στην ενότητα 3.3, στην τρέχουσα εφαρμογή, ένα $Specimen_i$ αναπαριστά κάθε πρότυπο P_i που είναι ένα φυσικό ανατομικό κομμάτι σε μία ιατρική εικόνα:

$$Specimen_i = \begin{pmatrix} SS:\left(D:[[\mu:[\text{Real}],\sigma:[\text{Real}]]_1^N\right) \\ MS:\left(pp:[\text{Real}],\ SV:[\text{Real}]\right) \end{pmatrix} \tag{3-13}$$

Σύμφωνα με αυτά, μία ιατρική εικόνα MI είναι ένα σύνθετο πρότυπο:

$$MI = \begin{pmatrix} SS : \{Specimen\}, \\ MS : \perp \end{pmatrix} \tag{3-14}$$

που αποτελείται από ένα σύνολο απλών προτύπων (δηλ. *specimens*).

Τα σχήματα υπολογισμού της συσταδοποίησης και ομοιότητας των προτύπων ακολουθούν τα σχήματα που περιγράφονται στην ενότητα 3.3.

Στην επόμενη ενότητα παρουσιάζονται τα αποτελέσματα της πειραματικής μελέτης.

3.4.2 Πειραματικά αποτελέσματα

Ένα πλήθος από πειράματα διεξήχθησαν με εικόνες ακτινογραφιών από το σύνολο δεδομένων IRMA (Image Retrieval in Medical Applications) (Lehmann, 2004), το οποίο συχνά χρησιμοποιείται ως αναφορά σε διαδικασίες ανάκτησης ιατρικών εικόνων. Περιέχει 10.000 ανώνυμες ακτινογραφίες που έχουν τυχαία επιλεχθεί από ασθενείς διαφόρων ηλικιών, φύλων και παθολογίας. Οι εικόνες έχουν κατηγοριοποιηθεί σε 116 κλάσεις σύμφωνα με τον κώδικα του IRMA (Lehmann et al., 2003).

Όλες οι ακτινογραφίες είναι σε 8-bit greyscale format και έχουν υποβαθμιστεί ώστε να «χωράνε» σε ένα 256×256-pixel bounding box διατηρώντας τις αρχικές αναλογίες της εικόνας. Από το διαθέσιμο σύνολο δεδομένων το 90% των εικόνων καταχωρίστηκαν στη βάση ενώ ένα υποσύνολο του 10% των εικόνων χρησιμοποιήθηκαν για επερωτήσεις στην βάση προτύπων. Κάθε εικόνα χωρίστηκε σε blocks με επικαλυπτόμενα sliding windows. Οι λεπτομέρειες της μεθόδου εξαγωγής των προτύπων περιλαμβάνουν μία τριών επιπέδων αποσύνθεση διπλά ορθογωνίων κυματιδίων (biorthogonal spline wavelet decomposition) κάθε block και την εκτίμηση των δύο πρώτων wavelet moments για κάθε μπάντα. Η διαδικασία αυτή έχει ως αποτέλεσμα έναν 20-διάστατο διάνυσμα χαρακτηριστικών για κάθε block.

Ο καθορισμός των παραμέτρων δειγματοληψίας βασίστηκε σε προκαταρκτικά πειράματα για την αναζήτηση της μεγαλύτερης μέσης απόστασης μεταξύ σύνθετων προτύπων *MI* από των διαφορετικών κατηγοριών που έχουν καταχωριστεί στη βάση. Οι παράμετροι δειγματοληψίας που δοκιμάστηκαν πριν από κάθε πείραμα CBIR περιλαμβάνουν sliding windows των 32×32, 64×64 και 128×128 pixels. Σε όλες τις περιπτώσεις, η μέγιστη μέση απόσταση

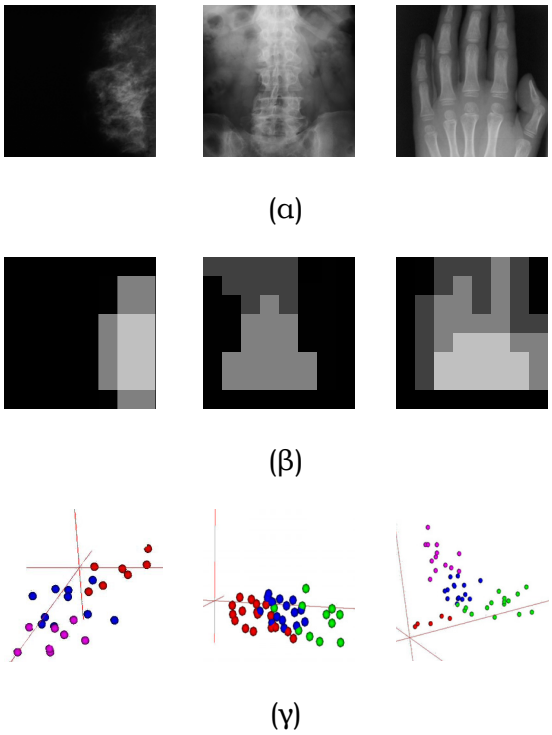
βρέθηκε με τα παράθυρα μεγέθους 64×64-pixels. Οι διαφοροποιήσεις στην επικάλυψη (0%, 25%, 50% και 75%) των block δειγματοληψίας δεν επηρέασε το αποτέλεσμα. Η αύξηση της επικάλυψης παρέχει καλύτερο εντοπισμό των προτύπων αλλά δημιουργεί πολλά ακόμα blocks δειγματοληψίας, επηρεάζοντας την αποτελεσματικότητα τόσο της εξαγωγής χαρακτηριστικών, όσο και τη διαδικασία της αναπαράστασης των προτύπων. Έτσι, το 50% της επικάλυψης, δηλαδή ένα βήμα των 32-pixel, χρησιμοποιήθηκε σαν ένας συμβιβασμός μεταξύ του εντοπισμού και της αποτελεσματικότητας.

Στη συνέχεια παρουσιάζουμε τα αποτελέσματα της αναπαράστασης των προτύπων μέσω συσταδοποίησης και μετρούμε την απόδοση του προτεινόμενου σχήματος σε όρους αποδοτικότητας και αποτελεσματικότητας.

Αναπαράσταση και δημιουργία προτύπων από τη συσταδοποίηση

Τα διανύσματα χαρακτηριστικών που εξάγονται από κάθε εικόνα συσταδοποιούνται χρησιμοποιώντας μία υλοποίηση του αλγορίθμου EM που είναι διαθέσιμη στο εργαλείο εξόρυξης γνώσης WEKA και χρησιμοποιώντας την τεχνική 10-fold cross-validation για να οριστεί ο αριθμός των συστάδων. Κάθε μία συστάδα αναπαρίσταται από ένα πρότυπο $Specimen_i$, $i = 1, ..., M$ (Εξίσωση (3-13)), και κάθε εικόνα αναπαρίσταται με ένα σύνθετο πρότυπο *MI* (Εξίσωση (3-14)). Στην Εικόνα 3-4α παρουσιάζονται τρεις ακτινογραφίες από θώρακα, κοιλιακή χώρα και χέρι (αριστερά προς τα δεξιά). Οι αντίστοιχες συστάδες που προέκυψαν από τις ακτινογραφίες αυτές φαίνονται στην Εικόνα 3-4b. Τα διαφορετικά επίπεδα του γκρι στην Εικόνα 3-4β δείχνουν τα διάφορα πρότυπα *specimen* που βρέθηκαν στις εικόνες. Η Εικόνα 3-4γ παρουσιάζει τις προβολές του 20-διάστατου διανύσματος χαρακτηριστικών σε ένα 3-διάστατο χώρο που κατασκευάστηκε σύμφωνα με την τεχνική προβολής διατήρησης του κέντρου (centroid-preserving) (Kopanakis and Theodoulidis, 2003).

Παρατηρείται ότι η συσταδοποίηση που προκύπτει είναι ιδιαίτερα ουσιώδης, αφού οι περιοχές του κάθε μέρους της ακτινογραφίας ξεχωρίζουν ιδιαίτερα. Ωστόσο, στα δάκτυλα ο αλγόριθμος ανάθεσε δύο δείγματα αντί για ένα, αλλά αυτό μπορεί να αποδοθεί στο μεγάλο μέγεθος του block δειγματοληψίας και σε σχέση με το κενό μεταξύ των δακτύλων.



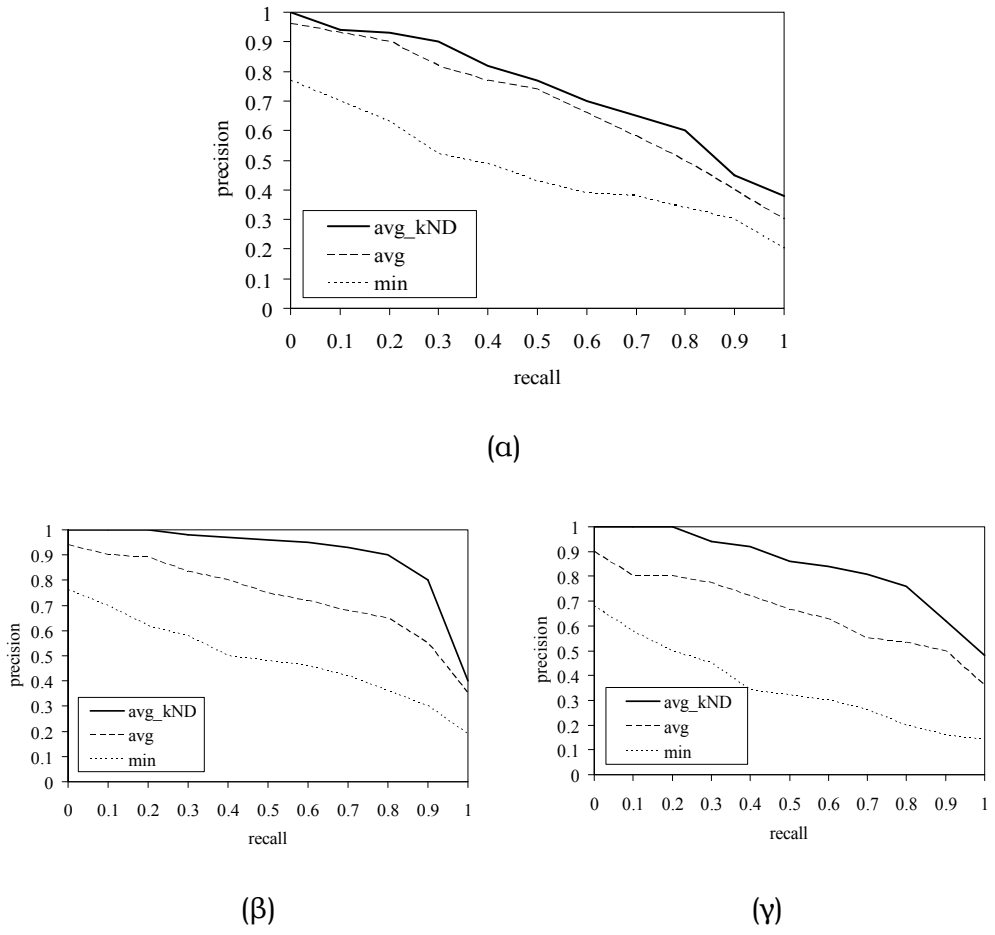
Εικόνα 3-4 (α) Αρχικές ακτινογραφίες, (β) Αποτέλεσμα συσταδοποίησης και (γ) 3-διάστατη απεικόνιση του χώρου των χαρακτηριστικών.

Αποδοτικότητα

Τα πρότυπα από τις καταχωρισμένες ακτινογραφίες χρησιμοποιούνται για να δημιουργήσουν μία βάση προτύπων (βλ. Εικόνα 3-3). Προκειμένου να αποτιμηθεί η αποδοτικότητα του προτεινόμενου σχήματος ομοιότητας προτύπων, ελέγχουμε τη δυνατότητά του να ανακτά εικόνες με τη χρήση των δημοφιλών μέτρων *recall* και *precision*, όπου recall ορίζεται ο λόγος των σχετικών (όμοιων) εικόνων που ανακτήθηκαν προς το συνολικό αριθμό σχετικών εικόνων στη βάση και precision ορίζεται ως ο λόγος των σχετικών εικόνων που ανακτήθηκαν προς το σύνολο των εικόνων που ανακτήθηκαν, σχετικές ή μη. Για τη σύγκριση με άλλα συστήματα/μεθοδολογίες ανάκτησης ιατρικών εικόνων γίνεται η χρήση ενός ενιαίου μέτρου σύγκρισης, του «Area Under the interpolated precision-recall Curve (AUC)» (Davis and Goadrich, 2006), δηλαδή του εμβαδού της περιοχής που σχηματίζεται από την καμπύλη precision-recall και τους 2 άξονες, presicion και recall.

Η προτεινόμενη μεθοδολογία ελέγχθηκε με χρήση των τριών συναθροιστικών συναρτήσεων g_{aggr}. Το αποτέλεσμα σε όρους μέσης precision vs. recall που υπολογίστηκε για όλες τις 116 κατηγορίες, φαίνονται στην Εικόνα 3-5α. Στην Εικόνα 3-5β και την Εικόνα 3-5γ παρουσιάζονται τα διαγράμματα precision vs.

recall για δύο ανεξάρτητες κατηγορίες του θώρακα και του κρανίου. Είναι φανερό ότι οι καλύτερη ανάκτηση γίνεται με χρήση της συνάρτησης *μέσος όρος των k κοντινότερων κατανομών* σ_{avg_kND}.



(α)

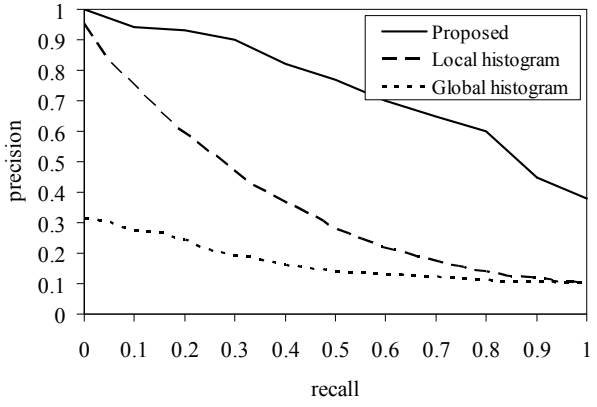
(β) (γ)

Εικόνα 3-5 Μέση precision vs. recall με χρήση των συναθροιστικών συναρτήσεων g_{avg_kND}, g_{avg} και g_{min} για την κατηγορία (α) όλα, (β) θώρακα, και (γ) κρανίου.

Η Εικόνα 3-5a δείχνει ότι για μία recall 90% η μέση precision που επιτυγχάνεται με χρήση της σ_{avg_kND} είναι σχεδόν 45% και το αντίστοιχο AUC είναι 74%. Αξίζει να σημειωθεί ότι αυτά τα αποτελέσματα μπορούν μόνο ελάχιστα να βελτιωθούν με πιο πυκνό σχήμα δειγματοληψίας. Σε σύγκριση με μία απλή μέθοδο που χρησιμοποιεί καθολικά ιστογράμματα επιπέδου του γκρι σαν χαρακτηριστικά και τη διατομή των ιστογραμμάτων σαν μέτρο απόστασης (Swain, M.J. and Ballard, 1991), η μέση precision για 90% recall είναι περίπου 10% και το αντίστοιχο AUC φτάνει μόνο το 17%. Το AUC που προέρχεται από το προτεινόμενο σχήμα με χρήση πληροφορίας ιστογράμματος

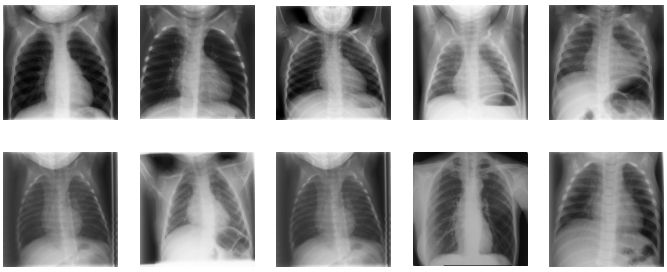
επιπέδου του γκρί φτάνει το 34%. Η αντίστοιχες καμπύλες precision vs. recall φαίνονται στην Εικόνα 3-6.

Η precision που αναφέρεται στο (Greenspan and Pinhas, 2007) για 90% recall φαίνεται να είναι συμβατή με αυτή στην προτεινόμενη προσέγγιση. Εντούτοις, το σύνολο δεδομένων από όπου υπολογίζεται είναι σημαντικά μικρότερο και αποτελείται από μόνο 1.500 ακτινογραφίες από 17 κατηγορίες. Προκειμένου να εξάγουμε συμβατά αποτελέσματα μεταξύ των δύο προσεγγίσεων έγινε ένα πείραμα ανάκτησης με το προτεινόμενο σχήμα σε ένα υποσύνολο των διαθέσιμων δεδομένων το οποίο δημιουργήθηκε σύμφωνα με τις οδηγίες που παρέχονται στο (Greenspan and Pinhas, 2007). Το AUC που υπολογίστηκε για την προτεινόμενη προσέγγιση στο υποσύνολο αυτό έφτασε το 78%, ενώ το αντίστοιχο από το (Greenspan and Pinhas, 2007) είναι περίπου 66%.

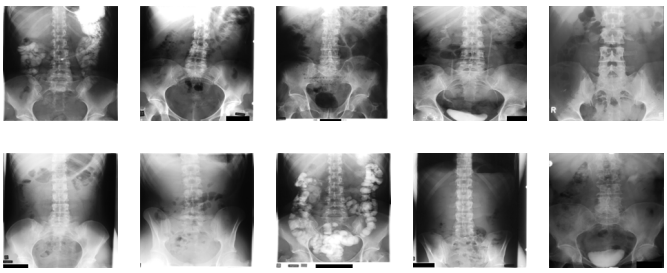


Εικόνα 3-6 Συγκριτικό διάγραμμα precision vs. recall.

Δύο παραδείγματα ανάκτησης με τη συνάρτηση $g_{\text{avg\_kND}}$ παρουσιάζονται στην Εικόνα 3-7. Η πρώτη εικόνα από κάθε ακολουθία είναι η εικόνα-επερώτησης και οι υπόλοιπες είναι οι εννέα εικόνες που ανακτήθηκαν. Η Εικόνα 3-7α δείχνει ότι όλες οι εικόνες που ανακτήθηκαν ανήκουν στην ίδια κατηγορία. Η Εικόνα 3-7β δείχνει ότι δύο από τις εικόνες που ανακτήθηκαν ανήκουν σε διαφορετική κλάση από την εικόνα-επερώτηση. Εντούτοις, η βασική διαφορά μεταξύ των δύο κατηγοριών είναι δύσκολο να διακριθεί και βρίσκεται στην περιοχή της λεκάνης (στο κάτω μέρος της κεντρικής εικόνας). Παρόμοιες παρατηρήσεις γίνονται και σε επερωτήσεις που γίνονται σε εικόνες από άλλες κατηγορίες.



(α)



(β)

Εικόνα 3-7 (α) μία επερώτηση για εννέα εικόνες θώρακα όμοιες με την πρώτη επάνω αριστερά επάνω. (1,1): όλες οι εικόνες που ανακτήθηκαν ανήκουν στην ίδια κατηγορία (β) μία επερώτηση για εννέα εικόνες της κοιλιακής/στομαχικής χώρας όμοιες με την επάνω αριστερή εικόνα (1,1): όλες οι εικόνες που ανακτήθηκαν ανήκουν στην ίδια κατηγορία εκτός από τις (1,4) και (2,5), που ανήκουν στην κατηγορία της κοιλιακής χώρας/ουροποιητικού συστήματος. (Η σημειολογία (*i*, *j*) δείχνει τη θέση της εικόνας στην *i* γραμμή, και την *j* στήλη στην εικόνα.)

Αποτελεσματικότητα

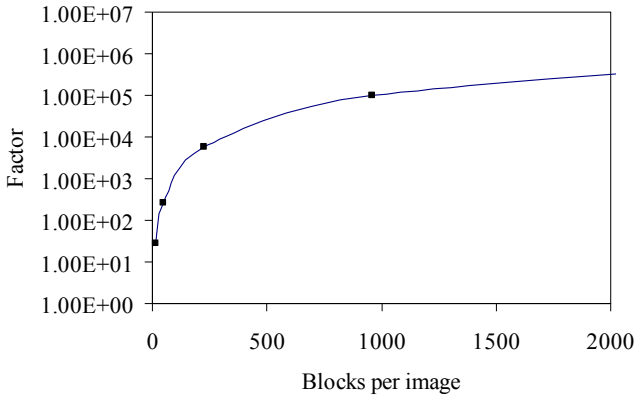
Σε αυτή την υποενότητα μετράμε την αποτελεσματικότητα της προτεινόμενης μεθόδου ανάκτησης ιατρικών εικόνων, σε σχέση με την απόδοση της συμβατικής μεθόδου, μέσω εξαντλητικής σύγκρισης όλων των διανυσμάτων χαρακτηριστικών.

Η συμβατική σύγκριση των διανυσμάτων είναι αντίστοιχη με τη σύγκριση των προτύπων στο προτεινόμενο σχήμα. Τα πειράματα έγιναν σε υπολογιστή Intel Pentium M1.6, με 1 GB RAM και σκληρό δίσκο 60 GB.

Έχουμε επιλέξει την σειριακή, εξαντλητική μέθοδο ως βάση για τη μέτρηση της μεθόδου μας, αφού άλλες κοινές μέθοδοι όπως είναι τα R-trees είναι ευαίσθητα σε μεγάλες διαστάσεις των διανυσμάτων χαρακτηριστικών, κάτι που όμως είναι σύνηθες στα CBIR συστήματα (π.χ. διάσταση της τάξης των 64 στο (Weber, 1998) και τουλάχιστον $2\times N=40$ στην περίπτωσή μας, όπου N είναι ο αριθμός των χαρακτηριστικών σε ένα πρότυπο). Η απόδοση τέτοιων προσεγγίσεων μειώνεται ταχέως όσο αυξάνει η διάσταση. Για παράδειγμα έχει δειχθεί ότι

ακόμα και για την μικρή διάσταση των 5, τα ευρετήρια τύπου R*-tree συμπεριφέρονται προβληματικά στην αναζήτηση ομοιότητας (Weber, 1998). Ο κύριος λόγος είναι ότι με την αύξηση της διάστασης η επικάλυψη στους εσωτερικούς κόμβους του δέντρου αυξάνεται και έτσι η ικανότητα διακριτοποίησης μειώνεται.

Ο παράγοντας αύξησης της ταχύτητας μεταξύ της συμβατικής και της προτεινόμενης προσέγγισης ως συνάρτηση του αριθμού των block ανά εικόνα παρουσιάζεται στην Εικόνα 3-8. Μπορεί να παρατηρηθεί ότι το πλεονέκτημα της προτεινόμενης προσέγγισης αυξάνει με τον αριθμό των block ανά εικόνα (δηλ. με την αύξηση του βήματος δειγματοληψίας), και για μερικές εκατοντάδες block ανά εικόνα απαιτείται τουλάχιστον 3 φορές λιγότερες συγκρίσεις από την συμβατική προσέγγιση.



Εικόνα 3-8 ο παράγοντας αύξησης μεταξύ της συμβατικής και της προτεινόμενης προσέγγισης ως συνάρτηση του αριθμού των block ανά εικόνα.

Από την άλλη μεριά, στο (Greenspan and Pinhas, 2007) αναφέρεται ένας παράγοντας αύξησης της τάξης του δύο σε σύγκριση με τη συμβατική προσέγγιση. Επιπλέον, στην ίδια μελέτη σημειώνεται ότι το πλαίσιο GMM-KL δεν είναι ικανό για να ανταποκριθεί σε ένα μεγάλο αριθμό διαδικασιών ανάκτησης εικόνων που ξεπερνά τις 6.000 εικόνες λόγω της αυξημένης υπολογιστικής απαίτησης που σχετίζεται με το μέτρο KL. Επιπλέον υπολογίζουμε το μέσο χρόνο επεξεργασίας (χρόνο CPU συν I/O) για τη σύγκριση ενός ζευγαριού εικόνων. Για το παραπάνω πείραμα το προτεινόμενο σχήμα ομοιότητας απαιτεί πάντα λιγότερο από 0.1 msec. Ο μέσος χρόνος που απαιτείται για να συγκλίνουν οι παράμεροι του mixture model σε μία σταθερά είναι σχεδόν σταθερός στα 0.22±0.04 sec.

3.5 Εφαρμογή ΙΙ: Σύγκριση συστάδων από εικόνες πολιτιστικής κληρονομιάς

Τα συστήματα Content-Based Image Retrieval (CBIR) για εικόνες πολιτιστικής κληρονομιάς είναι ένα αναπτυσσόμενο πεδίο έρευνας που απαιτεί συνεργασία του πολιτισμού και της τεχνολογίας (Chen et al., 2004). Η επερώτηση με βάση το παράδειγμα σε βάσεις δεδομένων (querying by example databases) πινάκων, γλυπτών, φωτογραφιών και κειμένων ιστορικής αξίας από διαφορετικούς πολιτισμούς, θα διευκόλυνε τόσο την εκπαίδευση όσο και την έρευνα και θα έκανε εφικτή την διερεύνηση εσω-πολιτισμικών και δια-πολιτισμικών εφαρμογών.

Προσφάτως, παρουσιάστηκαν μελέτες που στοχεύουν ειδικά στην ανάκτηση εικόνων πολιτιστικής κληρονομιάς. Οι περισσότερες από τις μελέτες αυτές βασίζονται σε χαρακτηριστικά χρώματος των εικόνων (Ardizzone et al., 2004), (Valle et al., 2006). Περισσότερο έξυπνες προσεγγίσεις περιλαμβάνουν τη χρήση των περιγραφέων κυματιδιακών παραμέτρων ανάλυσης (*wavelet domain feature descriptors)* σε συνδυασμό με τα mixtures of stochastic models για την ανάκτηση εικόνων Κινέζικης ζωγραφικής (Jia & Wang, 2004).

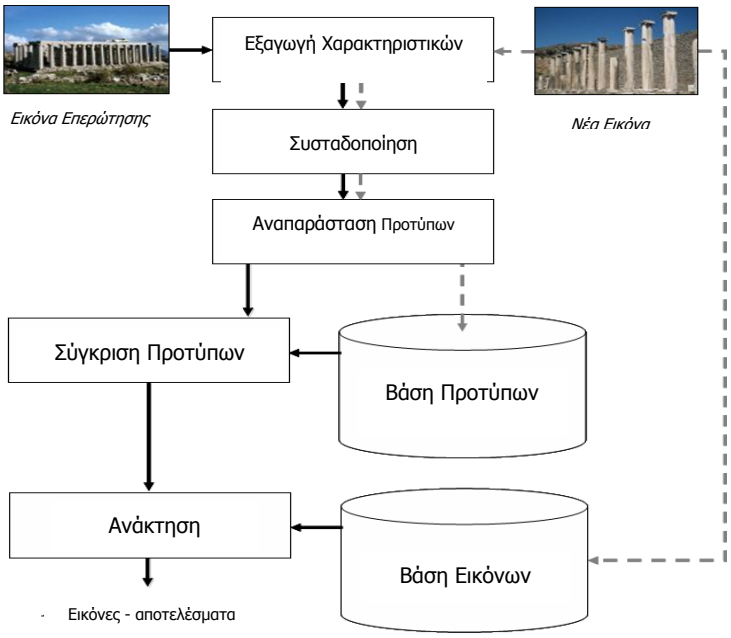
Κοινός τόπος για τα περισσότερα από τα συστήματα που αναφέρθηκαν παραπάνω είναι ότι η ανάκτηση βασίζεται σε μέτρα ομοιότητας που υπολογίζονται απευθείας πάνω στα χαρακτηριστικά χαμηλού επιπέδου των εικόνων ενώ χρησιμοποιούν πολυδιάστατες, συνήθως εξαντλητικές αναζητήσεις κοντινότερου γείτονα σε όλο το σύνολο των διαθέσιμων διανυσμάτων χαρακτηριστικών. Εντούτοις μία τέτοια προσέγγιση είναι ιδιαίτερα χρονοβόρα σε μεγάλες βάσεις εικόνων.

Όπως έχει προαναφερθεί η έρευνα στη βελτίωση της αποδοτικότητας της διαδικασίας ανάκτησης εικόνας έχει επικεντρωθεί κυρίως σε τεχνικές ευρετηριοποίησης των εικόνων με τη χρήση δομών δεδομένων, όπως τα R-trees (Faloutsos et al., 1994), (Petrakis & Faloutsos, 1997), δενδρικά ευρετήρια χαρακτηριστικών - feature index trees (Grosky & Mehrotra, 1990), εικονικά δέντρα ευρετηρίου - iconic index trees (Wu & Narasimhalu, 1994), και πλέγματα δέντρων - meshes of trees (Jeng & Hsiao, 2005). Άλλες προσεγγίσεις για την βελτίωση της αποδοτικότητας σχετίζονται με τη συσταδοποίηση των χαρακτηριστικών της εικόνας (Stehling et al. 2001), (Zhang R. & Zhang Z, 2002), και τη χρήση εναλλακτικών μέτρων ομοιότητας, συνήθως εξαρτώμενα

από σύνολα χαρακτηριστικών (Berman & Shapiro, 1997), (Stehling et al., 2002).

3.5.1 Η προτεινόμενη μεθοδολογία

Η προτεινόμενη προσέγγιση CBIR φαίνεται στην Εικόνα 3-9. Αποτελείται από τέσσερα βήματα: α) εξαγωγή χαρακτηριστικών από κάθε αποθηκευμένη εικόνα και της εικόνας-επερώτηση, β) συσταδοποίηση των διανυσμάτων χαρακτηριστικών ανά εικόνα, γ) αναπαράσταση των προτύπων (συστάδων) και δ) υπολογισμός των ομοιοτήτων των προτύπων. Η καταχώρηση μίας νέας εικόνας στη βάση απαιτεί τα τρία πρώτα βήματα της ανάκτησης (α, β, και γ).



Εικόνα 3-9 Περίγραμμα της προτεινόμενης προσέγγισης CBIR συστήματος βασισμένου στα πρότυπα. Τα συμπαγή βέλη δείχνουν τη ροή των δεδομένων για την ανάκτηση εικόνας ενώ τα διακεκομμένα βέλη δείχνουν τη ροή των δεδομένων για την καταχώρηση μίας νέας εικόνας.

Κάθε εικόνα που αποθηκεύεται στη βάση, καθώς και η εικόνα-επερώτησης, σκανάρονται με ένα κυλιόμενο παράθυρο καθορισμένου από το χρήστη μεγέθους και βήματος.

Το βήμα μπορεί να επιτρέπει την επικάλυψη των παραθύρων. Για κάθε παράθυρο N χαρακτηριστικά f_i, i=1,2,..,N υπολογίζονται για να σχηματίσουν ένα διάνυσμα χαρακτηριστικών F. Ο αριθμός των διανυσμάτων χαρακτηριστικών που παράγεται για κάθε εικόνα, εξαρτάται από το μέγεθος, τις διαστάσεις και το βήμα του κυλιόμενου παραθύρου.

Λόγω της φύσης των δεδομένων χρησιμοποιούμε τις Local Binary Pattern (LBP) κατανομές για χαρακτηριστικά. Τα LBP χαρακτηριστικά υπολογίζονται από το βεβαρημένο μέσο των τιμών των pixels σε μία μικρή περιοχή, στην οποία κάθε pixel εκλαμβάνεται ξεχωριστά.

Επιπλέον των χαρακτηριστικών αυτών, ένα ανεξάρτητο μέτρο τοπικής αντίθεσης (contrast) χρησιμοποιείται, σύμφωνα με το οποίο ο μέσος όρος των επιπέδων του γκρι που βρίσκονται κάτω από το κεντρικό pixel αφαιρείται από αυτό των επιπέδων του γκρι που βρίσκονται πάνω από το κεντρικό pixel (Ojala et al., 1996). Συγκριτικές μελέτες έχουν δείξει ότι η χρήση των χαρακτηριστικών LBP με κατανομές αντίθεσης μπορεί να αποδώσει μεγαλύτερη ακρίβεια κατηγοριοποίησης από τα χαρακτηριστικά Gabor και wavelet, με μικρότερο υπολογιστικό κόστος (Maenpaa & Pietikäinen, 2004), (Iakovidis et al., 2005).

Μετά την εξαγωγή των χαρακτηριστικών γίνεται η συσταδοποίηση με τη χρήση του αλγορίθμου Expectation Minimization (EM) (Dempster et al., 1997).

Σύμφωνα με το σχήμα που παρουσιάστηκε στην ενότητα 3.3, δοθείσας μίας συσταδοποίησης με M συστάδες C_i, i=1,2,..,M, μιας εικόνας ένα πρότυπο $object_i$ δημιουργείται για κάθε συστάδα C_i αναπαριστώντας ένα αντικείμενο της εικόνας:

$$object_i = \begin{pmatrix} SS : (D : [[mean : [\text{Real}], stdDev : [\text{Real}]]_1^N), \\ MS : (pp : \text{Real}) \end{pmatrix}$$

όπου *mean* και *stdDev* είναι η μέση τιμή και η απόκλιση της κατανομής D_j για κάθε ένα από τα N χαρακτηριστικά (j=1,2,..,N) στη συστάδα C_i, αντίστοιχα, και *pp* είναι η εκ των προτέρων πιθανότητα της C_i. Εδώ η εκ των προτέρων πιθανότητα ορίζεται ως το μέρος των διανυσμάτων χαρακτηριστικών της εικόνας που ανήκουν στη συστάδα C_i. Συνεπώς, η εκ των προτέρων πιθανότητα *pp* είναι ίση με το μέτρο υποστήριξης (support) που χρησιμοποιείται ευρέως σε μοντέλα εξόρυξης γνώσης. Στη δική μας περίπτωση, δίνει επιπλέον και μία ένδειξη για το μέγεθος του αντικειμένου.

Έτσι, μία εικόνα θεωρείται ένα σύνθετο πρότυπο που ορίζεται στο (3-15), και αποτελείται από ένα σύνολο απλών προτύπων κάθε ένα εκ των οποίων αναπαρίσταται από τη μέση τιμή και την τυπική απόκλιση μιας κατανομής.

$$image = \begin{pmatrix} SS : \{object\}, \\ MS : \perp \end{pmatrix} \tag{3-15}$$

Για τον ορισμό της ομοιότητας μεταξύ δύο εικόνων (δηλ. σύνθετων προτύπων), πρέπει να οριστεί η ομοιότητα μεταξύ των μερών της δομής (structure) και μέτρου (measure) μεταξύ των δύο συστάδων C_1 και C_2 (δηλ, απλών προτύπων). Η ομοιότητα ως απόσταση μεταξύ δύο εικόνων και οι λεπτομέρειες του υπολογισμού της αναλύονται παρακάτω.

Στην παρούσα εφαρμογή, χρησιμοποιείται μία λίγο διαφορετική προσέγγιση από αυτήν που περιγράφεται στην ενότητα 3.3. Οι διαφορές βρίσκονται στην απόσταση του μέρους του μέτρου, της δομής και στην τελική συναθροιστική συνάρτηση.

Η απόσταση μεταξύ των μερών του μέτρου (measure) των δύο συστάδων υπολογίζεται με την Ευκλείδεια απόσταση όπως στην Εξίσωση (3-16).

$$dis_{meas}(C_1, C_2) = \left| C_1.pp - C_2.pp \right| \tag{3-16}$$

Για την ομοιότητα του μέρους της δομής χρησιμοποιείται το μέτρο *Cohen's d* (Cohen, 1988) όπως ορίστηκε ήδη στην εξίσωση (3-6).

Για τη σύγκριση δύο συνόλων κατανομών (δηλ. δύο συστάδων C_1 and C_2) χρησιμοποιείται ο μέσος όρος των αποστάσεων των κατανομών για κάθε ζεύγος από τα N χαρακτηριστικά:

$$dis_{struct}(C_1, C_2) = \sum_{j=1}^{N} dis(D_j^1, D_j^2) \Big/ N \tag{3-17}$$

Οι αποστάσεις μεταξύ των dis_{meas} και dis_{struct} μεταξύ των συστάδων συναθροίζονται με την ακόλουθη συνάρτηση f_{aggr}, που δίνει το ίδιο βάρος στις δύο αποστάσεις, ενώ επιπλέον δίνει βάρος στη συνολική απόσταση μέσω των εκ των προτέρων πιθανοτήτων των συστάδων, για να ενισχυθούν οι όμοιες και παράλληλα μεγάλες συστάδες.

$$dis(C_1, C_2) = \frac{dis_{struct}(C_1, C_2) + dis_{meas}(C_1, C_2)}{2} \cdot (C_1.pp + C_2.pp) / 2 \tag{3-18}$$

Αφού έχει οριστεί η ομοιότητα μεταξύ δύο συστάδων (δηλ. απλών προτύπων), για να συγκριθούν δύο εικόνες I_1 και I_2 (δηλ. σύνθετα πρότυπα) απαιτείται να οριστεί η μέθοδος ταιριάσματος μεταξύ των διαφόρων συστάδων κάθε εικόνας. Αν και, όπως έχει ήδη συζητηθεί, αρκετοί διαφορετικοί τύποι ταιριάσματος μπορούν να εφαρμοστούν από το πλαίσιο PANDA (Bartolini et al., 2004), υιοθετούμε αυτόν που φαίνεται στην Εξίσωση (3-19), που επιτρέπει κάθε συστάδα από την πρώτη εικόνα να αντιστοιχηθεί με περισσότερες από μία συστάδες της δεύτερης εικόνας και το αντίστροφο.

$$dis(I_1, I_2) = \frac{1}{M^2}\left(\sum_{i=1}^{M}\sum_{j=1}^{M} dis\left(C_i^{I_1}, C_j^{I_2}\right)\right) \qquad (3\text{-}19)$$

3.5.2 Πειραματικά αποτελέσματα

Τα πειράματα στοχεύουν στο να δειχθεί η αποτελεσματικότητα της προτεινόμενης προσέγγισης του CBIR βασισμένου στα πρότυπα απέναντι στα συμβατικά CBIR συστήματα. Το σύνολο δεδομένων αποτελούν οι εικόνες πολιτιστικής κληρονομιάς που προέρχονται από τη βάση δεδομένων του Ιδρύματος Μείζονος Ελληνισμού (FMW, 2009). Οι εικόνες είναι πέντε κατηγοριών, αρχαία μνημεία, νομίσματα, πορτραίτα, πίνακες ζωγραφικής και μαρμάρινα γλυπτά και περιλαμβάνουν έγχρωμες και ασπρόμαυρες εικόνες διαφόρων μεγεθών και από διάφορες πηγές. Όλες οι εικόνες έχουν μετατραπεί σε 8-bit μορφοποίηση επιπέδου του γκρι και έχουν σμικρυνθεί σε μέγεθος 256×256.

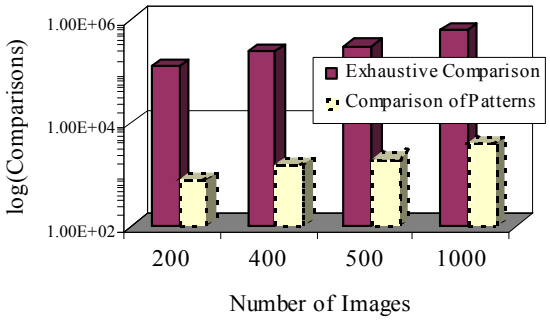
Εικόνα 3-10 Δείγματα εικόνων από τη βάση δεδομένων που χρησιμοποιήθηκε στα πειράματα.

5,000 περιοχές από τις εικόνες της βάσης επεξεργάστηκαν με τη χρήση παραθύρου 128×128-pixel με βήμα επικάλυψης 96-pixel. Τα διανύσματα χαρακτηριστικών που εξήχθησαν από κάθε εικόνα συσταδοποιήθηκαν με τον αλγόριθμο EM του εργαλείου εξόρυξης γνώσης WEKA (Witten et al., 2005).

Μία δυαδική συσταδοποίηση ακολουθήθηκε, αφού οι εικόνες περιέχουν ένα ή περισσότερα αντικείμενα του ιδίου είδους (πχ. ένα ή δύο νομίσματα) και φόντο.

