ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Προχωρημένες Τεχνικές Εξόρυξης Δεδομένων σε Νοσοκομειακές Βάσεις Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Θεόδωρος Μ. Παλτόγλου

Επιβλέπων : Δημήτριος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Προχωρημένες Τεχνικές Εξόρυξης Δεδομένων σε Νοσοκομειακές Βάσεις Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Θεόδωρος Μ. Παλτόγλου

Επιβλέπων : Δημήτριος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Κουτσούρης Καθηγητής Ε.Μ.Π.	Γεώργιος Ματσόπουλος Καθηγητής Ε.Μ.Π.	Δημήτριος Φωτιάδης Καθηγητής Ε.Μ.Π

Αθήνα, Σεπτέμβριος 2013

(Υπογραφή)

.....

Θεόδωρος Μ. Παλτόγλου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στην παρούσα πτυχιακή ασχολούμαστε με την εξαγωγή πληροφορίας από νοσοκομειακές βάσεις δεδομένων στις οποίες καταγράφονται αστοχίες ιατροτεχνολογικών προϊόντων δυνάμενες να προκαλέσουν δυσμενή περιστατικά ασθενών, εξαρτώμενων από τα προϊόντα. Στόχος είναι η εκπαίδευση ταξινομητών με τη βοήθεια των διαθέσιμων δεδομένων ώστε να γίνεται αυτόματη κατηγοριοποίηση των νέων αστοχιών ως προς το είδος της δυσλειτουργίας του προϊόντος στην οποία οφείλονται. Η δυσλειτουργία μπορεί να είναι του υλικού του προϊόντος (hardware) ή του λογισμικού (software). Στα πρώτα τρία κεφάλαια γίνεται αναφορά σε τεχνικές εξόρυξης δεδομένων, σε εξόρυξη δεδομένων από κείμενα καθώς και στο πρόγραμμα εξόρυξης πληροφορίας Weka με το οποίο θα γίνει η εκπαίδευση των ταξινομητών. Στο τελευταίο κεφάλαιο χρησιμοποιούμε διαφορετικούς αλγορίθμους για την εκπαίδευση ταξινομητών και συγκρίνουμε τις αποδόσεις των εξαγομένων μοντέλων.

Λέξεις - Κλειδιά

Εξόρυξη Δεδομένων, Δέντρα Ταξινόμησης, Στατιστική Μοντελοποίηση, Μηχανές Διανυσμάτων Υποστήριξης, Εξόρυξη Δεδομένων από Κείμενα, Αστοχίες Ιατροτεχνολογικών Προϊόντων.

Abstract

In the current project we use data mining techniques in medical databases which record different medical product errors capable of causing adverse patient events. The goal is to train classifiers using the available data for the automatic classification of newly recorded errors to the malfunction cause that are connected with. The malfunction is due to either product's hardware or software. The first three chapters describe data mining techniques, text mining techniques and contain a tutorial of the program WEKA with which the classifiers are going to be trained. In the last chapter we use different algorithms for the classifier training and compare the output models.

KeyWords

Data Mining, Classification Trees, Statistical Modeling, Support Vector Machines, Text Mining, Medical Product Errors.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 4: ΕΠΑΓΡΥΠΝΗΣΗ ΙΑΤΡΟΤΕΧΝΟΛΟΓΙΚΟΥ ΕΞΟΠΛΙΣΜΟΥ

ΚΕΦΑΛΑΙΟ 1: ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

1.1 Εισαγωγή

Ο άνθρωπος όταν δεν μπορεί να ερμηνεύσει αιτιοκρατικά ένα φαινόμενο, αξιοποιεί την εμπειρία του πάνω σε αυτό. Συγκεντρώνει δηλαδή όλες τις περιπτώσεις στις οποίες έχει παρατηρήσει το φαινόμενο και τις χρησιμοποιεί για να κάνει γενικεύσεις και να βγάλει συμπεράσματα σχετικά με αυτό. Προφανώς, ανάλογα με το βαθμό αντιπροσωπευτικότητας των συλλεχθέντων δεδομένων, ανάλογα δηλαδή με το αν το φαινόμενο που μελετά εμφανίζεται και σε περιπτώσεις που δεν έχει λάβει υπ'όψιν του ή δεν έχει συναντήσει ακόμα, αυτές οι γενικεύσεις και τα συμπεράσματα θα εμπεριέχουν σφάλμα.

Στην παρούσα πτυχιακή εργασία δε θα μας απασχολήσει το κατά πόσο αντιπροσωπευτικό είναι ένα οποιοδήποτε σύνολο δεδομένων, άρα υποθέτουμε ότι μπορούμε να βγάλουμε ασφαλή συμπεράσματα από αυτό. Όταν λοιπόν συγκεντρώσουμε τα δεδομένα που μας ενδιαφέρουν, προσπαθούμε συνήθως να τα αναπαραστήσουμε αφαιρετικά. Οι νέες δομές που προσπαθούμε να δημιουργήσουμε ονομάζονται **πρότυπα (patterns)**. Για παράδειγμα, έστω ότι έχουμε ως δεδομένα δύο θετικές και τρεις αρνητικές περιπτώσεις πελατών μιας τράπεζας που δανειοδοτήθηκαν και θέλουμε να συμπεράνουμε πότε κάποιος είναι καλός υποψήφιος για δανειοδότηση.

Πελάτης	Τρέχουσες Οφειλές	Εισόδημα	Παντρεμένος(η)	Χαρακτηρισμός
1	Υψηλές	Υψηλό	Ναι	Καλός(p)
2	Χαμηλές	Υψηλό	Όχι	Κακός(n)
3	Χαμηλές	Υψηλό	Ναι	Καλός(p)
4	Υψηλές	Χαμηλό	Ναι	Κακός(n)
5	Χαμηλές	Χαμηλό	Ναι	Κακός(n)

Οι δυνατές τιμές είναι:

- Τρέχουσες Οφειλές: Υψηλές, Χαμηλές
- Εισόδημα: Υψηλό, Χαμηλό
- Παντρεμένος(η): Ναι, Όχι
- Χαρακτηρισμός: Καλός (θετικό παράδειγμα), Κακός (αρνητικό παράδειγμα)

Με επισκόπηση των δεδομένων παρατηρούμε ότι ο καλός υποψήφιος για δανειοδότηση πρέπει να έχει σχετικά υψηλό εισόδημα και να είναι παντρεμένος ανεξάρτητα από τις τρέχουσες οφειλές του. Η παρατήρησή μας ικανοποιεί τις πέντε περιπτώσεις και αποτελεί ένα πρότυπο που εξάγεται από αυτές. Τα πρότυπα ή μοντέλα μπορούν να ανιχνευθούν και από ένα υπολογιστικό σύστημα με την

εκτέλεση κατάλληλων αλγορίθμων και είσοδο το σύνολο δεδομένων. Αυτή η διαδικασία ονομάζεται **μηχανική μάθηση (machine learning).** Από εδώ και πέρα θα χρησιμοποιείται συχνά ο όρος **σύνολο εκπαίδευσης** για τα δεδομένα εισόδου.

Έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα παρακάτω είδη:

- Μάθηση με επίβλεψη (supervised learning) ή μάθηση με παραδείγματα (learning from examples),
- Μάθηση χωρίς επίβλεψη (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation)

Στη μάθηση με επίβλεψη το σύστημα καλείται να "μάθει" μία έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί την περιγραφή ενός μοντέλου. Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Μάθηση με επίβλεψη

Στη μάθηση με επίβλεψη το σύστημα πρέπει να μάθει μια συνάρτηση που ονομάζεται **συνάρτηση στόχος (target function)** και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής (**κλάση**), που ονομάζεται **εξαρτημένη μεταβλητή** ή **μεταβλητή κλάσης**, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται **ανεξάρτητες μεταβλητές** ή **χαρακτηριστικά**.

Διακρίνονται δύο είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παρεμβολής.

- Η **ταξινόμηση (classification)** αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών) (π.χ ομάδα αίματος).
- Η **παρεμβολή (regression)** αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών (π.χ. πρόβλεψη ισοτιμίας νομισμάτων ή τιμής μετοχής).

1.2 Μάθηση Εννοιών (Concept Learning)

Η έννοια (concept) είναι ένα υποσύνολο αντικειμένων που ορίζονται σε σχέση με ένα μεγαλύτερο σύνολο. Η έννοια "πουλί" ορίζεται ως "το υποσύνολο των ζώων που έχουν φτερά". Εναλλακτικά, η έννοια είναι μια συνάρτηση που επιστρέφει μια λογική τιμή: *αληθής* για τα αντικείμενα ενός συνόλου που ανήκουν σε αυτή και *ψευδής* για αυτά που δεν ανήκουν.

Η δημιουργία ενός μοντέλου από ένα σύνολο δεδομένων καλείται και ως **επαγωγική μάθηση**. Η μάθηση εννοιών είναι ένα τυπικό παράδειγμα επαγωγικής μάθησης κατά την οποία το σύστημα τροφοδοτείται με παραδείγματα που ανήκουν (θετικά) ή όχι (αρνητικά) στη συγκεκριμένη έννοια. Ακολούθως πρέπει να παραχθεί κάποια γενικευμένη περιγραφή της έννοιας, δηλαδή να δημιουργηθεί ένα μοντέλο, ώστε να αποφασιστεί στη συνέχεια αν κάποια άγνωστη περίπτωση ανήκει

στην έννοια. Ο πιο γνωστός αλγόριθμος μάθησης εννοιών είναι ο αλγόριθμος απαλοιφής υποψηφίων.

Ο Αλγόριθμος Απαλοιφής Υποψηφίων (Candidate Elimination Algorithm):

Περιορίζει το χώρο αναζήτησης επιτελώντας γενικεύσεις και εξειδικεύσεις σε κάποιες αρχικές υποθέσεις (έννοιες) με βάση τα δεδομένα εκπαίδευσης. Διατηρεί δύο σύνολα, G και S που από κοινού περιγράφουν όλο το χώρο αναζήτησης και ορίζονται ως εξής:

- G : το σύνολο των πιο γενικών (maximally general) υποψηφίων υποθέσεων (δηλαδή εννοιών).
- S : το σύνολο των πιο εξειδικευμένων (maximally specific) υποψηφίων υποθέσεων.

Ο αλγόριθμος φαίνεται παρακάτω:

Αρχικοποίησε:
Το G στο σύνολο όλων των υποθέσεων.
Το S στο κενό σύνολο.
Για κάθε δεδομένο εκπαίδευσης x:
Αν το x είναι θετικό:
i) Διάγραψε τα μέλη του G που δεν ικανοποιούν το x.
ii) Για κάθε υπόθεση s∈S που δεν ικανοποιεί το x:
α) Διάγραψε την s από το S.
β) Πρόσθεσε στο S όλες τις ελάχιστες γενικεύσεις h της s,
έτσι ώστε κάθε υπόθεση h να ικανοποιεί το x και να
υπάρχει κάποια υπόθεση του G που να είναι πιο γενική.
γ) Διάγραψε από το S όποια υπόθεση είναι πιο γενική από
κάποια άλλη υπόθεση του S.
Αν το x είναι αρνητικό:
i) Διάγραψε τα μέλη του S που δεν ικανοποιούν το x.
ii) Για κάθε υπόθεση g∈G που δεν ικανοποιεί το x:
α) Διάγραψε την g από το G.
β) Πρόσθεσε στο G όλες τις ελάχιστες ειδικεύσεις h της g,
έτσι ώστε κάθε υπόθεση h να ικανοποιεί το x και να
υπάρχει κάποια υπόθεση του S που να είναι πιο ειδική.
γ) Διάγραψε από το G όποια υπόθεση είναι πιο ειδική από κάποια άλλη
υπόθεση του G.