Για κάθε συστάδα ένα πρότυπο $object_i$, i=1,2 ανατέθηκε και κάθε εικόνα αναπαραστάθηκε από ένα σύνθετο πρότυπο *image*. Η συλλογή των προτύπων που προέρχεται από εικόνες της βάσης δεδομένων χρησιμοποιήθηκε για να δημιουργηθεί η βάση προτύπων. Στη συνέχεια, εκτελέσθηκαν επερωτήσεις για τον έλεγχο της απόδοσης της προτεινόμενης προσέγγισης.

Η απόδοση αυτή σε σχέση με την εξαντλητική συμβατική προσέγγιση παρουσιάζεται στην Εικόνα 3-11, σε όρους αριθμού συγκρίσεων μεταξύ της επερώτησης και των δεδομένων της βάσης. Παρατηρείται ότι η προτεινόμενη προσέγγιση επιτυγχάνει περίπου 156 φορές λιγότερες συγκρίσεις.



Εικόνα 3-11. Αριθμός συγκρίσεων μεταξύ της εικόνας επερώτησης και των αποθηκευμένων εικόνων για την προτεινόμενη και τη συμβατική προσέγγιση.

Η απόδοση της ανάκτησης για το προτεινόμενο CBIR με χαρακτηριστικά τις κατανομές LBP και αντίθεσης υπολογίζεται στα 80.4%. Η αντίστοιχη απόδοση που επιτυγχάνεται με τη χρήση των 3-level Discrete Wavelet Transform (DWT) χαρακτηριστικών είναι 62.2%.

3.6 Σύνοψη

Στο κεφάλαιο αυτό περιγράψαμε την πολύ σημαντική διαδικασία της σύγκρισης των προτύπων εστιάζοντας στα πρότυπα συσταδοποίησης. Η σύγκριση προτύπων είναι μία προχωρημένη λειτουργία που βασίζεται στο μοντέλο αναπαράστασης των προτύπων και κυρίως στα μέρη της δομής και του μέτρου των προτύπων. Η αποτελεσματικότητα των αποτελεσμάτων της σύγκρισης των προτύπων εξαρτάται όχι μόνο από το σχήμα – τον ορισμό του προτύπου, αλλά και από την απόσταση/ομοιότητα και τις συναρτήσεις συνάθροισης που χρησιμοποιούνται.

Χρησιμοποιώντας το πλαίσιο PANDA για τη σύγκριση των προτύπων (Ntoutsi, 2008), ορίσαμε όλες τις απαραίτητες συναρτήσεις που χρειάζονται για τον υπολογισμό ομοιότητας των προτύπων που εξάγονται με τον αλγόριθμο ΕΜ και αφού τα πρότυπα αυτά περιγράφονται ως κατανομές, χρησιμοποιήθηκε η συνάρτηση σύγκρισης *Cohen's d* (Cohen, 1988).

Ορίσαμε μία μεθοδολογία για τη σύγκριση διαφόρων τύπων δεδομένων/αντικειμένων (εικόνων συγκεκριμένα), που περιλαμβάνει τέσσερα βήματα. Την εξαγωγή χαρακτηριστικών από τα δεδομένα, τη συσταδοποίησή τους, την δημιουργία προτύπων από τη συσταδοποίηση και τον υπολογισμό των ομοιοτήτων των προτύπων. Προκειμένου να συγκρίνουμε δύο αντικείμενα, πρέπει να συγκρίνουμε τις συστάδες (πρότυπα) που υπάρχουν σε αυτά.

Μία υψηλή τιμή ομοιότητας μεταξύ των συστάδων υπονοεί υψηλή ομοιότητα στα αρχικά αντικείμενα. Η ομοιότητα μεταξύ δύο προτύπων του ιδίου τύπου ορίζεται ως η συνδυασμένη ομοιότητα των μερών της δομής (structure) και του μέτρου ποιότητας (measure) των προτύπων. Μία συσταδοποίηση αποτελείται από ένα αριθμό από συστάδες/πρότυπα και η ομοιότητα μεταξύ δύο συσταδοποιήσεων ορίζεται με τη συνάθροιση των ομοιοτήτων των συστάδων που την αποτελούν. Ορίζοντας την ομοιότητα μεταξύ των συσταδοποιήσεων, βρίσκουμε την ομοιότητα των αρχικών αντικειμένων τα οποία αναπαριστούν οι συσταδοποιήσεις.

Επιπλέον, παρουσιάσαμε δύο περιπτώσεις πραγματικών περιπτώσεων σύγκρισης εικόνων μέσω της σύγκρισης των προτύπων. Και στις δύο περιπτώσεις ακολουθήθηκε η ίδια μεθοδολογία αλλά με διαφορετικές συναρτήσεις ομοιότητας και συνάθροισης. Τα πειραματικά αποτελέσματα έδειξαν ότι τόσο η μεθοδολογία όσο και οι συναρτήσεις που ορίστηκαν έχουν καλή απόδοση σε τέτοιες εφαρμογές.

4 Σύγκριση Προτύπων – Η Περίπτωση της Ασαφούς Συσταδοποίησης (Fuzzy Clustering)

Στο προηγούμενο κεφάλαιο αναφερθήκαμε στη σύγκριση διακριτών συστάδων. Στο παρόν κεφάλαιο αναφερόμαστε στη συσταδοποίηση ασαφών δεδομένων και τη σύγκριση των εξαγόμενων από αυτά προτύπων. Τα ασαφή δεδομένα έχουν ένα βαθμό συμμετοχής (membership) για τα χαρακτηριστικά τους, ενώ τα διαισθητικά ασαφή δεδομένα εκτός του βαθμού συμμετοχής, έχουν και ένα βαθμό μη-συμμετοχής (non-membership) για τα χαρακτηριστικά τους. Μία τρίτη τιμή, η διστακτικότητα (hesitancy) ορίζει το βαθμό αβεβαιότητας. Σε πολλές πραγματικές εφαρμογές η έννοια της αβεβαιότητας υπάρχει με διάφορους τρόπους. Ανακρίβεια στα δεδομένα λόγω αποκλίσεων από τη δειγματοληψία ή και την μέτρηση των τιμών, αβεβαιότητα στις επερωτήσεις και στα αποτελέσματά τους, εσκεμμένη ασάφεια για τη διατήρηση της ανωνυμίας κλπ. Η συσταδοποίηση διαισθητικά ασαφών δεδομένων παρέχει μία εξελιγμένη τεχνική για συσταδοποίηση τέτοιου είδους δεδομένων.

Στο κεφάλαιο αυτό παρουσιάζουμε τη θεωρία των διαισθητικά ασαφών συνόλων, ορίζουμε ένα μέτρο απόστασης για διαισθητικά ασαφή δεδομένα και παρουσιάζουμε μία τροποποιημένη έκδοση του αλγορίθμου Fuzzy C-Means (FCM) (Bezdek, et al., 1984) που εμπεριέχει το μέτρο αυτό. Παρέχουμε μία πειραματική μελέτη για συσταδοποίηση εικόνων που έχουν αναπαρασταθεί ως διαισθητικά ασαφή δεδομένα (με τη χρήση ασαφών ιστογραμμάτων). Η συσταδοποίηση χρησιμοποιείται για την κατηγοριοποίηση των εικόνων σε προκαθορισμένες κλάσεις.

Η έννοια του PBMS χρησιμοποιείται για την αναπαράσταση στη βάση προτύπων των ασαφών δεδομένων και του αποτελέσματος της συσταδοποίησης. Το αποτέλεσμα του αλγορίθμου διαισθητικής ασαφούς συσταδοποίησης, αποθηκεύονται στη βάση προτύπων για μελλοντική αναφορά και για την

κατηγοριοποίηση νέων εικόνων στις κλάσεις που έχουν ήδη οριστεί στη βάση προτύπων. Με τη χρήση του προτεινόμενου μέτρου ομοιότητας για τα διαισθητικά ασαφή δεδομένα, νέα αντικείμενα μπορούν εύκολα να κατηγοριοποιηθούν στις προκαθορισμένες κλάσεις.

4.1 Εισαγωγή

Οι προσεγγίσεις συσταδοποίησης που βασίζονται στην ασαφή λογική (Zadeh, 1965), όπως είναι η Fuzzy C-Means (FCM) (Bezdek, et al., 1984) και οι παραλλαγές του (Yong, 2004; Thitimajshima, 2000; Chumsamrong, et al., 2000) έχει αποδειχτεί ότι είναι ανταγωνιστικές των συμβατικών αλγορίθμων συσταδοποίησης, ειδικά σε πραγματικές εφαρμογές. Το συγκριτικό πλεονέκτημα αυτών των προσεγγίσεων είναι ότι δεν προϋποθέτουν ακριβή όρια μεταξύ των συστάδων, επιτρέποντας έτσι σε κάθε διάνυσμα χαρακτηριστικών να ανήκει σε διαφορετικές συστάδες με κάποιο βαθμό. Ο βαθμός συμμετοχής (membership) ενός διανύσματος χαρακτηριστικών σε μία συστάδα συνήθως θεωρείται μία συνάρτηση της απόστασης του διανύσματος από το κέντρο της συστάδας ή από κάποιο άλλο αντιπροσωπευτικό διάνυσμα στη συστάδα.

Ένα από τα βασικά θέματα που προκύπτουν από πραγματικές εφαρμογές συσταδοποίησης είναι η αβεβαιότητα που υπάρχει στα διανύσματα χαρακτηριστικών των δεδομένων. Επειδή οι τιμές των χαρακτηριστικών μπορεί να περιέχουν αβεβαιότητα λόγω των ανακριβών μετρήσεων και του θορύβου, η απόσταση που ορίζει τη συμμετοχή ενός διανύσματος χαρακτηριστικών σε μία συστάδα θα περιέχει επίσης αβεβαιότητα. Η πιθανότητα για ανάθεση λανθασμένων τιμών συμμετοχής σε μία συστάδα είναι για το λόγο αυτό αυξημένη. Οι προσεγγίσεις συσταδοποίησης που υπήρχαν μέχρι τώρα δεν χρησιμοποιούν πληροφορία σχετικά με την αβεβαιότητα στο επίπεδο των χαρακτηριστικών/ δεδομένων.

Στο κεφάλαιο αυτό παρουσιάζουμε μία τροποποίηση του αλγορίθμου FCM. Η καινοτόμος αυτή παραλλαγή του FCM αντιμετωπίζει τα δεδομένα σαν αναπαραστάσεις διαισθητικά ασαφών τιμών, δηλαδή, στοιχεία ενός διαισθητικά ασαφούς συνόλου. Τα διαισθητικά ασαφή σύνολα, (Atanassov, 1986, 1989, 1994a, 1994b, 1999) είναι γενικευμένα ασαφή σύνολα (Zadeh, 1965) που αντιμετωπίζουν τη διστακτικότητα που προέρχεται από ανακριβή πληροφορία (Vlachos and Sergiadis, 2006). Τα στοιχεία ενός διαισθητικά ασαφούς συνόλου

χαρακτηρίζονται από δύο τιμές που αναπαριστούν το πόσο ανήκουν και πόσο δεν ανήκουν τα στοιχεία αυτά στο σύνολο, αντίστοιχα.

Για παράδειγμα στο σύνολο $A=\{x, 0.4, 0.2\}$, το x αναπαριστά το στοιχείο του συνόλου, η τιμή 0.4 αναπαριστά τη συμμετοχή του στοιχείου x στο σύνολο και η τιμή 0.2 αναπαριστά τη μη-συμμετοχή του στοιχείου x στο σύνολο *A*. Παρατηρείται ότι το άθροισμα των τιμών αυτών (0.4+0.2) είναι μικρότερο της μονάδος, που σημαίνει ότι υπάρχει μία αβεβαιότητα με τιμή 0.4 (=1-0.4-0.2) για το αν το στοιχείο ανήκει ή όχι στο σύνολο A. Στη συνέχεια θα ακολουθήσει ανάλυση της θεωρίας καθώς και παραδείγματα σύγκρισης τέτοιων συνόλων.

Η πληθώρα και η σημαντικότητα των πιθανών εφαρμογών διαισθητικά ασαφών συνόλων έχει οδηγήσει πολλούς ερευνητές στο να προτείνουν πολλά και διαφορετικά είδη μετρικών ομοιότητας μεταξύ των συνόλων αυτών. Τέτοιες μετρικές έχουν προταθεί από τον Chen (1995, 1997) με τη μετρική S_C, από τους Hong & Kim (1999) με τη μετρική S_H, από τους by Fan & Zhangyan (2001) με τη S_L, και των Li et al. (2002) που πρότειναν τη μετρική S_O. Οι Dengfeng & Chuntian (2002) πρότειναν την μετρική S_{DC}, ο Mitchell (2003) πρότεινε μία τροποποίηση της μετρικής S_{DC}, τη μετρική S_{HB}, οι Zhizhen & Pengfei (2003) πρότειναν τρεις μετρικές S_e^p, S_s^p and S_h^p και τρεις ακόμα μετρικές έχουν προταθεί από τους Hung & Yang (2004), τις S_{HY}^1, S_{HY}^2, and S_{HY}^3. Οι Li et al. (2007) παρέχουν μία λεπτομερή σύγκριση των μετρικών αυτών, δείχνοντας τις αδυναμίες κάθε μίας.

Μερικές μετρικές, όπως είναι οι S_C , S_H , S_L , S_{HB} και οι S_{HY}^1 , S_{HY}^2 και S_{HY}^3 εστιάζουν στη συνάθροιση (aggregation) των διαφορών μεταξύ των τιμών της συμμετοχής (membership) και των διαφορών των τιμών της μη-συμμετοχής (non-membership) ενώ άλλες χρησιμοποιούν αποστάσεις όπως η Minkowski, για την μετρική S_{DC}, ή την Hausdorff, για τις μετρικές S_{HY}^1 , S_{HY}^2 και S_{HY}^3 προκειμένου να υπολογιστεί ο βαθμός ομοιότητας των ασαφών συνόλων. Οι μετρικές S_{DC}, S_s^P και S_h^P εστιάζουν επίσης στις διαφορές μεταξύ των τιμών συμμετοχής και μη-συμμετοχής.

Σχετικά με την αποτελεσματικότητα των μετρικών αυτών, μερικές από αυτές όπως οι S_C και S_{HY}^1 , S_{HY}^2 και S_{HY}^3 δεν ικανοποιούν τις ιδιότητες που πρέπει να έχει μία μετρική ομοιότητας ορισμένη για διαισθητικά ασαφή σύνολα, ενώ όλες οι παραπάνω μετρικές αποτυγχάνουν σε συγκεκριμένες περιπτώσεις όπως οι Li

et al. (2007) αναφέρουν με αντι-παραδείγματα. Η σύγκριση όλων των παραπάνω μετρικών μεταξύ τους και με την προτεινόμενη μετρική παρουσιάζεται στις παρακάτω ενότητες.

4.2 Συσταδοποίηση διαισθητικά ασαφών δεδομένων

Προκειμένου να αναπαρασταθούν τα διαισθητικά ασαφή δεδομένα και να μπορεί να οριστεί ένα μέτρο ομοιότητας για να χρησιμοποιηθεί στο ΣΔΒΠ (PBMS), γίνεται στις επόμενες ενότητες μία παρουσίαση της θεωρίας των ασαφών και διαισθητικά ασαφών συνόλων. Επιπλέον, ορίζονται τα μέτρα ομοιότητας και προτείνεται ένα κατάλληλο σχήμα συσταδοποίησης, ενώ τέλος ορίζεται η αναπαράσταση των ασαφών δεδομένων στη βάση προτύπων και αναφέρονται οι εφαρμογές που μπορούν να υποστηριχθούν από το PBMS.

4.2.1 Διαισθητικά Ασαφή Σύνολα

Οι θεωρητικές βάσεις των ασαφών και διαισθητικά ασαφών συνόλων περιγράφονται στο (Zadeh, 1965; Atanassov, 1986). Στην ενότητα αυτή γίνεται μία σύντομη περιγραφή των βασικών εννοιών που χρησιμοποιούνται στην παρούσα διατριβή.

Ορισμός 4-1 (Zadeh, 1965). Έστω ένα σύνολο *E*. Ένα ασαφές σύνολο ορισμένο στο *E* είναι ένα αντικείμενο $\tilde{A}$ της μορφής

$$\tilde{A} = \left\{ \left\langle x, \mu_{\tilde{A}}(x) \right\rangle \middle| x \in E \right\}$$

όπου $\mu_{\tilde{A}} : E \to [0,1]$ είναι ο βαθμός συμμετοχής (degree of membership) του στοιχείου $x \in E$ στο σύνολο $\tilde{A} \subset E$. Για κάθε στοιχείο $x \in E$, $0 \le \mu_{\tilde{A}}(x) \le 1$. ■

Ορισμός 4-2 (Atanassov, 1986; Atanassov, 1994). Ένα διαισθητικά ασαφές σύνολο *A* είναι ένα αντικείμενο της μορφής

$$A = \left\{ \left\langle x, \mu_A(x), \gamma_A(x) \right\rangle \middle| x \in E \right\}$$

όπου $\mu_A : E \to [0,1]$ και $\gamma_A : E \to [0,1]$ είναι ο βαθμός συμμετοχής (degree of membership) και μη-συμμετοχής, αντίστοιχα, του στοιχείου $x \in E$ στο σύνολο $A \subset E$. Για κάθε στοιχείο $x \in E$, ισχύει ότι $0 \le \mu_A(x) \le 1$, $0 \le \gamma_A(x) \le 1$ και

$$0 \le \mu_A(x) + \gamma_A(x) \le 1 \qquad (4\text{-}1)$$

Για κάθε $x \in E$, εάν $\gamma_A(x) = 1 - \mu_A(x)$, το A αναπαριστά ένα ασαφές σύνολο. Η συνάρτηση

$$\pi_A(x) = 1 - \mu_A(x) - \gamma_A(x)$$

αναπαριστά το βαθμό της *διστακτικότητας (hesitancy)* του στοιχείου $x \in E$ στο σύνολο $A \subset E$.

■

Για κάθε δύο διαισθητικά ασαφή σύνολα A και B ισχύουν οι παρακάτω λειτουργίες και σχέσεις (Atanassov, 1986; Atanassov 1994)

$$A \subset B \text{ iff } \forall x \in E, \mu_A(x) \le \mu_B(x) \text{ and } \gamma_A(x) \ge \gamma_B(x)$$

$$A = B \text{ iff } A \subset B \text{ and } B \subset A$$

$$A^C = \left\{ \langle x, \gamma_A(x), \mu_A(x) \rangle \middle| x \in E \right\}$$

$$A \cap B = \left\{ \langle x, \min\left(\mu_A(x), \mu_B(x)\right), \max\left(\gamma_A(x), \gamma_B(x)\right) \rangle \middle| x \in E \right\}$$

$$A \cup B = \left\{ \langle x, \max\left(\mu_A(x), \mu_B(x)\right), \min\left(\gamma_A(x), \gamma_B(x)\right) \rangle \middle| x \in E \right\}$$

$$A @ B = \left\{ \left\langle x, \frac{1}{2}\left(\mu_A(x) + \mu_B(x)\right), \frac{1}{2}\left(\gamma_A(x), \gamma_B(x)\right) \right\rangle \middle| x \in E \right\}$$

$$\underset{i=1}{\overset{n}{@}} A_i = \left\{ \left\langle x, \frac{1}{n}\left(\sum_{i=1}^{n} \mu_{A_i}(x)\right), \frac{1}{n}\left(\sum_{i=1}^{n} \gamma_{A_i}(x)\right) \right\rangle \middle| x \in E \right\}$$

■

Ορισμός 4-3 (Dengfeng and Chuntian, 2002). Έστω S μία αντιστοίχηση $\text{IFSs}(E) \times \text{IFSs}(E) \to [0,1]$, όπου το IFSs($E$) δηλώνει ένα σύνολο από όλα τα διαισθητικά ασαφή σύνολα στο E. Το $S(A, B)$ λέγεται ότι είναι ο βαθμός ομοιότητας μεταξύ του $A \in \text{IFSs}(E)$ και του $B \in \text{IFSs}(E)$, εάν το $S(A, B)$ ικανοποιεί τις παρακάτω συνθήκες:

P1. $S(A,B) \in [0,1]$

P2. $S(A,B) = 1 \Leftrightarrow A = B$

P3. $S(A,B) = S(B,A)$

P4. $S(A,C) \le S(A,B)$ and $S(A,C) \le S(B,C)$ if $A \subseteq B \subseteq C$, $C \in \text{IFSs}(E)$ ■

Η αναπαράσταση των δεδομένων πραγματικών εφαρμογών συσταδοποίησης με διαισθητικά ασαφή σύνολα είναι μία πρόκληση που δίνει τη δυνατότητα να

ερευνηθεί η αποτελεσματικότητα της θεωρίας διαισθητικά ασαφών συνόλων στην πράξη.

4.2.2 Μέτρα Σύγκρισης Διαισθητικά Ασαφών Συνόλων

Στην ενότητα αυτή προτείνουμε μία πρωτότυπη μετρική ομοιότητας μεταξύ διαισθητικά ασαφών συνόλων, βασισμένη στις τιμές συμμετοχής και μη-συμμετοχής των στοιχείων τους. Δοθέντος ενός διαισθητικά ασαφούς συνόλου A, ορίζουμε δύο ασαφή σύνολα, τα $M_A, \Gamma_A \in \mathcal{F}(E)$, όπου το $\mathcal{F}(E)$ είναι το σύνολο όλων των ασαφών υποσυνόλων ενός στοιχείου $x \in E$. Η συμμετοχή και μη-συμμετοχή των συνόλων αυτών ορίζεται ως $M_A = \{\mu_A(x)\},\ \Gamma_A = \{\gamma_A(x)\}\ \forall x \in E$. Έτσι, το A μπορεί να περιγραφεί σαν το ζευγάρι (M_A, Γ_A).

Ορισμός 4-5. Δεδομένων δύο διαισθητικά ασαφών συνόλων A=(M_A, Γ_A), B=(M_B, Γ_B), όπου $M_A, M_B, \Gamma_A, \Gamma_B \in \mathcal{F}(E)$, και δεδομένου του E ως ενός πεπερασμένου συνόλου E=$\{\chi_1, \chi_2, \ldots, \chi_n\}$, ορίζουμε τη μετρική ομοιότητας Z_1 μεταξύ των διαισθητικά ασαφών συνόλων A και B με την ακόλουθη εξίσωση:

$$Z_1(A, B) = \frac{z_1(M_A, M_B) + z_1(\Gamma_A, \Gamma_B)}{2} \tag{4-2}$$

όπου

$$z_1(A', B') = \begin{cases} \dfrac{\sum_{i=1}^{n} \min(A'(x_i), B'(x_i))}{\sum_{i=1}^{n} \max(A'(x_i), B'(x_i))}, & A' \cup B' \neq \emptyset \\ 1, & A' \cup B' = \emptyset \end{cases} \tag{4-3}$$

με $A', B' \in \mathcal{F}(E)$. ■

Προκειμένου να είναι αποδεκτό το Z_1 ως μετρική ομοιότητας, πρέπει να αποδειχθεί ότι το z_1 ικανοποιεί τις συνθήκες που ορίστηκαν στον Ορισμό 4-3. Οι συνθήκες P1, P2 και P3 είναι εύκολο να δειχθεί ότι ικανοποιούνται από το z_1. Παρακάτω δίνουμε την απόδειξη για τη συνθήκη P4.

Λήμμα. Για όλα τα $A', B', C' \in \mathcal{F}(E)$, όπου $\mathcal{F}(E)$ είναι το σύνολο όλων των ασαφών υποσυνόλων ενός στοιχείου $x \in E$ και θεωρώντας το E σαν ένα πεπερασμένο

σύνολο $E = \{x_1, x_2, ..., x_n\}$, εάν $A' \subseteq B' \subseteq C'$ τότε $z_1(A',C') \leq z_1(A',B')$ και $z_1(A',C') \leq z_1(B',C')$.

Απόδειξη: Το $A' \subseteq B' \subseteq C'$ συνεπάγεται ότι $A'(x_i) \leq B'(x_i) \leq C'(x_i) \forall x_i \in E$ και

$$z_1(A',C') = \frac{\sum_{i=1}^{n} \min(A'(x_i), C'(x_i))}{\sum_{i=1}^{n} \max(A'(x_i), C'(x_i))} = \frac{\sum_{i=1}^{n} A'(x_i)}{\sum_{i=1}^{n} C'(x_i)}, \quad z_1(A',B') = \frac{\sum_{i=1}^{n} \min(A'(x_i), B'(x_i))}{\sum_{i=1}^{n} \max(A'(x_i), B'(x_i))} = \frac{\sum_{i=1}^{n} A'(x_i)}{\sum_{i=1}^{n} B'(x_i)},$$

$$z_1(B',C') = \frac{\sum_{i=1}^{n} \min(B'(x_i), C'(x_i))}{\sum_{i=1}^{n} \max(B'(x_i), C'(x_i))} = \frac{\sum_{i=1}^{n} B'(x_i)}{\sum_{i=1}^{n} C'(x_i)}. \text{ Έτσι, } \frac{\sum_{i=1}^{n} A'(x_i)}{\sum_{i=1}^{n} C'(x_i)} \leq \frac{\sum_{i=1}^{n} A'(x_i)}{\sum_{i=1}^{n} B'(x_i)}, \quad \frac{\sum_{i=1}^{n} A'(x_i)}{\sum_{i=1}^{n} C'(x_i)} \leq \frac{\sum_{i=1}^{n} B'(x_i)}{\sum_{i=1}^{n} C'(x_i)}$$

οπότε, $z_1(A',C') \leq z_1(A',B')$ and $z_1(A',C') \leq z_1(B',C')$

Αφού $A, B, C \in IFSs(E)$ και $A \subseteq B \subseteq C$ έχουμε

$\mu_A(x) \leq \mu_B(x) \leq \mu_C(x)$ and $\gamma_A(x) \geq \gamma_B(x) \geq \gamma_C(x)$ $\forall x_i \in E, i = 1, 2, ..., n$,

Ως εκ τούτου, τα $z_1(M_A, M_B)$ και $z_1(\Gamma_A, \Gamma_B)$ ικανοποιούν τις συνθήκες P1-P4 και έτσι το Z_1 επίσης ικανοποιεί τις συνθήκες αυτές. Έτσι, το Z_1 είναι μία μετρική ομοιότητας. ■

Για την καλύτερη κατανόηση της προτεινόμενης μετρικής, ακολουθεί ένα απλό παράδειγμα.

Παράδειγμα. Θεωρούμε τρία σύνολα $A, B, C \in IFSs(E)$ με A={x, 0.4, 0.2}, B={x, 0.5, 0.3}, C={x, 0.5, 0.2} και θέλουμε να αποφασίσουμε εάν το B ή το C είναι πιο όμοιο προς το A. Με τη χρήση των Εξ. (4-2) και (4-3) υπολογίζουμε την ομοιότητα τόσο του B όσο και του C προς το σύνολο A.

$$Z_1(A,B) = \frac{\frac{0.4}{0.5} + \frac{0.2}{0.3}}{2} = 0.733, \quad Z_1(A,C) = \frac{\frac{0.4}{0.5} + \frac{0.2}{0.2}}{2} = 0.9$$

Έτσι, συμπεραίνουμε ότι το C είναι πιο όμοιο στο A απ' ότι το B.

Η προτεινόμενη μετρική ομοιότητας χρησιμοποιεί τη συνάθροιση των ελάχιστων και μέγιστων τιμών συμμετοχής με τις αντίστοιχες της μη-συμμετοχής. Αν και ο υπολογισμός της είναι απλός, η μετρική αυτή είναι ευαίσθητη σε μικρές αλλαγές των τιμών και αντιμετωπίζει αποτελεσματικά όλες τις περιπτώσεις που οι άλλες μετρικές αποτυγχάνουν. Οι περισσότερες μετρικές που αναφέρονται στην ενότητα 4.1 αποτυγχάνουν να υπολογίσουν τη σωστή τιμή για ορισμένες περιπτώσεις. Κάποιες από αυτές καταλήγουν σε μία τιμή 0 ή 1, κάτι που σημαίνει ότι τα προς σύγκριση σύνολα είναι είτε τελείως όμοια ή ανόμοια, ενώ είναι φανερό ότι κάτι τέτοιο δεν ισχύει, και κάποιες άλλες μετρικές καταλήγουν σε μία ψηλή τιμή για φανερά ανόμοια σύνολα. Πιο συγκεκριμένα ο Πίνακας 4-1 παρουσιάζει όλες τις περιπτώσεις που όρισαν οι Li et al. (2007) και για τις οποίες όλες οι μετρικές αποτυγχάνουν, καθώς και τον υπολογισμό της προτεινόμενης μετρικής για τις περιπτώσεις αυτές.

Στην περίπτωση (I) (Πίνακας 4-1) οι τιμές της μετρικής $S_C(A,B)$ και $S_{DC}(A,B)$ δείχνουν ότι τα *A* και *B* είναι τελείως όμοια. Στις περιπτώσεις (II) και (IV) κάποιες μετρικές καταλήγουν σε μάλλον υψηλές τιμές ομοιότητας· η προτεινόμενη μετρική δεν είναι τόσο οπτιμιστική. Επιπλέον, στην περίπτωση (IV) είναι φανερό ότι το σύνολο *A* είναι περισσότερο όμοιο προς το *C* από ότι το *B* (τα *A* και *C* έχουν την ίδια τιμή μη-συμμετοχής), κάτι που άλλες μετρικές δεν το λαμβάνουν υπόψη. Στην περίπτωση (III), ενώ το *B* και το *C* είναι τελείως διαφορετικά, οι μετρικές S_H, S_{HB}, S_e^p δίνουν τιμή 0.5. Αντιθέτως, στην περίπτωση (V) οι μετρικές $S_{HY}^1, S_{HY}^2, S_{HY}^3$ δίνουν τιμή ομοιότητας 0 ακόμα και εάν η τιμές μη-συμμετοχής του *A* και *B* είναι ίδιες, κάτι που υπονοεί ένα επίπεδο ομοιότητας μεταξύ των δύο συνόλων. Στις περιπτώσεις (VI) και (VII) οι μετρικές $S_{HY}^1, S_{HY}^2, S_{HY}^3$ δίνουν μία σχετικά ψηλή τιμή ομοιότητας και στην περίπτωση (VII) δεν αναγνωρίζουν ότι το *A* είναι περισσότερο όμοιο στο *C* από το *B*, λόγω της ίδιας τιμής μη-συμμετοχής του *A* και του *C*.

Πίνακας 4-1 Τιμές διαφόρων μετρικών ομοιότητας και της προτεινόμενης μετρικής, σε ιδιαίτερες περιπτώσεις σύγκρισης

Τύπος	**Μετρική**	**Περιπτώσεις**	**Τιμές μετρικών**	**Τιμή προτεινόμενης μετρικής**
I.	S_C, S_{DC}	$A=\{(x,0,0)\}$, $B=\{(x,0.5,0.5)\}$	$S_C(A,B)=S_{DC}(A,B)=1$	$Z_1=0$
II.	S_H, S_{HB}, S_e^p	$A=\{(x,0.3,0.3)\}$, $B=\{(x,0.4,0.4)\}$, $C=\{(x,0.3,0.4)\}$, $D=\{(x,0.4,0.3)\}$	$S_H(A,B)=S_{HB}(A,B)=S_e^p(A,B)=0.9$ $S_H(C,D)=S_{HB}(C,D)=S_e^p(C,D)=0.9$	$Z_1(A,B)=$ $Z_1(C,D)=0.75$
III.	S_H, S_{HB}, S_e^p	$A=\{(x,1,0)\}$, $B=\{(x,0,0)\}$, $C=\{(x,0.5,0.5)\}$	$S_H(A,B)=S_{HB}(A,B)=S_e^p(A,B)=0.5$ $S_H(B,C)=S_{HB}(B,C)=S_e^p(B,C)=0.5$	$Z_1(A,B)=0.5$, $Z_1(B,C)=0$
IV.	S_L και S_S^p	$A=\{(x,0.4,0.2)\}$, $B=\{(x,0.5,0.3)\}$, $C=\{(x,0.5,0.2)\}$	$S_L(A,B)=S_S^p(A,B)=0.95$ $S_L(A,C)=S_S^p(C,D)=0.95$	$Z_1(A,B)=0.73$ $Z_1(A,C)=0.9$
V.	$S_{HY}^1, S_{HY}^2, S_{HY}^3$	$A=\{(x,1,0)\}$, $B=\{(x,0,0)\}$	$S_{HY}^1(A,B)=S_{HY}^2(A,B)=S_{HY}^3(A,B)=0$	$Z_1(A,B)=0.5$
VI.	$S_{HY}^1, S_{HY}^2, S_{HY}^3$	$A=\{(x,0.3,0.3)\}$, $B=\{(x,0.4,0.4)\}$, $C=\{(x,0.3,0.4)\}$, $D=\{(x,0.4,0.3)\}$	$S_{HY}^1(A,B)=S_{HY}^1(C,D)=0.9$ $S_{HY}^2(A,B)=S_{HY}^2(C,D)=0.85$ $S_{HY}^3(A,B)=S_{HY}^3(C,D)=0.82$	$Z_1(A,B)=$ $Z_1(C,D)=0.75$
VII.	$S_{HY}^1, S_{HY}^2, S_{HY}^3$	$A=\{(x,0.4,0.2)\}$, $B=\{(x,0.5,0.3)\}$, $C=\{(x,0.5,0.2)\}$	$S_{HY}^1(A,B)=S_{HY}^1(A,C)=0.9$ $S_{HY}^2(A,B)=S_{HY}^2(A,C)=0.85$ $S_{HY}^3(A,B)=S_{HY}^3(A,C)=0.82$	$Z_1(A,B)=0.73$, $Z_1(A,C)=0.9$

Όλα τα παραπάνω δείχνουν την ικανότητα της προτεινόμενης μετρικής να αποδίδει σωστά διαισθητικές τιμές ομοιότητας, ενώ δεν αποτυγχάνει σε περιπτώσεις που οι άλλες μετρικές αποτυγχάνουν.

Επιπλέον, η προτεινόμενη μετρική είναι εύκολη στον υπολογισμό και δεν περιέχει εκθετικές ή άλλες συναρτήσεις που καθυστερούν σημαντικά τους υπολογισμούς.

4.2.3 Συσταδοποίηση Διαισθητικά Ασαφών Δεδομένων

Οι περισσότερες μέθοδοι συσταδοποίησης υποθέτουν ότι κάθε διάνυσμα δεδομένων ανήκει σε μία μόνο συστάδα. Αυτό είναι αποδεκτό εάν τα δεδομένα μπορούν να ομαδοποιηθούν σε συμπαγείς και καλώς διαχωρισμένες συστάδες. Εντούτοις, σε πραγματικές εφαρμογές, οι συστάδες μπορεί να επικαλύπτονται, κάτι που σημαίνει ότι ένα διάνυσμα δεδομένων μπορεί να ανήκει μερικώς σε περισσότερες από μία συστάδες. Σε μία τέτοια περίπτωση και σε όρους θεωρίας ασαφών συνόλων (Zadeh, 1965), ο βαθμός συμμετοχής ενός διανύσματος x_k στην i-στή συστάδα u_{ik} είναι μία τιμή στο διάστημα [0,1]. Ο Ruspini (1969) εισήγαγε την ιδέα αυτή την οποία αργότερα εφάρμοσε ο Dunn (1973) για να προτείνει μία μεθοδολογία συσταδοποίησης βασισμένη στην ελαχιστοποίηση μίας αντικειμενικής συνάρτησης (objective function). Στο (Bezdek, et al., 1984) ο Bezdek εισήγαγε τον αλγόριθμο Fuzzy C-Means (FCM) ο οποίος χρησιμοποιεί έναν βεβαρημένο δείκτη στις ασαφής συμμετοχές.

Ο FCM είναι ένας επαναληπτικός αλγόριθμος και ο στόχος του είναι να ανακαλύψει τα κέντρα των συστάδων που ελαχιστοποιούν μία συνάρτηση-κριτήριο, η οποία μετρά την ποιότητα της ασαφούς κατάτμηση.

Μία ασαφής κατάτμηση (fuzzy partition) συμβολίζεται με έναν (c x N)-διάστατο πίνακα U πραγματικών αριθμών $u_{ik} \in [0,1], \forall 1 \le i \le c$ και $1 \le k \le N$, όπου c και N είναι ο αριθμός των συστάδων και ο πληθάριθμος του διανύσματος χαρακτηριστικών, αντίστοιχα. Ο παρακάτω περιορισμός ισχύει για το u_{ik}:

$$\sum_{i=1}^{c} u_{ik} = 1, 0 < \sum_{k=1}^{N} u_{ik} < N \qquad (4\text{-}4)$$

Με δεδομένο αυτό, η αντικειμενική συνάρτηση του FCM έχει την εξής μορφή:

$$J_m(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m d_{ik}^2$$

όπου V είναι ένας (p x c)-διάστατος πίνακας που περιέχει τα c κέντρα, p είναι η διάσταση των δεδομένων, d_{ik} είναι μία A-norm για τον υπολογισμό της απόστασης μεταξύ του διανύσματος δεδομένων x_k και του κέντρου της συστάδας v_i, και $m \in [1, \infty)$ είναι ένας βεβαρημένος δείκτης. Η παράμετρος m ελέγχει την ασάφεια των συστάδων. Όταν το m προσεγγίζει το 1, ο FCM πραγματοποιεί μία διακριτή κατάτμηση (hard partitioning) όπως κάνει ο

αλγόριθμος k-means, ενώ όσο το m μεγαλώνει προς το άπειρο, η κατάτμηση γίνεται όσο δυνατόν πιο ασαφής. Δεν υπάρχει αναλυτική μεθοδολογία για τη βέλτιστη επιλογή της τιμής του m.

Οι Bezdek, Ehrlich και Full (1984) απέδειξαν ότι εάν το m και το c είναι σταθερές παράμετροι και τα $I_k, \tilde{I}_k$ είναι σύνολα ορισμένα ως:

$$\forall 1 \le k \le N, \begin{cases} I_k = \{ i \mid 1 \le i \le c; d_{ik} = 0 \}, \\ \tilde{I}_k = \{1,2,...,c\} \setminus I_k, \end{cases}$$

Τότε η $J_m(U,V)$ μπορεί να ελαχιστοποιηθεί μόνο εάν:

$$\underset{\substack{1 \le i \le c \\ i \le k \le N}}{\forall} \quad u_{ik} = \begin{cases} \dfrac{(d_{ik})^{\frac{2}{1-m}}}{\sum_{j=1}^{c} (d_{jk})^{\frac{2}{1-m}}}, & I_k = \varnothing, \\ \begin{cases} 0, & i \notin I_k \\ \sum_{i \in I_k} u_{ik} = 1, & i \in I_k, \end{cases} & I_k \neq \varnothing, \end{cases} \tag{4-5}$$

και

$$\underset{1 \le i \le c}{\forall} \quad v_i = \frac{\sum_{k=1}^{N} (u_{ik})^m x_k}{\sum_{k=1}^{N} (u_{ik})^m}. \tag{4-6}$$

Με την συνεχή-επαναληπτική τροποποίηση των κέντρων των συστάδων και των τιμών συμμετοχής για κάθε διάνυσμα χαρακτηριστικών, ο FCM συνεχώς μετακινεί τα κέντρα των συστάδων στη «σωστή» περιοχή εντός του συνόλου δεδομένων. Πιο αναλυτικά, ο αλγόριθμος που οδηγεί στη βέλτιστη κατάτμηση είναι ο αλγόριθμος Picard που περιγράφεται παρακάτω:

Αλγόριθμος 4-1. Αλγόριθμος FCM

Βήμα 1: καθορισμός του c ($1<c<N$), $m \in [1,\infty)$ και αρχικοποίηση του $V^{(0)}$, $j \leftarrow 1$,

Βήμα 2: υπολογισμός του πίνακα συμμετοχών $U^{(j)}$, με χρήση της εξίσωσης (4-5),

Βήμα 3: Ανανέωση του πίνακα των κέντρων $V^{(j)}$, με χρήση της εξίσωσης (4-6) και του $U^{(j)}$,

Βήμα 4: Εάν το $\left\|U^{j+1}-U^{j}\right\|_F > \varepsilon$ τότε $j \leftarrow j+1$ και επιστροφή στο Βήμα 2.

Η παράμετρος ε κάνει τον αλγόριθμο να συγκλίνει όταν η βελτίωση της ασαφούς κατάτμησης της τρέχουσας επανάληψης σε σχέση με αυτήν της προηγούμενης επανάληψης είναι χαμηλότερη από ένα κατώφλι, ενώ το $\|\cdot\|_F$ δηλώνει την Frobenious νόρμα.

Ο αλγόριθμος FCM ελαχιστοποιεί την διακύμανση μεταξύ των συστάδων, αλλά παρουσιάζει τα ίδια προβλήματα με τον k-means (MacQueen, 1967), καθώς δεν εγγυάται ότι συγκλίνει στη βέλτιστη λύση, ενώ το ελάχιστο που έχει υπολογιστεί είναι τοπικό και τα αποτελέσματα εξαρτώνται από την αρχική επιλογή των κέντρων.

Ο αλγόριθμος FCM προσπαθεί να κατατμήσει τα δεδομένα χρησιμοποιώντας τα διανύσματα χαρακτηριστικών των δεδομένων, αγνοώντας το γεγονός ότι αυτά τα διανύσματα μπορεί να περιέχουν και ποιοτικές πληροφορίες ανά χαρακτηριστικό. Για παράδειγμα, σύμφωνα με τη θεωρία των ασαφών διαισθητικά δεδομένων, ένα σημείο (δεδομένων) x_k δεν είναι απλά ένα p-διάστατο διάνυσμα $(x_{k_1},...,x_{k_p})$ ποσοτικής πληροφορίας, αλλά ένα p-διάστατο διάνυσμα από τριάδες $[(x_{k_1},\mu_{k_1},\gamma_{k_1}),...,(x_{k_p},\mu_{k_p},\gamma_{k_p})]$, όπου για κάθε x_{k_l} υπάρχει ποιοτική πληροφορία που δίνεται μέσω των τιμών της συμμετοχής μ_{k_l} και μη-συμμετοχής γ_{k_l} του τρέχοντος σημείου στο χαρακτηριστικό l. Είναι φανερό ότι ο FCM δε χρησιμοποιεί αυτή τη σημαντική πληροφορία. Στο σενάριο συσταδοποίησης εικόνων, ένα χαρακτηριστικό l μπορεί να αντιστοιχεί σε πληροφορία χρώματος. Θα ήταν σημαντικό πλεονέκτημα της μεθόδου συσταδοποίησης εάν μπορούσε να συμπεριλάβει τις τιμές συμμετοχής και μη-συμμετοχής, που σχετίζονται (για παράδειγμα) με το πόσο κόκκινο μπορεί να περιέχει η εικόνα, και με πόση σιγουριά ισχύει αυτό.

Ο κύριος λόγος που ο FCM δεν μπορεί να εκμεταλλευτεί αυτή την πληροφορία είναι ότι η συνάρτηση απόστασης λειτουργεί μόνο στα διανύσματα χαρακτηριστικών των δεδομένων (τις τιμές των δεδομένων) και όχι στα πληροφορία συμμετοχής και μη-συμμετοχής που συνοδεύει τα διανύσματα αυτά. Στην παρούσα διατριβή προτείνουμε μία διαφορετική προσέγγιση στην οποία αντικαθιστούμε τη συνάρτηση απόστασης με την μετρική απόστασης διαισθητικά ασαφών συνόλων που αναλύσαμε στην ενότητα 4.2.2.

Χρησιμοποιώντας την προτεινόμενη μετρική απόστασης η αντικειμενική συνάρτηση του FCM παίρνει την παρακάτω μορφή:

$$J_m^{IFS}(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m\left|x_k-v_i\right|_{IFS} \tag{4-7}$$

Η ελαχιστοποίηση της (4-7) μπορεί να επιτευχθεί όρο προς όρο:

$$J_m^{IFS}(U,V)=\sum_{k=1}^{N}\varphi_k(U) \tag{4-8}$$

όπου

$$\underset{1\le k\le N}{\forall}\quad \varphi_k(U)=\sum_{i=1}^{c}(u_{ik})^m\left|x_k-v_i\right|_{IFS} \tag{4-9}$$

Η συνάρτηση Lagrange της (4-9) με τους περιορισμούς από την εξίσωση (4-4) είναι:

$$\underset{1\le k\le N}{\forall}\quad \Phi_k(U,\lambda)=\sum_{i=1}^{c}(u_{ik})^m\left|x_k-v_i\right|_{IFS}-\lambda\left(\sum_{i=1}^{c}u_{ik}-1\right) \tag{4-10}$$

Όπου λ είναι ο πολλαπλασιαστής Lagrange. Ορίζοντας της μερικές παραγώγους της $\Phi_k(U,\lambda)$ στο μηδέν, έχουμε:

$$\underset{1\le k\le N}{\forall}\quad \frac{\partial\Phi_k(U,\lambda)}{\partial\lambda}=\sum_{i=1}^{c}u_{ik}-1=0 \tag{4-11}$$

και

$$\underset{\substack{1\le z\le c\\1\le k\le N}}{\forall}\quad \frac{\partial\Phi_k(U,\lambda)}{\partial u_{zk}}=m(u_{zk})^{m-1}\left|x_k-v_i\right|_{IFS}-\lambda=0 \tag{4-12}$$

Λύνοντας την εξίσωση (4-12) προς u_{zk} έχουμε:

$$u_{zk}=\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}\left(\left|x_k-v_z\right|_{IFS}\right)^{\frac{1}{1-m}} \tag{4-13}$$

Από τις (4-11) και (4-13) έχουμε:

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{j=1}^{c}\left(\left\|x_k - v_j\right\|_{IFS}\right)^{\frac{1}{1-m}}} \tag{4-14}$$

Συνδυάζοντας τις (4-13) και (4-14) έχουμε:

$$\underset{\substack{1\le z\le c\\ 1\le k\le N}}{\forall} \quad u_{zk} = \frac{\left(\left\|x_k - v_z\right\|_{IFS}\right)^{\frac{1}{1-m}}}{\sum_{j=1}^{c}\left(\left\|x_k - v_j\right\|_{IFS}\right)^{\frac{1}{1-m}}} \tag{4-15}$$

Ομοίως με την $J_m(U,V)$, η συνάρτηση $J_m^{IFS}(U,V)$ μπορεί να ελαχιστοποιηθεί μόνο εάν:

$$\underset{\substack{1\le i\le c\\ i\le k\le N}}{\forall} \quad u_{ik} = \begin{cases} \dfrac{\left(\left\|x_k - v_i\right\|_{IFS}\right)^{\frac{1}{1-m}}}{\sum_{j=1}^{c}\left(\left\|x_k - v_j\right\|_{IFS}\right)^{\frac{1}{1-m}}}, & I_k = \varnothing, \\ \begin{cases} 0, & i \notin I_k \\ \sum_{i\in I_k} u_{ik} = 1, & i \in I_k, \end{cases} & I_k \neq \varnothing, \end{cases} \tag{4-16}$$

Ενώ τα κέντρα των συστάδων υπολογίζονται με την εξίσωση (4-6).

Πρέπει να σημειωθεί ότι το u_{ik} αναφέρεται στη συμμετοχή του *k*-οστού διαισθητικά ασαφούς διανύσματος στην *i*-οστή συστάδα και δεν έχει σχέση με την εσωτερική τιμή διαισθητικά ασαφούς συμμετοχής του διανύσματος (τη συμμετοχή που αναφέρεται στα χαρακτηριστικά). Επιπλέον, επειδή η προτεινόμενη συνάρτηση απόστασης μεταξύ δύο διανυσμάτων υπολογίζεται μόνο χρησιμοποιώντας τις τιμές διαισθητικά ασαφούς συμμετοχής και μη-συμμετοχής των διανυσμάτων, μετά τον υπολογισμό των κέντρων των συστάδων στην εξίσωση (4-6) και πριν την επόμενη επανάληψη, όπου γίνεται η ανανέωση των τιμών συμμετοχής των u_{ik} στις νέες συστάδες, χρειάζεται η προσθήκη ενός βήματος που θα υπολογίζει τις τιμές της διαισθητικά ασαφούς συμμετοχής και μη-συμμετοχής των νέων (εικονικών) συστάδων. Με άλλα λόγια, είναι απαραίτητο να υπολογιστούν οι τιμές συμμετοχής μ_{i_l} και μη-συμμετοχής γ_{i_l} για κάθε χαρακτηριστικό *l* που αντιστοιχεί στην *l*-οστή διάσταση του *i*-οστού

κέντρου (συστάδας). Σε κάθε επανάληψη και για κάθε κέντρο συστάδας εξάγεται ο βαθμός συμμετοχής μ_{i_l} του κέντρου συστάδας v_i ως ο μέσος όρος των βαθμών συμμετοχής όλων των διαισθητικά ασαφών διανυσμάτων που ανήκουν στη συστάδα *i*. Ομοίως, εξάγεται και ο βαθμός μη-συμμετοχής γ_{i_l} .

Πιο τυπικά, εάν P_i είναι ένα σύνολο ορισμένο ως:

$$\underset{1\le i\le c}{\forall} \quad P_i = \left\{ k \,\middle|\, 1 \le k \le N;\, d_{ik} < d_{rk}, \forall 1 \le r \le N \wedge r \ne i \right\} \tag{4-17}$$

Τότε το διαισθητικά ασαφές σύνολο IFS_{v_i} για το κέντρο v_i ορίζεται ως:

$$\underset{1\le i\le c}{\forall} \quad IFS_{v_i} = \underset{\forall k \in P_i}{@} IFS_k \tag{4-18}$$

Από το (Atanassov, 1994) έχουμε:

$$\underset{1\le l\le p}{\forall} \quad \mu_{i_l} = \frac{\sum_{\forall k \in P_i} \mu_{k_l}}{|P_i|}, \;\; \nu_{i_l} = \frac{\sum_{\forall k \in P_i} \gamma_{k_l}}{|P_i|} \tag{4-19}$$

Σύμφωνα με τα παραπάνω, ο τροποποιημένος αλγόριθμος FCM που συσταδοποιεί διαισθητικά ασαφή δεδομένα περιγράφεται ως εξής:

Αλγόριθμος 4-2. Αλγόριθμος iFCM

Βήμα 1 Καθορισμός του *c* (1 < *c* < *N*), $m \in [1,\infty)$ και αρχικοποίηση του $V^{(0)}$ με την τυχαία επιλογή *c* διαισθητικά ασαφών διανυσμάτων, $j \leftarrow 1$,

Βήμα 2: Υπολογισμός του πίνακα συμμετοχών $U^{(j)}$, χρησιμοποιώντας την εξίσωση (4-16),

Βήμα 3: Ενημέρωση του πίνακα των κέντρων $V^{(j)}$, χρησιμοποιώντας την εξίσωση (4-6) και τον $U^{(j)}$, και υπολογισμός των βαθμών συμμετοχής και μη-συμμετοχής του $V^{(j)}$ χρησιμοποιώντας την εξίσωση (4-19)

Βήμα 4: Εάν $\left\| U^{j+1} - U^{j} \right\|_F > \varepsilon$ τότε $j \leftarrow j+1$ και επιστροφή στο Βήμα 2.

Συγκριτικά με τον απλό αλγόριθμο FCM, το νέο μοντέλο συσταδοποίησης εισάγει (α) μία διαφορετική τακτική αρχικοποίησης του πίνακα *V* αφού στην

περίπτωσή μας, τα διανύσματα κέντρου είναι διαισθητικά ασαφή διανύσματα (Βήμα 1), (β) ένα νέο τρόπο υπολογισμού του βαθμού συμμετοχής ενός διανύσματος στη συστάδα, ο οποίος λαμβάνει υπόψη τις τιμές συμμετοχής και μη-συμμετοχής των διαισθητικά ασαφών διανυσμάτων (Βήμα 2) και (γ) μία μέθοδο για την ενημέρωση του πίνακα V σε κάθε επανάληψη βασισμένη μόνο στη θεωρία των διαισθητικά ασαφών συνόλων (Βήμα 3).

Η πολυπλοκότητα του FCM είναι $O(n\ f\ c^2\ i)$ όπου n είναι ο αριθμός των σημείων δεδομένων, f είναι ο αριθμός των διαστάσεων, c είναι ο αριθμός των συστάδων και i ο αριθμός των επαναλήψεων (Hore et al., 2007). Η πολυπλοκότητα του προτεινόμενου αλγορίθμου iFCM δεν διαφέρει από αυτήν του FCM αφού δεν υπάρχουν νέα βήματα στη διαδικασία και τα υπάρχοντα βήματα δεν εισάγουν περισσότερη πολυπλοκότητα.

4.2.4 Αναπαράσταση Ασαφών Προτύπων στη Βάση-Προτύπων

Τα διαισθητικά ασαφή δεδομένα μπορούν να αναπαρασταθούν στη βάση προτύπων ως απλά πρότυπα, ενώ το αποτέλεσμα του αλγορίθμου iFCM μπορεί να αποθηκευτεί ως σύνθετα πρότυπα όπως αυτό περιγράφηκε στο κεφάλαιο 3. Συνεχίζοντας το παράδειγμα που χρησιμοποιούμε για τα δεδομένα των εικόνων, μία εικόνα μπορεί να αναπαρασταθεί ως σύνθετο πρότυπο ως εξής:

$$image = \begin{pmatrix} SS : \{object\}, \\ MS : \perp \end{pmatrix}$$

Σχετικά με την αναπαράσταση των κέντρων των συστάδων, πρέπει να οριστούν τα μέρη της δομής και του μέτρου των προτύπων (structure και measure) προκειμένου να μπορεί να πραγματοποιηθεί η σύγκριση μεταξύ διαφορετικών αντικειμένων. Τα δεδομένα που έχουν υποστεί επεξεργασία ασάφειας (στην περίπτωση του παραδείγματός μας μία εικόνα) αναπαρίστανται χρησιμοποιώντας τέσσερις τιμές. Την τιμή του χαρακτηριστικού (δεδομένα), την τιμή συμμετοχής, της μη-συμμετοχής και της διστακτικότητας. Η τελευταία τιμή, της διστακτικότητας, δε χρειάζεται να οριστεί αποκλειστικά, αφού μπορεί να υπολογιστεί από τις τιμές της συμμετοχής και μη-συμμετοχής, σύμφωνα με τη θεωρία των ασαφών συνόλων.

Το μέρος της δομής (structure component) για κάθε αντικείμενο περιγράφεται από το διάνυσμα τιμών των δεδομένων, X, ενώ το μέρος του μέτρου (measure

component) περιλαμβάνει τα διανύσματα των διαισθητικά ασαφών τιμών M για τις τιμές συμμετοχής και Γ για τις τιμές μη-συμμετοχής.

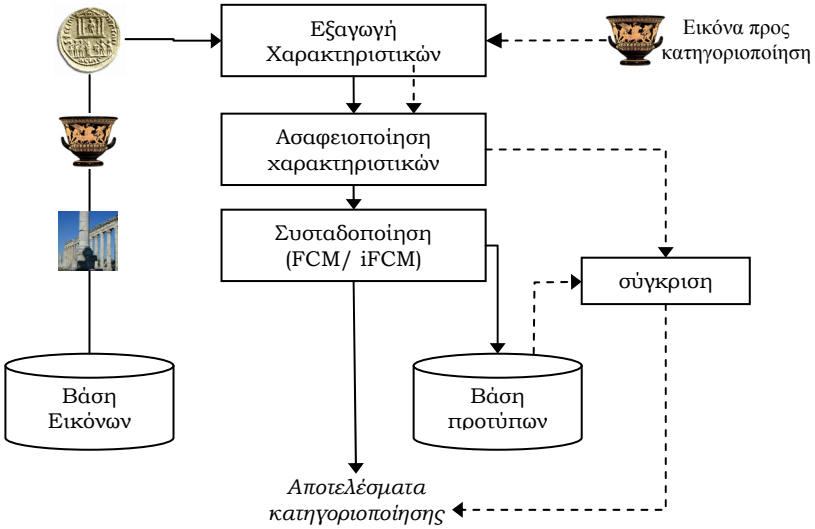
Έτσι θα είναι:

$$object_i = \begin{pmatrix} SS : (X : x : [\text{Real}]), \\ MS : [M : [\text{Real}], [\Gamma : [\text{Real}]] \end{pmatrix}$$

και

$$Cluster_i = \begin{pmatrix} SS : (X : x : [\text{Real}]), \\ MS : [M : [\text{Real}], [\Gamma : [\text{Real}]] \end{pmatrix}$$

Με την υποστήριξη της διαισθητικά ασαφούς συσταδοποίησης από το πλαίσιο του PBMS πολλές εφαρμογές μπορούν να υλοποιηθούν. Στην Εικόνα 4-1 παρουσιάζουμε τη μεθοδολογία της εφαρμογής κατηγοριοποίησης εικόνων σε κλάσεις. Ακολουθεί η περιγραφή της σχετικής μεθοδολογίας για την εφαρμογή αυτή.



Εικόνα 4-1 Κατηγοριοποίηση εικόνων με χρήση διαισθητικά ασαφούς συσταδοποίησης και της βάσης προτύπων

Τα εξαγόμενα χαρακτηριστικά από τις διαθέσιμες εικόνες ασαφειοποιούνται και ο αλγόριθμος iFCM χρησιμοποιείται για να συσταδοποιηθούν τα δεδομένα σε κλάσεις. Τα αποτελέσματα κατηγοριοποίησης αναπαριστούν τα κέντρα των συστάδων και αποθηκεύονται στην βάση προτύπων, χρησιμοποιώντας την αναπαράσταση του PBMS όπως φαίνεται παρακάτω. Η προτεινόμενη μεθοδολογία μπορεί επίσης να χρησιμοποιηθεί για να κατηγοριοποιηθούν οι εικόνες που δεν ανήκουν ήδη στην βάση των εικόνων. Στην περίπτωση αυτή, τα

χαρακτηριστικά της νέας εικόνας εξάγονται και εφαρμόζεται η ίδια μέθοδος ασαφειοποίησης. Το αποτέλεσμα συγκρίνεται κατόπιν, με τα κέντρα των συστάδων που είναι αποθηκευμένα στη βάση προτύπων, χρησιμοποιώντας το μέτρο ομοιότητας που έχει οριστεί στην ενότητα 4.2.2. Η νέα εικόνα θα κατηγοριοποιηθεί στη συστάδα που το κέντρο της είναι πιο όμοιο στην εικόνα σύμφωνα με το μέτρο ομοιότητας Z_1.

Το μέτρο ομοιότητας Z_1 μεταξύ των διαισθητικά ασαφών συνόλων A και B ορίζεται με την παρακάτω εξίσωση:

$$Z_1(A,B) = \frac{z_1(M_A, M_B) + z_1(\Gamma_A, \Gamma_B)}{2}$$

όπου

$$z_1(A',B') = \begin{cases} \dfrac{\sum_{i=1}^{n} \min(A'(x_i), B'(x_i))}{\sum_{i=1}^{n} \max(A'(x_i), B'(x_i))}, & A' \cup B' \neq \emptyset \\ 1, & A' \cup B' = \emptyset \end{cases}$$

με $A', B' \in \mathcal{F}(E)$.

Το μέτρο ομοιότητας ενσωματώνεται στο πλαίσιο PANDA και στην περίπτωση αυτή εφαρμόζεται μεταξύ των διανυσμάτων τιμών συμμετοχής και μη-συμμετοχής για κάθε αντικείμενο.

4.3 Εφαρμογή: Κατηγοριοποίηση Εικόνων με Διαισθητικά Ασαφή Συσταδοποίηση

Στην ενότητα αυτή παρουσιάζουμε μία πειραματική μελέτη συσταδοποίησης διαισθητικά ασαφών δεδομένων. Ορίζουμε μία διαισθητικά ασαφή αναπαράσταση των εικόνων και χρησιμοποιούμε την προτεινόμενη μετρική ομοιότητας για τη συσταδοποίησή τους. Στο διαθέσιμο σύνολο δεδομένων υπάρχει συγκεκριμένος αριθμός κλάσεων και έτσι, η συσταδοποίηση χρησιμοποιείται ως μέθοδος κατηγοριοποίησης (classification).

Η παρούσα μελέτη στοχεύει στην αξιολόγηση της προτεινόμενης μετρικής ομοιότητας και του αλγορίθμου iFCM γενικότερα. Η παρούσα εφαρμογή αυτή δεν συμπεριλαμβάνει την κατηγοριοποίηση νέων εικόνων με χρήση των αποτελεσμάτων της αρχικής κατηγοριοποίησης, καθότι αυτή η μεθοδολογία έχει παρουσιαστεί αναλυτικά στις ενότητες 3.4 και 3.5.

4.3.1 Διαισθητικά Ασαφής Αναπαράσταση των Δεδομένων

Η προτεινόμενη διαισθητικά ασαφής συσταδοποίηση απαιτεί κάθε στοιχείο δεδομένων x από ένα σύνολο τιμών E, να ανήκει σε ένα διαισθητικά ασαφές σύνολο $A \subset E$ με ένα βαθμό $\mu_A(x)$ και να μην ανήκει στο A με βαθμό $\gamma_A(x)$. Τα στοιχεία δεδομένων μπορεί να είναι οποιουδήποτε είδους. Για τις ανάγκες της παρούσας μελέτης που εστιάζει στη συσταδοποίηση δεδομένων εικόνων επεκτείνουμε τον ορισμό της διαισθητικά ασαφούς αναπαράστασης μίας ασπρόμαυρης εικόνας (Vlachos and Sergiadis, 2005), για την αναπαράσταση μίας έγχρωμης ψηφιακής εικόνας.

Ορισμός 4-6. Μία έγχρωμη ψηφιακή εικόνα P μεγέθους $a \times b$ pixels, που αποτελείται από ξ κανάλια P_k, k=1,2,…, ξ, ψηφιοποιημένη σε q επίπεδα κβαντοποίησης ανά κανάλι, αναπαρίσταται ως ένα διαισθητικά ασαφές σύνολο

$$\Phi = \left\{ \left\langle \theta_{ij}^k, \mu_\Phi(\theta_{ij}^k), \gamma_\Phi(\theta_{ij}^k) \right\rangle_k \middle| \theta_{ij}^k \in P_k,\ i = 1,2,...,a,\ j = 1,2,...,b,\ k = 1,2,...,\xi \right\} \tag{4-20}$$

όπου θ_{ij}^k είναι η τιμή του P_k στη θέση (i, j), ενώ $\mu_\Phi(\theta_{ij}^k)$ και $\gamma_\Phi(\theta_{ij}^k)$ ορίζουν τις τιμές της συμμετοχής και μη-συμμετοχής του θ_{ij}^k στο P_k, αντίστοιχα.

Ως συνάρτηση συμμετοχής $\mu_\Phi(\theta)$, ορίζουμε την πιθανότητα εμφάνισης του $\theta \in [0, q-1]$ σε ένα κανάλι της εικόνας

$$\mu_\Phi(\theta) = \frac{h(\theta)}{a \cdot b}, \quad \forall \theta \in [0, q-1] \tag{4-21}$$

όπου

$$h(\theta) = \left\| \left\{ (i, j) \in P_k \middle| \theta_{ij}^k = \theta;\ i = 1,...,a;\ j = 1,...,b,\ k = 1,2,...,\xi \right\} \right\|, \qquad \blacksquare$$

είναι ένα διακριτό ιστόγραμμα των τιμών των pixel σε ένα κανάλι, και $\|\cdot\|$ αναπαριστά τον πληθάριθμο του έγκλειστου συνόλου (enclosed set). Η πιθανοτική κατανομή που περιγράφεται με την εξίσωση (4-21) περιγράφει μία πρωτοβάθμια στατιστική αναπαράσταση του καναλιού της εικόνας, που είναι εύκολο να υπολογιστεί.

Δεδομένου ότι οι πραγματικές ψηφιακές εικόνες συνήθως περιέχουν θόρυβο διαφόρων πηγών και ανακρίβεια στις τιμές των καναλιών, ο βαθμός με τον οποίο ανήκει μία τιμή έντασης θ σε ένα κανάλι εικόνας όπως εκφράζεται με το

$\mu_{\Phi}(\theta)$ περιέχει αβεβαιότητα. Προκειμένου να μοντελοποιηθεί η περίπτωση αυτή, εισάγουμε έναν παράγοντα penalty $p(\theta)$ στο $\mu_{\Phi}(\theta)$ τέτοιο ώστε το θ να ανήκει λιγότερο στο κανάλι της εικόνας εάν το $h(\theta)$ αποκλίνει περισσότερο από το ασαφές ιστόγραμμα $\tilde{h}(\theta)$. Το ασαφές ιστόγραμμα, που προτάθηκε αρχικά από τους Jawahar και Ray (1996), ορίζεται ως

$$\tilde{h}(\theta) = \left\| \left\{ (i,j) \in P_k \left| \mu_{\tilde{\theta}}(\theta_{ij}^k);\ i = 1,...,a,\ j = 1,...,b,\ k = 1,2,..,\xi \right. \right\} \right\| \tag{4-22}$$

με

$$\mu_{\tilde{\theta}}(x) = \max\left(0,\ 1 - \frac{|x-\theta|}{\psi} \right) \tag{4-23}$$

όπου η παράμετρος ψ ελέγχει το διάστημα του ασαφούς αριθμού $\tilde{\theta}: R \to [0,1]$ που αναπαριστά ένα ασαφές επίπεδο έντασης θ. Αυτό σημαίνει ότι ένα pixel μίας δεδομένης τιμής καναλιού δε συνεισφέρει μόνο σε ένα συγκεκριμένο bin, αλλά επίσης και στο γειτονικό bin στο ιστόγραμμα. Έτσι, το ασαφές ιστόγραμμα γίνεται ομαλότερο και περισσότερο ανεκτικό στο θόρυβο από το αντίστοιχο διακριτό ιστόγραμμα, όσο αυξάνει το ψ.

Σύμφωνα με την προτεινόμενη σύνθεση η τιμή μη-συμμετοχής του θ στο κανάλι της εικόνας μπορεί να εκφραστεί ως

$$\gamma_{\Phi}(\theta) = 1 - \mu_{\Phi}(\theta) \cdot p(\theta) \tag{4-24}$$

Ο παράγοντας penalty $p(\theta)$ επιλέγεται να είναι αναλογικός στην διάσταση μεταξύ του διακριτού $h(\theta)$ και του ασαφούς ιστογράμματος $\tilde{h}(\theta)$, έτσι ώστε να ικανοποιεί την Εξ. (4-1)

$$p(\theta) = \lambda \cdot \frac{\left| h(\theta) - \tilde{h}(\theta) \right|}{\max_{\theta} \left(\left| h(\theta) - \tilde{h}(\theta) \right| \right)} \tag{4-25}$$

όπου $\lambda \in [0,1]$ είναι σταθερά και ο παρονομαστής χρησιμοποιείται για κανονικοποίηση. Η φυσική σημασία αυτής της συνάρτησης μη-συμμετοχής είναι ότι ο βαθμός που δεν ανήκει μία τιμή έντασης θ στο κανάλι μιας εικόνας αυξάνει κατά ένα παράγοντα που είναι ανάλογος στην τραχύτητα του διακριτού ιστογράμματος. Έτσι, όσο αυξάνει το επίπεδο θορύβου στο κανάλι, το διακριτό ιστόγραμμα γίνεται περισσότερο τραχύ (με μεγαλύτερα κομμάτια) και η διστακτικότητα στον καθορισμό της τιμής έντασης θ αυξάνει.

Η συμμετοχή και η μη-συμμετοχή που ορίζονται στις εξισώσεις (4-21) και (4-24) στις τιμές των καναλιών της εικόνας θεωρούνται διανύσματα χαρακτηριστικών.

Όπως φαίνεται στην ενότητα 4.2.4, η αναπαράσταση της συστάδας (του κέντρου της συστάδας) είναι:

$$Cluster_i = \begin{pmatrix} SS : (X : x : [\text{Real}]), \\ MS : [M : [\text{Real}], [\Gamma : [\text{Real}]] \end{pmatrix}$$

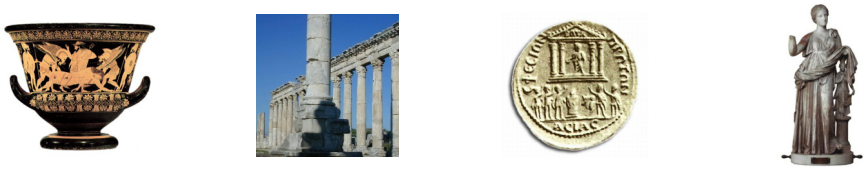
ενώ κάθε εικόνα είναι ένα σύνθετο πρότυπο αντικειμένων – πολυδιάστατων διανυσμάτων των χαρακτηριστικών της ασαφειοποιημένης εικόνας αναπαρίσταται ως:

$$object_i = \begin{pmatrix} SS : (X : x : [\text{Real}])^{\text{D}}, \\ MS : [M : [\text{Real}], [\Gamma : [\text{Real}]]^{\text{D}} \end{pmatrix}$$

Κάθε ασαφειοποιημένη εικόνα και τα κέντρα των συστάδων αποθηκεύονται στη βάση προτύπων για να είναι δυνατή η μελλοντική κατηγοριοποίηση.

4.3.2 Πειραματικά Αποτελέσματα

Για την αξιολόγηση της απόδοσης του προτεινόμενου αλγορίθμου συσταδοποίησης iFCM, σε σύγκριση με τον καθιερωμένο FCM, διεξήχθησαν αναλυτικά πειράματα. Το σενάριο εφαρμογής για την αξιολόγηση περιλαμβάνει συσταδοποίηση μίας συλλογής 400 εικόνων από τέσσερις διαφορετικές ισοπληθείς κλάσεις διαφορετικών χρωματικών θεμάτων και περιλαμβάνουν αμφορείς, αρχαία μνημεία, νομίσματα και αγάλματα (Εικόνα 4-2).



Κλάση Α **Κλάση Β** **Κλάση Γ** **Κλάση Δ**

(α) (β) (γ) (δ)

Εικόνα 4-2 Παραδείγματα εικόνων για τις τέσσερις κλάσεις που χρησιμοποιούνται για τα πειράματα, (α) αμφορείς, (β) αρχαία μνημεία, (γ) νομίσματα, και (δ) αγάλματα.

Οι εικόνες χορηγήθηκαν από το Ίδρυμα Μείζονος Ελληνισμού, το οποίο διατηρεί μία μεγάλη ψηφιακή βιβλιοθήκη κειμένων, εικόνων και πολυμεσικών

δεδομένων Ελληνικών ιστορικών αντικειμένων και τέχνης (FHW, 2009). Οι εικόνες είναι διαφορετικών μεγεθών και έχουν συλλεχθεί από διάφορες πηγές και έχουν ψηφιοποιηθεί σε 256 επίπεδα κβαντοποίησης ανά RGB κανάλι και έχουν σμικρυνθεί σε ένα τετράγωνο 256×256.

Η μεθοδολογία που ακολουθήθηκε στα πειράματα περιγράφεται στην Εικόνα 4-1.

Βασισμένοι στην παρατήρηση ότι το χρώμα είναι διαχωριστικό χαρακτηριστικό για τις περισσότερες από τις διαθέσιμες εικόνες, κάθε εικόνα αναπαρίσταται από ένα διαισθητικά ασαφές σύνολο σύμφωνα με την (4-20), χρησιμοποιώντας μόνο πληροφορία χρώματος ώστε να είναι προσεγγιστικά ανεξάρτητη από τις διαφορές τις έντασης. Προκειμένου να απο-συσχετιστεί η ένταση από την χρωματική πληροφορία, οι εικόνες μετασχηματίστηκαν στο χρωματικό χώρο $I_1I_2I_3$ σύμφωνα με την ακόλουθη εξίσωση (Ohta et al., 1980)

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 0.333 & 0.333 & 0.333 \\ 0.500 & 0.000 & -0.500 \\ -0.500 & 1.000 & -0.500 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{4.1}$$

Σε αυτόν το χρωματικό χώρο, η συνιστώσα I_1 περιγράφει την υψηλότερη αναλογία της συνολικής διακύμανσης και αναπαριστά την ένταση, ενώ οι I_2 και I_3 αντιστοιχούν στη δεύτερη και τρίτη μεγαλύτερη αναλογία, αντίστοιχα, και περιγράφουν χρωματική πληροφορία. Μία πολύ χρήσιμη ιδιότητα του χώρου αυτού είναι ότι οι περιοχές της εικόνας με διαφορετικό χρώμα μπορούν εύκολα να διαχωριστούν με απλές λειτουργίες κατωφλίου. Με άλλα λόγια, τα ιστογράμματα που παράγονται από τις τιμές χρωματικών συνιστωσών παρουσιάζουν κορυφές που αντιστοιχούν σε περιοχές διαφορετικών χρωμάτων.

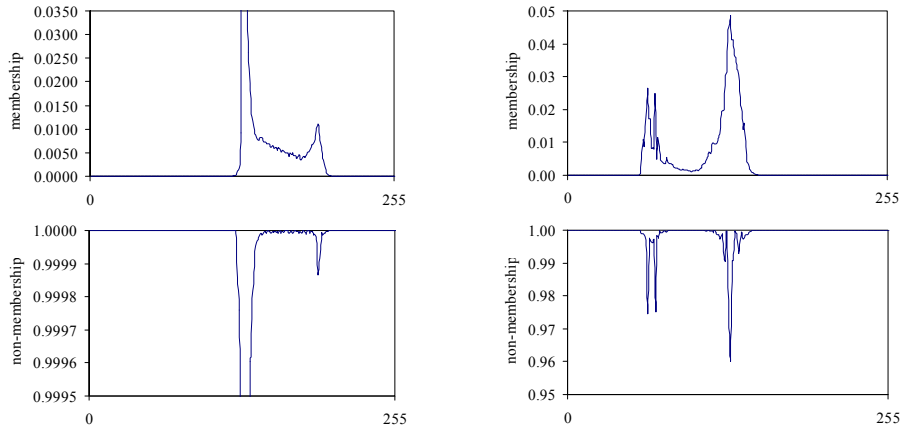
Από τις τρεις χρωματικές συνιστώσες $I_1I_2I_3$ επιλέγουμε την I_2 καθώς είναι η πιο διαχωριστική για τις χρωματικές περιοχές των διαθέσιμων εικόνων. Αυτό συνάδει με την εργασία (Ohta et al., 1980) που προτείνει ότι η διαχωριστική δύναμη της I_2 μπορεί ελάχιστα να βελτιωθεί με τη χρήση της I_3. Επιπλέον, παρατηρείται ότι τα κανάλια της εικόνας που αντιστοιχούν στη συνιστώσα I_3 παρουσιάζουν μικρό εύρος τιμών, με ιστόγραμμα που έχει μία κορυφή και ποικίλει ελάχιστα μεταξύ των εικόνων που ανήκουν σε διαφορετικές κλάσεις.

Παραδείγματα συναρτήσεων συμμετοχής και μη-συμμετοχής που χρησιμοποιούνται σε διαισθητικά ασαφή αναπαράσταση έγχρωμων εικόνων παρουσιάζονται στην Εικόνα 4-3. Οι τιμές των παραμέτρων που

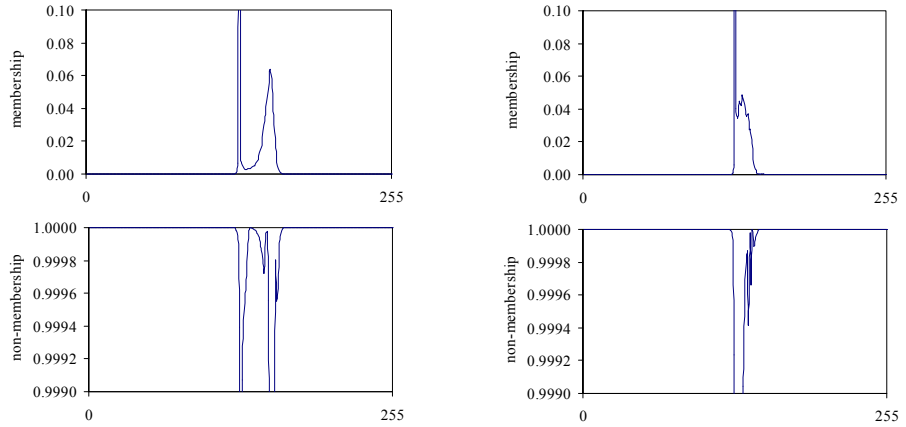
χρησιμοποιούνται στις εξισώσεις (4-23) - (4-25) για την εκτίμηση των συναρτήσεων συμμετοχής και μη-συμμετοχής είναι λ = 1 και ψ = 5. Οι οριζόντιοι άξονες αναπαριστούν τις τιμές του I_2 κανονικοποιημένες στο εύρος [0, 255], ενώ οι κάθετοι άξονες έχουν τροποποιημένη κλίμακα για να βελτιωθεί η οπτική εικόνα των γραφημάτων. Τα γραφήματα εστιάζουν στις περιοχές των συναρτήσεων συμμετοχής στις οποίες η διακύμανση είναι υψηλότερη. Οι γραμμές που αγγίζουν το πλαίσιο των γραφημάτων επεκτείνονται πέρα από τα όρια της ορατής περιοχής και ενώνονται σε κορυφή τιμών συμμετοχής και μη-συμμετοχής. Στις εικόνες 4-4α, 4-4γ και 4-4δ, οι ψηλότερες από τις δύο κορυφές αντιστοιχεί στις περιοχές άσπρου φόντου των εικόνων, ενώ οι χαμηλότερες κορυφές αντιστοιχούν στα απεικονιζόμενα αντικείμενα. Παρομοίως, στην Εικόνα 4-3β η υψηλότερη κορυφή αντιστοιχεί στο μάρμαρο των αρχαίων μνημείων και οι χαμηλότερη κορυφή αντιστοιχεί στην περιοχή του ουρανού. Σχετικά με τις συναρτήσεις μη-συμμετοχής, μία διαισθητική αναπαράσταση μπορεί να γίνει από τη συσχέτισή τους με τις αντίστοιχες συναρτήσεις συμμετοχής. Η συσχέτιση είναι συνήθως μικρότερη σε σημεία γύρω από τις κορυφές που αντιστοιχούν σε λιγότερο ομοιογενής περιοχές εικόνων. Για παράδειγμα στην Εικόνα 4-3β, η απόλυτη συσχέτιση μεταξύ των συναρτήσεων συμμετοχής και μη-συμμετοχής που υπολογίζεται για την περιοχή των αρχαίων μνημείων είναι 70%, ενώ για τις περιοχές του ουρανού είναι 82%. Παρομοίως, η απόλυτη συσχέτιση μεταξύ της συμμετοχής και μη-συμμετοχής στις εικόνες 4-4α, 4-4β και 4-4γ, για τις ομοιογενείς περιοχές άσπρου φόντου φτάνει το 96.5%. Διεξήχθησαν πειράματα με όλους τους δυνατούς συνδυασμούς κλάσεων, χρησιμοποιώντας α) τον προτεινόμενο αλγόριθμο με τα διαισθητικά ασαφή δεδομένα, β) τον αλγόριθμο FCM με δεδομένα διακριτού ιστογράμματος I_2, και γ) τον FCM με ασαφή δεδομένα ιστογράμματος I_2. Σε όλα τα πειράματα χρησιμοποιήθηκαν οι ίδιες παράμετροι (ε=0.00001, m=2.0) και συνθήκες αρχικοποίησης. Η απόδοση της συσταδοποίησης υπολογίστηκε με όρους ακρίβειας κατηγοριοποίησης, επαναλήψεων του αλγορίθμου και απόλυτου χρόνου εκτέλεσης. Η ακρίβεια κατηγοριοποίησης υπολογίζεται όπως στον αρχικό αλγόριθμο FCM, με την ανάθεση μίας εικόνας στη συστάδα με τον υψηλότερο βαθμό της τιμής συμμετοχής.

Τα πειράματα εκτελέστηκαν σε PC με Intel Pentium M στα 1.86 GHz, 512 MB RAM και 60 GB σκληρό δίσκο. Τα αποτελέσματα συνοψίζονται στην Εικόνα 4-4.

Η Εικόνα 4-4 δείχνει ότι σε όλα τα πειράματα η ακρίβεια που επιτυγχάνεται από τον προτεινόμενο αλγόριθμο είναι ψηλότερη από αυτήν που πετυχαίνεται από τον FCM για τέσσερις ή τρεις κλάσεις. Οι υψηλότερες ακρίβειες που επιτυγχάνονται από τον προτεινόμενο αλγόριθμο είναι 74.4% και 93.3% για τέσσερις και τρεις κλάσεις αντίστοιχα.