Έστω το σύνολο εκπαίδευσης των 5 υποψηφίων. Τα σύνολα G και S αρχικοποιούνται: G={(X,Y,Z)}, S={()}

1ος κύκλος: (#1 p) (Υψηλές, Υψηλό, Ναι)

G={(X,Y,Z)}

S={(Υψηλές, Υψηλό, Ναι)} *#θετικό παράδειγμα, ικανοποιείται από το G και γενικεύουμε ελάχιστα το S ώστε να το ικανοποιεί#*

2ος κύκλος: (#2 n) (Χαμηλές, Υψηλό, Όχι)

G={(Υψηλές,Υ,Ζ),(Χ,Χαμηλό,Ζ),(Χ,Υ,Ναι)}

S={(Υψηλές, Υψηλό, Ναι)} *#αρνητικό παράδειγμα που ικανοποιείται από το S , το G αντικαθίσταται από όλες τις δυνατές ελάχιστες εξειδικεύσεις που ικανοποιούν το παράδειγμα#*

3ος κύκλος: (#3 p) (Χαμηλές, Υψηλό, Ναι)

G={(Χ,Υ,Ναι)}

S={(Χ, Υψηλό, Ναι)} *#διαγράφονται τα μέλη του G που δεν ικανοποιούν το θετικό παράδειγμα και το S γενικεύεται ελαχίστως ώστε να συμπεριλάβει το παράδειγμα#*

4ος κύκλος: (#4 n) (Υψηλές, Χαμηλό,Ναι)

G={(Χαμηλές,Υ,Ναι),(Χ,Υψηλό,Ναι)}

S={(Χ, Υψηλό, Ναι)}

5ος κύκλος: (#5 n) (Χαμηλές, Χαμηλό,Ναι)

G={(Χ, Υψηλό,Ναι)}

S={(Χ, Υψηλό, Ναι)}

Αν όντως το σύνολο δεδομένων μπορεί να εκφραστεί με μόνο μία πρόταση, μετά το πέρας των παραδειγμάτων θα είναι G=S όπως στην προκειμένη περίπτωση αλλιώς G<>S.

Ο αλγόριθμος απαλιφής υποψηφίων αποτελεί μία συστηματική μέθοδο για την εξαγωγή προτύπων που αποτελούνται όμως μόνο από μία πρόταση. Στα περισσότερα σύνολα εκπαίδευσης ένα πρότυπο αποτελείται από παραπάνω από μία προτάσεις. Έστω το επόμενο σύνολο εκπαίδευσης, με χαρακτηριστικά {Outlook, Temperature, Humidity, Windy} και μεταβλητή κλάσης Play.

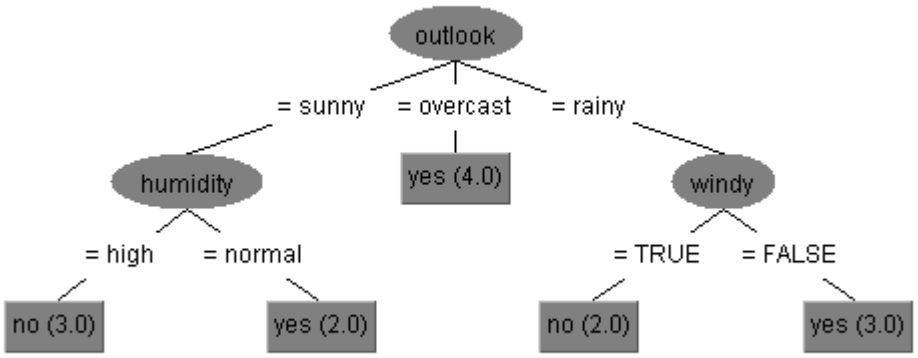
Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Ένα σύνολο προτάσεων που θα μπορούσε να εξαχθεί είναι το επόμενο:

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true    then play = no
If outlook = overcast                  then play = yes
If humidity = normal                   then play = yes
If none of the above                   then play = yes
```

Πρόκειται για ένα σύνολο προτασιακών κανόνων το οποίο καλύπτει όλα τα στιγμιότυπα του συνόλου εκπαίδευσης, μόνο όμως αν δούμε τους κανόνες με τη σειρά που φαίνονται παραπάνω. Η πρώτη πρόταση καλύπτει ένα μέρος από τα στιγμιότυπα του συνόλου εκπαίδευσης. Η δεύτερη πρόταση συμπεραίνεται από τα υπόλοιπα που δεν καλύπτονται από την πρώτη. Και αυτή με τη σειρά της καλύπτει κάποια στιγμιότυπα. Κάθε πρόταση συμπεραίνεται από στιγμιότυπα που δεν έχουν καλυφθεί από τις προηγούμενες. Αν θεωρήσουμε την τέταρτη πρόταση αγνοώντας τις προηγούμενες, θα διαπιστώσουμε ότι δεν καλύπτει όλα τα στιγμιότυπα με τιμή humidity=normal. Αν αποκλειστούν πρώτα όλα τα στιγμιότυπα των προηγούμενων κανόνων η πρόταση θα καλύπτει όλα όσα έχουν την εν λόγω τιμή. Μπορούν να εξαχθούν και άλλα σύνολα κανόνων.

Άλλη μορφή κανόνων που μπορεί να εξαχθεί είναι σε δενδροειδή μορφή:



If outlook=sunny and humidity=high then play=no
If outlook=sunny and humidity=normal then play=yes
If outlook=overcast then play=yes
If outlook=rainy and windy=true then play=no
If outlook=rainy and windy=false then play=yes

Το δέντρο κανόνων καλείται και **δέντρο απόφασης** ή **ταξινόμησης**. Οι αριθμοί σε κάθε τερματικό κόμβο είναι το πλήθος των στιγμιοτύπων που καλύπτει ο κάθε κανόνας.

Για την εύρεση τέτοιου είδους προτύπων εφαρμόζονται άλλου είδους αλγόριθμοι τους οποίους εξετάζουμε παρακάτω.

1.3 Δέντρα ταξινόμησης

Έστω ότι επιθυμούμε να εξάγουμε πρόταση με απλή επισκόπηση του συνόλου εκπαίδευσης. Η πρόταση θα αποτελείται από συγκεκριμένες τιμές χαρακτηριστικών.

Ποιο χαρακτηριστικό θα επιλέξουμε για αρχή και εν συνεχεία ποια τιμή του; Αν πάρουμε την τιμή *Χαμηλές* του χαρακτηριστικού *Τρέχουσες Οφειλές* βλέπουμε ότι συνδέεται με δύο κακούς χαρακτηρισμούς και έναν καλό. Αν το εισόδημά του είναι χαμηλό τότε είμαστε σίγουροι ότι είναι κακός υποψήφιος. Αν γνωρίζουμε ότι έχει υψηλές οφειλές, τότε σε μία περίπτωση είναι καλός και σε άλλη είναι κακός.

Κάθε τιμή οποιουδήποτε χαρακτηριστικού έχει μία διασπορά ως προς το σύνολο τιμών {Καλός, Κακός}. Οι χαμηλές οφειλές δεν συνεπάγονται μόνο καλό ή κακό χαρακτηρισμό αλλά σε δύο περιπτώσεις κακό και σε μία καλό. Όσο μεγαλύτερη είναι η διασπορά τόσο περισσότερο αναξιόπιστη είναι μία τιμή για την εξαγωγή συμπεράσματος.

Το μέγεθος της διασποράς της κάθε τιμής ως προς ένα σύνολο κλάσεων υπολογίζεται με τη βοήθεια του κριτηρίου της **εντροπίας πληροφορίας.** Η τιμή της δίνεται από τη σχέση

$$E(S_u) = - p_+ * \log_2(p_+) - p_- * \log_2(p_-)$$

όπου Su η τιμή u του χαρακτηριστικού A στο σύνολο στιγμιοτύπων S, p_+ το κλάσμα των θετικών παραδειγμάτων ως προς το πλήθος εγγραφών της u στο σύνολο S και p_- το κλάσμα των αρνητικών παραδειγμάτων. Γενικά για c διαφορετικές κλάσεις η εντροπία ορίζεται από τη σχέση

$$E(S_u) = - \sum_{i=1}^{c} p_i * \log_2(p_i)$$

όπου p_i το ποσοστό των παραδειγμάτων του S που ανήκουν στην κλάση i. Όσο μικρότερη είναι η εντροπία τόσο μεγαλύτερο είναι το **κέρδος πληροφορίας(gain)**, το οποίο ορίζεται για κάθε χαρακτηριστικό A από τη σχέση

$$G(S, A) = E(S) - \sum_{u \Xi Values(A)} \frac{|S_u|}{|S|} * E(S_u)$$

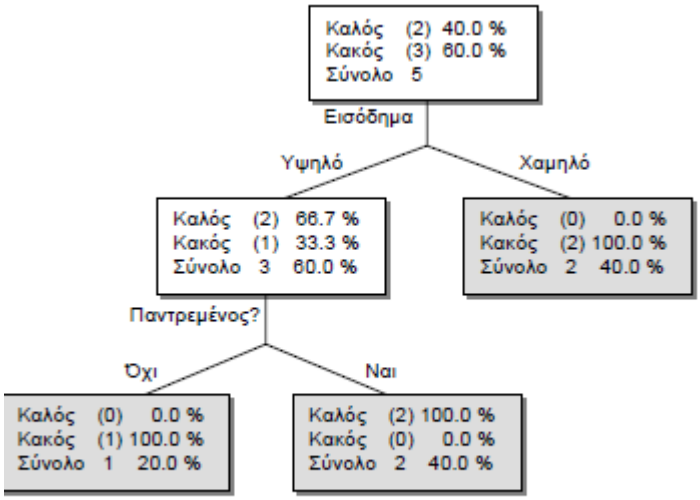
$|S_u|$ το πλήθος των στιγμιοτύπων με A=u. Παρακάτω φαίνεται ένα παράδειγμα υπολογισμού της εντροπίας κάθε τιμής ξεχωριστά

		Εισόδημα		Χρέος		Παντρεμένος	
	Σύνολα	Υψηλό	Χαμηλό	Υψηλό	Χαμηλό	Ναι	Όχι
S	5	3	2	2	3	4	1
+	2	2	0	1	1	2	0
-	3	1	2	1	2	2	1
p+	2/5	2/3	0	1/2	1/3	1/2	0
p-	3/5	1/3	1	1/2	2/3	1/2	1
E	0.97	0.92	0.00	1.00	0.92	1.00	0.00
Gain		0.42		0.02		0.17	

Ο αλγόριθμος δέντρων ταξινόμησης (**ID3**) εξάγει πρότυπα σε δενδροειδή μορφή.

1. Βρες την ανεξάρτητη μεταβλητή η οποία αν χρησιμοποιηθεί ως κριτήριο διαχωρισμού των δεδομένων εκπαίδευσης θα οδηγήσει σε κόμβους κατά το δυνατό διαφορετικούς σε σχέση με την εξαρτημένη μεταβλητή.
2. Κάνε το διαχωρισμό.
3. Επανέλαβε τη διαδικασία για κάθε έναν από τους κόμβους που προέκυψαν μέχρι να μην είναι δυνατός περαιτέρω διαχωρισμός.

Κατασκευάζει το δέντρο αναδρομικά κατά πλάτος επιλέγοντας αρχικά το πιο κατάλληλο χαρακτηριστικό στη ρίζα με βάση το κριτήριο μεγίστου κέρδους πληροφορίας. Στη συνέχεια για κάθε δυνατή τιμή του χαρακτηριστικού δημιουργούνται οι αντίστοιχοι απόγονοι της ρίζας σε καθέναν από τους οποίους αντιστοιχεί ένα μέρος από τα στιγμιότυπα του συνόλου εκπαίδευσης. Η διαδικασία επαναλαμβάνεται για κάθε νέο κόμβο με τα δεδομένα που ανήκουν σε αυτόν. Για παράδειγμα, στον πρώτο κύκλο εκπαίδευσης επιλέγεται ως αρχικό χαρακτηριστικό διαχωρισμού (ρίζα) το εισόδημα επειδή έχει το μεγαλύτερο κέρδος πληροφορίας. Όταν γίνεται ο διαχωρισμός επαναλαμβάνεται η διαδικασία για κάθε απόγονο ξεχωριστά με τα δικά του στιγμιότυπα ο καθένας: για τα στιγμιότυπα με υψηλό εισόδημα καταρτίζεται νέος πίνακας με εντροπίες των υπολοίπων τιμών του συνόλου δεδομένων με σκοπό να γίνει ο έλεγχος για περαιτέρω διαχωρισμό. Ακριβώς το ίδιο γίνεται και με τον έτερο απόγονο.



Μπορεί να διαπιστωθεί ότι ένας κόμβος είναι τερματικός και δεν μπορεί να διαχωριστεί περαιτέρω. Ένας τερματικός κόμβος έχει την ίδια κλάση σε όλα τα στιγμιότυπά του, είναι δηλαδή αμιγής. Άλλη συνθήκη τερματισμού είναι να μην υπάρχουν άλλα προς εξέταση χαρακτηριστικά.

1.4 Προτασιακοί κανόνες ταξινόμησης

Στον αλγόριθμο δέντρων ταξινόμησης η έξοδος ήταν ένα σύνολο κανόνων σε δενδροειδή μορφή. Σύνολο προτασιακών κανόνων μπορεί να παραχθεί και με άλλους αλγορίθμους όπως αυτός της σειριακής κάλυψης. Στον αλγόριθμο δέντρων ταξινόμησης η κατασκευή του δενδροειδούς συνόλου κανόνων γίνεται κατά πλάτος. Κάθε κλαδί του δέντρου, το οποίο αντιπροσωπεύει έναν τρέχοντα κανόνα, αναπτύσσεται παράλληλα με τα άλλα κλαδιά δηλαδή τους άλλους κανόνες. Ο αλγόριθμος σειριακής κάλυψης κατασκευάζει το σύνολο κανόνων κατά βάθος όχι όμως δενδροειδώς. Κάθε φορά επιλέγει τη βέλτιστη τιμή και όχι μόνο το χαρακτηριστικό ώστε να μη χρειάζεται διαχωρισμός. Πρώτα ολοκληρώνει έναν κανόνα και μετά συνεχίζει να ψάχνει για τον επόμενο στο εναπομείναν σύνολο

στιγμιοτύπων. Σε σχέση με τα δέντρα ταξινόμησης για την έναρξη ενός καινούριου κανόνα έχουμε να επεξεργαστούμε μικρότερα σύνολα στιγμιοτύπων.