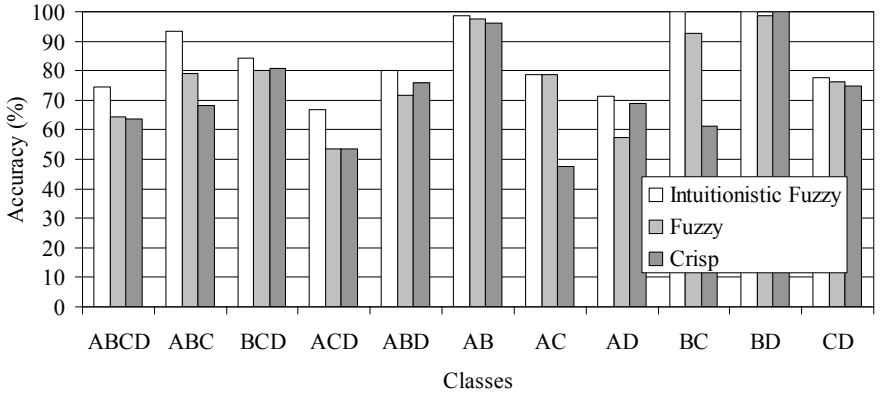


(α) (β)

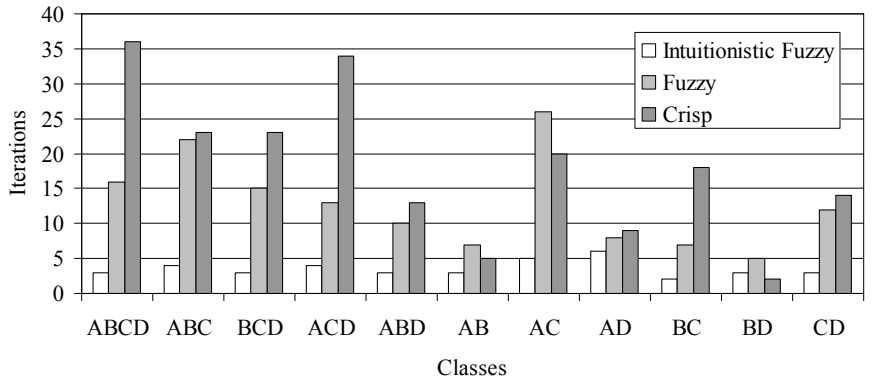


(γ) (δ)

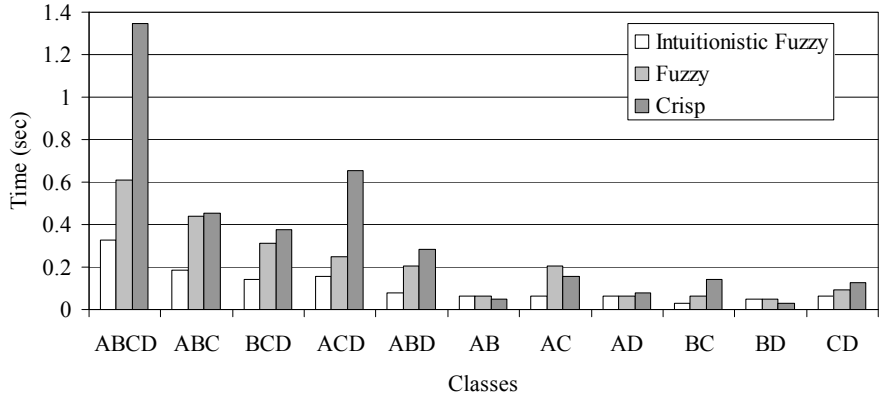
Εικόνα 4-3 Συναρτήσεις συμμετοχής και μη-συμμετοχής που αντιστοιχούν στις κλάσεις που παρουσιάζονται στην Εικόνα 4-2.



(α)



(β)



(γ)

Εικόνα 4-4 Συγκριτικά αποτελέσματα χρήσης του προτεινόμενου αλγορίθμου με διαισθητικά ασαφή δεδομένα, και χρήσης του FCM με διακριτά και με ασαφή δεδομένα ως είσοδο: (α) ακρίβεια κατηγοριοποίησης, (β) αριθμός επαναλήψεων που απαιτούνται για τη σύγκλιση των αλγορίθμων συσταδοποίησης, και (γ) χρόνος εκτέλεσης σε δευτερόλεπτα.

Αυτά τα ποσοστά μειώνονται στο 64.4% και 79.2%, στην περίπτωση της συσταδοποίησης με FCM με ασαφή δεδομένα. Τα αποτελέσματα των πειραμάτων με δεδομένα από δύο κλάσεις δείχνουν ότι η ακρίβεια του προτεινόμενου αλγορίθμου iFCM μπορεί να θεωρηθεί ίση ή μεγαλύτερη από αυτήν του αλγορίθμου FCM. Εντούτοις, αυτό μπορεί να αποδοθεί σε μία μικρότερη συμμετοχή των τιμών μη-συμμετοχής στην αναπαράσταση των εικόνων των συγκεκριμένων κλάσεων. Η μεγαλύτερη ακρίβεια που πετυχαίνεται κι από τους δύο αλγόριθμους έφτασε το 100% στις δύο περιπτώσεις (BC and BD).

Συγκρίνοντας τους δύο αλγορίθμους σε όρους αποδοτικότητας, η Εικόνα 4-4β και η Εικόνα 4-4γ δείχνουν ότι ο προτεινόμενος αλγόριθμος έχει πλεονέκτημα έναντι του FCM, καθώς απαιτεί λιγότερες επαναλήψεις και στις περισσότερες περιπτώσεις λιγότερο χρόνο για να συγκλίνει. Η μέση βελτίωση απόλυτης εκτέλεσης είναι 63±27%.

4.4 Σύνοψη

Στο κεφάλαιο αυτό, σε αντίθεση με το κεφάλαιο 3, εστιάζουμε στην διαισθητικά ασαφή συσταδοποίηση, που επίσης μπορεί να υποστηριχτεί από την έννοια του PBMS χρησιμοποιώντας την κατάλληλη αναπαράσταση. Περιγράψαμε μία παρόμοια εφαρμογή με αυτήν που παρουσιάστηκε στο κεφάλαιο 3 για την κατηγοριοποίηση εικόνων σε προκαθορισμένες κλάσεις. Παρουσιάστηκε επίσης λεπτομερώς μία πρωτότυπη μετρική σύγκρισης και ένα νέο αλγόριθμο Διαισθητικά Ασαφούς Συσταδοποίησης.

Οι διάφορες συσταδοποιήσεις ομαδοποιούν ένα σύνολο αντικειμένων σε ομάδες των οποίων τα μέλη είναι όμοια σύμφωνα με μία συνάρτηση ομοιότητας που ορίζεται στα χαμηλού επιπέδου χαρακτηριστικά, δεδομένου ότι οι τιμές τους δεν έχουν κάποιο είδος αβεβαιότητας. Επιπλέον, αυτές οι μέθοδοι θεωρούν ότι η ομοιότητα μετράται υπολογίζοντας μόνο το βαθμό με τον οποίο δύο οντότητες σχετίζονται, αγνοώντας τη διστακτικότητα που εισάγεται από το βαθμό που αυτές δεν σχετίζονται. Παρακινούμενοι από πραγματικές εφαρμογές προβλημάτων συσταδοποίησης, προτείναμε ένα σχήμα ασαφούς συσταδοποίησης δεδομένων που παράγονται στο πλαίσιο της θεωρίας διαισθητικά ασαφών συνόλων. Πιο συγκεκριμένα, παρουσιάσαμε μία πρωτότυπη παραλλαγή του αλγόριθμου συσταδοποίησης Fuzzy C-Means (FCM) που αντιμετωπίζει την αβεβαιότητα που παρουσιάζουν τα διανύσματα

χαρακτηριστικών λόγω ανακριβών μετρήσεων και θορύβου, και μία πρωτότυπη μετρική ομοιότητας μεταξύ διαισθητικά ασαφών συνόλων, η οποία ενσωματώνεται κατάλληλα στον αλγόριθμο συσταδοποίησης. Επίσης, παρουσιάσαμε μία διαισθητικά ασαφής αναπαράσταση έγχρωμων ψηφιακών εικόνων ως παράδειγμα διαισθητικά ασαφειοποίησης δεδομένων.

Για την αξιολόγηση της προσέγγισής μας, παρουσιάσαμε μία διαισθητικά ασαφειοποίηση έγχρωμων ψηφιακών εικόνων πάνω στις οποίες εφαρμόστηκε το προτεινόμενο σχήμα συσταδοποίησης. Τα πειραματικά αποτελέσματα του προτεινόμενου σχήματος δείχνουν ότι αυτό μπορεί να είναι πιο αποτελεσματικό και αποδοτικό από τον καθιερωμένο αλγόριθμο FCM, ειδικά όσο ο αριθμός των συστάδων αυξάνεται, δημιουργώντας έτσι καλές προοπτικές για διάφορες εφαρμογές. Η όλη διαδικασία διαισθητικά ασαφούς συσταδοποίησης υποστηρίζεται από το PBMS δίνοντας προχωρημένες δυνατότητες σε εφαρμογές κατηγοριοποίησης.

5 Άλλες Εφαρμογές του Συστήματος Διαχείρισης Βάσεων Προτύπων

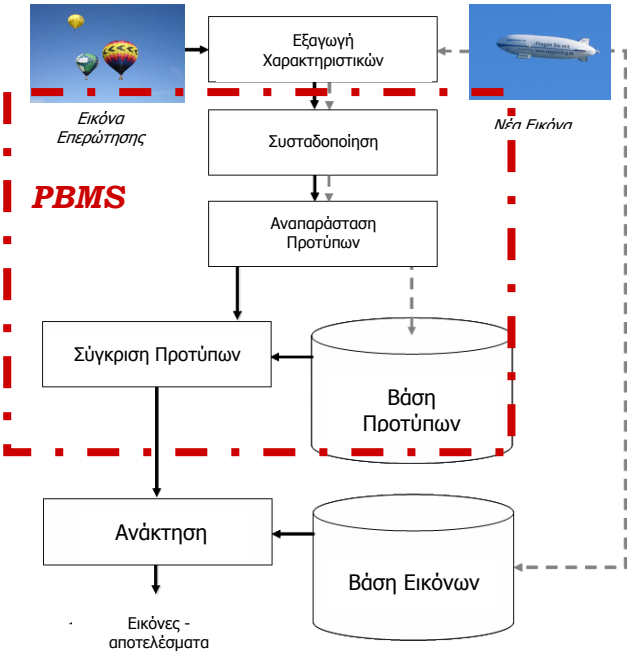
Στο κεφάλαιο παρουσιάζουμε πραγματικά προβλήματα εφαρμογών, όπως αυτό της ανάκτησης εικόνων με βάση το περιεχόμενο (CBIR), ή της κατηγοριοποίησης αστρονομικών δεδομένων – πιο συγκεκριμένα φασμάτων γαλαξιών.

Έχουμε ήδη παρουσιάσει εφαρμογές του PBMS για CBIR, για το λόγο αυτό απλά συνοψίζουμε την προσέγγιση μας. Στην περίπτωση της κατηγοριοποίησης των δεδομένων γαλαξιακού φάσματος, παρουσιάζουμε ένα πραγματικό πρόβλημα με το οποίο ασχοληθήκαμε σε συνεργασία με το τμήμα Αστροφυσικής του Πανεπιστημίου Αθηνών.

5.1 Εισαγωγή

Όπως αναφέρθηκε στις ενότητες 3.4 και 3.5, το PBMS μπορεί να χρησιμοποιηθεί ως ένα λειτουργικό μέρος ενός CBIR που με τη χρήση τεχνικών εξόρυξης γνώσεις, ανακτά όμοιες εικόνες. Η Εικόνα 5-1 δείχνει την προσέγγιση που προτείνεται στην ενότητα 3.5. Το τετράγωνο με τις διακεκομμένες κόκκινες γραμμές δείχνει το μέρος του συστήματος που μπορεί να αντικατασταθεί από το PBMS.

Όπως φαίνεται στην Εικόνα 5-1 το PBMS αντικαθιστά τον πυρήνα του CBIR. Κάθε εικόνα αναπαρίσταται από πρότυπα και έτσι η σύγκριση των προτύπων αντιστοιχεί σε σύγκριση των αντίστοιχων εικόνων. Τα πρότυπα αποθηκεύονται ως XML έγγραφα και η σύγκρισή τους είναι ευκολότερη και ταχύτερη.



Εικόνα 5-1 Περιγραφή της προσέγγισης CBIR βασισμένο στη βάση προτύπων και του μέρους που μπορεί να αντικατασταθεί από το PBMS.

5.2 Εφαρμογή του PBMS για την κατηγοριοποίηση Αστρουομικών Δεδομένων

Στα πλαίσια της συνεργασίας μας με το τμήμα Αστροφυσικής του Πανεπιστημίου Αθηνών αντιμετωπίσαμε το πρόβλημα της εύρεσης του καλύτερου μοντέλου κατηγοριοποίησης αστρονομικών δεδομένων. Αυτή η διαδικασία είναι μέρος του πακέτου εργασίας «Unresolved galaxy classifier» του προγράμματος GAIA (2009) της Ευρωπαϊκής Εταιρείας Διαστήματος. Ο στόχος του πακέτου εργασίας είναι «η μελέτη, ανάπτυξη και έλεγχος των αλγορίθμων που θα παρέχουν τις βέλτιστη κατηγοριοποίηση του γαλαξιακού φάσματος βάση της παραδοχής ότι το κάθε αντικείμενο ανήκει σε μία μόνο κλάση» (GAIA, 2009).

Χρησιμοποιώντας συνθετικά αλλά και πραγματικά δεδομένα (φάσματα γαλαξιών), διάφορα μοντέλα κατηγοριοποίησης πρέπει να δοκιμαστούν. Το καλύτερο μοντέλο θα χρησιμοποιηθεί τελικά σε ένα σύστημα που θα συλλέγει δεδομένα (φάσματα γαλαξιών) από το διαστημικό τηλεσκόπιο GAIA και θα κατηγοριοποιεί αυτόματα κάθε γαλαξία που παρατηρεί, σε προκαθορισμένες κλάσεις (τους γαλαξιακούς μορφολογικούς τύπους) εξοικονομώντας πολύτιμο χρόνο από τους ειδικούς επιστήμονες, αλλά και δίνοντας τη δυνατότητα προχωρημένης επεξεργασίας.

Προκειμένου να βρεθεί ο καλύτερος αλγόριθμος κατηγοριοποίησης, δηλαδή ο αλγόριθμος αλλά και οι παράμετροι που δίνουν τα πιο ακριβή αποτελέσματα, δεδομένου ενός συνόλου πραγματικών και συνθετικών δεδομένων για την εκπαίδευση και δοκιμή του αλγορίθμου, οι ειδικοί πρέπει να διεξάγουν έναν μεγάλο αριθμό πειραμάτων με διάφορους αλγορίθμους. Τα αποτελέσματα των πειραμάτων αυτών πρέπει να αξιολογηθούν και κατόπιν θα οριστεί το καλύτερο μοντέλο κατηγοριοποίησης. Στην εφαρμογή αυτή, η χρήση του PBMS μπορεί να υποστηρίξει την αποθήκευση, σύγκριση και εκμετάλλευση των αποτελεσμάτων αυτών με αποδοτικό τρόπο.

Στη συγκεκριμένη μελέτη, τρεις διαφορετικοί αλγόριθμοι χρησιμοποιήθηκαν, ο J.48 (μία παραλλαγή του C4.5 για το εργαλείο εξόρυξης WEKA), ο αλγόριθμος κατηγοριοποίησης Naïve Bayes από το εργαλείο WEKA και το μοντέλο Support Vector Machine από το εργαλείο R (R-project, 2009). Ο J4.8 επιλέχτηκε ως παραλλαγή του πολύ διαδεδομένου αλγορίθμου δέντρων απόφασης C4.5, ενώ ο αλγόριθμος Naïve bayes είναι ένας επίσης πολύ διαδεδομένος αλγόριθμος κατηγοριοποίησης που χρησιμοποιεί την έννοια της ανεξαρτησίας γνωρισμάτων. Τα μοντέλα SVM (Cristianini, 2000) έχουν ήδη χρησιμοποιηθεί από τους αστρονόμους του προγράμματος GAIA για τη διεξαγωγή πειραμάτων κατηγοριοποίησης.

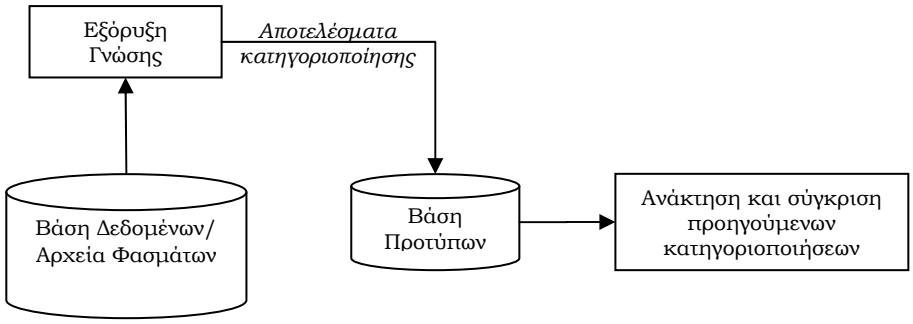
Χωρίς τη χρήση του PBMS οι ειδικοί επιστήμονες θα πρέπει να εξάγουν και να αποθηκεύουν ένα προς ένα τα αποτελέσματα των αλγορίθμων στο σύστημα αρχείων του υπολογιστή. Η σύγκριση των αποτελεσμάτων στην περίπτωση αυτή είναι μία χειροκίνητη διαδικασία στην οποία ο ειδικός ψάχνει τα αρχεία του συστήματος για να βρει αρχεία που περιέχουν τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης προκειμένου να τα συγκρίνει.

Η όλη διαδικασία απαιτεί ιδιαίτερη οργάνωση από τους χρήστες/ ειδικούς και είναι πολύ χρονοβόρα και πολύπλοκη. Οι χρήστες πρέπει να διεξάγουν τα πειράματα, να αποθηκεύσουν το αποτέλεσμα στο σύστημα αρχείων, σημειώνοντας παράλληλα σε κάποιο άλλο αρχείο τις παραμέτρους που χρησιμοποιήθηκαν για κάθε εκτέλεση των πειραμάτων, και άλλα μεταδεδομένα όπως είναι το σύνολο δεδομένων που χρησιμοποιήθηκε ή ο χρόνος εκτέλεσης του αλγορίθμου. Μετά την κατηγοριοποίηση, οι ειδικοί πρέπει να αξιολογήσουν ένα προς ένα τα αποτελέσματα και να τα συγκρίνουν με όλα τα υπόλοιπα με βάση τα αρχεία που έχουν αποθηκεύσει, μία χρονοβόρα και επίπονη διαδικασία.

Χρησιμοποιώντας το PBMS η διαδικασία απλοποιείται. Ο χρήστης μπορεί να εκτελέσει τον αλγόριθμο κατηγοριοποίησης του εργαλείου εξόρυξης γνώσης, και το αποτέλεσμα μαζί με τα απαιτούμενα μεταδεδομένα αποθηκεύονται στη βάση προτύπων με ένα όνομα που ορίζει και μπορεί να ανακαλέσει εύκολα ο χρήστης. Σε οποιαδήποτε στιγμή ο χρήστης μπορεί να ανακαλέσει ένα συγκεκριμένο αποτέλεσμα κατηγοριοποίησης κάνοντας την κατάλληλη επερώτηση για να ανακτήσει επίσης την ακρίβεια του προτύπου ή του μοντέλου ή τα μεταδεδομένα της εκτέλεσης.

Κάποιες περισσότερο περίπλοκες επερωτήσεις μπορούν επίσης να υποστηριχθούν, όπως:

- Ανάκτηση του αλγορίθμου και των παραμέτρων εκτέλεσης που δίνουν το καλύτερο αποτέλεσμα κατηγοριοποίησης για το σύνολο δεδομένων “A”.
- Ανάκτηση του συνόλου δεδομένων που έχει τη χειρότερη ακρίβεια κατηγοριοποίησης όταν χρησιμοποιείται ο αλγόριθμος naïve bayes.



Εικόνα 5-2 Χρήση του PBMS για την εκτέλεση πολλαπλών πειραμάτων κατηγοριοποίησης

Η Εικόνα 5-2 παρουσιάζει τον τρόπο που πολλαπλά πειράματα κατηγοριοποίησης μπορεί να διεξαχθούν με τη χρήση μίας pattern-base για να αποθηκευτούν κάθε φόρα τα αποτελέσματα της κατηγοριοποίησης και για την ανάκτησή τους προκειμένου να συγκριθούν με άλλα νεότερα πειράματα (νεότερες εκτελέσεις των αλγορίθμων στα ίδια δεδομένα). Όλα τα βήματα της παραπάνω διαδικασίας μπορούν να εκτελεστούν χρησιμοποιώντας το PBMS.

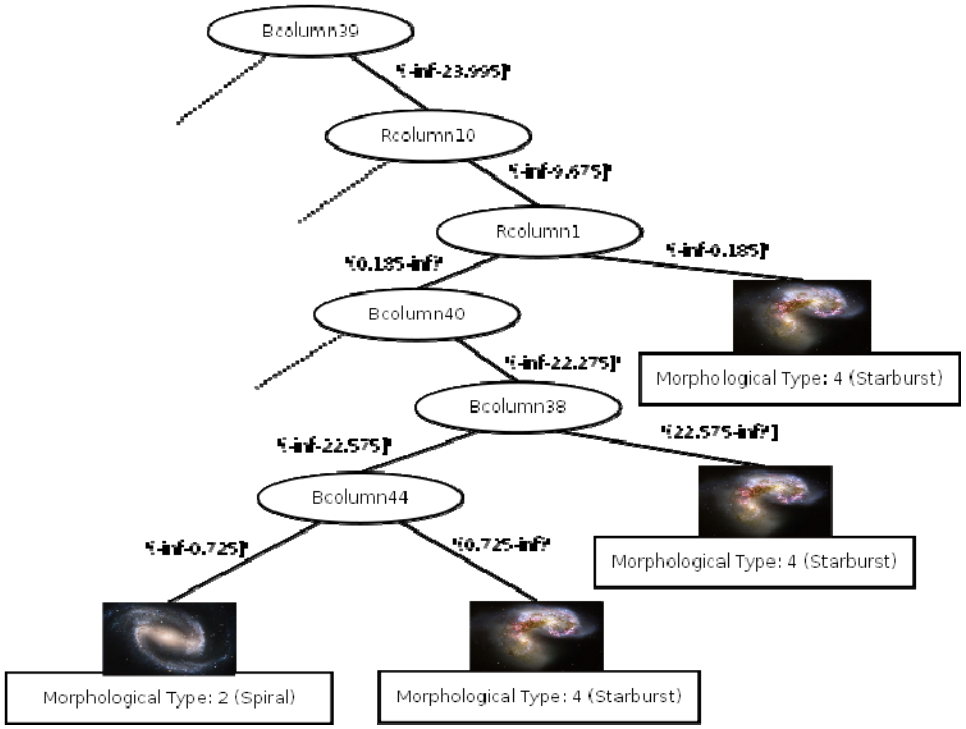
Προκειμένου να χρησιμοποιηθεί όμως το PBMS για τα πειράματα αυτά, πρέπει να οριστεί ένα μοντέλο XML που θα περιγράφει το αποτέλεσμα των αλγορίθμων κατηγοριοποίησης. Αυτό απαιτείται ιδιαίτερα αν το PBMS δεν υποστηρίζει ήδη τους συγκεκριμένους αλγορίθμους και βεβαίως θα είναι διαθέσιμα για μελλοντικές εφαρμογές.

Τα δέντρα απόφασης που παράγονται από τους αλγορίθμους κατηγοριοποίησης αναπαρίστανται με τη χρήση των μοντέλων που παρουσιάζονται στην ενότητα 2.3 ως εξής:

aPath =
(SS : [(ValueFrom: Real, ValueTo: Real)]$_1^N$,
MS: sup: Real)

aDecisionTree =
(SS : {Path},
MS: ⊥)

Τα αποτελέσματα των τριών αλγορίθμων κατηγοριοποίησης που αποθηκεύονται στη βάση προτύπων, μπορούν εύκολα να συγκριθούν μέσω του PBMS. Οι ειδικοί μπορούν εύκολα να αποφασίσουν κατόπιν ποιο μοντέλο θα χρησιμοποιήσουν. Στην Εικόνα 5-3 παρουσιάζεται ένα μέρος του δέντρου κατηγοριοποίησης. Το γνώρισμα κατηγοριοποίησης είναι ο μορφολογικός τύπος.



Εικόνα 5-3 Μέρος του δέντρου κατηγοριοποίησης κατασκευασμένο από τον αλγόριθμο J4.8, που δείχνει τις στήλες B (Blue spectrum area) και R (Red spectrum area) και τις διαφορετικές κλάσεις ανάλογα με τις τιμές του φάσματος.

Οι τέσσερις μορφολογικοί τύποι στους οποίους μπορούν να κατηγοριοποιηθούν οι γαλαξίες είναι: Early, Spiral, Irregular και Starburst. Εικόνες από τον κάθε γαλαξιακό τύπο παρουσιάζονται παρακάτω.

Εικόνα 5-4 Γαλαξιακός τύπος Early

Εικόνα 5-5 Γαλαξιακός τύπος Spiral

Εικόνα 5-6 Γαλαξιακός τύπος Irregular

Εικόνα 5-7 Γαλαξιακός τύπος Starburst

Εκτελέστηκαν πειράματα χρησιμοποιώντας τους αλγορίθμους J4.8 και Naïve Bayes του εργαλείου WEKA και τα αποτελέσματα συγκρίθηκαν με αυτά του καλύτερου μοντέλου SVM που προέκυπτε από το εργαλείο R ενώ όλα τα αποτελέσματα αποθηκεύονται σε μία κοινή βάση προτύπων σύμφωνα με το PBMS μοντέλο.

Πριν την εκτέλεση των πειραμάτων και την αξιολόγηση των αλγορίθμων κατηγοριοποίησης, τα φασματικά δεδομένα έπρεπε να διακριτοποιηθούν σε bins ίσου μεγέθους και ίσης συχνότητας. Ο αριθμός των bins είναι επίσης προς μελέτη και είναι μέρος των πειραμάτων με τιμές από 2 έως 10.

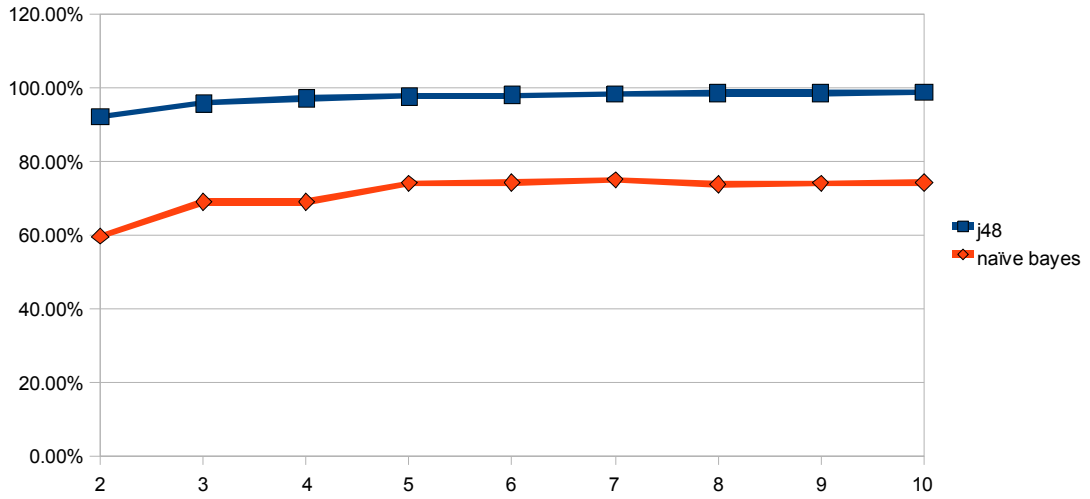
Πίνακας 5-1 Οι διάφορες περιπτώσεις πειραμάτων κατηγοριοποίησης

Number of bins	
2 to 10	
Discretization method	
equal width	equal frequency
Classification algorithm	
J48	Naive Bayes

Ο Πίνακας 5-1 περιγράφει όλες τις διαφορετικές περιπτώσεις πειραμάτων. Σε κάθε περίπτωση οι παρακάτω παράμετροι λαμβάνονται υπόψη:

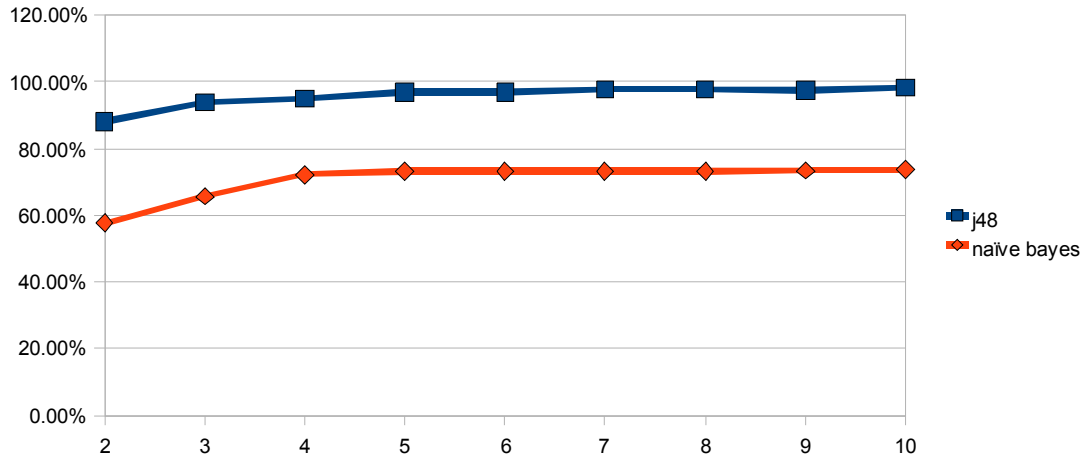
1. Ο αριθμός των bins των διακριτοποιημένων δεδομένων.
2. Η μέθοδος διακριτοποίησης (ίσο μέγεθος ή ίση συχνότητα για κάθε bin).
3. Ο αλγόριθμος κατηγοριοποίησης (Naive Bayes, J48).

Οι παρακάτω εικόνες (διαγράμματα) παρουσιάζουν τα αποτελέσματα ακρίβειας κατηγοριοποίησης για τους δύο αλγορίθμους για 2 έως 10 bins. Συγκεκριμένα η Εικόνα 5-8 παρουσιάζει την κατηγοριοποίηση για τα χωρισμένα δεδομένα με ίση συχνότητα bins, ενώ η Εικόνα 5-9 παρουσιάζει την κατηγοριοποίηση για τα χωρισμένα με ίσου μεγέθους bins. Φαίνεται και στα δύο διαγράμματα ότι ο J.48 αποδίδει πολύ καλύτερα αγγίζοντας το 100% της επιτυχίας, δίνοντας ωστόσο λίγο καλύτερα αποτελέσματα στα δεδομένα χωρισμένα σε bins ίσης συχνότητας.



Εικόνα 5-8 Αποτελέσματα κατηγοριοποίησης για τον αλγόριθμο J4.8 και τον Naive Bayes χρησιμοποιώντας bins διακριτοποίησης ίσης συχνότητας

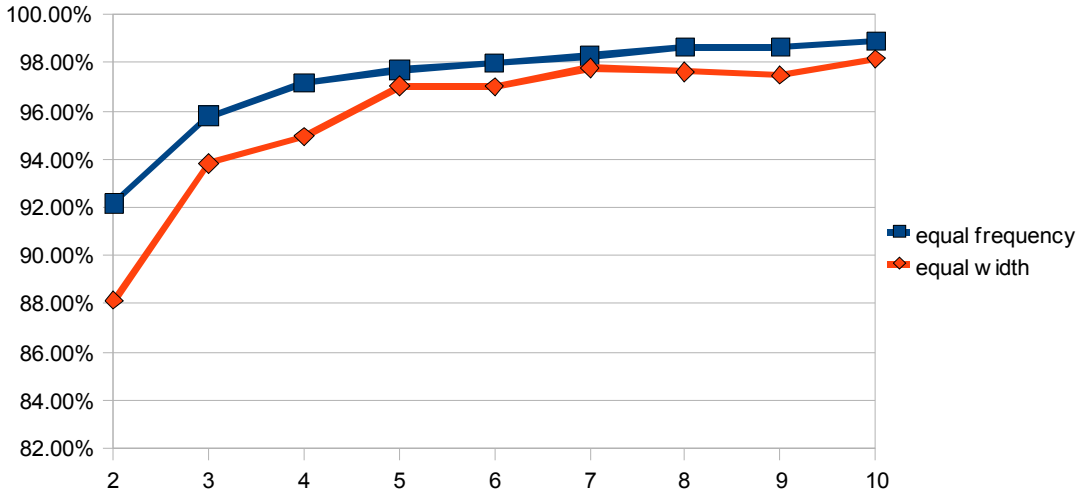
Ο αλγόριθμος J4.8 πάντα αποδίδει καλύτερα, έχοντας ακρίβεια κατηγοριοποίησης τουλάχιστον 95% και ένα μέσο όρο 97.25%, ενώ ο naïve bayes παρουσιάζει ένα μέγιστο μόνο 75.01%. Και στις δύο περιπτώσεις ωστόσο, η μέγιστη ακρίβεια επιτυγχάνεται για τη διακριτοποίηση των 10 bins.



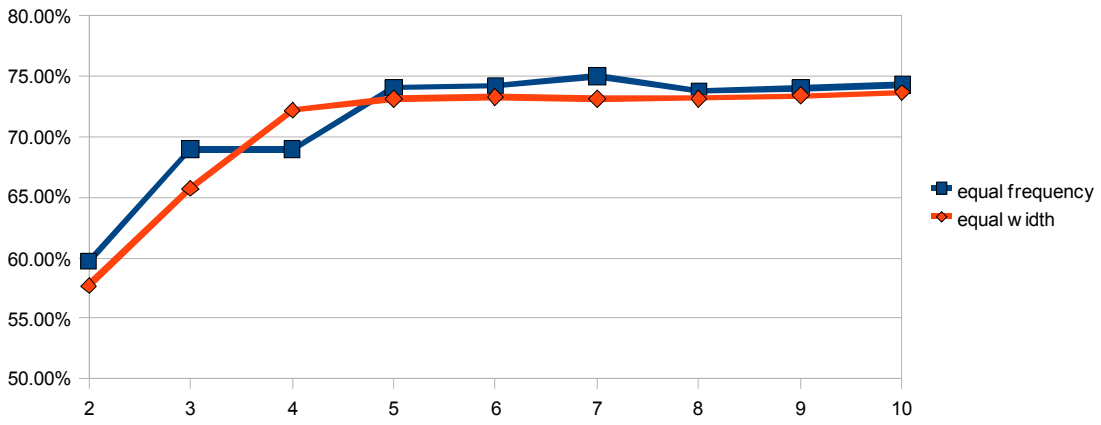
Εικόνα 5-9 Αποτελέσματα κατηγοριοποίησης για τον αλγόριθμο J4.8 και τον Naive Bayes χρησιμοποιώντας bins διακριτοποίησης ίσου μεγέθους

Στη μέθοδο διακριτοποίησης ίσου μεγέθους ο αλγόριθμος J4.8 αποδίδει επίσης καλύτερα έχοντας ένα μέσο όρο 95.78% σε αντίθεση με το 70.68% που πετυχαίνει ο naïve bayes. Η μέγιστη ακρίβεια κατηγοριοποίησης για τον J4.8 είναι στην περίπτωση αυτή 97.79% και για τον Naïve Bayes είναι 73.66%.

Συγκρίνοντας την ακρίβεια συσταδοποίησης για κάθε αλγόριθμο ξεχωριστά και για τις δύο μεθόδους διακριτοποίησης ίσου μεγέθους και ίσης συχνότητας, συμπεραίνουμε ότι η μέθοδος διακριτοποίησης ίσης συχνότητας δίνει καλύτερα αποτελέσματα και για τους δύο αλγορίθμους όπως φαίνεται στις παρακάτω εικόνες (Εικόνα 5-10 και Εικόνα 5-11).

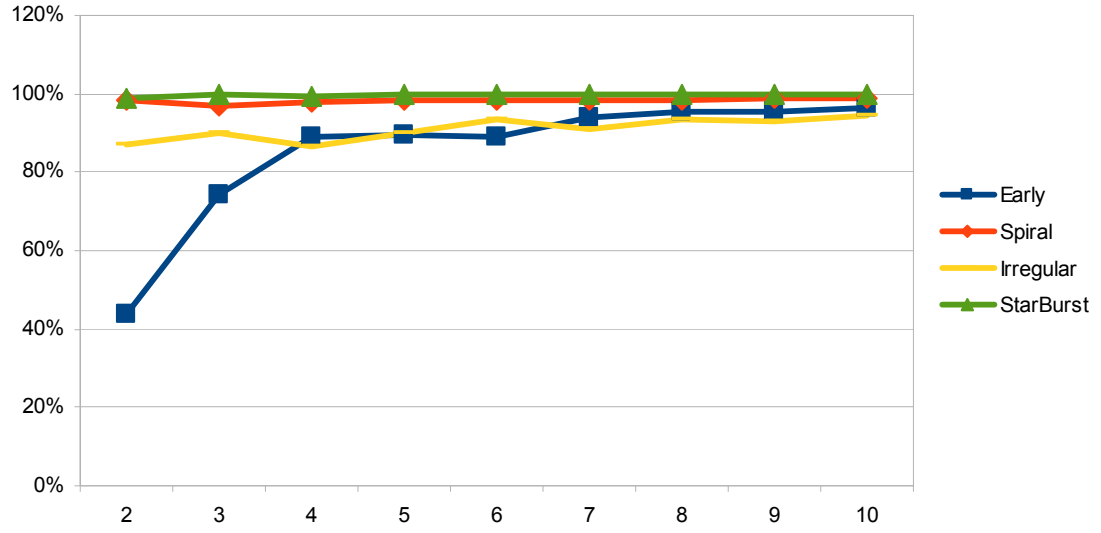


Εικόνα 5-10 Αποτελέσματα κατηγοριοποίησης για τον J4.8 συγκρίνοντας τις μεθόδους κατηγοριοποίησης ίσου μεγέθους και ίσης συχνότητας



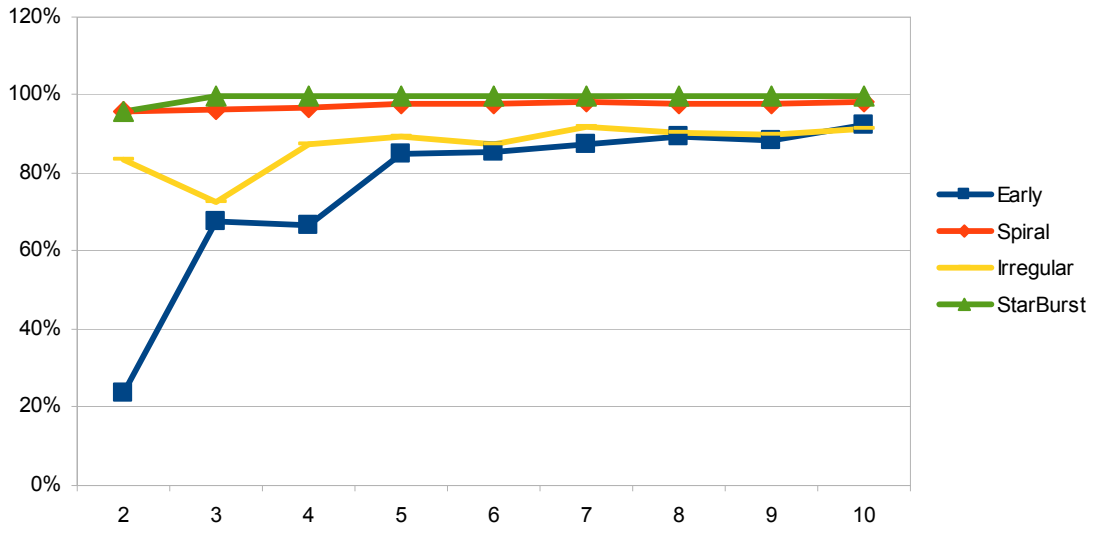
Εικόνα 5-11 Αποτελέσματα κατηγοριοποίησης για τον Naïve Bayes συγκρίνοντας τις μεθόδους κατηγοριοποίησης ίσου μεγέθους και ίσης συχνότητας

Επιπλέον, εκτελέστηκαν πειράματα για το *recall* και των δύο αλγορίθμων και για τους τέσσερις μορφολογικούς γαλαξιακούς τύπους. Το recall είναι ο λόγος που εκφράζει τον αριθμό των ορθά κατηγοριοποιημένων γαλαξιών σε μία κλάση (μορφολογικό τύπο) προς τον συνολικό αριθμό των γαλαξιών που ανήκουν στην κλάση αυτή. Στην Εικόνα 5-12 παρουσιάζεται το recall για όλους τους μορφολογικούς τύπους, στην περίπτωση του αλγορίθμου J4.8 και τη διακριτοποίηση ίσης συχνότητας. Για τους γαλαξιακούς τύπους Spiral και Starburst το recall είναι πολύ υψηλό για όλους τους αριθμούς των bins. Για τον τύπο Irregular, επιτυγχάνεται ένας μέσος όρος recall 90% ενώ στον τύπο early, το ποσοστό του recall είναι χαμηλό για δύο και τρία bins αλλά αυξάνεται but it raises στις περιπτώσεις των έξι ή περισσότερων bins φτάνοντας στο 92%.