```
Αλγόριθμος Σειριακής Κάλυψης
1. Αρχικοποίησε το Σύνολο_Κανόνων με το κενό σύνολο.
2. Μάθε_έναν_Κανόνα(Εξαρτημένη_Μεταβλητή,Μεταβλητές,Παραδείγματα).
3. Αν ο Κανόνας ικανοποιεί το Κριτήριο Απόδοσης:
   3α. Αφαίρεσε τα θετικά παραδείγματα που κάλυψε ο Κανόνας αυτός.
   3β. Πρόσθεσε τον Κανόνα στο Σύνολο_Κανόνων.
4. Επανέλαβε από το 2, όσο ικανοποιείται το Κριτήριο Απόδοσης.
```

```
Συνάρτηση Μάθε_έναν_Κανόνα (Εξαρτημένη_Μεταβλητή,Χαρακτηριστικά,Παραδείγματα)
//Αναζήτηση "Γενικό προς Ειδικό" (General-to-Specific Search)
1. Έστω η βέλτιστη υπόθεση (αρχικά, ο πιο γενικός κανόνας)
   που ταιριάζει με όλα τα παραδείγματα του συνόλου εκπαίδευσης.
2. Επανέλαβε, όσο υπάρχουν υποψήφιες υποθέσεις:
   Εξειδίκευσε τη βέλτιστη υπόθεση, προσθέτοντας το ζεύγος
   χαρακτηριστικού-τιμής που βελτιστοποιεί το κριτήριο απόδοσης.
```

Κάθε φορά η βέλτιστη υπόθεση επιλέγεται με κριτήριο την εντροπία πληροφορίας αλλά μπορεί να χρησιμοποιηθούν και άλλα κριτήρια όπως η σχετική συχνότητα ή ο m εκτιμητής ακριβείας.

1.5 Στατιστική μοντελοποίηση

Στη στατιστική μοντελοποίηση κάθε χαρακτηριστικό και η μεταβλητή κλάσης του συνόλου εκπαίδευσης εκλαμβάνονται ως τυχαίες μεταβλητές οι οποίες παίρνουν τιμές που εμφανίζονται στο σύνολο εκπαίδευσης. Θα ασχοληθούμε με δύο αλγορίθμους: Τον Naïve Bayes, ο οποίος αγνοεί τις συσχετίσεις μεταξύ των χαρακτηριστικών και τον BayesNet, ο οποίος τις λαμβάνει υπ'όψιν. Πριν όμως μιλήσουμε για τους δύο αλγορίθμους ταξινόμησης, ακολουθεί μία εισαγωγική ενότητα πάνω στη θεωρία στην οποία βασίζονται, αυτή των μπεϋζιανών δικτύων.

1.5.1 Μπεϋζιανά δίκτυα (Bayesian networks)

Παραμετροποίηση υπό συνθήκη

Ας θεωρήσουμε το παράδειγμα μιας εταιρείας που θέλει να προσλάβει έναν φοιτητή. Κριτήριο της εταιρείας είναι η *εξυπνάδα* και η *εξέταση* των υποψηφίων. Δεν μπορεί όμως να ελέγξει απ'ευθείας καθέναν από αυτούς. Γι'αυτό το λόγο τα δύο αυτά κριτήρια αντιστοιχούν σε δύο τυχαίες μεταβλητές E και S με τιμές val(E)={e^0,e^1}, val(S)={s^0,s^1}. Οι τιμές e^0 και e^1 είναι υψηλή και χαμηλή νοημοσύνη

όπως και οι τιμές s^0 kai s^1, καλή συμπεριφορά και κακή συμπεριφορά. Έστω ότι η κοινού κατανομή πιθανότητας είναι η παρακάτω

N	E	P(E,N)
ν^0	ε^0	0.665
ν^0	ε^1	0.035
ν^1	ε^0	0.06
ν^1	ε^1	0.24

Σύμφωνα με τον κανόνα του Bayes $P(N,\Delta)=P(N)*P(\Delta|N)$.
Επομένως η από κοινού κατανομή των δύο μεταβλητών μπορεί να αναπαρασταθεί και από τους εξείς δύο πίνακες, καθένας από τους οποίους αντιπροσωπεύει τις πιθανότητες P(N) και $P(\Delta|N)$.

ν^0	ν^1
0.7	0.3

E	ε^0	ε^1
ν^0	0.95	0.05
ν^1	0.2	0.8

Η εξαρτημένη πιθανότητα (**Conditional Probability Distribution-CPD)** $P(\Delta|N)$ αντιπροσωπεύει την πιθανότητα ο φοιτητής να γράψει καλά στην εξέταση στις περιπτώσεις που έχει χαμηλή νοημοσύνη (ν^0) ή υψηλή νοημοσύνη (ν^1). Έτσι, παρατηρούμε ότι είναι πολύ πιθανόν να μην πάει καλά αν έχει χαμηλή νοημοσύνη ($P(\varepsilon^0|\nu^0)=0.95$) και πολύ πιθανόν να πάει καλά αν έχει υψηλή ($P(\varepsilon^1|\nu^1)=0.8$). Το γνωστό γεγονός N καλείται **ένδειξη** (**evidence**).

Η P(N) είναι η **περιθώρια κατανομή** της N. Η P(E|N) υπολογίζεται από τη σχέση $P(E|N)=P(E,N)/P(N)$

ΤΟ ΝΑΪVE BAYES ΜΟΝΤΕΛΟ

Συνεχίζοντας το παράδειγμά μας, υποθέτουμε ότι η εταιρεία γνωρίζει και την πρόοδο Π του φοιτητή μέσα στο εξάμηνο. Σε αυτήν την περίπτωση έχουμε πλέον από κοινού κατανομή των μεταβλητών N, Δ, Π. Οι N και Δ παίρνουν τις τιμές όπως πριν και η Π τις τιμές π^1, π^2, π^3 δηλαδή A, B, C (καλή, μέτρια, κακή πρόοδος). Συνολικά η από κοινού κατανομή θα έχει δώδεκα πιθανότητες.

Παρατηρούμε στο γράφο ότι η νοημοσύνη του μαθητή σχετίζεται με την πρόοδο και το διαγώνισμα αλλά και ότι το διαγώνισμα σχετίζεται με την πρόοδο: Αν υποθέσουμε ότι ο φοιτητής έγραψε καλά στο διαγώνισμα δεξιότητας (δ^1), τότε είναι πιθανόν να έχει πρόοδο A (π^1) στο εξάμηνο. Ορίζεται λοιπόν η δεσμευμένη πιθανότητα $P(\pi^1|\delta^1)$, και ας μην υπάρχει ακμή στον γράφο από τη μεταβλητή Δ στην Π!

Ωστόσο, μπορούμε να εντοπίσουμε και μια περίπτωση ανεξαρτησίας τυχαίων μεταβλητών υπό συνθήκη. Αν γνωρίζουμε ότι ο φοιτητής έχει υψηλή νοημοσύνη, η πληροφορία ότι έγραψε καλά στο διαγώνισμα είναι πλέον περιττή για να προβλέψουμε την πρόοδό του. Με άλλα λόγια $P(\pi|\nu^1,\delta^1)=P(\pi|\nu^1)$. Γενικότερα λέμε ότι η από κοινού κατανομή των N, Π, Δ περιέχει το ζεύγος ανεξάρτητων

μεταβλητών (Π,Δ) με γνωστή τη Ν και γράφουμε $P'(\Pi \wedge \Delta|N)$. Αν η τιμή της Ν γνωστή, Π και Δ ανεξάρτητες ενώ αν Ν άγνωστη Π και Δ σχετίζονται.

Η από κοινού κατανομή πιθανότητας των τριών μεταβλητών είναι

$$P(N,\Pi,\Delta)=P(N)*P(\Pi,\Delta|N)=P(N)*P(\Pi|N)*P(\Delta|N)$$

Υποθέτουμε ότι οι δεσμευμένες πιθανότητες P(N) και P(Δ|N) έχουν τις ίδιες τιμές με προηγουμένως και επιπλέον θεωρούμε για τη νέα πιθανότητα P(Π|N) τον ακόλουθο πίνακα τιμών:

N	π^1	π^2	π^3
ν^0	0.2	0.34	0.46
ν^1	0.74	0.17	0.09

Επομένως, $P(\nu^1, \delta^1, \pi^2)=P(\nu^1)*P(\delta^1|\nu^1)*P(\pi^2|\nu^1)= 0.3*0.8*0.17=0.0408$

Κάθε δεσμευμένη πιθανότητα στο γράφο καλείται και **τοπική δεσμευμένη πιθανότητα (local CPD)**. Με τη βοήθεια των τοπικών δεσμευμένων πιθανοτήτων καταφέραμε να αναπαραστήσουμε την κατανομή με συνολικά 2+4+6=12 παραμέτρους αντί για 2*4*6=48, αν παίρναμε κάθε γεγονός που ορίζεται από τις τρεις τυχαίες μεταβλητές. Με αυτήν την αναπαράσταση άλλωστε μπορούμε να υπολογίσουμε την πιθανότητα οποιουδήποτε γεγονότος των Ν, Π, Δ θελήσουμε.

Το γενικό μοντέλο