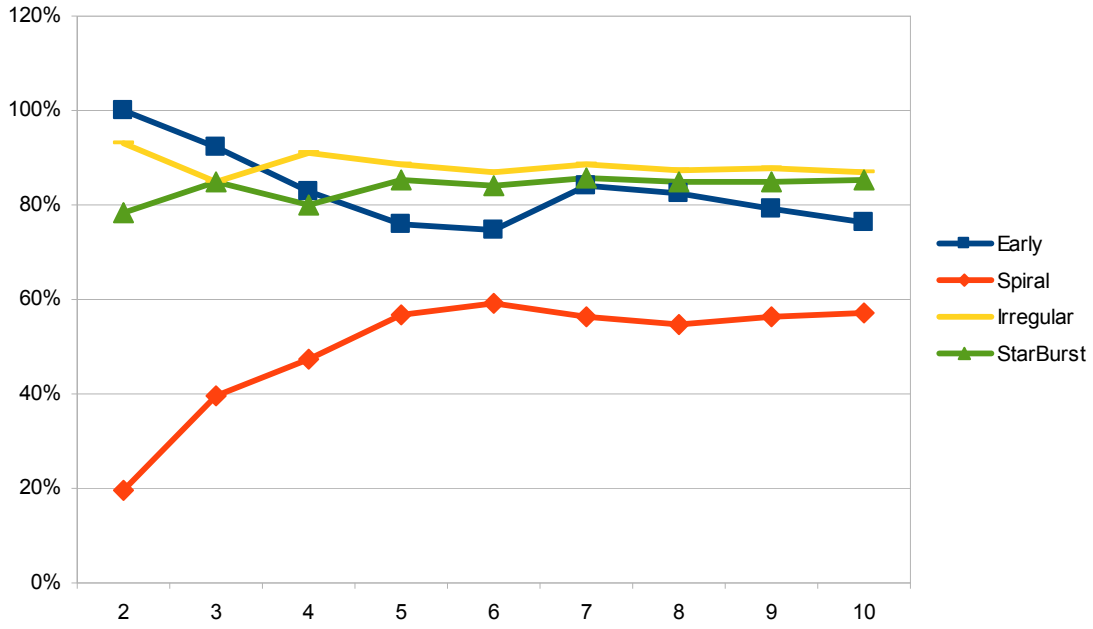
Εικόνα 5-12 Ο λόγος recall όλων των μορφολογικών τύπων για τον τύπο αλγόριθμο J4.8 και για τη μέθοδο διακριτοποίησης ίσης συχνότητας.

Η Εικόνα 5-13 παρουσιάζει τα ποσοστά του recall για όλους τους μορφολογικούς τύπους, στην περίπτωση του αλγορίθμου J4.8 και για τη μέθοδο διακριτοποίησης ίσου μεγέθους. Όπως στην περίπτωση της περίπτωσης ίσης συχνότητας, ο J4.8 καταφέρνει να κατηγοριοποιήσει με μεγάλη επιτυχία τους γαλαξίες τύπου Spiral και Starbursts. Το recall για τον τύπο Irregular ποικίλει από 72.6%, για τρία bins έως 91.7 % για επτά bins. Σχετικά με το recall του τύπου Early, αυτό είναι ακόμα πιο χαμηλό από την περίπτωση της ίσης συχνότητας αλλά αυξάνεται για πέντε ή περισσότερα bins.

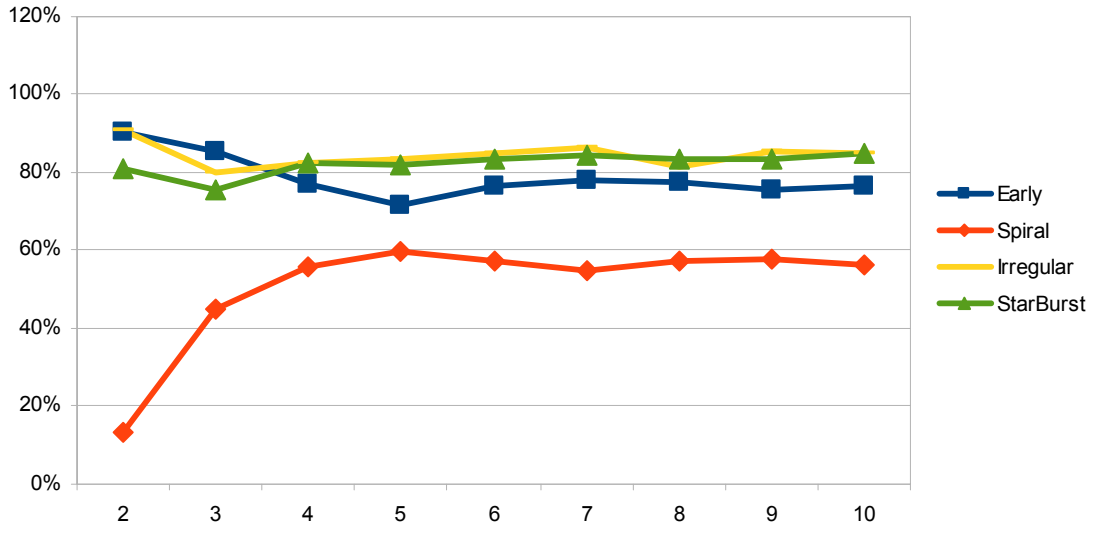


Εικόνα 5-13 Ο λόγος recall για όλους τους μορφολογικούς τύπους για τον αλγόριθμο J4.8 και για τη μέθοδο διακριτοποίησης ίσου μεγέθους.

Αντίστοιχα, η Εικόνα 5-14 και η Εικόνα 5-15 παρουσιάζουν τα ποσοστά του recall για όλους τους γαλαξιακούς τύπους όταν η κατηγοριοποίηση γίνεται με τον αλγόριθμο Naïve Bayes και για τις δύο μεθόδους διακριτοποίησης. Το Recall για όλους τους τύπους είναι αρκετά χαμηλό, κάτι που δείχνει την αδυναμία του αλγορίθμου Naïve Bayes στο να κατηγοριοποιήσει σωστά τους γαλαξίες στους σωστούς τύπους.

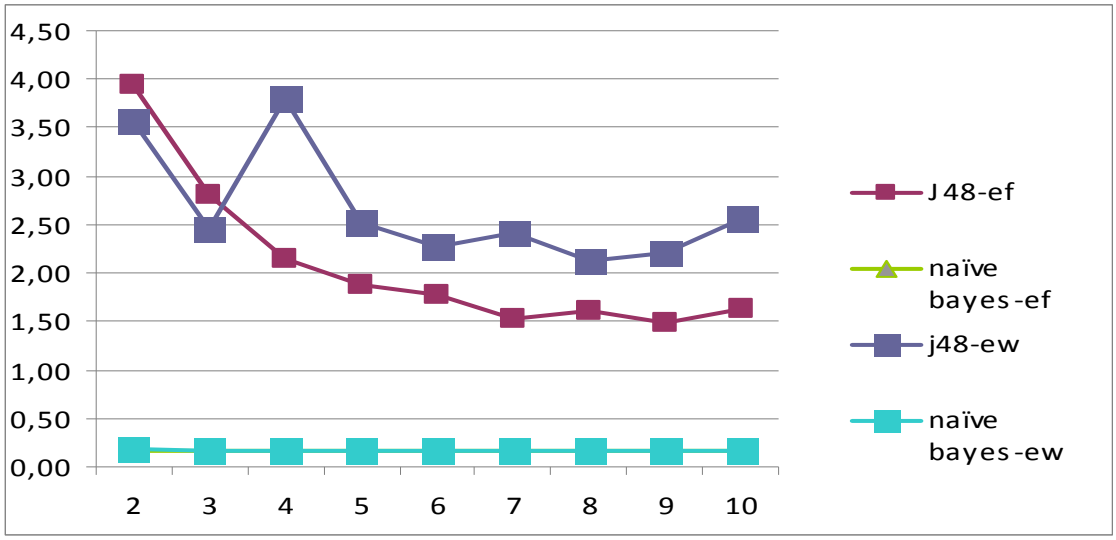


Εικόνα 5-14 Ο λόγος recall για όλους τους μορφολογικούς τύπους για τον αλγόριθμο Naïve Bayes και για τη μέθοδο διακριτοποίησης ίσης συχνότητας.



Εικόνα 5-15 Ο λόγος recall για όλους τους μορφολογικούς τύπους για τον αλγόριθμο Naïve Bayes και για τη μέθοδο διακριτοποίησης ίσου μεγέθους.

Όλα τα πειράματα έγιναν σε PC Pentium M 2 GHZ PC με 1 Gbyte RAM. Στην Εικόνα 5-16 παρουσιάζεται ο χρόνος εκτέλεσης (σε δευτερόλεπτα) για τους δύο αλγορίθμους και για τις δύο μεθόδους διακριτοποίησης για δύο έως δέκα bins.



Εικόνα 5-16 Χρόνος εκτέλεσης για τον J4.8 και Naive Bayes για τις μεθόδους διακριτοποίησης ίσης συχνότητας και ίσου μεγάθους

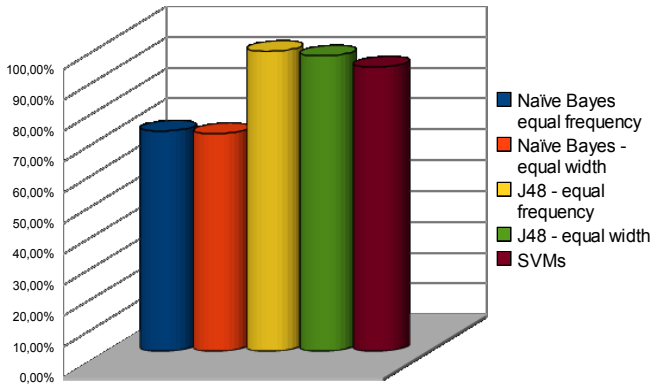
Ο αλγόριθμος Naïve Bayes είναι πολύ πιο γρήγορος από τον J4.8, ενώ ο J4.8 είναι γρηγορότερος όταν η μέθοδος διακριτοποίησης είναι αυτή της ίσης συχνότητας, εκτός των περιπτώσεων των δύο και τριών bins.

Το αποτέλεσμα και η ακρίβεια των αλγορίθμων που παρουσιάστηκαν συγκρίθηκαν με αυτά της καλύτερης μεθόδου κατηγοριοποίησης SVM που έγινε με το εργαλείο R από τους αστρονόμους. Το μοντέλο SVM πέτυχε μία ακρίβεια κατηγοριοποίησης της τάξης του 92.2%. ο Πίνακας 5-2 παρουσιάζει την ακρίβεια κατηγοριοποίησης για όλους τους αλγορίθμους και τις παραλλαγές των πειραμάτων.

Πίνακας 5-2 Ακρίβεια κατηγοριοποίησης των τριών αλγορίθμων και των πιθανών παραλλαγών των πειραμάτων

Αλγόριθμος – μέθοδος διακριτοποίησης	**Συνολική ακρίβεια κατηγοριοποίησης**	**Ακρίβεια για τον τύπο Early**	**Ακρίβεια για τον τύπο Spiral**	**Ακρίβεια για τον τύπο Irregular**	**Ακρίβεια για τον τύπο Starburst**
Naive Bayes - equal frequency	71,43%	83,05%	49,63%	88,35%	83,75%
Naive Bayes- equal width	70,60%	78,75%	50,85%	84,46%	82,38%
J48 - equal frequency	97,25%	85,16%	98,25%	91,01%	99,60%
J48 - equal width	95,78%	76,18%	97,48%	87,08%	99,38%
SVM	92,20%	-	-	-	-

Είναι φανερό ότι ο αλγόριθμος J4.8 αποδίδει καλύτερα στην μέθοδο διακριτοποίησης ίσης συχνότητας. Το γεγονός ότι είναι αρκετά πιο αργός από τον Naïve Bayes δεν είναι σημαντικό στην παρούσα εφαρμογή καθότι η κατηγοριοποίηση θα γίνεται off-line. Στην Εικόνα 5-17 φαίνεται η ακρίβεια κατηγοριοποίησης για όλους τους αλγορίθμους δείχνοντας την υπεροχή του J.48 έναντι των υπολοίπων, ενώ η κατηγοριοποίηση με την SVM μέθοδο έχει σχεδόν εξίσου καλά αποτελέσματα.



Εικόνα 5-17 Ακρίβεια κατηγοριοποίησης για όλους τους αλγορίθμους

Όλα τα πειράματα και το αποτέλεσμα των αλγορίθμων κατηγοριοποίησης καθώς και τα δέντρα απόφασης που εξήχθησαν θα χρησιμοποιηθούν στο έργο GAIA.

Σε εφαρμογές που πολλά πειράματα εξόρυξης γνώσης πρέπει να εκτελεστούν με διάφορους αλγορίθμους και παραμέτρους το PBMS παρέχει ένα ισχυρό εργαλείο καθώς όλα τα αποτελέσματα μπορούν να αποθηκευτούν, να ανακτηθούν και να συγκριθούν μέσα από το PBMS με ολοκληρωμένο και διαφανές προς το χρήστη τρόπο καθιστώντας πολύ πιο εύκολη τη διαδικασία επιλογής της καλύτερης μεθόδου κατηγοριοποίησης/ συσταδοποίησης κλπ.

5.3 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάσαμε μία εφαρμογή του PBMS για την υποστήριξη της κατηγοριοποίησης αστρονομικών δεδομένων, δείχνοντας την προστιθέμενη αξία για τους ειδικούς του πεδίου της αστρονομίας. Τα αποτελέσματα της εφαρμογής αυτής θα χρησιμοποιηθούν στο έργο GAIA της ESA, για την αυτόματη κατηγοριοποίηση των γαλαξιών που θα παρατηρούνται από το νέο διαστημικό τηλεσκόπιο.

6 PatternMiner – ένα Πρωτόλειο Σύστημα Διαχείρισης Βάσεων Προτύπων

6.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζουμε ένα πρωτόλειο PBMS (που το ονομάζουμε PatterMiner), ενώ επίσης μελετάμε το θέμα της αξιολόγησης των προτύπων που εξάγονται από διαδικασίες εξόρυξης γνώσης προκειμένου να επεκτείνουμε το PBMS ώστε να περιλαμβάνει και το βήμα της αξιολόγησης των προτύπων.

Το σύστημα PatternMiner είναι ένα ολοκληρωμένο περιβάλλον για τη διαχείριση και εξόρυξη των προτύπων και περιλαμβάνει λειτουργίες για όλο τον κύκλο ζωής των προτύπων.

Προχωρώντας ένα βήμα από την εξόρυξη, αποθήκευση και σύγκριση των προτύπων, μελετάμε το πρόβλημα της αξιολόγησης των προτύπων με χρήση οντολογιών για να διευκολυνθεί η δύσκολη και χρονοβόρα εργασία των ειδικών για την αξιολόγηση των προτύπων που εξάγονται από μεγάλες βάσεις δεδομένων με τεχνικές εξόρυξης γνώσης. Περιγράφουμε αναλυτικά το πρόβλημα και παρουσιάζουμε μία πρώιμη μελέτη αξιολόγησης στην οποία χρησιμοποιούμε οντολογίες του πεδίου εφαρμογής για το φιλτράρισμα των κανόνων συσχέτισης που εξάγονται από σεισμολογικά δεδομένα.

6.2 Το PBMS PatternMiner

Σε αυτή την ενότητα παρουσιάζουμε το PatternMiner, ένα πρωτόλειο Σύστημα Διαχείρισης Βάσεων Προτύπων, που είναι βασισμένο στη θεωρία που περιγράφηκε στα προηγούμενα κεφάλαια. Το PatternMiner βασίζεται σε μία XML βάση προτύπων και χρησιμοποιεί XML έγγραφα για την αναπαράσταση των προτύπων και τη γλώσσα επερωτήσεων XQuery για την ανάκτησή τους. Η αναπαράσταση των προτύπων βασίζεται στο λογικό μοντέλο του πλαισίου

PANDA, που επίσης χρησιμοποιείται για διαδικασίες σύγκρισης των προτύπων. Για τη φυσική αναπαράσταση σε XML έγγραφα χρησιμοποιούνται PMML σχήματα με ειδικές προσθήκες στοιχείων. Το PatternMiner είναι ανοικτού κώδικα και χρησιμοποιεί επίσης την μηχανή εξόρυξης γνώσης του WEKA για την αρχική εξαγωγή των προτύπων.

Η αρχιτεκτονική και μία περιγραφή του συστήματος, καθώς επίσης αναφέρονται και πιθανές εφαρμογές του, παρουσιάζονται παρακάτω.

Το Pattern-Miner, είναι ένα ολοκληρωμένο περιβάλλον για την εξαγωγή και διαχείριση προτύπων με λειτουργίες σε όλο τον κύκλο ζωής τους, από τη δημιουργία τους (με τη χρήση τεχνικών εξόρυξης γνώσης) μέχρι την αποθήκευση και ανάκτησή τους, με έμφαση επιπλέον στη σύγκριση των προτύπων και σε λειτουργίες meta-mining, πάνω στα ήδη εξαγχθέντα πρότυπα. Αυτό είναι σε αντίθεση με τα υπάρχοντα συστήματα που αντιμετωπίζουν συγκεκριμένα μόνο ζητήματα της διαχείρισης των προτύπων, κυρίως την αναπαράσταση και την απλή αποθήκευση. Η σύγκριση των προτύπων (σύγκριση των αποτελεσμάτων της διαδικασίας εξόρυξης γνώσης) και η διαδικασία meta-mining είναι λειτουργίες υψηλού επιπέδου που μπορούν να εφαρμοστούν σε διάφορες εφαρμογές, από τη διαχείριση αλλαγών σε βάσεις δεδομένων μέχρι τη σύγκριση και ανάκτηση εικόνων. Το Pattern-Miner μπορεί ακόμα να εντοπίσει αλλαγές συσταδοποιήσεων που έχουν εξαχθεί από δυναμικά δεδομένα και έτσι παρέχει πληροφόρηση για το σύνολο δεδομένων και υποστηρίζει στρατηγικές αποφάσεις χωρίς να υπάρχει το πρόβλημα διαλειτουργικότητας ή ασυμβατότητας συστημάτων, όπως θα υπήρχαν στην περίπτωση που χρησιμοποιούνταν διαφορετικές εφαρμογές για κάθε διαφορετική λειτουργία. Το PatternMiner ενσωματώνει τα διαφορετικά εργαλεία που απαιτούνται για τη διαχείριση προτύπων με διαφανή προς το χρήστη τρόπο.

Το PatternMiner βασίζεται στο πλαίσιο PANDA για τον ορισμό των προτύπων, των τύπων προτύπων και χρησιμοποιεί ειδικά εμπλουτισμένα σχήματα PMML για την υλοποίηση των ορισμών αυτών, προσφέροντας μεγάλη διαλειτουργικότητα με τα συστήματα που υποστηρίζουν την PMML.

Το PatternMiner, που παρουσιάζεται παρακάτω, χρησιμοποιεί τεχνολογίες ανοικτού κώδικα, είναι ανεξάρτητο από τη μηχανή/εργαλείο εξόρυξης γνώσης και χρησιμοποιεί XML για την αποθήκευση προτύπων στη βάση προτύπων.

Επιπλέον, δεν υπάρχει PBMS ή άλλο σύστημα που να υποστηρίζει διαδικασίες και λειτουργίες όπως το προτεινόμενο PatternMiner. Μία περιγραφή και παρουσίαση όλων των σχετικών προσεγγίσεων μπορεί να βρεθεί στο (Catania & Maddalena, 2006).

6.2.1 Τεχνολογία και Απαιτήσεις Υλοποίησης

Το PatternMiner, ως ενιαίο περιβάλλον, πρέπει να είναι διαφανές προς το χρήση και να «κρύβει» όλα τα επιμέρους τμήματα που ενσωματώνει και διασυνδέει, παρέχοντας όλη όμως τη λειτουργικότητα στον χρήση μέσω κατάλληλης διεπαφής.

Το PatternMiner χρησιμοποιεί λογισμικό ανοικτού κώδικα και είναι εύκολα αναβαθμίσιμο και επεκτάσιμο.

Στην ενότητα αυτή περιγράφονται οι τεχνολογίες που επιλέχθηκαν για την υλοποίηση και οι ειδικές απαιτήσεις που λήφθηκαν υπόψη και αντιμετωπίστηκαν.

Γλώσσα Προγραμματισμού

Η γλώσσα προγραμματισμού JAVA χρησιμοποιήθηκε για την υλοποίηση του PatternMiner για τους παρακάτω λόγους.

Το εργαλείο εξόρυξης γνώσης WEKA είναι ανοικτού κώδικα υλοποιημένο σε JAVA και για το λόγο αυτό είναι πιθανό να χρησιμοποιηθούν συγκεκριμένα τμήματά του, για τη φόρτωση των αρχείων δεδομένων, ή για την εκτέλεση των απαιτούμενων αλγορίθμων και φίλτρων.

Επιπλέον, το PatternMiner πρέπει να έχει μία φιλική διεπαφή. Το JAVA 2 GUI παρέχει ένα εργαλείο για την υλοποίηση σύνθετων διεπαφών.

Η μηχανή εξόρυξης γνώσης

Η μηχανή για την εξόρυξη γνώσης είναι υπεύθυνη για την εξαγωγή των προτύπων σύμφωνα με τα κριτήρια που έχει θέσει ο χρήστης όπως η επιλογή συγκεκριμένων συνόλων δεδομένων, την προεπεξεργασία τους, τον αλγόριθμο εξόρυξης και τις παραμέτρους του. Για τη λειτουργία αυτή χρησιμοποιείται το WEKA αφού είναι ανοικτού κώδικα και παρέχει μεγάλη ποικιλία αλγορίθμων εξόρυξης γνώσης (συμπεριλαμβανομένων της κατηγοριοποίησης, της συσταδοποίησης και της εξαγωγής κανόνων συσχέτισης) όπως επίσης και πολλών δυνατοτήτων προεπεξεργασίας των δεδομένων. Εκτός της GUI έκδοσης

του προγράμματος, η έκδοση γραμμής εντολών επιτρέπει τη φόρτωση αρχείων και την εκτέλεση αλγορίθμων από κάθε άλλο πρόγραμμα που χρησιμοποιεί το συγκεκριμένο API. Το εργαλείο εξόρυξης WEKA έχει δοκιμαστεί από πολλούς χρήστες και είναι ιδιαίτερα αξιόπιστο. Επιπλέον, κάθε άλλο εργαλείο εξόρυξης γνώσης θα μπορούσε να χρησιμοποιηθεί όσο το αποτέλεσμά του παρέχεται σε συγκεκριμένη PMML μορφή.

Σχήμα/ μοντέλο αναπαράστασης προτύπων

Για το θέμα της αναπαράστασης των προτύπων στη βάση προτύπων έχουν χρησιμοποιηθεί XML έγγραφα αφού αυτά έχουν καλύτερη απόδοση στις λειτουργίες των προτύπων από άλλες προσεγγίσεις (Kotsifakos et al., 2005) και για το θέμα της σύγκρισης των προτύπων έχει επιλεγεί το πλαίσιο PANDA (Ntoutsi, 2008).

Στην ενότητα 2.4.3 περιγράψαμε ένα XML σχήμα που υποστηρίζει το μοντέλο αναπαράστασης προτύπων του PANDA.

Για λόγους συμβατότητας με άλλα συστήματα βάσεων δεδομένων προσαρμόσαμε το μοντέλο της PMML για να υποστηρίζει την αναπαράσταση του PANDA.

Το μοντέλο για τα πρότυπα που προτάθηκε από το πλαίσιο PANDA βασίζεται στα στοιχεία pt = (n, ss, ds, ms, f). Το μοντέλο PMML για κάθε τύπο προτύπου πρέπει να εμπλουτιστεί με στοιχεία μεταδεδομένων για να εμπεριέχει και τα πέντε μέρη του μοντέλου προτύπων καθώς και άλλες πληροφορίες όπως είναι ο αλγόριθμος και οι παράμετροι που χρησιμοποιούνται για την εξαγωγή των προτύπων, την χρονική στιγμή της εκτέλεσής τους κλπ.

Στην υποενότητα αυτή γίνεται μία πιο λεπτομερής περιγραφή της χρήσης της PMML στο σύστημα patternMiner. Επιπλέον παρουσιάζονται οι κατάλληλες και απαραίτητες προσθήκες που γίνονται στο PMML σχήμα για να υποστηρίζεται η αναπαράσταση του μοντέλου PANDA.

Η PMML υποστηρίζει μόνο συγκεκριμένους και προκαθορισμένους τύπους προτύπων (ή models όπως ονομάζονται). Τα μοντέλα που υποστηρίζονται στην έκδοση 3.2 (PMML, 2009) είναι: Κανόνες Συσχέτισης, Συστάδες, Δέντρα, Νευρωνικά Δίκτυα, Χρονοσειρές και περισσότερο σύνθετοι τύποι όπως Κείμενο και Support Vector Machines. Με την PMML μπορούν να αναπαρασταθούν κάποια μέτρα ποιότητας. Επιπλέον, αποθηκεύεται και η σχέση μεταξύ των

προτύπων και του υποσυνόλου των αρχικών δεδομένων (τα οποία αναπαριστούν τα πρότυπα) καθώς και ο χρόνος της εξαγωγής τους.

Η δομή της PMML περιλαμβάνει:

1 *Header*. Περιλαμβάνει γενικές πληροφορίες για το πρότυπο όπως είναι η εφαρμογή που το δημιούργησε η ημερομηνία και ώρα της δημιουργίας του και μία μικρή περιγραφή.

2 *Data Dictionary*. Ορίζει τα γνωρίσματα των δεδομένων εισόδου για τα πρότυπα, τον τύπο τους και το πεδίο τιμών τους.

3 *Transformation Dictionary*. Η PMML ορίζει διάφορα είδη απλών μετασχηματισμών δεδομένων:

Normalization: αντιστοίχιση τιμών σε αριθμούς, με διακριτή ή συνεχή είσοδο.

Discretization: αντιστοίχιση συνεχών τιμών σε διακριτές.

Value mapping: αντιστοίχιση διακριτών τιμών σε συνεχείς.

Functions: εξάγει μία τιμή εφαρμόζοντας μία συνάρτηση σε μία ή περισσότερες παραμέτρους.

Aggregation: συνοψίζει ή συναθροίζει σύνολα τιμών, π.χ., υπολογισμός του μέσου όρου.

Το Transformation Dictionary είναι προαιρετικό πεδίο.

4 **Model*. Όπου * είναι το όνομα της τεχνικής εξόρυξης. Ορίζει τη συγκεκριμένη πληροφορία για κάθε πρότυπο όπως η τεχνική εξόρυξης και τον αλγόριθμο που χρησιμοποιείται στην εξαγωγή των προτύπων, τα γνωρίσματα των αρχικών δεδομένων που χρησιμοποιούνται ως είσοδος, και άλλες πληροφορίες σχετικές με τον τύπο του προτύπου, όπως τα συχνά στοιχειοσύνολα και έναν κανόνα συσχέτισης, ή τις συστάδες και τα χαρακτηριστικά τους για τη συσταδοποίηση.

4.1 *Mining Schema*. Κάθε μοντέλο περιλαμβάνει ένα mining schema (σχήμα εξόρυξης), το οποίο αναφέρει τα πεδία που χρησιμοποιούνται στο μοντέλο. Αυτά τα πεδία είναι υποσύνολα ετών πεδίων στο Data Dictionary. Το mining schema περιέχει πληροφορία που είναι συγκεκριμένη για ένα μοντέλο, ενώ το data dictionary περιέχει ορισμούς δεδομένων που δεν διαφέρουν από μοντέλο σε μοντέλο. Για

παράδειγμα, το Mining Schema καθορίζει τον τύπο χρήσης ενός γνωρίσματος, το οποίο μπορεί να είναι *active* (σαν είσοδος στο μοντέλο), *predicted* (σαν έξοδος του μοντέλου), ή *supplementary* (περιέχοντας περιγραφική πληροφορία αδιάφορη στο μοντέλο).

4.2 *Model Statistics*. Το Model Statistics περιέχει βασικά στατιστικά με μία μεταβλητή σχετικά με το μοντέλο, όπως είναι το ελάχιστο, το μέγιστο, η τυπική απόκλιση ο μέσος κλπ. των αριθμητικών γνωρισμάτων.

Η PMML υποστηρίζει την έννοια του Model Composition (σύνθεση μοντέλων). Απλά μοντέλα μπορούν να χρησιμοποιηθούν σαν μετασχηματισμοί. Η PMML προσφέρει τη δυνατότητα να συνδυαστούν πολλά απλά μοντέλα σε ένα νέο, χρησιμοποιώντας το κάθε ξεχωριστό μοντέλο σαν θεμέλιους λίθους. Αυτό μπορεί να καταλήξει σε μοντέλα που χρησιμοποιούνται σε σειρά(ακολουθία), όπου το αποτέλεσμα κάθε μοντέλου είναι είσοδος για το επόμενο. Αυτή η προσέγγιση που καλείται ακολουθιακά μοντέλα, δεν είναι χρήσιμη μόνο για την κατασκευή πιο σύνθετων μοντέλων, αλλά μπορεί επίσης να χρησιμοποιηθεί και για την προετοιμασία των δεδομένων. Μία ακόμα μορφή σύνθεσης μοντέλων υποστηρίζεται: το αποτέλεσμα ενός μοντέλου μπορεί να χρησιμοποιηθεί για να επιλεχθεί ποιο μοντέλο θα εφαρμοστεί στη συνέχεια. Για παράδειγμα, ένα δέντρο απόφασης μπορεί πλέον να έχει ένα ενσωματωμένο regression μοντέλο σε κάθε κόμβο-φύλλο.

Τόσο το ακολουθιακό μοντέλο όσο και το μοντέλο επιλογής μπορούν να συνδυαστούν για να παράγουν αρκετά πολύπλοκα πρότυπα.

Η PMML υποστηρίζει συναρτήσεις που μπορεί να χρησιμοποιηθούν για την προπαρασκευή των δεδομένων όπως προκαθορισμένες συναρτήσεις για απλές αριθμητικές λειτουργίες όπως άθροισμα, διαφορά, πολ/μός, τετραγωνική ρίζα, λογάριθμος κλπ, για αριθμητικά πεδία όπως και συναρτήσεις για διαχείριση συμβολοσειρών.

Στην PMML υπάρχει επίσης ένας μηχανισμός για την επαλήθευση των μοντέλων ο οποίος αυξάνει τη συμβατότητα των μοντέλων μεταξύ διαφορετικών εφαρμογών που υποστηρίζουν την PMML. Αυτό το μοντέλο επαλήθευσης παρέχει ένα μηχανισμό διάθεσης ενός δοκιμαστικού συνόλου δεδομένων με τα αποτελέσματα ώστε οποιοσδήποτε χρησιμοποιεί PMML να μπορεί να διαπιστώσει εάν την έχει υλοποιήσει σωστά. Αυτό μπορεί να κάνει πολύ εύκολη

και διαφανή προς το χρήστη την ανταλλαγή μοντέλων και να τους πληροφορεί για τα προβλήματα συμβατότητας που μπορεί να παρουσιαστούν.

Εκτός από την προκαθορισμένη πληροφορία που αποθηκεύει η PMML για κάθε μοντέλο, κατά τη διάρκεια της εισαγωγής των μοντέλων στην XML βάση προτύπων, τα παρακάτω επιπλέον πεδία μεταδεδομένων έχουν προστεθεί:

- dateCreated: η ημερομηνία και ώρα που τα πρότυπα εισήχθησαν στη βάση προτύπων.
- dataFileName: το όνομα του αρχείου (με όλη τη διαδρομή στο δίσκο - path) που περιέχει τα αρχικά δεδομένα
- modelName: το όνομα του προτύπου όπως το ορίζει ο χρήστης.

Σε κάθε PMML έγγραφο μία ετικέτα “extension” έχει δημιουργηθεί που περιέχει την κατάλληλη πληροφορία για τα πρότυπα συσταδοποίησης. Πιο συγκεκριμένα:

- Η εκ των προτέρων πιθανότητα (Prior Probability) και η τιμή διασποράς (scatter value) της συστάδας, με όνομα extension «Prior probability» και «Scatter value» αντίστοιχα.
- Το ποσοστό των δεδομένων που ανήκουν σε κάθε συστάδα ενός σύνθετου προτύπου, με όνομα extension «Clustered Instances».

Μέχρι το χρόνο που γράφτηκε η παρούσα διατριβή, μία νεότερη έκδοση της PMML ανακοινώθηκε.

Η έκδοση 4.0 της PMML έχει επιπλέον τα παρακάτω νέα χαρακτηριστικά:

- υποστηρίζει μοντέλα χρονοσειρών

- βελτιωμένη υποστήριξη για προεπεξεργασία δεδομένων, που βοηθάει στην απλοποίηση της δημιουργίας μοντέλων

- νέα μοντέλα όπως survival models

- υποστήριξη για επιπλέον πληροφορία για τα μοντέλα, που ονομάζεται model explanation, που περιέχει πληροφορία για οπτικοποίηση, ποιότητα του μοντέλου, διαγράμματα gains και lift, confusion matrix, και σχετική πληροφορία.

Η νέα αυτή έκδοση της PMML είναι μία μεγάλη ανανέωση της έκδοσης 3.2, που παρουσιάστηκε τον Μάιο του 2007.

Σύστημα αποθήκευσης και ανάκτησης προτύπων

Όπως έχει παρουσιαστεί στην ενότητα 2.4 το μοντέλο της XML είναι το περισσότερο κατάλληλο για την αναπαράσταση και διαχείριση των προτύπων. Για το λόγο αυτό μία εγγενής (native) XML βάση δεδομένων θα ήταν η καλύτερη επιλογή για την αποθήκευση των προτύπων και θα έχει τα παρακάτω πλεονεκτήματα:

- Τα XML δεδομένα εισάγονται στη βάση χωρίς την ανάγκη επιπλέον επεξεργασίας. Τα πρότυπα αποθηκεύονται απευθείας ως XML έγγραφα.
- Κάθε χαρακτήρας (συμπεριλαμβανομένου του κενού και άλλων εδικών χαρακτήρων) των XML εγγράφων, παραμένουν αναλλοίωτοι μετά την εισαγωγή στη βάση.
- Οι επερωτήσεις στην XML βάση επιστρέφουν όλο το έγγραφο ή μέρος αυτού, διατηρώντας την ιεραρχική δομή των εγγράφων.

Επιπλέον, η ανταλλαγή των δεδομένων είναι πολύ πιο εύκολη και δεν απαιτεί μετασχηματισμό των εγγράφων σε διαφορετικές δομές.

Ως εγγενή XML βάση δεδομένων επιλέξαμε την ORACLE Berkeley DB XML. Η Berkeley DB XML αποθηκεύει XML έγγραφα σε λογικές ομάδες που ονομάζονται "*Containers*", που αντιστοιχούν στα "*Collections*" σε άλλες XML βάσεις δεδομένων. Οι χρήστες μπορούν να ορίσουν διάφορες ιδιότητες για κάθε container, συμπεριλαμβανόμενης της επιλογής για επικύρωση των εγγράφων, αποθήκευση ολόκληρων των εγγράφων ή μέρος τους και τη δημιουργία ευρετηρίου.

Στην εφαρμογή μας τα πρότυπα ομαδοποιούνται βάση την τεχνική εξόρυξης με τα οποία έχουν εξαχθεί. Έτσι, υπάρχουν τρία βασικά "containers": AssociationRules.dbxml, Clustering.dbxml και Trees.dbxml για κανόνες συσχέτισης, συστάδες και δέντρα απόφασης, αντίστοιχα.

Η Berkeley DB XML μπορεί ακόμα να αποθηκεύσει έγγραφα που δεν είναι XML όπως επίσης και μεταδεδομένα XML εγγράφων. Τα μεταδεδομένα είναι ορισμένα από το χρήστη ζευγάρια "property-value" (ιδιότητα-τιμή) και μπορούν να ανακτηθούν ως στοιχεία παιδιά (child elements) του στοιχείου ρίζα, ενώ δεν εμφανίζονται στα αποθηκευμένα XML έγγραφα.

Η Berkeley DB XML υποστηρίζει τη γλώσσα XQuery για τη δημιουργία και εκτέλεση επερωτήσεων στη βάση δεδομένων και υποστηρίζει ευρετήρια για την επιτάχυνση της εκτέλεσης των επερωτήσεων.

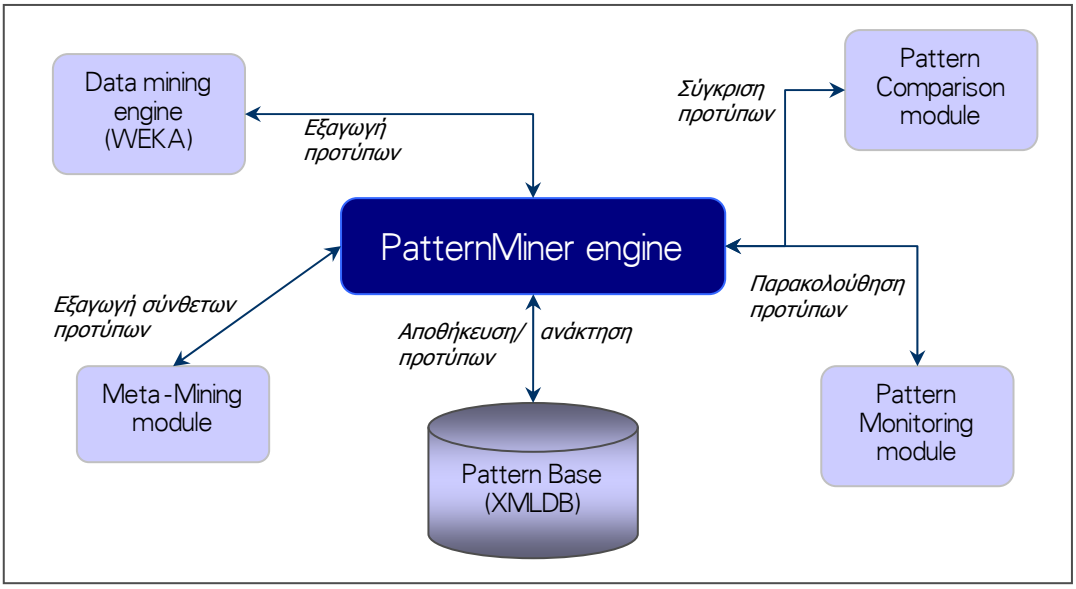
Τέλος η Berkeley DB XML παρέχει και περιβάλλον γραμμής εντολών και παρέχει APIs για C++, Java, Tcl, Perl, Python και PHP. Επίσης, υπάρχουν διαθέσιμα APIs ανεπτυγμένα από άλλους προμηθευτές για άλλες γλώσσες. Υποστηρίζει UNIX, Linux, Windows και Mac OS X.

6.2.2 Αρχιτεκτονική Συστήματος

Η αρχιτεκτονική του PatternMiner παρουσιάζεται στην Εικόνα 6-1. Στο κέντρο του συστήματος βρίσκεται η *PatternMiner engine* που είναι υπεύθυνη για την επικοινωνία των διάφορων μερών του συστήματος (Data Mining engine, Pattern Base, Pattern Comparison module, Meta-mining module) και επίσης παρέχει και τη διεπαφή προς το χρήστη.