Το **μοντέλο Naïve Bayes** υποθέτει ότι κάθε στιγμιότυπο ανήκει σε μία από τις $\{c_1, c_2,\ldots, c_k\}$ κλάσεις μιας μεταβλητής κλάσεως C. Στο παράδειγμά μας η μεταβλητή κλάσης είναι η Ν η οποία έχει δύο κατηγορίες, τους φοιτητές χαμηλής νοημοσύνης και τους φοιτητές υψηλής νοημοσύνης.

Το μοντέλο περιλαμβάνει επισης έναν αριθμό χαρακτηριστικών $X_1,X_2,\ldots,X_\kappa$ τα οποία θεωρούνται ενδείξεις. Η υπόθεση του «αφελούς» Bayes λέει ότι τα χαρακτηριστικά είναι ανεξάρτητα δοθείσης της τιμής της κλάσεως. Δηλαδή
$(X_i \wedge \mathbf{X}_{-i}|C)$ για κάθε i, όπου $\mathbf{X}_{-i}=\{X_1,\ldots,X_k\}-\{X_i\}$.

Η κατανομή δίνεται από τον τύπο

$$P(C,X_1,\ldots,X_n)=P(C)\prod_{i=1}^{n} P(X_i|C)$$

Το μοντέλο αφελώς αμελεί τις συσχετίσεις μεταξύ των χαρακτηριστικών, εξ'ού και ο χαρακτηρισμός. Σίγουρα η κατανομή είναι εσφαλμένη αν δεν λάβουμε υπ'όψιν τις συσχετίσεις. Το Naïve Bayes μοντέλο όμως, παρά τις υπεραπλουστευμένες παραδοχές που κάνει, διακρίνεται για την απλότητά του και τον μικρό αριθμό παραμέτρων που απαιτούνται. Χρησιμοποιείται συχνά στην ταξινόμηση αποφασίζοντας την κλάση στην οποία ανήκει ένα στιγμιότυπο του συνόλου των χαρακτηριστικών με τον υπολογισμό της δεσμευμένης πιθανότητας το στιγμιότυπο να ανήκει στην c^1 ή της πιθανότητας να ανήκει στη c^2. Παρακάτω φαίνεται ο λόγος των δύο δεσμευμένων πιθανοτήτων

$$\frac{P(C=c^1|x_1,\ldots,x_n)}{P(C=c^2|x_1,\ldots,x_n)} = \frac{P(C=c^1)}{P(C=c^2)} * \prod_{i=1}^{n} \frac{P(x_i|C=c^1)}{P(x_i|C=c^2)}$$

Το μοντέλο χρησιμοποιήθηκε αρχικά στην ιατρική διάγνωση, στην οποία διαφορετικές τιμές της μεταβλητής ταξινόμησης αντιπροσωπεύουν διαφορετικές ασθένειες που μπορεί να έχει ο ασθενής. Οι μεταβλητές που έχουμε για ενδείξεις αντιπροσωπεύουν συμπτώματα ή αποτελέσματα εξετάσεων. Βέβαια, και πάλι έχουν γίνει μη αποδεκτές στην πραγματικότητα παραδοχές όπως ότι ο ασθενής δεν μπορεί να έχει παραπάνω από μία ασθένεια και ότι γνωρίζοντας την ασθένειά του τα συμπτώματα δεν επιρρεάζουν το ένα το άλλο, αλλά συμβαίνουν ανεξάρτητα.

Όπως προείπαμε οι παραδοχές του naïve Bayes οδηγούν σε σφάλματα στη διάγνωση. Συγκεκριμένα, το μοντέλο τείνει να υπερεκτιμήσει την επιρροή μιας ένδειξης στην τιμή της πιθανότητας, μετρώντας την παραπάνω από μία φορές. Για παράδειγμα, η υπέρταση και η παχυσαρκία είναι σοβαρές ενδείξεις καρδιακής ασθένειας. Αυτές οι δύο ενδείξεις στην πραγματικότητα έχουν μεγάλο βαθμό συσχέτισης. Στην προηγούμενη εξίσωση όμως δύο όροι στο γινόμενο αντιστοιχούν ένας σε κάθε ένδειξη, και είναι σαν να μετράμε την ίδια ένδειξη δύο φορές. Πράγματι, έρευνες έχουν δείξει ότι όσο αυξάνεται ο αριθμός των χαρακτηριστικών, τόσο μειώνεται η δύναμη του Naïve Bayes και το σφάλμα εντοπιζόταν στην υπόθεση μηδαμινής συσχέτισής τους. Το φαινόμενο αυτό οδήγησε στην ανάγκη χρήσης μπεϋζιανών δικτύων με τις απαραίτητες συσχετίσεις μεταξύ των χαρακτηριστικών. Ένα παράδειγμα που αξιοποιεί όχι πλήρως αλλά μερικώς τις συσχετίσεις είναι η χρήση μπεϋζιανών δικτύων για συμπερασμό πάνω σε εκφράσεις γονιδιωμάτων [12]. Γενκά, έχουν αναπτυχθεί διάφορα συστήματα που στηρίζονται στο συμπερασμό με μπεϋζιανά δίκτυα για διαγνωστικούς σκοπούς [13].

Μπεϋζιανά δίκτυα

Όπως και ο Naïve Bayes, έτσι και τα μπεϋζιανά δίκτυα αξιοποιούν τις τοπικές δεσμευμένες πιθανότητες για μία πιο συμπυκνωμένη και φυσική αναπαράσταση. Πυρήνας των μπεϋζιανών δικτύων είναι ένας κατευθυνόμενος ακυκλικός γράφος (**direct acyclic graph- dag**), του οποίου οι κόμβοι είναι χαρακτηριστικά (τυχαίες μεταβλητές) και οι ακμές συσχετίσεις μεταξύ των χαρακτηριστικών.

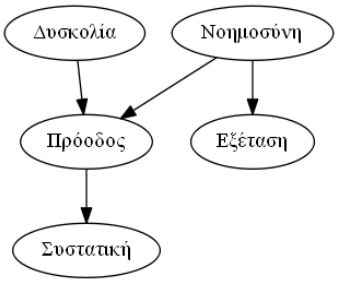
Παρουσίαση του μοντέλου

Ας θεωρήσουμε πάλι το παράδειγμα του φοιτητή, με κάπως διαφορετικό σενάριο. Η πρόοδός του πλέον δεν εξαρτάται μόνο από τη νοημοσύνη του αλλά και από τη δυσκολία των μαθημάτων, η οποία έχει δύο τιμές, δύσκολα και εύκολα, val(Δ)={δ^0,δ^1}. Ο φοιτητής ζητά συστατική επιστολή από τον καθηγητή. Ο καθηγητής είναι αφηρημένος και συνήθως δε θυμάται τους φοιτητές του, παρά μόνο τους αξιολογεί από την καταγεγραμμένη πορεία τους μέσα στο εξάμηνο (πρόοδος). Η συστατική επιστολή εκλαμβάνεται ως τυχαία μεταβλητή που παίρνει

δύο τιμές: val(Σ)={σ^0=αρνητική γνώμη του καθηγητή για τον φοιτητή, σ^1=θετική γνώμη του καθηγητή για τον φοιτητή}. Έχουμε λοιπόν τον παρακάτω γράφο

δ^0	δ^1
0.6	0.4

ν^0	ν^1
0.7	0.3



	π^1	π^2	π^3
ν^0,δ^0	0.3	0.4	0.3
ν^0,δ^1	0.05	0.25	0.7
ν^1,δ^0	0.9	0.08	0.02
ν^1,δ^1	0.5	0.3	0.2

	ε^0	ε^1
ν^0	0.95	0.05
ν^1	0.2	0.8

	σ^0	σ^1
π^1	0.1	0.9
π^2	0.4	0.6
π^3	0.99	0.01

Τα τοπικά μοντέλα δεσμευμένων μεταβλητών αποτελούν ένα συστατικό στοιχείο της αναπαράστασης με μπεϋζιανούς γράφους και δείχνουν την εξάρτηση του εκάστοτε κόμβου από τους γονείς του. Για κάθε τυχαία μεταβλητή, όπως βλέπουμε, ορίζεται δεσμευμένη πιθανότητα η οποία μας δίνει την κατανομή της μεταβλητής για κάθε γεγονός των μεταβλητών που χρησιμοποιούνται ως ενδείξεις. Το δίκτυο μαζί με τις τοπικές δεσμευμένες πιθανότητες αποτελούν το **μπεϋζιανό δίκτυο Bayes**. Χρησιμοποιούμε το συμβολισμό $B^{\text{φοιτητής}}$ για το δίκτυο του συγκεκριμένου παραδείγματος.

Πώς χρησιμοποιούμε το δίκτυο για να προσδιορίσουμε την κατανομή; Ας θεωρήσουμε το στιγμιότυπο $\nu^1,\delta^0,\pi^2,\varepsilon^1,\sigma^0$. Η πιθανότητά του υπολογίζεται από το γινόμενο των πιθανοτήτων των γεγονότων που το αποτελούν.

$P(\nu^1,\delta^0,\pi^2,\varepsilon^1,\sigma^0)=P(\nu^1)*P(\delta^0)*P(\pi^2|\nu^1,\delta^0)*P(\varepsilon^1|\nu^1)*P(\sigma^1|\pi^2)=0.3*0.6*0.08*0.8*0.4=0.004608$

Γενικά έχουμε P(Ν,Δ,Π,Ε,Σ)=P(Ν)*P(Δ)*P(Π|Ν,Δ)*P(Ε|Ν)*P(Σ|Π).
Ο τύπος καλείται ***κανόνας αλυσίδας για τα μπεϋζιανά δίκτυα*** *(βλέπε και Aji, S. M. and R. J. McEliece (2000). The generalized distributive law.* IEEE Trans. Information Theory 46, *325-343)*.

.

Συμπερασμοί μέσα από μπεϋζιανά δίκτυα

Όπως αναφέρθηκε πρωτύτερα, από μία από κοινού κατανομή P_B (με τον τρόπο αναπαράστασης των τοπικών δεσμευμένων μεταβλητών τον οποίο χρησιμοποιούν τα δίκτυα) είναι δυνατόν να υπολογιστεί η πιθανότητα $P_B(\boldsymbol{Y = y} | \boldsymbol{E = e})$ ενός γεγονότος y δεδομένων των ενδείξεων e. Στον πίνακα της τοπικής δεσμευμένης πιθανότητας της μεταβλητής Y υπολογίζουμε νέες δεσμευμένες πιθανότητες οι οποίες περιέχουν μόνο τις ενδείξεις e. Με ποιόν τρόπο; Για κάθε e^i των ενδείξεων E αθροίζουμε τις πιθανότητες στον πίνακα που περιέχουν την e^i και αντιστοιχούν στην y. Στη συνέχεια διαιρούμε με το άθροισμα όλων των πιθανοτήτων του πίνακα ώστε το αποτέλεσμα να κανονικοποιηθεί, και να ισχύει

$$P(Y = y|E = e) + P(Y = \text{NOT } y|E = e)=1$$

δηλαδή το άθροισμα των πιθανοτήτων του y και του συμπληρωματικού γεγονότος του y να κάνει μονάδα. Ουσιαστικά υπολογίζουμε περιθώριες κατανομές πιθανοτήτων των ενδείξεων του συνόλου E.

Παράδειγμα, ας θεωρήσουμε έναν συγκεκριμένο φοιτητή για τον οποίο θέλουμε να κάνουμε συμπερασμό με τη βοήθεια του μοντέλου μας. Έστω ότι θέλουμε να μάθουμε πόσο πιθανό είναι να έχει καλή συστατική επιστολή (σ^1) στο μάθημα Οικονομικά. Δεν έχουμε καμμιά άλλη πληροφορία (ένδειξη) στη διάθεσή μας. Από τον γράφο έχουμε την πιθανότητα Ρ(Σ|Π). Έχουμε λοιπόν τις παρακάτω περιθώριες πιθανότητες ως προς Π

$$P_{\mu\eta\text{-}\kappa\alpha\nu\sigma\nu}(\sigma^1) = P(\sigma^1|\pi^1) + P(\sigma^1|\pi^2) + P(\sigma^1|\pi^3) = 1.51$$
$$P_{\mu\eta\text{-}\kappa\alpha\nu\sigma\nu}(\sigma^0) = P(\sigma^0|\pi^1) + P(\sigma^0|\pi^2) + P(\sigma^0|\pi^3) = 1.49$$

Το άθροισμα των δύο πιθανοτήτων πρέπει να κάνει 1. Άρα

$$P(\sigma^1)= \frac{1.51}{1.51 + 1.49} = 0.503$$

Βλέπουμε ότι από τη στιγμή που δεν ξέρουμε τίποτα για τον φοιτητή, είναι εξίσου πιθανό να έχει καλή ή κακή συστατική. Απόλυτα λογικό. Ίσως αν είχαμε περισσότερες ενδείξεις για τον φοιτητή, να μπορούσαμε να αποφανθούμε καλύτερα για το γεγονός που μας ενδιαφέρει. Έστω λοιπόν ότι έχουμε την ένδειξη $\Pi=\pi^1$, δηλαδή ότι είχε καλή πρόοδο στο μάθημα. Ποια η δεσμευμένη πιθανότητα $P(\sigma^1|\pi^1)$;

$P(\sigma^1|\pi^1)= \frac{0.9}{0.9 + 0.1} = 0.9$, άρα αν έχει καλή πρόοδο είναι σχεδόν σίγουρο ότι θα έχει και καλή συστατική. Μπορούμε να υπολογίσουμε οποιαδήποτε δεσμευμένη πιθανότητα ή από κοινού κατανομής πιθανότητας θελήσουμε, ακόμα και πιθανότητες όπως Ρ(*Νοημοσύνη*|*Συστατική*). Αυτή η αλληλουχία γεγονότων δεν υποστηρίζεται από το γράφο και γι'αυτό χρησιμοποιούμε τον τύπο

Ρ(*Νοημοσύνη*|*Συστατική*)=Ρ(*Νοημοσύνη*,*Συστατική*)/Ρ(*Συστατική*) Ο υπολογισμός του αριθμητή γίνεται πάλι υπλογίζοντας τις απαραίτητες περιθώριες κατανομές και ακολούθως εφαρμόζοντας τον κανόνα της αλυσίδας.

Βασικές ανεξαρτησίες στα μπεϋζιανά δίκτυα

Στο παράδειγμα του φοιτητή οι ακμές του γράφου αντιπροσωπεύουν απ'ευθείας εξάρτηση. Παρατηρώντας το γράφο, κάποιος θα μπορούσε να ισχυριστεί ότι το περιεχόμενο της συστατικής εξαρτάται μόνο από την πρόοδο του φοιτητή στο μάθημα, μόνο και μόνο επειδή ο μοναδικός γονιός του κόμβου Σ είναι ο Π. Ο ισχυρισμός ότι ένας οποιοσδήποτε κόμβος εξαρτάται μόνο από τους γονείς του είναι λάθος.

Από τον γράφο παρατηρούμε ότι (Σ⫫ Ν,Δ,Ε|Π), δηλαδή ότι από τη στιγμή που γνωρίζουμε την πρόοδο του φοιτητή, οι πεποιθήσεις μας για τη συστατική του δεν επιρρεάζονται από κανέναν από τους υπόλοιπους κόμβους. ΄Ομοια, ο βαθμός της εξέτασης αν έχουμε ένδειξη για τη νοημοσύνη δεν εξαρτάται από κανέναν άλλο κόμβο του γράφου, (Ε⫫ Δ,Π,Σ|Ν). Αυτοί οι δύο κόμβοι, αν έχουμε ενδείξεις για τους γονείς τους, εξαρτώνται μόνο από αυτούς και από κανέναν άλλο κόμβο.

Ας θεωρήσουμε τώρα τον κόμβο Π. Με τη λογική των προαναφερθέντων κόμβων θα μπορούσαμε να παρασυρθούμε και να ισχυριστούμε ότι δεδομένων ενδείξεων των γονιών του δεν εξαρτάται από κανέναν άλλο κόμβο του γράφου. Αν πάρουμε τις ενδείξεις $ν^1$, $δ^1$ σημαίνει ότι έχουμε να κάνουμε με έναν έξυπνο φοιτητή σε ένα δύσκολο μάθημα. Ο κόμβος Π είναι ανεξάρτητος σε αυτήν την περίπτωση από τον Σ; Σίγουρα όχι. Αν έχουμε ως ένδειξη την τιμή $σ^1$ (ο φοιτητής έχει καλή συστατική), τότε η πιθανότητα $P(π^1)$ αυξάνεται:

$$P(π^1|ν^1,δ^1,σ^1)>P(π^1|ν^1,δ^1).$$

Αυτό φαίνεται και από τους υπολογισμούς, όπου η πρώτη πιθανότητα έχει τιμή 0.712 και η δεύτερη 0.5.

Άρα, βλέπουμε ότι δεν περιμένουμε από οποιονδήποτε κόμβο να είναι ανεξάρτητος των υπολοίπων κόμβων δεδομένων των γονιών του, μπορεί να εξαρτάται από τους απογόνους του. Υπάρχει περίπτωση να εξαρτάται από άλλους κόμβους; Μπορούμε να περιμένουμε ο Π να εξαρτάται από τον Ε δεδομένων των Ν και Δ; Η απάντηση είναι όχι. Από τη στιγμή που ξέρουμε ότι ο φοιτητής έχει υψηλή νοημοσύνη, ο βαθμός της εξέτασής του δε μας δίνει επιπλέον πληροφορίες που να μας βοηθούν να προβλέψουμε την πρόοδό του. Άρα θα θέλαμε μια ιδιότητα τέτοια ώστε: (Π⫫ Ε|Ν,Δ).

Μένει μόνο να δούμε τι γίνεται με τις μεταβλητές Ν και Δ, οι οποίες δεν έχουν γονείς στον γράφο. Έτσι, ενώ πριν ψάχναμε για ανεξαρτησίες δοθέντων των γονιών ενός κόμβου, τώρα ψάχνουμε για περιθώριες ανεξαρτησίες. Ο Ν όπως είπαμε εξαρτάται από τους απογόνους του Π, Ε και Σ. Ο μόνος μη απόγονός του είναι ο Δ. Άρα περιμένουμε ότι (Ν⫫ Δ). Στα προηγούμενα παραδείγματα ωστόσο, γνώση της της δυσκολίας του μαθήματος επιρρέαζε δραστικά τις πεποιθήσεις μας για τη νοημοσύνη του φοιτητή. Βέβαια, αυτό συνέβαινε παρουσία ενδείξεων σχετικών με την πρόοδο. Με άλλα λόγια, παρουσιάζαμε την εξάρτηση ανάμεσα στους Ν και Δ δοθέντος του Π. Το φαινόμενο αυτό είναι πολύ σημαντικό και θα επανέλθουμε παρακάτω.