Εξαγωγή Προτύπων (Pattern extraction): το κομμάτι αυτό του συστήματος είναι υπεύθυνο για την εξαγωγή των προτύπων σύμφωνα με τα κριτήρια που έχει ορίσει ο χρήστης, όπως η επιλογή του συνόλου αρχικών δεδομένων, την προεπεξεργασία των δεδομένων, τους αλγορίθμους εξόρυξης και τις παραμέτρους τους. Για την εξαγωγή των προτύπων χρησιμοποιούμε το εργαλείο *WEKA* (Witten and Frank, 2005), ως λογισμικό ανοικτού κώδικα που προσφέρει ένα μεγάλο εύρος αλγορίθμων για διάφορες εργασίες εξόρυξης γνώσης (κατηγοριοποίηση, συσταδοποίηση και κανόνες συσχέτισης) αλλά και πολλές δυνατότητες προεπεξεργασίας δεδομένων.

Αναπαράσταση Προτύπων: Η αναπαράσταση των προτύπων δεν είναι μία απλή διαδικασία κυρίως επειδή υπάρχει μία μεγάλη ποικιλία διαφορετικών τύπων προτύπων (δέντρα απόφασης, συστάδες κλπ) διαφορετικής πολυπλοκότητας. Η ανάγκη για αναπαράσταση των προτύπων στην KDD έχει αναγνωριστεί από την ακαδημαϊκή και επιχειρησιακή κοινότητα και διάφορες προσεγγίσεις έχουν προταθεί. Στο PatternMiner, όπως περιγράφηκε στην προηγούμενη ενότητα, χρησιμοποιείται το PMML standard για την αναπαράσταση των προτύπων, αλλά εμπλουτισμένο ώστε να ταιριάζει απόλυτα με τις αρχές του πλαισίου PANDA, και έτσι, η έξοδος από τη *Data Mining engine* μετατρέπεται σε PMML.



Εικόνα 6-1 Η αρχιτεκτονική του PatternMiner

Αποθήκευση προτύπων (Pattern storage): Από τη στιγμή που τα πρότυπα αναπαρίστανται με XML έγγραφα, χρησιμοποιείται μία εγγενής XML βάση δεδομένων για την αποθήκευσή τους στη βάση προτύπων *Pattern Base*. Συγκεκριμένα, χρησιμοποιείται το σύστημα ανοικτού κώδικα *Berkeley DBXML* (Oracle Corp. Berkeley DB XML), που είναι μία επέκταση της Berkeley DB με την προσθήκη ενός XML parser, XML ευρετηρίων και της XQuery γλώσσας επερωτήσεων.

Ανάκτηση προτύπων (Pattern querying): Το PatternMiner παρέχει ένα βασικό περιβάλλον για ανάκτηση προτύπων από τη βάση προτύπων. Ο χρήστης ορίζει το σύνολο των προτύπων που πρόκειται να επερωτηθεί και υποβάλει την επερώτηση στη γλώσσα XQuery language (Xquery 1.0, 2003). Σχετικά με τα υποστηριζόμενα ήδη επερωτήσεων, ο χρήστης μπορεί να ανακτήσει είτε ολόκληρο το πρότυπο είτε ένα μέρος του (το μέρος της δομής ή του μέτρου ποιότητας) και φυσικά, να υποβάλει περιορισμούς στα μέρη αυτά. Το PatternMiner δημιουργεί την κατάλληλη σύνδεση με τη βάση προτύπων και επιστρέφει τα αποτελέσματα της ανάκτησης στο χρήστη. Τα αποτελέσματα παρουσιάζονται στην οθόνη αλλά αποθηκεύονται και στο σύστημα αρχείων.

Σύγκριση προτύπων: Μία από τις περισσότερο σημαντικές λειτουργίες σχετικές με τα πρότυπα είναι αυτή της σύγκρισης. Ο ορισμός μετρικών ομοιότητας/ απόστασης μπορεί να χρησιμοποιηθούν για να εκφραστούν επερωτήσεις ομοιότητας, συμπεριλαμβανομένων των *k-nearest neighbor queries* (δηλ. εύρεση των k-περισσότερο όμοιων προτύπων προς ένα δοθέν

πρότυπο) και *range queries* (δηλ. εύρεση των περισσότερο όμοιων προτύπων προς ένα δοθέν, μέσα σε ένα συγκεκριμένο διάστημα). Για τις λειτουργίες ομοιότητας χρησιμοποιείται το πλαίσιο PANDA (Bartolini et al., 2004; Ntoutsi et al., 2007) το οποίο περιέχει τη μεθοδολογία σύγκρισης των απλών και σύνθετων προτύπων καθώς και ένα πλήθος σχετικών συναρτήσεων για τη σύγκριση διαφόρων τύπων προτύπων. Στο πλαίσιο αυτό έχουν προστεθεί και οι μετρικές απόστασης που προτάθηκαν σε αυτή τη διατριβή και περιγράφονται σε προηγούμενα κεφάλαια και που χρησιμοποιήθηκαν στα πειράματα που διεξήχθησαν. Η σύγκριση χρησιμοποιεί τα μέρη της δομής και μέτρου ποιότητας των προτύπων ενώ ο χρήστης ορίζει το ποια πρότυπα θα συγκριθούν και με ποιο τρόπο. Το αποτέλεσμα της σύγκρισης παρουσιάζονται στο χρήστη με μία αναφορά για το πώς τα πρότυπα ταιριάστηκαν.

Meta-mining:

Λόγω του αυξημένου όγκου προτύπων που εξάγονται από τα αρχικά δεδομένα, έχουν προταθεί προσεγγίσεις που εφαρμόζουν τεχνικές εξόρυξης δεδομένων στα πρότυπα που έχουν ήδη εξαχθεί για να εξαχθεί ακόμα πιο συμπαγής πληροφορία. Το κομμάτι του *Meta-mining* δέχεται ως είσοδο ένα σύνολο διαφορετικών αποτελεσμάτων συσταδοποίησης που έχουν εξαχθεί από το ίδιο σύνολο αρχικών δεδομένων (με διαφορετικούς αλγορίθμους ή παραμέτρους) ή από διαφορετικά σύνολα δεδομένων και εφαρμόζουν τεχνικές εξόρυξης για να εξαχθούν τα λεγόμενα μετα-πρότυπα (*meta-patterns*). Προς το παρόν το κομμάτι του meta-mining εστιάζει στη μετα-συσταδοποίηση (Caruana et al., 2006), δηλαδή στην ομαδοποίηση συστάδων. Ο χρήστης έχει πλήρη έλεγχο της συσταδοποίησης με την επιλογή των συναρτήσεων ομοιότητας και του αλγορίθμου συσταδοποίησης.

Παρακολούθηση εξέλιξης προτύπων:

Η σύγκριση των προτύπων μπορεί να χρησιμοποιηθεί για την παρακολούθηση των αλλαγών των προτύπων που εξάγονται από ένα δυναμικό περιβάλλον (Spiliopoulou et al., 2006). Ενώ το PatternMiner είναι ένα εργαλείο για τη διαχείριση όλων των τύπων προτύπων, προς το παρόν έχει υλοποιηθεί η τεχνική για την παρακολούθηση συστάδων που βασίζεται στη θεωρία και στον αλγόριθμο που περιγράφεται στο (Spiliopoulou et al., 2006). Σε αυτή την προσέγγιση, η μετάβαση των συστάδων που εξάγονται από ένα δυναμικό σύνολο δεδομένων παρακολουθείται και μοντελοποιείται. Η συσταδοποίηση γίνεται σε συγκεκριμένα χρονικά διαστήματα και μία συνάρτηση μπορεί να

χρησιμοποιηθεί για να αποδίδει βάρη σε όλες ή κάποιες από τις προηγούμενες εγγραφές. Το σύνολο των χαρακτηριστικών που χρησιμοποιούνται για τη συσταδοποίηση μπορεί επίσης να αλλάξει κατά την περίοδο παρατήρησης, επιτρέποντας έτσι την εισαγωγή νέων ή τον αποκλεισμό προηγούμενων χαρακτηριστικών (γνωρισμάτων). Οι μεταβάσεις μπορούν να εντοπιστούν ακόμα και όταν αλλάζουν τα γνωρίσματα αυτά. Έννοιες όπως ταίριασμα συστάδας (cluster match), επικάλυψη συστάδων (cluster overlap), μετάβαση συστάδων (cluster transition) και διάρκεια ζωής (lifetime) συστάδας είναι βασικές στην παρακολούθηση εξέλιξης των συστάδων. Το κομμάτι αυτό του συστήματος χρησιμοποιεί συσταδοποιήσεις που είναι αποθηκευμένες στη βάση προτύπων και επίσης χρησιμοποιεί τις δυνατότητες επερώτησης και σύγκρισης προτύπων του συστήματος.

6.2.3 Παρουσίαση Λειτουργιών του PatternMiner

Στην ενότητα αυτή γίνεται μία συνοπτική παρουσίαση του συστήματος PatternMiner, όπως έγινε και στα (Kotsifakos et al., 2008a, 2008b).

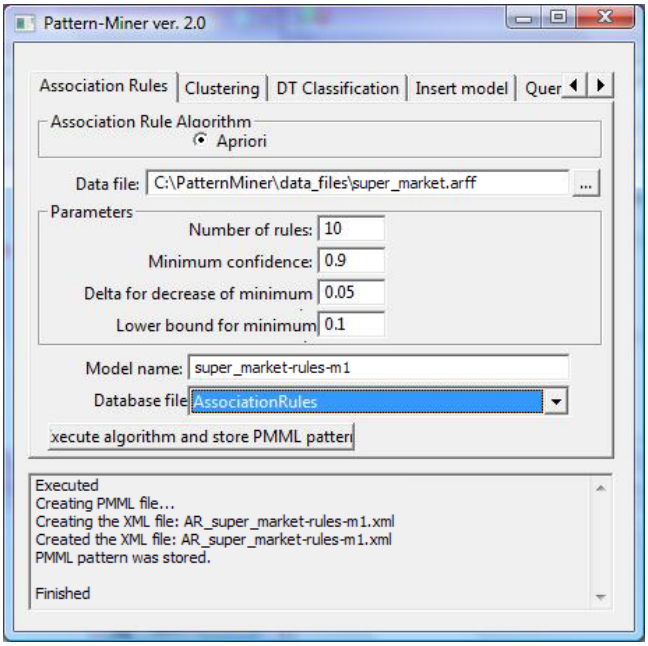
6.2.3.1 Παρουσίαση συστήματος

Για να γίνει κατανοητός ο τρόπος χρήσης του εργαλείου PatternMiner, θεωρούμε το παράδειγμα ενός supermarket σαν μία απλή μελέτη περίπτωση και το διευθυντή του σα χρήστη του συστήματος. Εκτός άλλων τύπων προτύπων, ο διευθυντής ενδιαφέρεται να ανακαλύψει τα προϊόντα που οι πελάτες αγοράζουν μαζί, δηλαδή πρότυπα κανόνων συσχέτισης. Εκτός του να γνωρίζει τις συσχετίσεις προϊόντων κάθε μήνα, ο διευθυντής επίσης θέλει να γνωρίζει και αν αυτές οι συσχετίσεις αλλάζουν από μήνα σε μήνα: αν υπάρχουν νέες συσχετίσεις, να κάποιες άλλες παύουν να ισχύουν, αν άλλες συσχετίσεις έγιναν ισχυρότερες (μεγαλύτερο confidence) ή λιγότερο ισχυρές. Επιπλέον, θέλει να ανακαλύψει ομάδες μηνών με παρόμοιες συσχετίσεις, ώστε να αποφασίσει για μία στρατηγική για κάθε ομάδα αντί για κάθε μήνα. Αυτή η διαδικασία περιλαμβάνει την αποθήκευση των προτύπων που εξάγονται για κάθε μήνα, τη σύγκριση και τη λειτουργία meta-mining σε αυτά. Τα υπάρχοντα εργαλεία εξόρυξης γνώσης από δεδομένα δεν αντιμετωπίζουν επαρκώς όλα αυτά τα θέματα. Αντίθετα το PatternMiner παρέχει στο χρήστη όλη αυτή την πληροφορία με απλό και διαφανή τρόπο. Παρακάτω περιγράφεται πώς

χρησιμοποιείται κάθε κομμάτι του συστήματος για το συγκεκριμένο παράδειγμα.

Εξαγωγή και αποθήκευση προτύπων:

Ο χρήστης ορίζει την πηγή των δεδομένων, τον αλγόριθμο εξόρυξης και τις παραμέτρους του, για παράδειγμα στην περίπτωσή μας, τη βάση δεδομένων του supermarket, τον αλγόριθμο κανόνων συσχέτισης (π.χ. Apriori) και τις παραμέτρους του minimum support και confidence. Η εξαγωγή γίνεται από το κομμάτι του Data Mining engine και τα αποτελέσματα μετατρέπονται σε μορφή PMML (με τις κατάλληλες προσθήκες) και αποθηκεύονται σε ένα ορισμένο από το χρήστη container στην XML βάση προτύπων (και επίσης σε ένα αρχείο στο δίσκο). Στην Εικόνα 6-2 φαίνεται η οθόνη εξαγωγής και αποθήκευσης προτύπων για τους κανόνες συσχέτισης. Με τη χρήση της PMML η ανταλλαγή των προτύπων μεταξύ διαφόρων εφαρμογών επιτυγχάνεται εύκολα, χωρίς τη χρήση ειδικών εργαλείων.



Εικόνα 6-2 Η οθόνη εξαγωγής κανόνων συσχέτισης

Επερωτήσεις προτύπων: Ο χρήστης ορίζει το σύνολο των προτύπων από τα οποία θα γίνει η ανάκτηση και την ίδια την επερώτηση στη γλώσσα Xquery. Η μηχανή του Pattern-Miner δημιουργεί τη σύνδεση με τη βάση προτύπων, εκτελεί την επερώτηση και επιστρέφει το αποτέλεσμα στο χρήστη (ενώ την αποθηκεύει και σε κάποιο αρχείο). Ένα παράδειγμα επερώτησης φαίνεται στην Εικόνα 6-3, και παρουσιάζεται σε φυσική γλώσσα και σε γλώσσα Xquery.

Query (natural language):

Ανάκτηση των κανόνων συσχέτισης από το σύνολο δεδομένων super_market που έχουν τιμή υποστήριξης (support) μεγαλύτερη από 0.2.

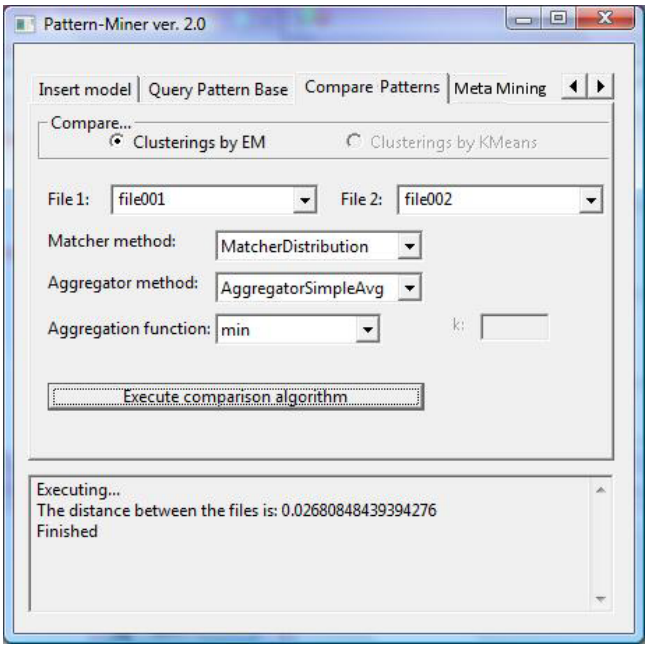
Query (XQuery):

```
declare namespace a = "http://www.dmg.org/PMML-3_1";

collection ("AssociationRules.dbxml")

[dbxml:metadata ("dbxml:dataFileName")=

"C:\Pattern-Miner\data_files\ supper_market.arff"]

/a:PMML /a:AssociationModel /a:AssociationRule
[@support>0.2]
```

Εικόνα 6-3 Παράδειγμα επερώτησης σε φυσική γλώσσα και στη γλώσσα XQuery

Σύγκριση προτύπων: Ο χρήστης ορίζει τα πρότυπα που θα συγκριθούν και τις παραμέτρους της σύγκρισης. Στο παράδειγμά μας, ο διευθυντής θέλει να συγκρίνει τους κανόνες συσχέτισης που εξάγονται από τα δεδομένα του supermarket των δύο τελευταίων μηνών, προκειμένου να διερευνήσει αν και κατά πόσο έχει αλλάξει η αγοραστική προτίμηση των πελατών. Τα πρότυπα εξάγονται από τη βάση προτύπων, ο χρήστης επιλέγει την κατάλληλη συνάρτηση σύγκρισης από τις διαθέσιμες στο πλαίσιο PANDA. Ανάλογα με την εφαρμογή ο χρήστης επιλέγει την κατάλληλη συνάρτηση σύγκρισης από τις διάφορες διαθέσιμες στο πλαίσιο PANDA για να πετύχει τα καλύτερα αποτελέσματα στην εφαρμογή. Τα αποτελέσματα επιστρέφουν στο χρήστη/ διευθυντή ο οποίος μπορεί να δει τις αλλαγές στα πρότυπα των πωλήσεων και να αποφασίσει αν οι αλλαγές είναι αναμενόμενες (σύμφωνα με τη στρατηγική της επιχείρησης ή την εποχική τάση) ή όχι (δείχνοντας κάποια μη προβλέψιμη συμπεριφορά). Σύμφωνα με τα αποτελέσματα, ο διευθυντής αποφασίζει για μελλοντικές στρατηγικές σχετικά ίσως με προσφορές, τις προμήθειες κλπ.

Ο διευθυντής μπορεί επίσης να εξάγει συστάδες πελατών βασισμένες στις αγοραστικές συνήθειες ή τα δημογραφικά τους στοιχεία. Συγκρίνοντας τέτοια πρότυπα μπορεί να ανακαλυφθούν ετήσια πρότυπα αγοραστικής συμπεριφοράς και έτσι ο διευθυντής να αποφασίσει για τις προμήθειες. Στην Εικόνα 6-4 φαίνεται η καρτέλα σύγκρισης προτύπων.

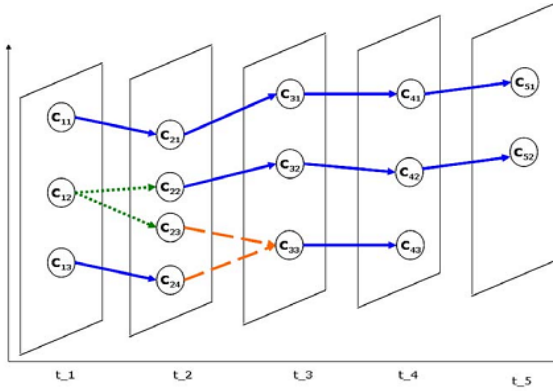


Εικόνα 6-4 Η καρτέλα σύγκρισης προτύπων στο PatternMiner

Meta-mining: Ο χρήστης ορίζει το σύνολο των προτύπων που θα χρησιμοποιηθεί ως είσοδο για το μέρος του Meta-mining module (για παράδειγμα, σύνολα κανόνων συσχέτισης που εξήχθησαν κάθε μήνα του 2007), επιλέγει τον αλγόριθμο συσταδοποίησης και τις παραμέτρους του, καθώς επίσης και το μέτρο ομοιότητας μεταξύ των κανόνων. Τα σύνολα αυτά συσταδοποιούνται σε ομάδες όμοιων κανόνων (πχ. οι κανόνες του Μαρτίου και του Απριλίου μπορεί να μπουν στην ίδια ομάδα, αφού αναδεικνύουν παρόμοια αγοραστική συμπεριφορά), η οποία συσταδοποίηση μπορεί να αποθηκευτεί στη βάση προτύπων για μελλοντική χρήση. Ο διευθυντής μπορεί να χρησιμοποιήσει τα αποτελέσματα για να αποφασίσει παρόμοιες στρατηγικές για τους μήνες που ανήκουν στην ίδια συστάδα.

Παρακολούθηση εξέλιξης των προτύπων: Ο χρήστης ορίζει το σύνολο των δεδομένων από τα οποία εξήχθησαν οι συστάδες. Μία λίστα με όλες τις συσταδοποιήσης που έχουν γίνει σε αυτά τα δεδομένα εμφανίζεται, ταξινομημένη με χρονολογική σειρά. Ο διευθυντής του supermarket επιθυμεί να παρατηρήσει τα προφίλ των πελατών στη διάρκεια του χρόνου. Επιλέγοντας το κατάλληλο αρχείο δεδομένων (supermarket.arff), το Pattern-Miner επιστρέφει όλες τις διαφορετικές συσταδοποιήσεις καθώς και τον αλγόριθμο συσταδοποίησης που χρησιμοποιήθηκε και την ώρα της εκτέλεσης. Ο διευθυντής επιλέγει δύο ή περισσότερες συσταδοποιήσεις και εκτελεί τη διαδικασία παρακολούθησης συστάδων. Αυτή η διαδικασία καταλήγει σε ένα

πίνακα που δείχνει τις συστάδες της πρώτης συσταδοποίησης και τις αλλαγές τους στο χρόνο (νέες συστάδες εμφανίζονται, άλλες παύουν να υπάρχουν, συρρικνώνονται ή επεκτείνονται κλπ). Το αποτέλεσμα παρουσιάζεται στο διάγραμμα στην Εικόνα 6-5 (Spiliopoulou et al., 2006).



Εικόνα 6-5 Γραφική αναπαράσταση της διαδικασίας παρακολούθησης εξέλιξης συστάδων

6.2.3.2 Συζήτηση

Το PatternMiner είναι ένα ολοκληρωμένο περιβάλλον διαχείρισης προτύπων που υποστηρίζει όλο τον κύκλο ζωής των προτύπων από τη δημιουργία τους και την ανάκτησή τους και περιλαμβάνει σύνθετες λειτουργίες όπως αυτή της σύγκρισης και του meta-mining. Το PatternMiner για κάθε μέρος του χρησιμοποιεί τεχνολογίες ανοικτού κώδικα ενώ όλα τα μέρη του έχουν υλοποιηθεί σε JAVA.

Εντούτοις, μπορεί να γίνουν κάποιες βελτιώσεις: πρώτον, κάθε μέρος του συστήματος μπορεί να βελτιωθεί. Για παράδειγμα, το μέρος της ανάκτησης μπορεί να υποστηρίζει περισσότερους τύπους επερωτήσεων προτύπων, όπως οι επερωτήσεις πλησιέστερου γείτονα (k-nearest neighbor) και οι επερωτήσεις εύρους τιμών (range queries) και επίσης η διαδικασία ανάκτησης μπορεί να γίνει πιο αποτελεσματική με χρήση κατάλληλων ευρετηρίων. Επιπλέον, το κομμάτι του *Meta-mining* και της διαδικασίας παρακολούθησης των συστάδων (*cluster monitoring*) μπορεί να επεκταθεί για να υποστηρίζει περισσότερους τύπους προτύπων, όπως δέντρα απόφασης, κανόνες συσχέτισης και ακολουθίες.

Εκτός από την εφαρμογή που περιγράφηκε, άλλες πιθανές εφαρμογές είναι η αξιολόγηση προτύπων, η παρακολούθηση αλλαγών σε βάσεις δεδομένων, η

σύγκριση προτύπων που έχουν εξαχθεί από διαφορετικές βάσεις σε ένα κατανεμημένο περιβάλλον κλπ.

6.3 Επέκταση του PBMS για την Υποστήριξη Αξιολόγησης των Προτύπων με Χρήση Οντολογιών

Η εξόρυξη γνώσης ως μέρος της γενικότερης διαδικασίας ανακάλυψης γνώσης από δεδομένα (*Knowledge Discovery from Data* - KDD) χρησιμοποιεί τα αρχικά δεδομένα για να εξάγει χρήσιμα συμπεράσματα για αυτά (βήμα Data Mining στην Εικόνα 1-1). Τα συμπεράσματα αυτά με τη μορφή προτύπων αξιολογούνται από τους ειδικούς κάθε φορά του πεδίου εφαρμογής για να εξεταστεί τόσο η εγκυρότητά τους όσο και η σημαντικότητά τους. Η συμβολή των ειδικών στη διαδικασία αυτή είναι ιδιαίτερα κρίσιμη. Η γνώση τους χρησιμοποιείται στα αρχικά στάδια της προετοιμασίας των δεδομένων (δηλ. να αποφασίσουν για τον καθαρισμό και προετοιμασία των δεδομένων) και στην επιλογή των κατάλληλων παραμέτρων στους αλγορίθμους εξόρυξης γνώσης. Επιπλέον, η συμβολή τους είναι απαραίτητη στην αξιολόγηση και επεξήγηση των αποτελεσμάτων, των εξαγόμενων προτύπων που οδηγεί στη σχετική με τον τομέα γνώση (Fayyad et al., 1996).

Στην ουσία, τα πρότυπα που έχουν εξαχθεί χρησιμοποιούνται από τους ειδικούς του τομέα για να ανακαλύψουν νέες συσχετίσεις στα δεδομένα, να επικυρώσουν θεωρίες και να ανακαλύψουν γενικότερα κρυμμένη γνώση που θα οδηγήσει σε νέα πειράματα και θεωρίες. Εντούτοις, κάποια από τα εξαγόμενα πρότυπα θεωρούνται τετριμμένα και κάποια άλλα ασήμαντα, αφού για τον συγκεκριμένο τομέα έρευνας αναφέρονται σε υπάρχουσα γνώση ή σε λανθασμένα συμπεράσματα. Για να αξιολογήσουν τα εξαγόμενα πρότυπα οι ειδικοί έχουν ορίσει διάφορα αντικειμενικά ή υποκειμενικά μέτρα σημαντικότητας βασισμένα σε στατιστικές κυρίως ιδιότητες των προτύπων. Ωστόσο η διαδικασία ανάλυσης και εξέτασης της χρησιμότητας και σημαντικότητας των προτύπων είναι μία χρονοβόρα και δύσκολη διαδικασία (Piatetsky-Shapiro, 2000).

Το θέμα στην περίπτωση αυτή σχετίζεται με την ενσωμάτωση της ήδη υπάρχουσας γνώσης στη διαδικασία της εξόρυξης γνώσης, και ειδικά στη φάση της αξιολόγησης των προτύπων. Διάφορα στατιστικά μέτρα έχουν προταθεί για την αξιολόγηση των προτύπων (Piatetsky-Shapiro, 1991; Freitas, 1999; Silberschatz & Tuzhilin, 1996; Piatetsky-Shapiro, & Matheus, 1994). Αυτά τα

μέτρα εφαρμόζονται πριν ή κατά τη διάρκεια της διαδικασίας της αξιολόγησης. Στην πρώτη περίπτωση, χρησιμοποιούνται για να μειώσουν τον αριθμό των προτύπων που θα εξαχθούν και για να επιταχύνουν τη διαδικασία, ενώ στη δεύτερη περίπτωση χρησιμοποιούνται για να καθαρίσουν τα πρότυπα που θεωρούνται ασήμαντα.

Ωστόσο δεν υπάρχει μέτρο για την αξιολόγηση των προτύπων τόσο αποτελεσματικό όσο η γνώση του ειδικού της περιοχής. Οι ειδικοί σε κάθε περιοχή μπορούν να αποφασίσουν ποια πρότυπα θεωρούνται ασήμαντα και ποιά όχι, αν είναι σπάνιες περιπτώσεις προτύπων ή είναι στατιστικός θόρυβος, με βάση τη γνώση και την εμπειρία τους (Pohle, 2003). Προκειμένου να αυτοματοποιηθεί η διαδικασία της αξιολόγησης θα πρέπει να ενσωματωθεί σε αυτή, η ίδια η γνώση της περιοχής (domain knowledge). Η γνώση αυτή μπορεί να αναπαρασταθεί από τις οντολογίες (Pohle, 2003). Οι οντολογίες είναι η αναπαράσταση της γνώσης γύρω από μία περιοχή με τη χρήση συγκεκριμένης ορολογίας και κανόνων/ συσχετίσεων μεταξύ των όρων. Έτσι ορίζουν απόλυτα και σαφώς τη γνώση μιας περιοχής που έχουν αποκτήσει οι ερευνητές/ ειδικοί του χώρου αυτού (Gruber, 1993).

Η χρήση της γνώσης της περιοχής ενδιαφέροντος εκφρασμένης με οντολογίες θεωρούμε ότι μπορεί να χρησιμοποιηθεί ως φίλτρο για τη φάση της αξιολόγησης στη διαδικασία KDD. Τα πρότυπα που εξάγονται από τους αλγορίθμους εξόρυξης γνώσης πρώτα θα αξιολογούνται σύμφωνα με την οντολογία. Αυτά που είναι αντίθετα με τη γνώση που παρέχεται από την οντολογία (που θα ονομάζονται στο εξής "*noisy*") θα μαρκάρονται ως πιθανώς άκυρα (invalid). Τα αποδεκτά πρότυπα θα αξιολογούνται και από τους ειδικούς και, αν αναπαριστούν χρήσιμη γνώση, η οντολογία μπορεί να ανανεωθεί για να τα συμπεριλάβει (όπως επίσης μπορεί η ίδια η οντολογία να τροποποιηθεί με την αφαίρεση/ προσθήκη σχέσεων, συσχετίσεων κλπ). Στην περίπτωση αυτή δίνεται προτεραιότητα στα πρότυπα που θεωρούνται ενδιαφέροντα, και που δεν έρχονται σε σύγκρουση με τη μέχρι τώρα γνώση και θεωρία. Με τον τρόπο αυτό επιτυγχάνεται μείωση του χρόνου εκτέλεσης των αλγορίθμων αλλά και του χρόνου και κόπου που απαιτείται από τον ειδικό για την αξιολόγηση. Τα "noisy" πρότυπα σημειώνονται και δεν διαγράφονται παρά μόνο αν το θελήσει ο ειδικός. Έτσι μειώνεται ο κίνδυνος να χαθεί χρήσιμη νέα πληροφορία.

Στις επόμενες ενότητες παρουσιάζουμε τα προβλήματα και τις προτάσεις μας για την ενσωμάτωση των οντολογιών στη διαδικασία της εξόρυξης, χρησιμοποιώντας για παράδειγμα το χώρο της σεισμολογίας.

6.3.1 Εξόρυξη Προτύπων με Χρήση Γνώσης της Περιοχής

Μέχρι πρόσφατα, αν και η σημασία της διαχείρισης γνώσης ήταν γνωστή, δεν είχε γίνει μεγάλη έρευνα για την έξυπνη ανάλυση των προτύπων και την ενσωμάτωση της ήδη υπάρχουσας γνώσης με τη νέα (Pohle, 2003). Ελάχιστες προσεγγίσεις υπάρχουν για τη χρήση της γνώσης της περιοχής στη διαδικασία εξόρυξης προτύπων από βάσεις δεδομένων. Η γνώση αυτή μπορεί να εφαρμοστεί σε τρία επίπεδα. Στο βήμα προετοιμασίας των δεδομένων, κατά τη διάρκεια της διαδικασίας εξόρυξης (χρήση της γνώσης αυτής από τον αλγόριθμο εξόρυξης για να αποφασιστεί το επόμενο βήμα) και στο βήμα μετά την εξόρυξη (για την αξιολόγηση των προτύπων).

Σχετικά με το πρώτο επίπεδο, οι Chen et al. (2003) προτείνουν τη χρήση οντολογίας σαν ιεραρχία για να προετοιμάσουν τα δημογραφικά δεδομένα για εξόρυξη κανόνων συσχέτισης. Σε κάποιες εγγραφές της βάσης των δημογραφικών δεδομένων υπάρχουν τιμές από ένα χαμηλότερο επίπεδο της ιεραρχίας ενώ σε άλλες στην ίδια στήλη/γνώρισμα, υπάρχουν τιμές από ένα ψηλότερο επίπεδο (για παράδειγμα η τιμή “basketball” και η τιμή “recreation sports” που βρίσκονται σε διαφορετικά επίπεδα της ιεραρχίας «ενδιαφερόντων». Με την αντικατάσταση των τιμών ενός χαμηλότερου επιπέδου με αυτές ενός υψηλότερου (raising), οι συγγραφείς δείχνουν ότι η υποστήριξη (support) αυξάνεται και έτσι ανακαλύπτονται περισσότεροι κανόνες.

Πολλές εργασίες υπάρχουν σχετικές με το πώς διάφορα μέτρα ενδιαφέροντος (interestingness measures), αντικειμενικά ή υποκειμενικά, χρησιμοποιούνται για την αξιολόγηση των εξαγόμενων προτύπων. Τα αντικειμενικά μέτρα βασίζονται σε στατιστικές συναρτήσεις. Στο (Piatetsky-Shapiro, 2000) ορίζονται οι βασικές αρχές των αντικειμενικών μέτρων ενδιαφέροντος, ενώ στο (Freitas, 1999) γίνεται μία σύγκριση των αντικειμενικών μέτρων. Σε αντίθεση με τα αντικειμενικά μέτρα, τα υποκειμενικά προσπαθούν να λάβουν υπόψη ξεχωριστές συνθήκες του (ανθρώπου) αναλυτή. Μία γενική συζήτηση γίνεται στο (Silberschatz & Tuzhilin, 1996), ενώ στο (Piatetsky-Shapiro, & Matheus, 1994) και στο (Padmanabhan, & Tuzhilin, 1998) γίνεται προσπάθεια αντιμετώπισης του προβλήματος. Όλες αυτές οι προσεγγίσεις παρέχουν έναν

τρόπο για την αξιολόγηση των προτύπων αλλά δεν κάνουν χρήση της γνώσης της περιοχής (domain knowledge).

Υπάρχουν επίσης κάποιες προσπάθειες χρήσης της γνώσης αυτής για τη βελτίωση της αξιολόγησης των προτύπων. Η γνώση αυτή, στη μορφή ιεραρχιών εννοιών μπορεί να χρησιμοποιηθεί για να βελτιώσει τα αποτελέσματα της εξόρυξης γνώσης από τον παγκόσμιο ιστό (Pohle & Spiliopoulou, 2002), ενώ ένα σύστημα ανάλυσης ενδιαφέροντος που απαιτεί από το χρήστη να εκφράσει διάφορους τύπους της υπάρχουσας γνώσης σε όρους μίας ειδικής γλώσσας παρουσιάζεται στο (Liu et al., 2000). Αυτές οι προσεγγίσεις χρησιμοποιούν τη γνώση της περιοχής αλλά το μειονέκτημά τους είναι ότι απαιτούν από το χρήστη να την δώσει αυτή σε μία συγκεκριμένη μορφή, ανάλογα κάθε φορά την εφαρμογή.

Προκειμένου να ενσωματωθεί η γνώση της περιοχής στη διαδικασία εξόρυξης γνώσης και να επιτραπεί η ανταλλαγή θεμελιώδους μοντέλων σε διάφορες περιοχές, είναι απαραίτητη η χρήση οντολογιών (Maedche, Motiket al., 2003). Μία εφαρμογή που χρησιμοποιεί οντολογίες πριν, κατά τη διάρκεια και μετά τη διαδικασία εξόρυξης γνώσης παρουσιάζεται από τους Hotho et al. (2002), στην οποία οι οντολογίες συνδυάζονται με τεχνολογίες εξαγωγής πληροφορίας (Information Extraction) για να βελτιωθούν οι αλγόριθμοι εξαγωγής γνώσης από κείμενα (text mining) και η ερμηνεία των προτύπων.

Η μεθοδολογία που προτείνουμε χρησιμοποιεί οντολογίες για τη βελτίωση του βήματος της αξιολόγησης των προτύπων και της ανάκτησης από τη βάση προτύπων. Κατά το βήμα της αξιολόγησης το σύστημα βασίζεται στην παρεχόμενη οντολογία και στις παραμέτρους που έχουν οριστεί από το χρήστη, φιλτράρει τα πρότυπα και τα μαρκάρει ως “noisy” εάν αντιφάσκουν στη γνώση της περιοχής. Οι ειδικοί στη συνέχεια μπορούν είτε να τους απορρίψουν είτε να τους αξιολογήσουν. Το σύστημα επιπλέον χρησιμοποιεί το μηχανισμό φιλτραρίσματος για να αποτρέψει έναν λιγότερο έμπειρο χρήστη να κάνει επερωτήσεις στη βάση για κάποια “noisy” πρότυπα.

6.3.2 Περιγραφή του Προβλήματος

Η ανάγκη για ενσωμάτωση της γνώσης της περιοχής στη διαδικασία εξόρυξης γνώσης μπορεί να φανεί από διάφορα παραδείγματα, ωστόσο οι εφαρμογές που σχετίζονται με επιστημονικά δεδομένα είναι περισσότερο χαρακτηριστικές καθότι οι ειδικοί στις περιοχές αυτές γνωρίζουν τα δεδομένα σε μεγάλη

λεπτομέρεια (Fayyad et al., 1996). Στην ενότητα αυτή θα παρουσιάσουμε μία πραγματική περίπτωση εξαγωγής σεισμολογικών δεδομένων για να δείξουμε τη χρήση ενός PBMS και των οντολογιών σε ένα περισσότερο ενοποιημένο περιβάλλον για διαχείριση και αξιολόγηση προτύπων.

6.3.2.1 Η Περίπτωση Σεισμολογικών Δεδομένων

Έστω ότι υπάρχει μία βάση δεδομένων που περιέχει ιστορικά δεδομένα σχετικά με σεισμολογικά γεγονότα (Theodoridis et al., 2004). Μια τέτοια βάση θα περιέχει πληροφορία σχετικά με το γεγονός (μέγεθος, γεωγραφικές συντεταγμένες, χρονοσφραγίδα, βάθος του σεισμού), καταγεγραμμένες μακροσεισμικές πληροφορίες (πχ. ένταση) από τις περιοχές που έχουν επηρεαστεί από το σεισμό. Επιπλέον, η βάση αυτή περιέχει δημογραφικά στοιχεία και διοικητικές πληροφορίες για περιοχές και χώρες καθώς επίσης και λεπτομέρειες για τη μορφολογία του εδάφους και τα σχετικά με το σεισμό ρήγματα (Theodoridis, Marketos, & Kalogeras, 2004).

Οι σεισμολόγοι χρησιμοποιούν τη βάση για να αποθηκεύσουν δεδομένα, μία αποθήκη δεδομένων (data warehouse) για να συναθροίζουν και να αναλύουν τα δεδομένα αυτά, μία βάση γνώσης (knowledge base) για να αποθηκεύουν σχετικά αρχεία που έχουν συλλεχθεί από διάφορες πηγές και ένα εργαλείο για να ορίζουν οντολογίες σχετικές με τον επιστημονικό χώρο.

Επιπλέον, ενδιαφέρονται για την ανακάλυψη κρυμμένης γνώσης στα δεδομένα. Τα πρότυπα που παράγονται από τη διαδικασία εξόρυξης γνώσης, αξιολογούνται και αποθηκεύονται σε ένα ΣΔΒΠ. Είναι φανερό ότι εάν όλα τα παραπάνω δεν είναι ενοποιημένα σε μία κοινή εφαρμογή, δε θα μπορεί να αξιοποιηθεί το μέγιστο της αξίας τους. Οι ερευνητές θέλουν να μπορούν να κάνουν ερωτήσεις για όλα αυτά πιθανώς χωρίς να γνωρίζουν ποιο από τα εργαλεία που αναφέρθηκαν πρέπει να χρησιμοποιηθεί. Τέτοιες ερωτήσεις μπορεί να είναι:

- *ερώτηση 1*: ποιο είναι το μέγιστο βάθος και το μέγιστο βάθος για τους σεισμούς που συνέβησαν στην Βόρεια Αδριατική Θάλασσα (ή σε οποιαδήποτε άλλη περιοχή) για τη δεκαετία 1994-2004.

- *ερώτηση 2*: υπάρχουν πληροφορίες για τη μέγιστη καταγεγραμμένη ένταση όταν είναι γνωστό ότι το βάθους του επίκεντρου είναι 60 km και το έδαφος της περιοχής χαρακτηρίζεται βραχώδες;

- *ερώτηση 3*: να βρεθούν ομοιότητες σε μετασεισμικές ακολουθίες για σεισμούς που έγιναν στην Ελλάδα το 2004.

Η ερώτηση 1 μπορεί εύκολα να απαντηθεί από μία αποθήκη δεδομένων χρησιμοποιώντας τη συνάρτηση μέγιστου στα κατάλληλα σεισμολογικά δεδομένα.