Για τη μεταβλητή Δ, ο Ν και Ε είναι μη απόγονοι. να θυμήσουμε ότι αν (Ν⫫ Δ) τότε και (Δ⫫ Ν). Η μεταβλητή Ε αυξάνει την πεποίθησή μας για τη νοημοσύνη του

φοιτητή, αλλά γνωρίζοντας ότι ο φοιτητής είναι έξυπνος ή όχι δεν επιρρεάζεται η πεποίθησή μας για τη δυσκολία του μαθήματος. Έτσι έχουμε ότι (Δ ⊥ Ν,Ε).

Μετά από όλα αυτά μπορούμε να βγάλουμε το εξής συμπέρασμα: από τη στιγμή που γνωρίζουμε την τιμή των γονέων της μεταβλητής που εξετάζουμε δεν υπάρχει πληροφορία σχετιζόμενη με προγόνους των γονέων η οποιά να επιρρεάζει την πεποίθησή μας για αυτήν. Ωστόσο, γνώση των απογόνων της μπορεί να επιρρεάσει τις πεποιθήσεις μας για αυτή μέσω μιας διαδικασίας **συμπερασμού ενδείξεων (evidential reasoning)**.

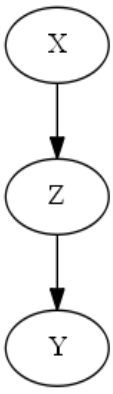
Ανεξαρτησίες στους γράφους

Όπως έχουμε δει μέχρι τώρα, ο γράφος G περιέχει ένα σύνολο από ανεξαρτησίες υπό συνθήκη $I_l(G)$. Μία κατανομή P η οποία παραγοντοποιείται σύμφωνα με το γράφο G ικανοποιεί το $I_l(G)$.

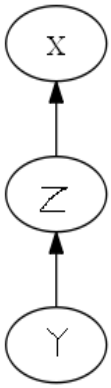
D-Διαχωρισμός

Σε αυτό το σημείο θα προσπαθήσουμε να κατανοήσουμε πώς μία ένδειξη που σχετίζεται με τη μεταβλητή X μπορεί να αλλάξει την πεποίθησή μας για την Y, δεδομένου ενός συνόλου ενδείξεων Z.

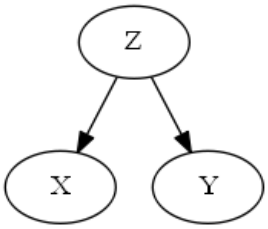
Απ'ευθείας σύνδεση. Ξεκινάμε με αυτήν την απλή περίπτωση στην οποία, οι X και Y συνδέονται μέσω της ακμής X->Y.



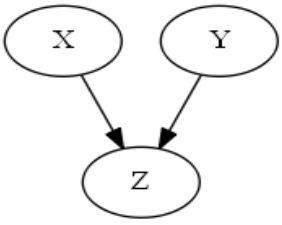
Εικόνα 1



Εικόνα 2



Εικόνα 3



Εικόνα 4

Μη απ'ευθείας σύνδεση. Έστω η απλούστερη περίπτωση στην οποία παρεμβάλλεται ένας κόμβος Z μεταξύ των X και Y. Υπάρχουν τέσσερις δυνατές περιπτώσεις για το μονοπάτι που μπορεί να προκύψει.

Μη απ'ευθείας αιτιοκρατική επίδραση (causal effect-εικόνα 1) Έστω το μονοπάτι X->Z->Y και ότι παρατηρούμε τη X. Οι δυνατές τιμές της Y (περιπτώσεις)

καθορίζονται από τις τιμές της Z, και οι δυνατές της Z από τις παρατηρειθήσες της X. Άρα και οι δυνατές της Y καθορίζονται από τις παρατηρειθήσες της X (μεταβατική ιδιότητα). Από τον κανόνα της αλυσίδας ισχύει P(Y|X)=P(Y|Z)*P(Z|X) Αν παρατηρήσουμε τη Z οι δυνατές τιμές της Y καθορίζονται από τις παρατηρειθήσες της Z. Οι παρατηρειθήσες της Z είναι υποσύνολο των δυνατών Z οι οποίες με τη σειρά τους καθορίζονται από την X. Άρα η γνώση της X δε μας προσφέρει τίποτα παραπάνω για την εκτίμηση της Y και μπορούμε να πούμε (X ⫫ Y|Z).

Ας θεωρήσουμε από το γνωστό παράδειγμα το μονοπάτι Ν->Π->Σ. Ας ξεκινήσουμε με την περίπτωση όπου η Π δεν έχει παρατηρηθεί. Αν παρατηρήσουμε ότι ο φοιτητής είναι έξυπνος, τότε είναι πολύ πιθανόν να έχει Α στην πρόοδό του και ακολούθως με βάση την πρόοδό του μια καλή συστατική επιστολή. Άρα η μεταβλητή Ν επιρρεάζει τη Σ μέσω της Π. Ας υποθέσουμε τώρα ότι η Π παρατηρείται. Τώρα πια δεν επιρρεάζει η γνώση της Ν τη μεταβλητή Σ και γενικά, το X δεν επιρρεάζει το Y αν το Z παρατηρείται.

Μη απ'ευθείας επιρροή ενδείξεων (evidential effect-εικόνα 2) Σε αυτή την περίπτωση το μονοπάτι είναι το X<-Z<-Y. Όπως ακριβώς και στην προηγούμενη περίπτωση, αν παρατηρήσουμε την X, η Y εξαρτάται από την X ενώ αν παρατηρήσουμε την Z τότε (X ⫫ Y|Z). Δηλαδή το μονοπάτι ενδείξεων έχει ακριβώς την ίδια συμπεριφορά με το αιτιοκρατικό μονοπάτι. Να θυμήσουμε ότι σε ένα μονοπάτι X->Y, αν γνωρίζουμε την Y τότε η Y επιρρρεάζει την X και ισχύει ο τύπος $P(X|Y)=\frac{P(X,Y)}{P(Y)}$.

Στο παράδειγμα του φοιτητή έχουμε μια αλυσίδα Ν->Π->Σ. Έχουμε δει ότι παρατήρηση μιας καλής συστατικής επιστολής μας κάνει να πιθανολογούμε για αυξγμένη νοημοσύνη του φοιτητή. Στην περίπτωση όμως που έχουμε ένδειξη για την πρόοδο, η συστατική δεν μας παρέχει παραπάνω πληροφορίες για τη νοημοσύνη και είναι περιττή γνώση.

Κοινή αιτία (common cause-εικόνα 3) Έτσι και εδώ η X επιρρεάζει την Y και γνώση της Z συνεπάγεται ανεξαρτησία της Y από την X.

Κοινή επίδραση (common effect-εικόνα 4) Σε όλες τις προαναφερόμενες περιπτώσεις υπάρχει μια κονή λογική: Αν γνωρίζουμε το X τότε το X επιρρεάζει το Y και αν γνωρίζουμε το Z τότε (X ⫫ Y|Z). Αυτήν την τελευταία περίπτωση όμως πρέπει να την ξανασκεφτούμε περισσότερο. Στο παράδειγμα του φοιτητή οι μεταβλητές Δ και Ν είναι γονείς της Π. Όταν η Π δεν παρατηρείται, Δ και Ν είναι ανεξάρτητες. Αυτό το συμπέρασμα βγαίνει και από τις τοπικές δεσμευμένες πιθανότητες του γράφου. Δεν μπορεί λοιπόν να υπάρξει ροή της επίδρασης στο μονοπάτι X->Z<-Y από την X στην Y αν η Z δεν παρατηρείται.

Ας θεωρήσουμε τώρα ότι η Z παρατηρείται ή ότι $\Pi=\pi^3$ (κακή). Με κατάλληλους υπολογισμούς στο γράφο μπορούμε να δούμε ότι η μειώνεται κατά πολύ η πιθανότητα να είναι έξυπνος ο φοιτητής. Αν όμως έχουμε ως επιπλέον ένδειξη ότι

το μάθημα είναι δύσκολο, η πιθανότητα να είναι έξυπνος αυξάνεται. Επομένως δεδομένης της ενδείξεως $\Pi=\pi^3$, οι μεταβλητές Ν και Δ σχετίζονται! Αν είχαμε ως ένδειξη ότι η συστατική είναι κακή ($\Sigma=\sigma^0$), η πιθανότητα της κακής προόδου θα ήταν μεγάλη και οι μετεβλητές Δ,Ν συσχετίσημες.

Ένα πρόβλημα που μπορεί να προκύψει στα μπεϋζιανά δίκτυα έναι η εύρεση των καταλλήλων ενδείξεων ώστε να μεγιστοποιείται η πιθανότητα ενός γεγονότος (maximum a posteriori -MAP). Το πρόβλημα είναι γενικά NP-hard και έχουν προταθεί διάφοροι αλγοριθμοι προσεγγίσεώς του [9, 10, 11, 12].

1.5.2 Naïve Bayes στην ταξινόμηση και σύγκριση με BayesNet (ονομαστικά και αριθμητικά χαρακτηριστικά)

Έστω ένα σύνολο τυχαίων μεταβλητών $\{X_1, X_2, \ldots, X_n\}$. Αν αγνοήσουμε τις συσχετίσεις μεταξύ τους και θεωρήσουμε τους συνδυασμούς $X_i \wedge X_j$ για οποιαδήποτε i<>j δηλαδή το σύνολο των ζευγών ανεξαρτήτων μεταβλητών, τότε η από κοινού κατανομή $P(X_1,X_2,\ldots,X_n)$ δίνεται από τη σχέση

$$P(X_1, X_2, \ldots, X_n) = P(X_1) * P(X_2) * \ldots * P(X_n)$$

Χρησιμοποιείται στην ταξινόμηση από τον ταξινομητή **Naive Bayes**. Ας υποθέσουμε ότι $c=\{c_1,c_2\}$ είναι μια μεταβλητή ταξινόμησης με δύο τιμές και $E=\{E_1, E_2,\ldots,E_n\}$ το σύνολο των χαρακτηριστικών του συνόλου εκπαίδευσης, τα οποία παίρνουν διακριτές τιμές. Σκοπός του Naive Bayes είναι να ταξινομήσει ένα καινούριο παράδειγμα e του E που δεν ανήκει στο σύνολο εκπαίδευσης, όποιος δηλαδή είναι και ο σκοπός των προαναφερόμενων αλγορίθμων ανίχνευσης προτύπων. Η δεσμευμένη πιθανότητα $P(c=c_1|E_1=e_1,E_2=e_2,\ldots,E_n=e_n)$ δίνεται από τη σχέση

$$P(c = c_1 | E_1 = e_1, E_2 = e_2, \ldots, E_n = e_n) = \frac{P(c = c_1) * P(E_1 = e_1 | c = c_1) * P(E_2 = e_2 | c = c_1) * \ldots * P(E_n = e_n | c = c_1)}{P(E_1 = e_1) * P(E_2 = e_2) * \ldots * P(E_n = e_n)}$$

(*βλέπε και [6])*

Οι πιθανότητες του δεύτερου μέλους υπολογίζονται από τις συχνότητες εμφάνισης της κάθε τιμής $e_1, e_2,\ldots, e_n$ στο σύνολο εκπαίδευσης. Αντίστοιχα υπολογίζεται και η $P(c_2| E_1,E_2,\ldots,E_n)$. Στη συνέχεια κάθε πιθανότητα διαιρείται με το άθροισμα $P(c_1| E_1,E_2,\ldots,E_n) + P(c_2| E_1,E_2,\ldots,E_n)$ ώστε οι δύο πιθανότητες να έχουν άθροισμα 1 (**κανονικοποίηση-normalization**). Όποια πιθανότητα είναι μεγαλύτερη, στην αντίστοιχη τιμή της μεταβλητής ταξινόμησης κατατάσσεται και το παράδειγμα. Για παράδειγμα, ακολουθεί το επόμενο σύνολο εκπαίδευσης

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Ο πίνακας με τις συχνότητες εμφάνισης των τιμών του κάθε χαρακτηριστικού φαίνεται παρακάτω

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Έστω το νέο γεγονός e = {Sunny, cool, high, true}

P(play=yes,E=e)= 9/14*2/9*3/9*3/9*3/9= 0.0053
P(play=no,E=e)=0.0206

$$P(play=yes|e)=\frac{\frac{P(play=yes,E=e)}{P(E=e)}}{\frac{P(play=yes,E=e)}{P(E=e)}+\frac{P(play=no,E=e)}{P(E=e)}}=\frac{0.0053}{0.0053+0.0206}=20.5\%$$

$$P(play=no|e)=\frac{0.0206}{0.0053+0.0206}=79.5\%$$

Στους παραπάνω υπολογισμούς υποθέσαμε ότι δεν υπάρχουν συσχετίσεις μεταξύ των χαρακτηριστικών. Για την ακρίβεια υπάρχουν στην πραγματικότητα αλλά τις αμελήσαμε. Τι σημαίνει ότι αμελούμε τις συσχετίσεις; Σημαίνει ότι αν γνωρίζουμε την τιμή ενός χαρακτηριστικού E_1, τα υπόλοιπα χαρακτηριστικά E_2,E_3,E_4 μπορούν να πάρουν οποιαδήποτε από το σύνολο των δυνατών τιμών τους δίχως να επιρρεάζει η γνώση της τιμής του E_1. Αν θεωρούσαμε τις συσχετίσεις, οι δεσμευμένες πιθανότητες υπολογίζονται από το σύνολο δεδομένων. Για παράδειγμα,

P(hot |sunny) = 2/5,
P(high |hot, sunny)= 2/2 =1,
P(false |high, hot, sunny)= 1/2

Άρα η πιθανότητα να συμβεί ένα νέο στιγμιότυπο, με συνδυασμό τιμών που δεν υπάρχει στο σύνολο εκπαίδευσης, θα ήταν μηδενική, όπως επακόλουθα και η πιθανότητα να ανήκει σε οποιαδήποτε κλάση. Ο Naïve Bayes αποφεύγει αυτό το πρόβλημα. Αδιαφορεί για τις συσχετίσεις μεταξύ των χαρακτηριστικών και το μόνο που τον νοιάζει είναι να ταξινομήσει σωστά ένα οποιοδήποτε στιγμιότυπο!

Τα καταφέρνει άραγε; Η χρήση του έχει δείξει ότι αποτελεί ένα πολύ ισχυρό εργαλείο ταξινόμησης, που σε πολλά και μεγάλα σύνολα εκπαίδευσης έχει πολύ καλή απόδοση και μάλιστα καλύτερη από εξεζητημένους αλγορίθμους ανίχνευσης προτύπων. Υπάρχουν όμως και σύνολα δεδομένων στα οποία οι αφελείς παραδοχές του οδηγούν σε μη αποδεκτά ποσοστά σφαλμάτων.

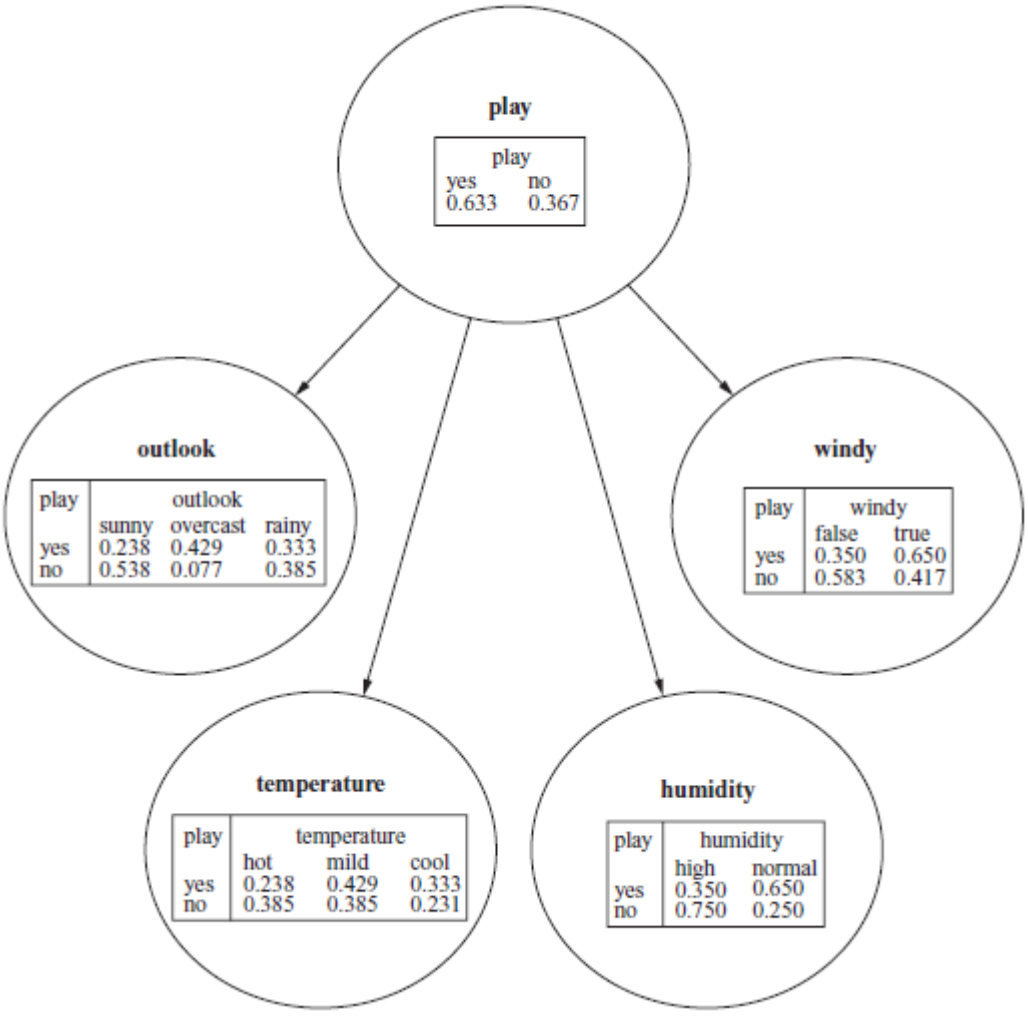
Υπάρχει όμως και ένα θέμα στην εφαρμογή του αλγορίθμου. Όπως μπορεί να παρατηρήσει κάποιος από τον πίνακα συχνοτήτων η τιμή outcast του weather δεν εμφανίζεται ποτέ με την τιμή no της μεταβλητής ταξινόμησης. Αυτό σημαίνει ότι αν outlook=overcast, η τιμή της πιθανότητας P(play=no|outlook=overcast) θα είναι μηδενική ανεξάρτητα από τις τιμές των υπόλοιπων χαρακτηριστικών. Ό,τι τιμές και να έχουν τα υπόλοιπα χαρακτηριστικά, αν outlook = overcast, δεν θα παίξουν κανένα ρόλο και η τιμή της πιθανότητας θα είναι μηδέν. Το γεγονός αυτό δε μας αρέσει καθόλου.

Επιθυμούμε λοιπόν να αλλάξουμε λίγο τους μετρητές ώστε να γίνουν περισσότερο ομοιόμορφες οι κατανομές πιθανοτήτων των στιγμιοτύπων που περιέχουν outlook=overcast. Για το σκοπό αυτό εφαρμόζουμε τη μέθοδο διόρθωσης Laplace *(Laplace estimator)*, προς τιμήν του Γάλλου μαθηματικού Laplace. Η μέθοδος προτείνει να πάμε στον πίνακα μετρητών και να αυξήσουμε όλους τους μετρητές των τιμών των χαρακτηριστικών ανά κλάση κατά 1. Έτσι, στη στήλη no του outcast οι πιθανότητες 3/5, 0/5, 1/5 γίνονται 4/8, 1/8 και 2/8 αντίστοιχα. Οι πιθανότητες 2/9, 4/9, 3/9 γίνονται 3/12, 5/12, 4/12, το ίδιο και με τις τιμές των υπολοίπων δεσμευμένων πιθανοτήτων.Ουσιαστικά αρχικοποιούμε όλους τους μετρητές στη μονάδα.

Στην πραγματικότητα υφίστανται συσχετίσεις μεταξύ των χαρακτηριστικών ενός συνόλου εκπαίδευσης με την έννοια που αναφέρθηκε παραπάνω, δηλαδή ότι οι τιμές των χαρακτηριστικών σχηματίζουν συγκεκριμένους συνδυασμούς, αυτούς του

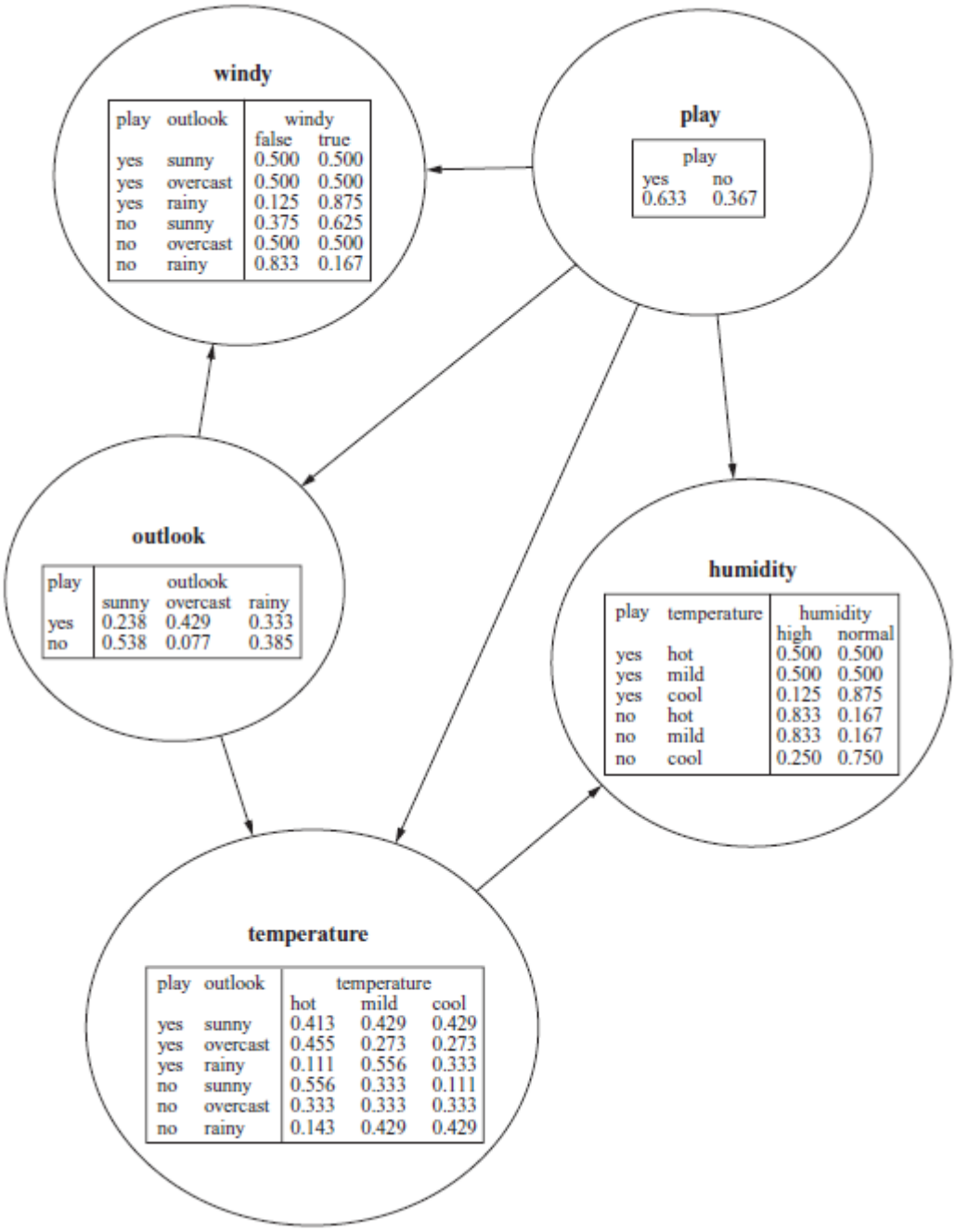