Η ερώτηση 2 επίσης μπορεί εύκολα να απαντηθεί με τη χρήση ενός δέντρου απόφασης. Στην περίπτωση που ένα τέτοιο δέντρο απόφασης (πρότυπο) δεν έχει ήδη καταχωριστεί στο PBMS τότε ο κατάλληλος αλγόριθμος κατηγοριοποίησης μπορεί να εφαρμοστεί. Η ερώτηση 3 προσφέρει περισσότερες προκλήσεις αφού απαιτεί την ενσωμάτωση πιο προχωρημένης γνώσης: α) τον ορισμό της μετασεισμικής ακολουθίας και β) τον ορισμό του μέτρου ομοιότητας δύο μετασεισμικών ακολουθιών από τον ειδικό.

Είναι φανερό ότι η ερώτηση 3 απαιτεί αρκετή προπαρασκευή δεδομένων από τον ειδικό της περιοχής (σεισμολόγο) σε συνεργασία με κάποιον αναλυτή βάσεων δεδομένων. Οι ιεραρχίες και οι κανόνες σχετικά με έννοιες και δεδομένα της σεισμολογίας πρέπει να οριστούν πριν να εφαρμοστεί ο αλγόριθμος εξόρυξης. Επιπλέον, ακόμα και όταν πρότυπα παράγονται και αποθηκεύονται στο PBMS, κάποια μετέπειτα επεξεργασία είναι απαραίτητη για την εξαγωγή της πληροφορίας. Αν ο σεισμολόγος είχε ήδη αναπαραστήσει την απαιτούμενη γνώση με χρήση οντολογιών, η ενσωμάτωσή της στο PBMS θα έλυνε τα παραπάνω.

Από την άλλη μεριά, όμως, μπορεί να γίνουν ερωτήσεις όπως

- ερώτηση 4: να βρεθούν συσχετισμοί μεταξύ του μεγέθους του σεισμού και της μέσης θερμοκρασίας στην περιοχή γύρω από το επίκεντρο του σεισμού ·

- ερώτηση 5: να βρεθούν πιθανές συσχετίσεις μεταξύ του μεγέθους του σεισμού και της εποχής του έτους·

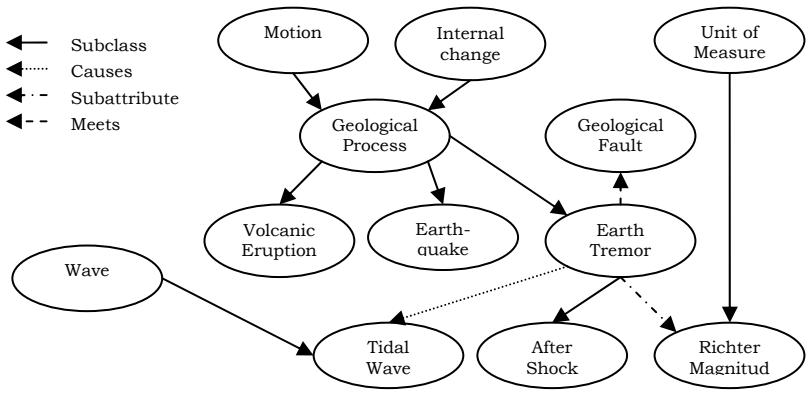
από κάποιον μη ειδικό χρήστη, των οποίων οι όποιες απαντήσεις είναι σημασιολογικά λανθασμένες. Παρότι οι διαδικασίες εξόρυξης γνώσης θα επιστρέψουν σχετικές απαντήσεις, ο ειδικός της περιοχής θα τις απορρίψει (για παράδειγμα, η επιστήμη της σεισμολογίας δεν αναγνωρίζει καμία σχέση μεταξύ του μεγέθους του σεισμού και της θερμοκρασίας της επιφάνειας ή της εποχής του χρόνου). Εντούτοις, η μηχανή εξόρυξης γνώσης θα μπορούσε να εξάγει

αποτελέσματα με τέτοιες συσχετίσεις και ένας ειδικός σίγουρα θα τα απέρριπτε εκ των υστέρων.

Παρόλα αυτά, ένα τέτοιο φιλτράρισμα γίνεται χειροκίνητα στις μέρες μας. Ακριβώς αυτή είναι η συνεισφορά ενός PBMS που ενσωματώνει οντολογίες. Να φιλτράρει τα "noisy" πρότυπα αποτελεσματικά χωρίς την ανάγκη προ-επεξεργασίας και με εγγυημένη απόδοση με χρήση του φίλτρου.

6.3.2.2 Αποτύπωση της Γνώσης της Περιοχής με Χρήση Οντολογιών

Για τα συγκεκριμένα δεδομένα και εφαρμογή (σεισμολογικά δεδομένα), χρησιμοποιείται η οντολογία Suggested Upper Merged Ontology-SUMO (IEEE Standard Upper Ontology) (Niles & Pease, 2001), και συγκεκριμένα η Mid-Level οντολογία (Niles & Terry, 2004) της οποίας ένα υποσύνολο φαίνεται στο παρακάτω σχήμα.



Εικόνα 6-6 Υποσύνολο της οντολογίας SUMO για σεισμολογία

Προφανώς η παραπάνω εικόνα δεν αναπαριστά το ευρύτερο σύνολο αναφοράς, αλλά είναι μέρος μόνο της οντολογίας που σχετίζεται με τη γεωλογία, μέρος που αφορά σε έννοιες σχετικές με σεισμολογία.

Σχετικά με την εξαγωγή κανόνων συσχέτισης, ένας γενικός κανόνας που μπορεί να εφαρμοστεί για την αξιολόγηση προτύπων με οντολογίες είναι ότι τα πρότυπα πρέπει να συσχετίζουν γνωρίσματα που να ανήκουν στην ίδια κλάση ή υποκλάση της ίδιας κλάσης. Λογικά, οι συσχετίσεις γνωρισμάτων που ανήκουν σε διαφορετικές κλάσεις ή σε κλάσεις που είναι σε αρκετή απόσταση μεταξύ τους (στο γράφημα της οντολογίας) μπορεί να οδηγήσουν σε λανθασμένα συμπεράσματα, συσχετίσεις άσχετων δηλαδή γνωρισμάτων. Η απόσταση των

ακμών και άλλες προσεγγίσεις έχουν ήδη προταθεί για την αναζήτηση σημασιολογικής ομοιότητας μεταξύ αντικείμενων σε μία οντολογία. Τέτοια μέτρα μπορεί να χρησιμοποιηθούν για να βρεθεί η σχέση μεταξύ δύο γνωρισμάτων. Αυτό σημαίνει ότι ο χρήστης μπορεί να επιλέξει το επίπεδο συσχετισμού μεταξύ γνωρισμάτων, ορίζοντας τη μέγιστη απόσταση που μία κλάση μπορεί να έχει από μία άλλη στο γράφημα της οντολογίας για να θεωρηθούν σχετικές.

Η εργασία ορισμού των κανόνων που θα χρησιμοποιηθούν για να φιλτράρουν τα πρότυπα που εξάγονται απαιτεί τη μελέτη της οντολογίας, τη μελέτη του τύπου προτύπου και των αποτελεσμάτων που περιμένει ο χρήστης. Τα μέρη της οντολογίας είναι κλάσεις, γνωρίσματα και σχέσεις μεταξύ τους. Οι κλάσεις έχουν υποκλάσεις και κάθε κλάση έχει έναν αριθμό γνωρισμάτων. Συνήθως οι κλάσεις για σχετικές έννοιες ανήκουν στην ίδια κλάση γονέα ενώ μη σχετικές έννοιες βρίσκονται κάτω από διαφορετικές κλάσεις. Το διάγραμμα κλάσεων και υποκλάσεων ορίζει ένα είδος ιεραρχίας με διάφορα επίπεδα λεπτομέρειας. Για παράδειγμα, οι κλάσεις “VolcanicEruption” και “Earhtremor” (Εικόνα 6-6) βρίσκονται κάτω από το ίδιο επίπεδο ενώ οι υποκλάσεις “volcanicGasRelease” και “AfterShock” βρίσκονται κάτω από διαφορετικό επίπεδο.

Καθώς κάθε πρότυπο έχει διαφορετική δομή, φίλτρα για κάθε τύπο προτύπου πρέπει να οριστούν. Συγκεκριμένα για τον τύπο κανόνα συσχέτισης (Association Rule) ορίζουμε το αντίστοιχο φίλτρο Association Rule Filter. Κάθε μέρος του κανόνα περιέχει γνωρίσματα (depth, magnitude κλπ) που σχετίζονται στο σχεσιακό μοντέλο, αλλά επιπλέον σχετίζονται με κάποιο τρόπο στην οντολογία. Έτσι, μπορούμε να ορίσουμε για κάθε κανόνα ένα μέτρο απόστασης μεταξύ της βασικής κλάσης σεισμών (*earth tremor*) και των κόμβων των γνωρισμάτων που περιέχονται στον κανόνα. Όσο μικρότερη είναι η απόσταση, τόσο περισσότερο τα γνωρίσματα σχετίζονται σημασιολογικά. Στην πραγματικότητα μπορούν να οριστούν δύο προσεγγίσεις για τη μέτρηση αυτής της απόστασης: στη λεγόμενη “*Risky*” προσέγγιση, θεωρούμε μία μέγιστη απόσταση μεταξύ των κόμβων των γνωρισμάτων και της κύριας κλάσης, ενώ στη “*Non-Risky*” προσέγγιση, θεωρούμε την ελάχιστη απόσταση μεταξύ τους. Είναι φανερό ότι τα γνωρίσματα της κλάσης του σεισμού έχουν απόσταση distance=0 και έτσι δεν συμπεριλαμβάνονται στον υπολογισμό.

Ο χρήστης επιλέγει το επίπεδο της σημασιολογικής σχέσης με τον ορισμό της μέγιστης απόστασης των κόμβων από τη βασική κλάση. Για παράδειγμα

κάποιος μπορεί να ενδιαφέρεται στο να βρει συσχετίσεις όχι μόνο μεταξύ των γνωρισμάτων του σεισμού αλλά και μεταξύ αυτών και των γεωλογικών ρηγμάτων (geological faults). Έτσι, το επίπεδο της σημασιολογικής σχέσης πρέπει να αυξηθεί για να περιλαμβάνει τους κατάλληλους κόμβους.

Με την παραπάνω διαδικασία, δημιουργείται ένα υπογράφημα της οντολογίας που περιέχει τα σχετικά γνωρίσματα. Τα γνωρίσματα των παραγόμενων κανόνων αξιολογούνται με βάση αυτό το υπογράφημα. Εάν όλα τα γνωρίσματα συμπεριλαμβάνονται σε αυτό τότε ο κανόνας που τα περιέχει θεωρείται έγκυρος. Σε αντίθετη περίπτωση ο κανόνας μαρκάρεται ως “noisy”. Οι κανόνες αυτοί είτε απορρίπτονται κατευθείαν είτε ελέγχονται από τον ειδικό καθότι κάποιοι μπορούν να οδηγήσουν σε ενδιαφέρουσες συσχετίσεις, άγνωστες προηγουμένως, και ο ειδικός μπορεί ενδεχομένως να αναθεωρήσει την οντολογία.

6.3.3 Προκαταρκτική Μελέτη Αποτίμησης

Στην ενότητα αυτή χρησιμοποιούμε το παράδειγμα της σεισμολογίας και της οντολογίας που περιγράφηκε στην ενότητα 6.3.2.2 για να περιγράψουμε τον τρόπο λειτουργίας του ενοποιημένου συστήματος. Το σύστημα κάνει ένα έλεγχο εγκυρότητας πριν τη διαδικασία εξόρυξης για να ελεγχθεί αν οι παράμετροι του χρήστη έχουν νόημα. Για παράδειγμα, ο χρήστης μπορεί να ζητήσει να βρεθούν συσχετίσεις μεταξύ των γνωρισμάτων “magnitude” και “date” ενός σεισμού, κάτι που δεν είναι αποδεκτό όπως αναφέρθηκε στην ενότητα 6.3.2.1. το σύστημα στην περίπτωση αυτή θα προτείνει στο χρήστη να αλλάξει τις παραμέτρους. Εάν ο χρήστης δε δηλώσει εξαρχής συγκεκριμένα γνωρίσματα το σύστημα θα αναζητήσει γενικώς κανόνες αλλά στην παραγωγή των συχνών στοιχειοσυνόλων θα απορρίψει αυτά που περιέχουν γνωρίσματα που δε σχετίζονται στην οντολογία. Με τον τρόπο αυτό, το χρονοβόρο μέρος της εύρεσης συχνών στοιχειοσυνόλων θα βελτιωθεί και δε θα παραχθούν κανόνες με μη-συσχετισμένα γνωρίσματα. Ωστόσο αυτό δεν είναι πάντα επιθυμητό καθότι μπορεί κάποιοι ενδιαφέροντες κανόνες να μην παραχθούν. Στην περίπτωση αυτή ο χρήστης θα πρέπει να αποφασίσει για αυτούς. Έτσι δίνεται η επιλογή στο χρήστη είτε να αφήσει στο σύστημα να απορρίψει αυτόματα αυτούς τους κανόνες είτε να τους μαρκάρει “noisy” για επιπλέον έλεγχο. Στην τελευταία περίπτωση ο χρήστης αποφασίζει ποιοι κανόνες είναι χρήσιμοι και θα αποθηκευτούν στη βάση προτύπων.

Μία άλλη περίπτωση είναι όταν ο χρήστης ανακτά κανόνες από τη βάση προτύπων όπως για παράδειγμα «ανάκτηση των κανόνων που περιέχουν τα γνωρίσματα “season” και “depth” και η υποστήριξη του κανόνα είναι πάνω από 0.3”. Τέτοιοι κανόνες δεν είναι έγκυροι και το σύστημα θα ενημερώσει το χρήστη (πριν προχωρήσει στην αναζήτηση στη βάση) ότι είναι μάλλον απίθανο να βρεθούν τέτοιοι κανόνες στη βάση.

Στα πρώτα πειράματά μας εκτελέσαμε τον αλγόριθμο Apriori που είναι υλοποιημένος στο WEKA (Witten & Frank, 2005) για να εξάγουμε κανόνες από πραγματικά μακροσεισμικά δεδομένα που συλλέχθηκαν από το Γεωδυναμικό Ινστιτούτο του Εθνικού Αστεροσκοπείου Αθηνών και αποτελούν τη βάση δεδομένων του εργαλείου Seismo-Surfer (Seismo-Surfer). Γνωρίσματα όπως earthquake depth, intensity, site, date και season of the year είναι μερικά από τα γνωρίσματα του πίνακα που περιέχει 10336 εγγραφές για σεισμούς στον Ελλαδικό χώρο τον 20ο αιώνα. Ο Πίνακας 6-1 αναφέρει 25 από τους 70 κανόνες που εξήχθησαν με τον Apriori με confidence threshold = 30% και support threshold = 10%.

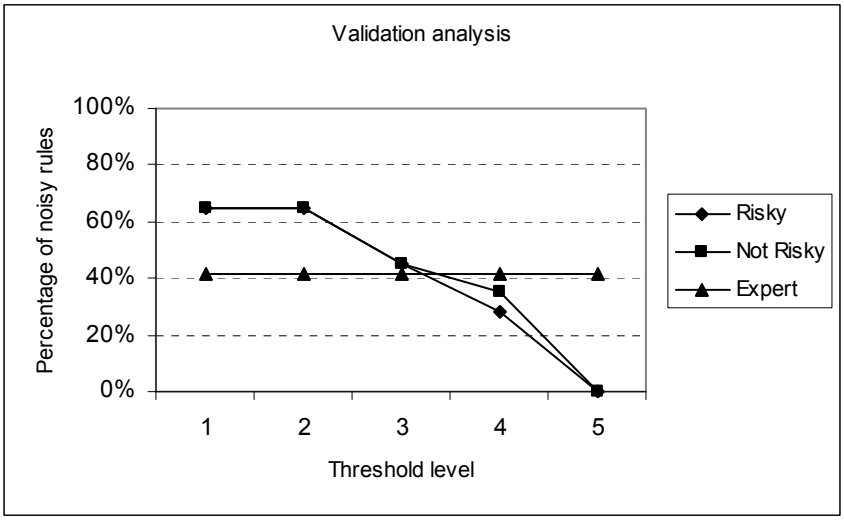
Πίνακας 6-1 κανόνες συσχέτισης από σεισμολογικά δεδομένα

id	Association Rule	Conf.	Supp.
1	intensity≥5 → distance≤80	74%	19%
2	weekDay=Tuesday, 11≤depth≤20 → season=Summer	71%	10%
3	weekDay=Tuesday → season=Summer	71%	17%
4	weekDay=Monday → season=Spring	68%	10%
5	season=Summer → 11≤depth≤20	65%	21%
6	weekDay=Saturday → 21≤depth≤50	62%	12%
7	depth≥50 → season=Spring	60%	11%
8	distance≥150 → intensity≤3	59%	15%
9	weekDay=Tuesday, season=Summer → 11≤depth≤20	57%	10%
10	weekDay=Tuesday → 11≤depth≤20	57%	14%
11	11≤depth≤20 → season=Summer	57%	21%
12	season=Autumn → 11≤depth≤20	55%	14%
13	season=Summer → weekDay=Tuesday	54%	17%
14	intensity≤3 → distance≥150	54%	15%

15	distance≤80 → intensity≥5	52%	19%
16	distance≥150 → 1000<population≤4000	48%	13%
17	3<intensity≤4 → 80<distance<150	48%	15%
18	season=Summer, 11≤depth≤20 → weekDay=Tuesday	48%	10%
19	weekDay=Tuesday → 1000<population≤4000	46%	11%
20	season=Spring → 21≤depth≤50	46%	14%
21	intensity≤3 → 1000<population≤4000	46%	13%
22	21≤depth≤50 → season=Spring	45%	14%
23	500<population≤1000 → distance≤80	43%	11%
24	season=Spring → 1000<population≤4000	43%	13%
25	80<distance<150 → 1000<population≤4000	43%	15%

Από τους 25 αυτούς κανόνες, ο ειδικός της περιοχής χαρακτήρισε ενδιαφέροντες μόνο τους 5 (ids 1, 8, 14, 15, 17) και τους υπόλοιπους τους χαρακτήρισε «θόρυβο» επειδή περιγράφουν συσχετίσεις μεταξύ δεδομένων που δεν έχουν πραγματική συσχέτιση στο χώρο της σεισμολογίας. Το σύστημα χρειάζεται μία παράμετρο κατωφλιού για να χαρακτηρίσει τους κανόνες ως «θόρυβο». Η απόσταση αυτή είναι η απόσταση του μονοπατιού από την κεντρική έννοια/κλάση “earth tremor”. Όταν αυτό οριστεί, το σύστημα ανακτά το υπο-γράφο της οντολογίας με τους κόμβους που απέχουν όσο το κατώφλι από τον κεντρικό κόμβο “earth tremor”. Όποιος κανόνας περιλαμβάνει χαρακτηριστικά που βρίσκονται στο γράφο αυτό θεωρείται σχετικός, αλλιώς θεωρείται «θόρυβος».

Για τον εντοπισμό του καλύτερου κατωφλιού στη συγκεκριμένη εφαρμογή, δοκιμάστηκαν τιμές από 1 έως 5 και τα αποτελέσματα συγκρίθηκαν με αυτά του ειδικού της περιοχής. Το αποτέλεσμα φαίνεται στο παρακάτω διάγραμμα (Εικόνα 6-7).



Εικόνα 6-7 Κατώφλι και κανόνες που απορρίφθηκαν από το σύστημα και το σεισμολόγο

Σύμφωνα με το παραπάνω πείραμα συμπεραίνεται ότι το κατώφλι με αριθμό 3 ταιριάζει στην επιλογή του ειδικού. Πρέπει να σημειωθεί ωστόσο ότι παρότι ίσος αριθμός κανόνων απορρίφθηκαν από το σύστημα και τον ειδικό, αυτό δεν εξασφαλίζει ότι είναι οι ίδιοι οι κανόνες αυτοί που απορρίφθηκαν.

Με τη διαδικασία που περιγράφηκε παραπάνω, μπορούμε να μετρήσουμε το ποσοστό των κανόνων που θα μαρκαριστούν ως “noisy” από το σύστημα και από τον ειδικό, αλλά δεν γνωρίζουμε εάν αυτοί οι κανόνες είναι οι ίδιοι. Δηλαδή, εάν οι κανόνες που μαρκαρίστηκαν από το σύστημα είναι οι ίδιοι με αυτούς που μάρκαρε ο ειδικός (το precision δηλαδή). Ενώ στο συγκεκριμένο πείραμα υπήρχε πλήρη ταύτιση, δεν είναι βέβαιο ότι θα ισχύει αυτό και σε άλλη περίπτωση.

6.3.4 Συζήτηση

Η χρήση οντολογιών στη διαδικασία της εξόρυξης γνώσης αποτελεί ένα σύγχρονο πεδίο έρευνας και μπορεί να έχει πολλές εφαρμογές. Πέρα από τις σχετικές με τη γεωλογία επιστήμες, σε κάθε πεδίο στο οποίο υπάρχει μία καλώς ορισμένη οντολογία μπορεί να χρησιμοποιηθεί το ενοποιημένο πλαίσιο για να βελτιωθεί η διαδικασία KDD. Για παράδειγμα στον τομέα των αγορών B2B η εύρεση συσχετίσεων μεταξύ προϊόντων είναι πιο αποτελεσματική όταν χρησιμοποιούνται οι ιεραρχίες που ορίζονται από την οντολογία προϊόντων. Αν και μέχρι στιγμής δεν υπάρχει μία παγκοσμίως αποδεκτή οντολογία προϊόντων,

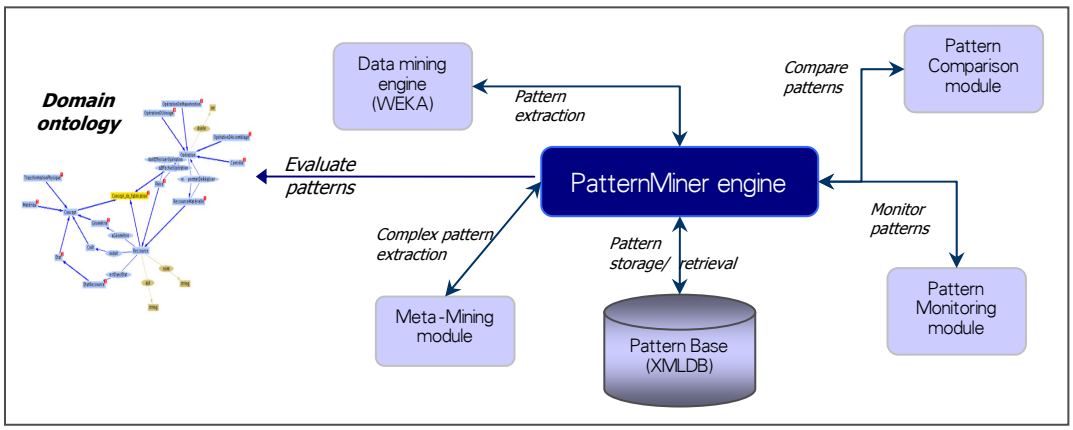
έχουν γίνει προσπάθειες για να ενοποιηθούν οι διάφορες σχετικές οντολογίες (Omelayenko, 2000).

Προκειμένου να είναι δυνατή η χρήση οντολογιών στη διαδικασία KDD και να είναι τα αποτελέσματα διαθέσιμα στους ειδικούς της περιοχής, οι οντολογίες πρέπει να ορίζονται με ενιαίο τρόπο. Υπάρχουν αρκετές προσπάθειες για το ταίριασμα οντολογιών (Doan, Madhavan, Domingos, & Halevy, 2003) και την ενοποίηση οντολογιών (Cui, Jones, & O'Brien, 2002), (Pinto & Martins, 2001) και αυτό δείχνει την ανάγκη για ένα standard για τη δημιουργία οντολογιών. Στην περίπτωση αυτή η ανταλλαγή και σύγκριση οντολογιών που περιγράφουν διαφορετικές περιοχές θα είναι δυνατή. Μέχρι τώρα έχουν αναπτυχθεί μόνο λίγες οντολογίες και λίγα σχετικά εργαλεία.

6.3.5 Επέκταση του PatternMiner για την Υποστήριξη Αξιολόγησης Προτύπων

Η επέκταση του συστήματος που προτείνουμε παρέχει τόσο σε μη έμπειρους όσο και σε έμπειρους χρήστες λειτουργικότητα για αποδοτική διαχείριση των προτύπων και αξιολόγηση των προτύπων με χρήση οντολογιών. Το σύστημα είναι σε θέση να αξιολογήσει τα πρότυπα πριν, κατά τη διάρκεια και μετά τη διαδικασία εξόρυξης, καθώς και κάθε φορά που ο χρήστης θέτει μία επερώτηση στη βάση προτύπων.

Η αρχιτεκτονική του επεκτεταμένου συστήματος παρουσιάζεται στην Εικόνα 6-8. Η επέκταση επικύρωσης με οντολογίες δεν είναι ενσωματωμένη στο PatternMiner και για το λόγο αυτό φαίνεται ξεχωριστά.



Εικόνα 6-8 Αρχιτεκτονική του επεκτεταμένου PatternMiner με την προσθήκη της οντολογίας

Ανεξάρτητα από τη μηχανή εξόρυξης, το PBMS αποθηκεύει τα εξαγόμενα πρότυπα στην XML βάση προτύπων. Το μοντέλο προτύπων που χρησιμοποιείται, εμπλουτίζεται για να υποστηρίξει την αξιολόγηση και τις σημασιολογικά σχετικές κλάσεις προτύπων. Το μοντέλο αυτό ορίζει τέσσερις λογικές έννοιες: *Pattern type*, *pattern*, *class* και *superclass.*

Πιο συγκεκριμένα, κάθε τύπος προτύπου περιέχει μεταδεδομένα σχετικά με:

- τον αλγόριθμο εξόρυξης που εφαρμόζεται για την εξόρυξη των προτύπων και τις παραμέτρους,
- την ώρα και ημερομηνία εκτέλεσης της διαδικασίας της εξόρυξης,
- την περίοδο εγκυρότητας,
- την πηγή των δεδομένων,
- τη συνάρτηση αντιστοίχησης και, τέλος,
- την πληροφορία σχετικά με τη δομή και τα μέτρα ποιότητας των προτύπων.

Τα πρότυπα είναι στιγμιότυπα των τύπων προτύπων. Στην XML αρχιτεκτονική που προτείνουμε, οι τύποι προτύπων είναι XML σχήματα για ένα πρότυπο (XML έγγραφο). Στην παρούσα υλοποίηση χρησιμοποιείται ένα XML σχήμα πιο κοντά στο PANDA πλαίσιο αντί του εμπλουτισμένου PMML μοντέλου. Αυτό δείχνει την ευελιξία του συστήματος καθώς δεν περιορίζει τη χρήση συγκεκριμένων PMML εγγράφων αλλά επιτρέπει κάθε καλώς-ορισμένο σχήμα προτύπου που έχει τις βασικές απαιτήσεις του μοντέλου PANDA.

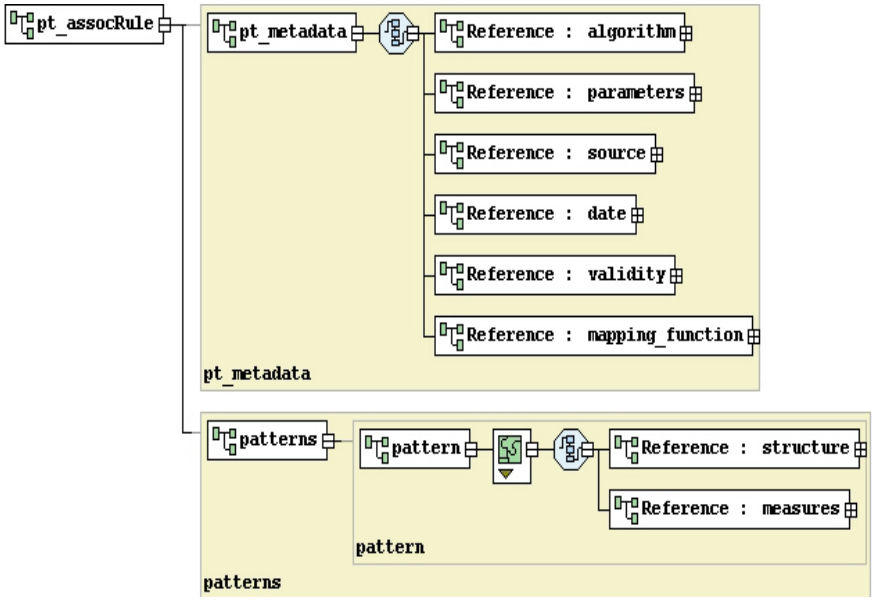
Το έγγραφο του προτύπου περιέχει μεταδεδομένα σχετικά με τη διαδικασία εξόρυξης καθώς και τα πρότυπα που εξήχθησαν από αυτήν. Για παράδειγμα ένα πρότυπο κανόνα συσχέτισης και ο αντίστοιχος τύπος προτύπου, παρουσιάζεται στην Εικόνα 6-9 και την Εικόνα 6-10, αντίστοιχα.

```
<pt_assocRule xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" name="assocRule"
pt_descr="association rules" pt_id="1" xsi:noNamespaceSchemaLocation="pt_assocRule.xsd">
        <pt_metadata>
                <algorithm>apriori</algorithm>
        <parameters>min_support=0.1,min_conf=0.4,rules=10</parameters>
                <source>select * from earthquakes</source>
                <date>2006/04/12 13:03:34</date>
                <validity>2006/06/12 13:03:34</validity>
                <mapping_function>{{'depth', 'magnitude', 'season'} ⊆ transaction}
</mapping_function>     </pt_metadata>
        <patterns>
                <pattern p_id="1">
                        <structure>
                                <body>
                                        <attrib>depth</attrib>
                                        <attrib_value>0-1</attrib_value>
                                </body>
                                <head>
                                        <attrib>magnitude</attrib>
                                        <attrib_value>(3,4]</attrib_value>
                                </head>
                        </structure>
                        <measures>
                                <measure_name>support</measure_name>
                                <measure_value>0.18</measure_value>
                                <measure_name>confidence</measure_name>
                                <measure_value>0.67</measure_value>
                        </measures>
                </pattern>
                <pattern p_id="2">
                        <structure>
                                <body>
                                        <attrib>season</attrib>
                                        <attrib_value>Autumn</attrib_value>
                                </body>
                                <head>
                                        <attrib>magnitude</attrib>
                                        <attrib_value>(3-4]</attrib_value>
                                </head>
                        </structure>
                        <measures>
                                <measure_name>support</measure_name>
                                <measure_value>0.18</measure_value>
                                <measure_name>confidence</measure_name>
                                <measure_value>0.58</measure_value>
                        </measures>
                </pattern>      </patterns> </pt_assocRule>
```

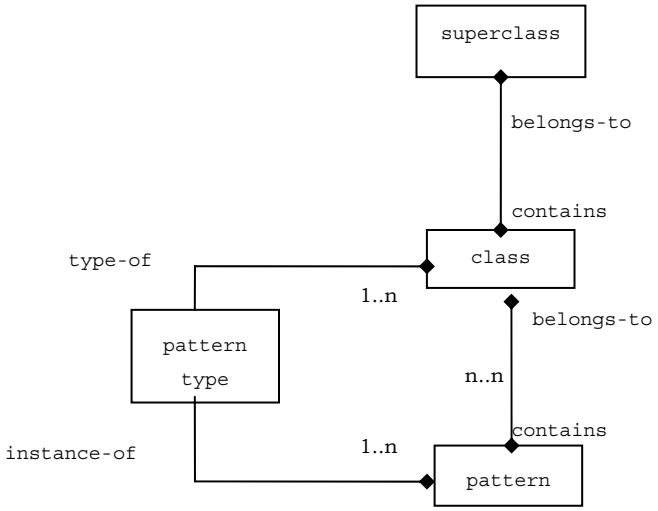
Εικόνα 6-9 Παράδειγμα XML προτύπου κανόνων

Εκτός από τις έννοιες του τύπου προτύπου και του προτύπου, η έννοια της κλάσης ορίζεται ως ένα σύνολο σημασιολογικά συσχετιζόμενων προτύπων του ίδιου τύπου. Μία κλάση ορίζεται από το χρήστη για να ομαδοποιήσει πρότυπα που έχουν κοινό νόημα και ανήκουν σε συγκεκριμένο τύπο προτύπων. Κάθε πρότυπο μπορεί να ανήκει σε μία ή περισσότερες κλάσεις. Για παράδειγμα, ο χρήστης μπορεί να ορίσει μία κλάση που να περιέχει κανόνες συσχέτισης σχετικούς με τη σεισμική δραστηριότητα του καλοκαιριού του 2003. Αυτή η κλάση μπορεί να περιέχει πρότυπα που ανήκουν σε αποτελέσματα διαφορετικών εκτελέσεων των αλγορίθμων εξόρυξης αλλά να έχουν κοινή

σημασία για το χρήστη. Η Εικόνα 6-11 παρουσιάζει το λογικό μοντέλο της βάσης προτύπων.



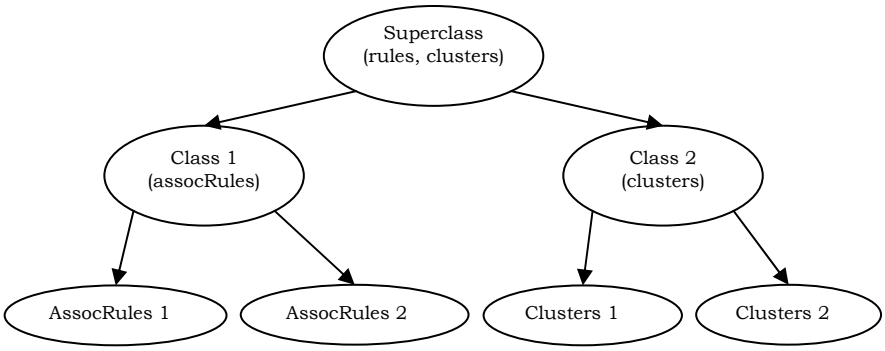
Εικόνα 6-10 Διάγραμμα XSD προτύπου Κανόνα Συσχέτισης



Εικόνα 6-11 Λογικό μοντέλο βάσης προτύπων

Επιπλέον, ορίζεται η έννοια της υπερκλάσης (*superclass*), που είναι ένα σύνολο από κλάσεις, πιθανώς διαφορετικών τύπων προτύπων. Έτσι, πρότυπα που ανήκουν σε διαφορετικούς τύπους μπορούν να ομαδοποιηθούν. Για παράδειγμα, ο χρήστης μπορεί να θέλει να ομαδοποιήσει όλους τους κανόνες συσχέτισης τους σχετικούς με τη σεισμική δραστηριότητα του καλοκαιριού του 2003 και τις συστάδες των ρηγμάτων που έδωσαν σεισμούς μεγέθους M>3 κατά την ίδια χρονική περίοδο. Ο σύνδεσμος μεταξύ των δύο τύπων θα μπορούσε να

είναι το μέγεθος των σεισμών. Με άλλα λόγια, ενδιαφερόμαστε να μελετήσουμε τη σχέση μεταξύ των σεισμών και των γεωλογικών ρηγμάτων, έτσι είναι απαραίτητη η ομαδοποίηση των κλάσεων διαφορετικών τύπων.



Εικόνα 6-12 Σχέση Κλάσης και Υποκλάσεων

Οι οντολογίες αποθηκεύονται σε εξωτερικά αρχεία και είναι γραμμένες σε OWL (Horrocks & Patel-Schneider, 2003).

Χρησιμοποιώντας το PatternMiner και οντολογίες του πεδίου ενδιαφέροντος καλύπτεται όλη η διαδικασία KDD. Τα εξαγόμενα πρότυπα μπορούν να αξιολογηθούν πριν την αποθήκευσή τους στη βάση προτύπων (ή σε οποιαδήποτε στιγμή μετά), ενώ και οι επερωτήσεις που υποβάλλονται από το χρήστη μπορούν να αξιολογηθούν πριν εκτελεστούν στη βάση προτύπων. Οι επερωτήσεις που περιέχουν έννοιες που δε σχετίζονται στην οντολογία, απορρίπτονται και ο χρήστης ενημερώνεται, μειώνοντας έτσι το χρόνο επεξεργασίας και το φόρτο του συστήματος.

6.4 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάσαμε το PatternMiner, ένα πρωτόλειο PBMS ανοικτού κώδικα, που ολοκληρώνει τις διαδικασίες εξόρυξης γνώσης, αποθήκευσης και διαχείρισης των εξαγόμενων προτύπων, παρέχοντας στο χρήστη και προχωρημένες δυνατότητες όπως η σύγκριση και η παρακολούθηση της εξέλιξης των προτύπων στο χρόνο. Περιγράφηκε η αρχιτεκτονική του συστήματος καθώς και τα διάφορα μέρη του ενώ παρουσιάστηκε και μία επίδειξη των βασικών του λειτουργιών παρέχοντας εικόνες από τις οθόνες του συστήματος.

Επιπλέον παρουσιάσαμε μία επέκτασή του για να ενσωματωθεί η γνώση του πεδίου εφαρμογής (domain knowledge) για τη φάση της αξιολόγησης των

προτύπων από τον ειδικό (domain expert), με τη χρήση οντολογιών. Περιγράφηκαν τα προβλήματα που παρουσιάζονται και ο τρόπος για να επιτευχθεί αυτή η ολοκλήρωση στο σύστημα PBMS, το οποίο θα αποτελεί ένα ισχυρό εργαλείο για το χρήστη, για τη διαχείριση και αξιολόγηση των προτύπων.

Όπως φάνηκε από τα σενάρια εφαρμογών που παρουσιάστηκαν, το PatternMiner μπορεί να παρέχει ένα ολοκληρωμένο περιβάλλον για την εξαγωγή, αποθήκευση και σύγκριση προτύπων οποιουδήποτε τύπου, εξοικονομώντας πολύτιμο χρόνο από τους ειδικούς – τελικούς χρήστες.

7 Συμπεράσματα

Στο κεφάλαιο αυτό συνοψίζουμε τη συνεισφορά της διατριβής και αναφέρουμε θέματα μελλοντικής έρευνας.

7.1 Συνεισφορά της Διατριβής

Λόγω του μεγάλου όγκου δεδομένων που συλλέγεται στις μέρες μας και αποθηκεύεται σε βάσεις δεδομένων διαφόρων πεδίων και εφαρμογών, οι εφαρμογές και τεχνικές εξόρυξης γνώσης από δεδομένα χρησιμοποιούνται πολύ συχνά για την ανακάλυψη κρυμμένης πληροφορίας, ομάδων και συσχετίσεων δεδομένων. Ο αριθμός των προτύπων που εξάγεται από τις βάσεις αυτές αυξάνεται επίσης ιδιαίτερα και σε πολλές περιπτώσεις η διαχείρισή τους δεν είναι απλή διαδικασία. Οι τελικοί χρήστες δεν μπορούν να αντιμετωπίσουν όλους τους διαφορετικούς τύπους προτύπων που παράγονται από μία ποικιλία λογισμικού σε ετερογενείς πηγές δεδομένων.

Αντιμετωπίζοντας αυτή την πρόκληση, μελετάμε τη διαχείριση των προτύπων σε ένα Σύστημα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ - Pattern Base Management System (PBMS)). Ένα ΣΔΒΠ χειρίζεται τα πρότυπα όπως ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) χειρίζεται τα απλά δεδομένα, χρησιμοποιώντας μία βάση προτύπων και μία γλώσσα επερωτήσεων προσανατολισμένη στην αναπαράσταση των προτύπων. Τα πρότυπα αποτελούν μια συμπαγή πλούσια σε σημασιολογία αναπαράσταση των αρχικών δεδομένων και μπορεί να είναι απλά ή σύνθετα (που ορίζονται πάνω σε απλά πρότυπα). Η ποικιλία των υπαρχόντων προτύπων είναι μεγάλη αλλά όλα τα πρότυπα μοιράζονται κοινά χαρακτηριστικά στον τρόπο που ορίζονται. Η ενιαία διαχείριση των προτύπων με προχωρημένες λειτουργίες πάνω στα πρότυπα, όπως η σύγκρισή τους, οδηγεί σε πολλές ενδιαφέρουσες εφαρμογές. Συγκεκριμένα, η σύγκριση συστάδων, στην περίπτωση της διατριβής αυτής, μπορεί να χρησιμοποιηθεί για να κατηγοριοποιήσει ή να ανακτήσει εικόνες στα

πλαίσια ενός συστήματος ανάκτησης εικόνων με βάση το περιεχόμενο (Content-Based Image Retrieval – CBIR).

Ένα επίσης σημαντικό θέμα σχετικό με τα πρότυπα που έχουν εξαχθεί αυτόματα είναι η αξιολόγησή τους, καθώς δεν είναι όλα τα πρότυπα σημαντικά και ενδιαφέροντα για τους χρήστες. Η αξιολόγηση των προτύπων είναι μία ενδιαφέρουσα αλλά και δύσκολη εργασία. Ωστόσο, με τη χρήση ενός ενοποιημένου συστήματος διαχείρισης προτύπων και τη χρήση οντολογιών που περιγράφουν τη γνώση της περιοχής (domain knowledge ontologies) αυτή η εργασία μπορεί να υποστηριχθεί από λογισμικό.

Στην παρούσα διατριβή αντιμετωπίσαμε τα παραπάνω σχετικά με τη διαχείριση των προτύπων προβλήματα. Πιο συγκεκριμένα:

- Μελετήσαμε το πιο κατάλληλο μοντέλο αναπαράστασης για μία βάση προτύπων, βασισμένη στον ορισμό των προτύπων του έργου PANDA. Μέσα από μία ποιοτική αξιολόγηση τριών μοντέλων βάσεων δεδομένων, του σχεσιακού, του αντικειμενο-σχεσιακού και του ημι-δομημένου (XML) μοντέλου, συμπεραίνουμε ότι το μοντέλο XML είναι πιο κατάλληλο για μία βάση προτύπων καθώς, ανάμεσα σε άλλα χαρακτηριστικά, είναι ευέλικτο, επεκτάσιμο και παρέχει αποτελεσματικότητα στις επερωτήσεις.
- Μελετήσαμε το θέμα της σύγκρισης διακριτών συστάδων, ορίζοντας νέες μετρικές ομοιότητας για συστάδες βασισμένες στην πυκνότητα που παράγονται από τον αλγόριθμο EM. Ορίσαμε μία μεθοδολογία για τη σύγκριση διαφόρων τύπων δεδομένων/ αντικειμένων (π.χ. εικόνων) που περιλαμβάνει τέσσερα βήματα: εξαγωγή χαρακτηριστικών από τα δεδομένα, συσταδοποίηση των χαρακτηριστικών αυτών, αναπαράσταση των εξαγόμενων προτύπων και υπολογισμό των ομοιοτήτων μεταξύ των προτύπων. Αξιολογήσαμε τις μετρικές και τη μεθοδολογία αυτή σε εφαρμογές CBIR με πολύ ικανοποιητικά αποτελέσματα.
- Επίσης μελετήθηκε η σύγκριση προτύπων ασαφών συστάδων και συγκεκριμένα αυτά της διαισθητικά ασαφούς συσταδοποίησης. Πιο συγκεκριμένα, παρουσιάσαμε μία πρωτότυπη παραλλαγή του γνωστού αλγορίθμου συσταδοποίησης Fuzzy C-Means (FCM), τον iFCM (intuitionistic FCM) που αντιμετωπίζει την αβεβαιότητα στο χώρο των διανυσμάτων χαρακτηριστικών των δεδομένων λόγω ανακριβών

μετρήσεων και θορύβου. Ορίσαμε επίσης μία πρωτότυπη μετρική ομοιότητας μεταξύ των διαισθητικά ασαφών συνόλων που ενσωματώθηκε κατάλληλα στον νέο αλγόριθμο συσταδοποίησης.

- Παρουσιάσαμε επίσης μία διαισθητικά ασαφή αναπαράσταση του χαρακτηριστικού του χρώματος για έγχρωμες εικόνες ως παράδειγμα για τη διαισθητικά ασαφειοποίηση των δεδομένων. Για την αξιολόγηση της προσέγγισής μας περιγράψαμε μία διαισθητική ασαφειοποίηση έγχρωμων εικόνων και εφαρμόσαμε την προτεινόμενη μεθοδολογία σε αυτά. Τα πειραματικά αποτελέσματα του προτεινόμενου σχήματος δείχνουν ότι μπορεί να είναι ιδιαίτερα αποδοτική και πιο αποτελεσματική από άλλους αλγόριθμους FCM, ειδικά όταν ο αριθμός των συστάδων αυξαίνεται, ανοίγοντας έτσι νέες προοπτικές για διάφορες εφαρμογές.
- Παρουσιάσαμε το PatternMiner, ένα πρωτόλειο PBMS ανοικτού κώδικα που παρέχει ένα ολοκληρωμένο περιβάλλον για τη διαχείριση των προτύπων χρησιμοποιώντας μέρη για την εξαγωγή, την αποθήκευση, την ανάκτηση και τη σύγκριση των προτύπων. Για την εξαγωγή των προτύπων επιλέχθηκε το σύστημα εξόρυξης γνώσης WEKA. Ειδικά εμπλουτισμένα PMML έγγραφα χρησιμοποιούνται για την αποθήκευση των προτύπων σε μία XMLDB βάση προτύπων και το επεκτεταμένο πλαίσιο σύγκρισης προτύπων PANDA χρησιμοποιείται για την υποστήριξη των λειτουργιών σύγκρισης.
- Μελετήσαμε τη χρήση οντολογιών για την υποστήριξη του βήματος της αξιολόγησης των προτύπων στη διαδικασία ανακάλυψης γνώσης. Οι οντολογίες αναπαριστούν τη γνώση του πεδίου εφαρμογής και με τον τρόπο αυτό μπορεί να χρησιμοποιηθούν για την αξιολόγηση των εξαγόμενων προτύπων. Προτείναμε μία μεθοδολογία που υποστηρίζει αυτή την εργασία και παρουσιάσαμε μία προκαταρκτική μελέτη περίπτωσης ενώ αναλύσαμε τον τρόπο που το PatternMiner μπορεί να επεκταθεί για να υποστηρίζει την διαδικασία αξιολόγησης προτύπων με χρήση οντολογιών.

7.2 Μελλοντική έρευνα

Μέσα από τη μελέτη της διαχείρισης, σύγκρισης και αξιολόγησης των προτύπων, προκύπτουν διάφορες προκλήσεις και προβληματισμοί για περαιτέρω έρευνα. Πιο συγκεκριμένα:

- Μελλοντικές προοπτικές της έρευνας που παρουσιάστηκε στο κεφάλαιο 3, σε συνδυασμό με το σχήμα αξιολόγησης περιλαμβάνουν την ολοκλήρωση του προτεινόμενου σχήματος με τεχνικές εξαγωγής πληροφορίας από οντολογίες και εξόρυξη γνώσης για την ανάκτηση ιατρικών εικόνων από ετερογενής πηγές. Με την αποθήκευση των πλούσιων σημασιολογικά προτύπων μαζί με τα χαρακτηριστικά χαμηλού επιπέδου με ένα ενιαίο τρόπο σύμφωνα με το πλαίσιο PANDA θα είναι εφικτή η επέκταση των μεθοδολογιών CBIR με τεχνικές αναπαράστασης γνώσης για σημασιολογική επεξεργασία και ανάλυση.
- Μελλοντικές προοπτικές της έρευνας που παρουσιάστηκε στο κεφάλαιο 4 περιλαμβάνουν τη συστηματική αξιολόγηση του προτεινόμενου σχήματος με άλλα σχήματα συσταδοποίησης για τη συσταδοποίηση διαφόρων ειδών συνόλων δεδομένων αφού γίνει η αναπαράστασή τους με όρους της θεωρίας των ασαφών συνόλων. Μέχρι τώρα δεν υπάρχουν σύνολα διαισθητικά ασαφών δεδομένων ευρέως διαθέσιμα για τον έλεγχο των αντίστοιχων αλγορίθμων. Ένα ιδιαίτερο θέμα είναι η βελτίωση της προτεινόμενης μεθοδολογίας συσταδοποίησης για να λαμβάνει υπόψη όχι μόνο τις τιμές συμμετοχής, αλλά και αυτές της μη-συμμετοχής ενός διανύσματος δεδομένων σε μία συστάδα.
- Σχετικά με το πρωτόλειο σύστημα PatternMiner, νέα μέρη (components) μπορεί να προστεθούν, όπως ένα για την οπτικοποίηση των αποτελεσμάτων, και να προστεθούν κι άλλοι τύποι προτύπων για το κομμάτι της παρακολούθησης της εξέλιξης των προτύπων στο χρόνο.
- Η ενσωμάτωση οντολογιών στη διαδικασία εξόρυξης δεν είναι απλή διαδικασία και πολλά σχετικά θέματα πρέπει να αντιμετωπιστούν. Το βασικότερο ζήτημα είναι η ανομοιομορφία των διαφόρων οντολογιών που κατασκευάζουν τόσο οι επιστήμονες όσο και οι εταιρείες με βάση τις δικές τους ξεχωριστές ανάγκες αντί να δημιουργούνται οντολογίες παγκοσμίως αποδεκτές. Υπάρχει ένας μεγάλος αριθμός γλωσσών για οντολογίες, οι περισσότερες από τις οποίες έχουν σχεδιαστεί για το

σημασιολογικό ιστό όπως οι RDF (Beckett, 2004), SHOE (Luke & Heflin, 2000), DAML, DAML+OIL (Harmelen et al., 2001), OWL (McGuinness, & Harmelen, 2005). Νέες οντολογίες δημιουργούνται για διάφορα πεδία και εφαρμογές χωρίς κεντρική καθοδήγηση και κοινή συμφωνία. Αυτό γίνεται ακόμα πιο σύνθετο καθώς οι τελευταίες μελέτες δείχνουν σημασιολογικές και συντακτικές ασυμβατότητες μεταξύ αυτών των γλωσσών, ειδικά μεταξύ των DAML+OIL και OWL (Horrocks & Patel-Schneider, 2003) (Patel-Schneiderand & Fensel, 2002). Για το λόγο αυτό, η κατασκευή ενός συστήματος που χρησιμοποιεί οντολογίες στη διαδικασία εξόρυξης γνώσης απαιτεί την επιλογή της κατάλληλης γλώσσας οντολογιών.

- Ένα ακόμα θεωρητικό ζήτημα είναι η αξιολόγηση προτύπων διαφόρων τύπων με χρήση οντολογιών. Δεν είναι απλό να οριστούν κανόνες με βάση την οντολογία που να εφαρμόζονται σε όλα τα πρότυπα. Προς το παρόν έχουμε προτείνει κανόνες και φίλτρα για πρότυπα κανόνων συσχέτισης αλλά ανάλογα την εφαρμογή, θα πρέπει να οριστούν φίλτρα για κάθε τύπο προτύπου προκειμένου το ολοκληρωμένο σύστημα να υποστηρίζει την πλειοψηφία των εφαρμογών.

8 Αναφορές

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (pp. 207-216). Washington, DC: ACM.

Ardizzone E., Chella A., Pirrone R., Gambino O., (2004). An image retrieval system for artistic database on cultural heritage. Atti Conferenza Italiana sui Sistemi Intelligenti, (CISI 2004), Perugia, Italia, 2004.

Atanassov, K.T. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, Vol. 20, pp. 87–96.

Atanassov, K.T. (1989). More on intuitionistic fuzzy sets. Fuzzy Sets Systems, Vol. 33, pp. 37–45.

Atanassov, K.T. (1994). New operations defined over the intuitionistic fuzzy sets. Fuzzy Sets and Systems, Vol. 61, pp. 137–142.

Atanassov, K.T. (1994). Operators over interval valued intuitionistic fuzzy sets. Fuzzy Sets Systems, Vol. 64, pp. 159–174.

Atanassov, K.T. (1999). Intuitionistic Fuzzy Sets: Theory and Applications. Studies in Fuzziness and Soft Computing, Vol. 35, Physica-Verlag, Heidelberg.

Bartolini I., Ciaccia P., Ntoutsi I., Patella M., and Theodoridis Y. (2004). A Unified and Flexible Framework for Comparing Simple and Complex Patterns. In Proceedings of 8th Eur. Conf. on Principles and Practice of Knowledge Discovery in Database, PKDD'04, Pizza, Italy. pp. 496-499, 2004.

Beckett, D. (2004). RDF/XML Syntax Specification (Revised), W3C Recommendation, 2004. http://www.w3.org/TR/rdf-syntax-grammar/.

Berman A., and Shapiro L. G. (1997). Efficient Image Retrieval with Multiple Distance Measures. In Proceedings of SPIE Conf. on Storage and Retrieval for Image and Video Databases, pp. 12-21, 1997.

Bezdek, J.C., Ehrlich, R., and Full, W. (1984). FCM: the Fuzzy c-Means clustering algorithm. Computers and Geosciences, Vol. 10, pp. 191-203.

Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, NewYork, 1981.

Bongard, F.S., Sue, D.Y. (2002). Current Critical Care: Diagnosis and Treatment 2nd Ed. McGraw-Hill/Appleton and Lange, 2002.

CINQ (Consortium on Discovering Knowledge with Inductive Queries). (2001). http://www.cinq-project.org.

CWM (Common Warehouse Model) (2001) homepage. http://www.omg.org/cwm.

Cai, W., Feng, D. D., Fulton, R. (2000). Content Based Retrieval of Dynamic PET Functional Images. IEEE Trans. Inf. Tech. Biomed., vol. 4, no. 2, pp. 152-158, 2000.

Carson, C., Belongie, S., Greenspan, H., and Malik, J. (2002). Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no.8, pp. 1026-1038, 2002.

Caruana R., Elhawary, M., Nguyen, N., and Smith, C. (2006). Meta Clustering. In Proceedings of ICDM 2006.

Catania B., & Maddalena A. (2006). Pattern Management: Practice and Challenges. In J. Darmont & O. Boussaid, (Eds). Processing and Managing Complex Data for Decision Support. Idea Group Publishing.

Catania B., Maddalena A., & Mazza M. (2005). PSYCHO: A Prototype System for Pattern Management. In Proceedings of the International Conference on Very Large Data Bases 2005.

Chen C.C, Wactlar H., Wang J., and Kiernan K. (2004). Digital Imagery for Significant Cultural and Historical Materials - An Emerging Research Field Bridging People, Culture, and Technologies. Int. J. Digital Libraries, 2004.

Chen X., Zhou X., Scherl R., & Geller J. (2003). Using an interest ontology for improved support in rule mining. In DaWaK 2003. pp. 320-329.

Chen, S.M. (1995). Measures of similarity between vague sets. Fuzzy Sets Systems, Vol. 74 No. 2, pp. 217–223.

Chen, S.M. (1997). Similarity measures between vague sets and between elements. IEEE Trans. Syst. Man Cybernet, Vol. 27, No. 1, pp. 153–158.

Chumsamrong, W., Thitimajshima, P. and Rangsanseri, Y. (2000). Synthetic Aperture Radar (SAR) Image Segmentation Using a New Modified Fuzzy C-Means Algorithm. Geoscience and Remote Sensing Symposium, IGARSS 2000. IEEE 2000 International, Vol. 2, pp. 624 – 626.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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

Cui Z., Jones D., & O'Brien P. (2002). Semantic B2B Integration: Issues in Ontology-based Approaches. ACM SIGMOD Record archive Vol. 31, Issue 1 (March 2002) SPECIAL ISSUE: Data management issues in electronic commerce table of contents. pp. 43–48

DMG - PMML, http://www.dmg.org/pmml-v3-1.html.

Davis, J. and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In Proceedings of 23rd Int. Conf. on Machine Learning (ICML), Pittsburgh, USA, 2006, pp. 233-240.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, vol. 39, no.1, pp.1–38, 1977.

Dengfeng, L., Chuntian, C. (2002). New similarity measure of intuitionistic fuzzy sets and application to pattern recognitions. Pattern Recognition Letters, Vol. 23, pp. 221–225.

Deselaers T., Keysers D., and Ney H. (2004). FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. LNCS 3491, pp. 688-698, 2004.

Deselaers, T., Keysers, D., Ney, H. (2004). Features for Image Retrieval-A Quantitative Comparison. In Proceedings of 26th DAGM Symp., LNCS, pp. 228-236, 2004.

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2003). Ontology Matching: A Machine Learning Approach. In S. Staab and R. Studer (Eds), Handbook on Ontologies in Information Systems. Springer-Velag.

Dowd, S.B., Wilson, B.G. (1995). Encyclopedia of Radiographic Positioning: Volume 2, Saunders, 1995.

Dunn, J.C. (1973). A Fuzzy Relative of the ISODATA process and its Use in Detecting Compact Well-Separated Cluster. Journal Cybernetics Vol. 3, No. 3, pp. 32-57.

El-Naqa, I., Yang, Y., Galatsanos, N.P., Nishikawa, R.M., Wernick, M.N. (2004). A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans. Med Imaging, vol. 23, no. 10, pp. 1233-1244, Oct. 2004.

FCD/fcd-datamining-2001-05.pdf, 2001.

FHW, Foundation of the Hellenic World, http://www.fhw.gr/index_en.html.

Faloutsos C., Barber R., Flickner M., Hafner J., Niblack W., Petkovic D., and Equitz W., (1994), Efficient and Effective Querying by Image Content, J. Intell. Inf. Systems, vol. 3, pp. 231-262, 1994.

Fan, L., Zhangyan, X. (2001). Similarity measures between vague sets. J. Software Vol. 12, No.6, pp. 922–927 (in Chinese).

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery, an overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, (Eds.), Advances in Knowledge Discovery and Data Mining. (pp. 1–30). Menlo Park, Calif. AAAI/MIT Press.

Fayyad, U., Haussler, D., & Stolorz, P. (1996). Mining scientific data. Communications of the ACM, Vol. 39, Issue 11, 51-57.

Freitas, A.A. (1999). On rule interestingness measures. Knowledge-Based Systems, Vol. 12. number 5-6. pp. 309–315, October 1999. Elsevier.

GAIA, ESA project. http://www.esa.int/science/gaia 2009.

Greenspan, H., Pinhas, A.T. (2007). Medical Image Categorization and Retrieval for PACS Using the GMM-KL Framework. IEEE Trans. Inf. Tech. Biomed., vol.11, no.2, pp.190-202, Mar. 2007.

Grosky W.I. and Mehrotra R. (1990). Indexed-Based Object Recognition in Pictorial Data Management. Computer Vision, Graphics, Image Processing, vol. 52, pp. 416-436, 1990.

Gruber, T. R. (1993). A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220.

Guarino, N., & Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. In N. Mars (Ed.), Towards very large knowledge bases: Knowledge building and knowledge sharing (pp.25-32). Amsterdam, IOS Press.

Hampapur A., Gupta A., Horowitz B., Shu C. F., Fuller C., Bach J., Gorkani M., and Jain R. (1997). Virage video engine. In Proceedings ofSPIE: Storage and Retrieval for Image and Video Databases V, pp. 188-197, San Jose, February 1997.

Harmelen, F.V., Patel-Schneider, P.F., & Horrocks, I. (2001). Reference Description of the DAML+ OIL Ontology Markup Langauge. http://www.daml.org/2001/03/daml+oil-index.html.

Hartigan, J. A. (1975). Clustering algorithms. New York: John Wiley, 1975.

Hong, D.H., Kim, C. (1999). A note on similarity measures between vague sets and between elements. Inform. Science Vol. 115, pp. 83–96.

Hore, P., Hall, L.O. and Goldgof, D.B. (2007). Single Pass Fuzzy C-means. IEEE International Conference on Fuzzy Systems, London, 2007.

Horrocks, I., & Patel-Schneider, P.F. (2003). Three theses of representation in the semantic web. In Proceedings of the Twelfth International Conference on World Wide Web.

Hotho, A., Maedche, A., Staab, S., & Zacharias, V. (2002). On knowledgeable unsupervised text mining. In Proceedings of the DaimlerChrysler Workshop on Text Mining, Ulm, April 26–27 2002. Springer.

Hung, W.-L., Yang, M.-S. (2004). Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance. Pattern Recognition Lett. Vol. 25, pp. 1603–1611.

ISO SQL/MM Part 6. http://www.sql-99.org/SC32/WG4/Progression_Documents/ (2001)

Iakovidis, D.K., Maroulis D.E., Karkanis S.A. (2005). A Comparative Study of Color-Texture Image Features. In Proceedings of IEEE Int. Workshop on Systems, Signal and Image Processing (IWSSIP), Halkida, Greece, 2005, pp. 205-209.

Iakovidis, D.K., Kotsifakos, E.E., Pelekis, N., Karanikas, H., Kopanakis, I., and Theodoridis, Y. (2007). Pattern-Based Retrieval of Cultural Heritage Images. 11th Panhellenic Conference on Informatics (PCI'2007), Patras, Greece, 2007.

Iakovidis, D.K., Pelekis, N., Karanikas, H., Kotsifakos, E.E., Kopanakis, I., and Theodoridis, Y. (2006). A Pattern Similarity Scheme for Medical Image Retrieval, ITAB 2006, Proc of the 7th Annual IEEE Conf on International Technology Applications in Biomedicine, Ioannina, Greece.

Iakovidis, D.K., Pelekis, N., Karanikas, H., Kotsifakos, E.E., Kopanakis, I., and Theodoridis, Y. (2009). A Pattern Similarity Scheme for Medical Image

Retrieval. IEEE Transactions on Information Technology in Biomedicine Journal, Volume: 13 Issue: 4, July 2009.

Iakovidis, D.K., Pelekis, N., Kotsifakos, E.E. and Kopanakis, I. (2008). Intuitionistic Fuzzy Clustering with Applications in Computer Vision. In the Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS'08), LNCS 5259, pp. 764–774, Juan-les-Pins, France, 2008

Iqbal Q., and Aggarwal J. K. (2002). CIRES: A System for Content-based Retrieval in Digital Image Libraries. In Proceedings of Int. Conf. Control, Automation, Robotics and Vision (ICARCV), pp. 205-210, December 2-5, 2002.

Java Data Mining API homepage (2003), http://www.jcp.org/jsr/detail/73.prt.

Jawahar, C.V., Ray, A. K. (1996). Fuzzy statistics of digital images. Pattern Recognition Letters, Vol. 17, pp. 541–546.

Jeng W.-M., and Hsiao J.-H. (2005). An Efficient Content Based Image Retrieval System Using the Mesh-of-Trees Architecture. J. Inf. Science Eng., vol. 21, pp. 797-808, 2005.

Jia L. and Wang J. Z. (2004). Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models. IEEE Trans. on Image Processing., vol.12, no.2, pp., 2004.

Karkanis, S.A., Iakovidis, D.K., Maroulis, D.E., Karras, D.A. and Tzivras, M. (2003). Computer Aided Tumor Detection in Endoscopic Video using Color Wavelet Features. IEEE Trans. Inf. Tech.Biomed., vol. 7, pp. 141-152, 2003.

Kopanakis, I. and Theodoulidis, B. (2003). Visual Data Mining Modelling Techniques for The Visualization of Mining Outcomes. J. Visual Languages and Computing, Special Issue on Visual Data Mining, vol. 14, no.6, pp. 543-589, 2003.

Kotsifakos, E.E., Marketos, G., and Theodoridis, Y. (2007). A framework for integrating ontologies and pattern-bases. In Nigro, H.O., Cisaro, S.G. and

Xodo, D. (eds), Data Mining with Ontologies: Implementations, Findings,and Frameworks. Idea Group Inc., Hershey, 2007.

Kotsifakos, E.E., Ntoutsi, I. and Theodoridis, Y. (2005). Database Support for Data Mining Patterns. 10th Panhellenic Conference on Informatics (PCI'2005), Volos, Greece, 2005. Advances in Informatics - Springer-Verlag LNCS #3746,2005

Kotsifakos, E.E., Ntoutsi, I., Vrahoritis, Y., and Theodoridis Y. (2008). Monitoring Patterns through an Integrated Management and Mining Tool. In Proceedings of Machine Learning and Knowledge Discovery in Databases, European Conference (ECML/PKDD'08), Antwerp, Belgium, 2008.

Kotsifakos, E.E., Ntoutsi, I., Vrahoritis, Y., and Theodoridis Y. (2008) PATTERN-MINER: Integrated Management and Mining over Data Mining Models. In Proceedings of 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'08), Las Vegas, USA, 2008.

Kwak, D.-M., Kim, B.-S., Yoon, O.-K., Park, C.-H., Won, J.-U., Park, K.-H. (2002). Content-Based Ultrasound Image Retrieval Using a Coarse to Fine Approach. Annals of the New York Academy of Sciences, vol. 980, pp.212-224, 2002.

Laaksonen J., Koskela M., Laakso S., and Oja E. (2000). PicSOM - Content-Based Image Retrieval with Self-Organizing Maps.Pattern Recognition Letters, vol. 21, pp. 1199-1207, 2000.

Lehmann, T.M., Guld, M.O., Thies, C., Plodowski, B., Keysers, D., Ott, B., Schubert, H. (2004). IRMA - Content-based image retrieval in medical applications. In Proceedings of 14th World Congress on Medical Informatics (Medinfo), IOS Press, Amsterdam, vol. 2, pp. 842-848, 2004.

Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B. (2003). The IRMA Code for Unique Classification of Medical Images. In Proc SPIE 2003, vol. 5033, pp. 109–17, 2003.

Li, Y., Olson D.L., Qin Z. (2007). Similarity measures between vague sets: A comparative analysis. Pattern Recognition Letters Vol. 28, pp. 278-285

Li, Y., Zhongxian, C., Degin,Y. (2002). Similarity measures between vague sets and vague entropy. J.Computer Sci. Vol. 29, No.12, pp. 129–132.

Lin, C.-Y., Yin, J.-X., Gao, X., Chen, J.-Y. and Qin, P. (2006). A Semantic Modelling Approach for Medical Image Semantic Retrieval Using Hybrid Bayesian Networks. In Proceedings of 6th Int. Conference on Intelligent Systems Design and Applications (ISDA), pp. 482-487, 2006.

Linacre, J.M. (1996). Overlapping Normal Distributions. Rasch Measurement Transactions, vol. 10, no. 1 pp. 487-488.

Littau, D. (2003). Using Low-Memory Approximations to Cluster Very Large Data Sets. In Proceedings of 3rd SIAM Int. Conf. on Data Mining (SDM), 2003.

Liu, B., Hsu W., Chen S., & Ma, Y. (2000). Analyzing the subjective interestingness of association rules. IEEE Intelligent Systems, 15(5):47-55.

Luke, S., & Heflin, J. (2000). SHOE 1.01 Proposed Specification, SHOE Project, http://www.cs.umd.edu/projects/plus/SHOE/spec.html.

Ma W.Y., and Manjunath B.S. (1999). Netra: A Toolbox for Navigating Large Image Databases,.Multimedia System, vol. 7, pp. 184-198, 1999.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.

Maedche, A., Motik, B., Stojanovic, L., Studer, R., & Volz, R. (2003). Ontologies for enterprise knowledge management. IEEE Intelligent Systems, 18(2):26–33, March/April 2003.

Maenpaa, T., and Pietikainen, M. (2004). Classification with color and texture: Jointly or separately?, Pattern Recognition, Vol. 37, No. 8, pp. 1629–1640.

Mallat, S. (1999). A Wavelet Tour of Signal Processing. Acad. Press, 2nd ed., 1999.

Maroulis, D.E., Savelonas, M., Iakovidis, D.K., Karkanis, S.A., Dimitropoulos, N. (2007). Variable Background Active Contour Model for Computer-Aided Delineation of Nodules in Thyroid Ultrasound Images. IEEE Trans. Inf. Tech. Biomed., vol. 11, no. 5, pp. 537-543, 2007.

McGuinness, D.L., & Harmelen, F.V. (2005). OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features/ (current Feb. 2005).

Mitchell, H.B. (2003). On the Dengfeng–Chuntian similarity measure and its application to pattern recognition. Pattern Recognition Lett. Vol. 24, pp. 3101–3104.

Muller, H., Michoux, N., Bandon, D., Geissbuhler, A. (2004). A Review of Content-Based Image Retrieval Systems in Medicine - Clinical Benefits and Future Directions. Int.J.of Med.Informatics, vol.73, pp.1-23, 2004.

Niles, I., & Pease, A. (2001). Toward a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).

Niles, Ian & Terry, Allan. (2004). The MILO: A general-purpose, mid-level ontology. In 2004 International Conference on Information and Knowledge Engineering (IKE'04).

Ntoutsi, I. (2008). Similarity Issues in Data Mining – Methodologies and Techniques. PhD thesis, University of Piraeus, June 2008.

Ntoutsi, I., Pelekis, N., and Theodoridis, Y. (2007). Pattern Comparison in Data Mining: a survey. In D.Taniar, editor, Research and Trends in Data Mining Technologies and Applications (Advances in Data Warehousing and Mining), pages 86 – 120. Idea Group Publishing, 2007.

Ohta, Y., Kanade, T., Sakai, T. (1980). Color information for region segmentation. ComputerVision, Graphics, and Image Processing, Vol. 13, pp. 222-241.

Ojala T., Pietikainen M., Harwood D. (1996). A Comparative Study of Texture Measures with Classification based on Feature Distributions. Pattern Recognition, vol. 29, 1996, pp. 51-59.

Omelayenko B. (2000). Integration of Product Ontologies for B2B Marketplaces: A Preview. In ACM SIGecom Exchanges. Vol. 2, issue 1, pp. 19-25.

Oracle Corp. Berkeley DB XML. Available at http://www.oracle.com/database/berkeley-db/xml/index.html

PANDA (Patterns for Next-generation Database Systems). (2001). project homepage. http://dke.cti.gr/panda.

PBMS (2006) homepage. http://www.pbms.org.

PMML (Predictive Model Markup Language) (2009). http://www.dmg.org/v4-0/GeneralStructure.html.

PQL, Information Discovery Data Mining Suite. http://www.patternwarehouse.com/ dmsuite.htm.

Padmanabhan B. & Tuzhilin, A. (1998). A belief-driven method for discovering unexpected patterns. In Proceedings of the International Conference on Knowledge Discovery and Data Mining, pages 94–100, August 1998.

Patel-Schneiderand, P.F., & Fensel, D. (2002). Layering the semantic web: Problems and Directions. In Proceedings of the 1st International Semantic Web Conference. LNCS 2342, Springer.

Pelekis, N., Iakovidis, D.K., Kotsifakos, E.E, Karanikas, H., and Kopanakis, I. (2007). Intuitionistic Fuzzy Clustering to Information Retrieval from Cultural Databases. 22nd European Conference on Operational Research, EURO XXII, Prague, 2007.

Pelekis, N., Iakovidis, D.K., Kotsifakos, E.E. and Kopanakis, I. (2008). Fuzzy Clustering of Intuitionistic Fuzzy Data. International Journal of Business Intelligence and Data Mining, 3(1), 45-65, 2008

Petrakis E. G.M., Faloutsos C. (1997). Similarity Searching in Medical Image Databases. IEEE Trans. Knowl. Data Eng., vol. 9, no. 3, pp. 435-447, 1997.

Piatetsky-Shapiro G. & Matheus C.J. (1994). The interestingness of deviations. In Proceedings of KDD-94: AAAI-94 Knowledge Discovery in Databases Workshop, pages 25–36. AAAI Press, July 1994.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro & W.J. Frawley (Eds). Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press, Cambridge, MA.

Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases: 10 years after. SIGKDD Explorations, Vol. 1, no 2. pp. 59–61, January 2000.

Pinto, H.S., & Martins, J.P. (2001) A methodology for ontology integration. 1st International conference on Knowledge Captur. Pp 131-138.

Pohle, C. & Spiliopoulou, M. (2002). Building and exploiting ad hoc concept hierarchies for web log analysis. In Y. Kambayashi, W. Winiwarter, & M. Arikawa (Eds). Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2002, volume 2454 of Lecture Notes in Computer Science, pages 83–93, Aix en Provence, France, September 4–6 2002. Springer-Verlag.

Pohle, C. (2003). Integrating and updating domain knowledge with data mining. VLDB PhD Workshop.

R-project, (2009). The R-tool. http://www.r-project.org/ 2009

Rahman, Md. M., Bhattacharya, P. and Desai, B.C. (2007). A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques With Relevance Feedback. IEEE Trans. Inf. Tech. Biomed., vol. 11, no. 1, pp. 58-67, Jan. 2007.

Rizzi, S., Bertino, E., Catania, B., Golfarelli, M., Halkidi, M., Terrovitis, M., Vassiliadis, P., Vazirgiannis, M., & Vrahnos, E. (2003). Towards a logical model for patterns. Proceedings of ER'03 conference, Chicago, IL, USA, 2003..

Ruspini, E. H. (1969). A New Approach to Clustering. Information Control Vol. 15, No. 1, pp. 22-32.

SUMO, Suggested Upper Merged Ontology (2009) http://www.ontologyportal.org/.

Schnorrenberg, F., Pattichis, C. S., Schizas, C. N., Kyriacou, K. (2000). Content-Based Retrieval of Breast Cancer Biopsy Slides. Technology and Health Care, vol. 8, pp. 291-297, 2000.

Seismo-Surfer, A WebGIS application for integrating, visualizing and analyzing seismic data. http://www.seismo.gr.

Setia, L., Teynor, A., Halawani, A. and Burkhardt, H. (2006). Image Classification using Cluster-Cooccurrence Matrices of Local Relational Features. In Proceedings of 8th ACM Int. Workshop on Mult. Inf. Retrieval, 2006.

Shyu, C. R., Brodley, C. E., Kak, A. C., Kosaka, A., Aisen, A. M., and Broderick, L. S. (1999). ASSERT: A Physician-in-the-loop Content-Based Image Retrieval System for HRCT Image Databases. Computer Vision and Image Understanding, pp. 111-131, 1999.

Silberschatz A. & Tuzhilin, A. (1996). What makes patterns interesting. In knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering, 8(6):970–974, December 1996.

Smith J.R., and Chang S.F. (1996). Visualseek: A Fully Automated Content-Based Image Query System. In Proceedings of ACM Int'l Multimedia Conf., pp. 87-98, 1996.

Spiliopoulou M., Ntoutsi, I., Theodoridis, Y., and Schult, R. (2006). MONIC: Modelling and monitoring cluster transitions. KDD, 2006.

Spiliopoulou M., and Roddick J. F. (2000). Higher Order Mining: Modellingand Mining the Results of Knowledge Discovery. In Proceedings of Int. Conf. Data Mining, vol. 2, pp. 309-320, 2000.

Stehling R.O., Falcao A.X., and Nascimento M.A. (2001). An Adaptive and Efficient Clustering-Based Approach for Content-Based Image Retrieval in Image Databases. In Proceedings of Int. Symp. Database Eng. & App. (IDEAS 01), pp. 56-365, 2001.

Stehling R.O., Nascimento M.A., and Falcao A.X. (2002). A Compact and Efficient Image Retrieval Approach based on Border/Interior Pixel Classification. In Proceedings of 11th Int. Conf. Inf. Knowledge Man. (CIKM'02), ACM Press, NY, pp. 102-109.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. Journal of the Royal Statistical Society, B-36, 111–147 1974.

Stonebraker, M. (1997). Object-Relational DBMS: The Next Wave. Informix Software, CA Feb, 1997.

Stonebraker, M., Brown, P., and Moore, D. (1999). Object-Relational DBMSs: Tracking the Next Great Wave 2e. Morgan-Kauffman Publishers. San Francisco, 1999.

Swain, M.J. and Ballard, D.H. (1991). Color Indexing. Int. J. Computer Vision, Vol. 7, No. 1, pp. 11–32, Nov. 1991.

Terrovitis , P., Skiadopoulos, S., Bertino, E., Catania, B., Maddalena, A., and Rizzi, S. (2007). Modeling and language support for the management of pattern-bases, Data Knowl. Eng. 62, 2 (Aug. 2007).

Terrovitis M., Vassiliadis P., Skiadopoulos S., Bertino E., Catania B., Maddalena A. (2004). Modelling and Language Support for the Management of Pattern-Bases. In Proceedings of SSDBM Conference, Santorini, Greece, 2004.

Theodoridis, S. and Koutroumbas K. (2006). Pattern Recognition, Elsevier.

Theodoridis, Y., Marketos, G., & Kalogeras, I.S. (2004). Collecting and Mining Seismic Data in Greek Territory - The Seismo-Surfer Tool. In Proceedings of 7th Panhellenic Geographical Conference of the Hellenic Geographical Association (HGA'04), Mytilene, Lesvos, Greece.

Theodoridis, Y., Vazirgiannis, M., Vassiliadis, P., Catania, B., and Rizzi, S. (2003) A Manifesto for Pattern Bases, PANDA TR-2003-03. Available at http://www.pbms.org/papers/TR-2003-03.pdf

Thitimajshima, P. (2000). A New Modified Fuzzy C-Means Algorithm for Multispectral Satellite Images Segmentation. Geoscience and Remote

Sensing Symposium, IGARSS 2000. IEEE 2000 International, Vol. 4, pp. 1684 – 1686.

Valle E., Cord M., and Philipp-Foliguet S. (2006). Content-Based Retrieval Of Images For Cultural Institutions Using Local Descriptors, Int. Conf. on Geometric Modeling and Imaging, London, 2006.

Vazirgiannis M., Halkidi M., Tsatsaronis G., Vrachnos E. (2003). A Survey on Pattern Application Domains and Pattern Management Approaches. PANDA Technical Report TR-2003-01, 2003. Available at http://dke.cti.gr/panda.

Veltcamp R., and Tanase M. (2000). Content--Based Image Retrieval Systems: A Survey. Tech. Report UU-CS-2000-34, Dept. of Comp. Sci., Utrecht Univ., 2000.

Vlachos, I.K. and Sergiadis G.D. (2005). Towards Intuitionistic Fuzzy Image Processing. Proceedings of the International Conference on Computational Intelligence for Modelling. Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), Vol. 1, pp. 2-7.

Vlachos, I.K. and Sergiadis, G.D. (2006). Intuitionistic fuzzy information – Applications to pattern recognition. Pattern Recognition Letters, Vol. 28, pp. 197–206.

Wang J.Z., Li J., and Wiederhold G. (2001). SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries. IEEE Trans. Patt. Anal. Machine Intell., vol. 23, no. 9, pp. 947-963, Sept. 2001.

Wang, J.Z., Wiederhold, G., Firschein, O. and Wei, S.X. (1998). Content-Based Image Indexing and Searching Using Daubechies' Wavelets. Int. Journal of Digital Libraries, vol. 1, no. 4, pp. 311–328, 1998.

Wang, J.Z. (2001). Wavelets and imaging informatics: A review of the literature. Jour. of Biomedical Informatics, vol. 34, pp. 129-141, 2001.

Weber, R., Schek, H.-J. and Blott, S. (1998). A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional

Spaces. In Proceedings of Very Large Data Bases Conference (VLDB), pp. 194-205 1998.

Witten I. H., and Frank E. (2005). Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Wu J.K., and Narasimhalu A.D. (1994). Identifying Faces Using Multiple Retrievals. IEEE Multimedia, vol. 1, no. 2, pp. 20-38, 1994.

XQuery 1.0 An XML Query Language. *XQuery 1.0 An XML Query Language*. W3C Working Draft 12 November 2003. http://www.w3.org/TR/2003/WD-xquery-20031112

Yaoa, J., Antani, S., Longb, R., Thoma, G. and Zhanga, Z. (2006). Automatic Medical Image Annotation and Retrieval Using SECC. In Proceedings of 19th Int. Symp. Computer-Based Medical Systems (CBMS), Utah, June 2006.

Yixin Chen, Wang, J.Z., Krovetz, R. (2005). CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning. IEEE Trans. on Image Processing, vol. 14, no. 8, pp. 1187- 1201, Aug. 2005.

Yong, Y., Chongxun, Z., Pan, L. (2004). A Novel Fuzzy C-Means Clustering Algorithm for Image Thresholding. Measurement Science Review, Vol. 4, Sec. 1.

Zadeh, L.A. (1965). Fuzzy sets. Information Control Vol. 8, pp. 338–356.

Zhang R., Zhang Z.M. (2002). A Clustering Based Approach to Efficient Image Retrieval. In Proceedings of 14th IEEE Int. Conf. Tools Artif. Intel. (ICTAI'02), p. 339, 2002.

Zheng, L., Wetzel, A.W., Gilbertson, J., Becich, M.J. (2003). Design and Analysis of a Content-Based Pathology Image Retrieval System IEEE Trans.Inf, Tech. Biomed., vol.7, no.4, pp.249-255, 2003.

Zhizhen, L., Pengfei, S. (2003). Similarity measures on intuitionistic fuzzy sets. Pattern Recognition Lett. Vol. 24, pp. 2687–2693.