συνόλου εκπαίδευσης. Αν λοιπόν λάβουμε τελικά υπ' όψιν τις συσχετίσεις; Το σίγουρο είναι ότι ένα οποιοδήποτε στιγμιότυπο του συνόλου εκπαίδευσης θα ταξινομείται σωστά με 100% βεβαιότητα, ενώ ο Naïve Bayes μπορεί να έχει για παράδειγμα 80% βεβαιότητα (και στις δύο περιπτώσεις ο σκοπός επιτυγχάνεται, γίνεται σωστά η ταξινόμηση!). Από την άλλη, η πιθανότητα ενός νέου στιγμιοτύπου να ανήκει σε μια οποιαδήποτε κλάση είναι μηδέν επειδή η πιθανότητα εμφάνισής του είναι μηδέν, με τον Naïve Bayes όμως να κάνει κανονικά την ταξινόμηση. Προφανώς στη θεώρηση των συσχετίσεων το γεγονός της μηδενικής πιθανότητας εμφάνισης των νέων στιγμιοτύπων αποτελεί εμπόδιο (**bug**) στην ταξινόμησή τους. Για να λυθεί το πρόβλημα απλά προστίθεται μια μικρή ποσότητα στον αριθμητή και η ίδια στον παρονομαστή. Αν έχουμε μία πιθανότητα P(X|E)=0/k, απλά την κάνουμε P(X|E)=λ/(λ+k).

Ας δούμε τώρα συνολικά έναν Naïve και έναν BayesNet γράφο όπως αυτοί βγαίνουν από το σύνολο δεδομένων καιρού.



Στην εικόνα φαίνεται ένας Naïve γράφος. Οι τιμές των πιθανοτήτων έχουν προκύψει με τη διόρθωση Laplace για την αποφυγή του προβλήματος μηδενικής

εμφάνισης του outcast με την κλάση play. Παρακάτω φαίνεται ο γράφος του αλγορίθμου Bayes Net



Οι απαραίτητες διορθώσεις για την αποφυγή του προβλήματος μηδενικής συχνότητας το οφειλόμενο στη λήψη των συσχετίσεων έχουν γίνει. Για παράδειγμα η πιθανότητα P(humidity=normal|play=yes, temperature=cool) με απλή παρατήρηση των στιγμιοτύπων είναι 3/3=1 και η συμπληρωματική της 0/3=0. Οπότε απλά προσθέτουμε 0,5 στον αριθμητή της καθεμιάς, άρα συνολικά έχουμε μία δυνατή περίπτωση παραπάνω, και τελικά οι πιθανότητες γίνονται (3+0,5)/(3+1)=0.875 και (0+0.5)/(3+1)=0,125 αντιστοίχως. Να σημειωθεί σε αυτό το

σημείο ότι έχει γίνει κάποιο λάθος στις πιθανότητες του windy: Στις γραμμή yes rainy οι πιθανότητες πρέπει να τοποθετηθούν ανάποδα, το ίδιο και στη γραμμή no rainy.

Θα μπορούσε να αναρωτηθεί κάποιος για ποιο λόγο να μη θεωρούσαμε όλες τις δυνατές συσχετίσεις. Δε θα είχαμε καλύτερα αποτελέσματα; Είναι γνωστό ότι όσες περισσότερες συσχετίσεις θεωρούμε, τόσο λιγότερες είναι οι δυνατές περιπτώσεις ενός ενδεχομένου. Επειδή όμως κατά πάσα πιθανότητα θα παρουσιαστεί το πρόβλημα μηδενικής συχνότητας, τι θα ήταν προτιμότερο, να έχουμε να διορθώσουμε μια τιμή 0/2 ή μια τιμή 0/11; Στην πρώτη περίπτωση με τις περισσότερες συσχετίσεις η τιμή θα γινόταν (0+0,5)/(5+1)=0,083 οπότε θα είχαμε κατανομή 0,167-0,833 και στη δεύτερη περίπτωση (0+0,5)/(11+1)=0,0417 με κατανομή 0,0417-0,958. Δηλαδή η δεύτερη περισσότερο naïve περίπτωση δίνει καλύτερα αποτελέσματα από την πρώτη! Ιδού λοιπόν γιατί αγνοούνται κάποιοιες συσχετίσεις στο BayesNet γράφο. Τίθεται δηλαδή και ένα θέμα επιλογής του δικτύου με το βέλτιστο αριθμό συσχετίσεων ώστε ο αλγόριθμος BayesNet να είναι πράγματι καλύτερης ή ίσης δυναμικότητας με το Naïve Bayes, θέμα το οποίο δε θα αναλυθεί στην παρούσα πτυχιακή.

Μέχρι στιγμής η ταξινόμηση περιελάμβανε μόνο χαρακτηριστικά των οποίων οι τιμές είναι απλά συμβάντα και δεν αντιπροσωπεύουν ποσότητες. Τα χαρακτηριστικά αυτά καλούνται **ονομαστικά (nominal).** Nominal χαρακτηριστικό είναι το `outlook ={ sunny, overcast, rainy }`. Τα χαρακτηριστικά όμως μπορούν να παίρνουν για τιμές και αριθμούς οι οποίοι εκφράζουν ποσότητες. Τότε τα χαρακτηριστικά ονομάζονται **αριθμητικά (numeric)**. Το σύνολο δεδομένων του καιρού θα μπορούσε να δινόταν ως εξής:

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Τα χαρακτηριστικά Temperature και Humidity πλέον εκφράζονται στις μονάδες μέτρησής τους (π.χ η θερμοκρασία σε βαθμούς Κελσίου). Σε αυτήν την περίπτωση προσαρμόζεται σε κάθε χαρακτηριστικό μία κανονική κατανομή για κάθε μία από τις κλάσεις { yes, no}. Η μέση τιμή και η τυπική απόκλιση καθορίζονται από τις τιμές

των αριθμητικών χαρακτηριστικών στο σύνολο εκπαίδευσης. Έτσι μπορούμε να καταρτίσουμε τον επόμενο πίνακα

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std. dev.*	6.2	7.9	*std. dev.*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

Η πιθανότητα μιας τιμής , λόγου χάριν Temperature=83|play=yes, υπολογίζεται από το όριο

$\lim_{e \to 0} \int_{83-e}^{83+e} N(x = 83 | m = 73, s = 6,2)dx$, δηλαδή πρακτικά 2ε*N(x=83). Όταν όμως θελήσουμε να υπολογίσουμε την πιθανότητα ενός στιγμιοτύπου που να περιλαμβάνει τη συγκεκριμένη τιμή Temperature, μπορούμε να αγνοήσουμε τον όρο 2ε, να χρησιμοποιήσουμε μόνο τον όρο N(x=83), και στη συνέχεια να κάνουμε τις απαιτούμενες κανονικοποιήσεις. Έστω ότι έχουμε το παρακάτω στιγμιότυπο

Outlook	Temperature	Humidity	Windy	Play
Sunny	66	90	true	?

Είναι, $f(\text{temperature} = 66 | \text{yes}) = \frac{1}{\sqrt{2\pi}\,6,2} e^{-\frac{(66-73)^2}{2 \cdot 6,2^2}} = 0,0340$

$f(\text{humidity} = 90 | \text{yes}) = 0,0221$

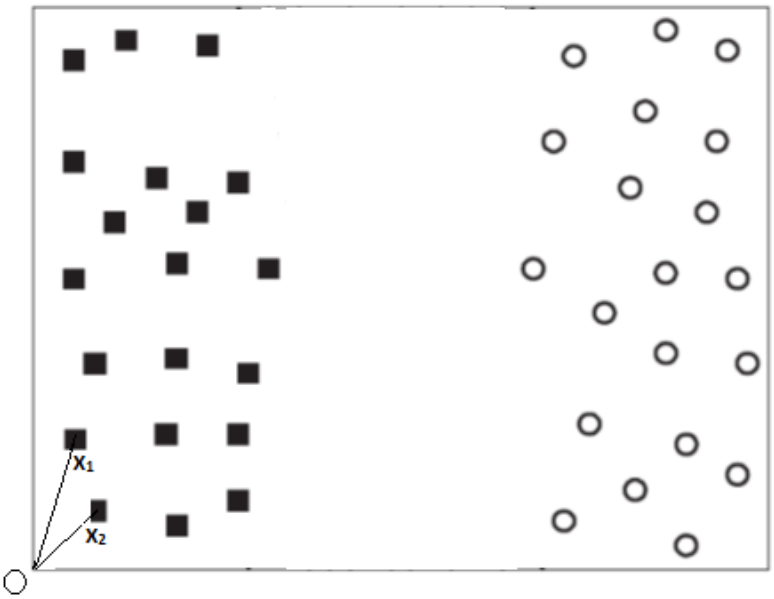
P(play=yes,E=e)= 2/9*0,0340*0,0221*3/9*9/14=0,000036

P(play=no,E=e)=3/5*0,0279*0,0381*3/5*5/14=0.000137. Κανονικοποιούμε και υπολογίζουμε την επιθυμητή κατανομή. Βέβαια, αγνοήσαμε την διόρθωση Laplace που πρέπει να γίνει σε ορισμένες nominal τιμές απλά για τις ανάγκες του παραδείγματος. Σαφέστατα η σωστή κατανομή προκύπτει ύστερα από διόρθωση Laplace.

1.6 Γραμμικές μηχανές διανυσμάτων υποστήριξης

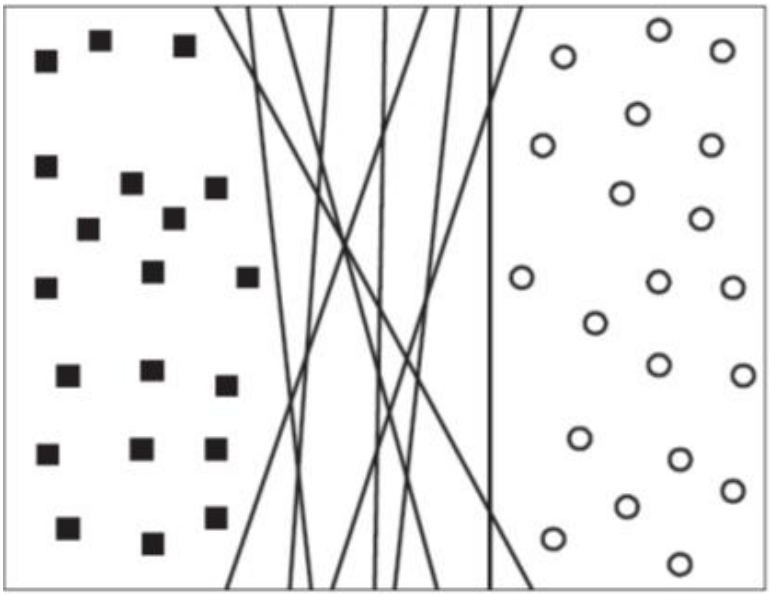
1.6.1 Όριο απόφασης και υπερεπίπεδα μεγίστου περιθωρίου

Έστω ένα διάνυσμα αριθμητικών χαρακτηριστικών $X = \{x_1, x_2,\ldots,x_n\}$ και ένα σύνολο στιγμιοτύπων με τα συγκεκριμένα χαρακτηριστικά. Κάθε στιγμιότυπο αντιστοιχεί σε ένα διάνυσμα X_i με τις συνιστώσες του να είναι αριθμοί. Ας υποθέσουμε αρχικά ότι, σε ένα πρόβλημα ταξινόμησης δύο τιμών, τα διανύσματα των στιγμιοτύπων που ανήκουν στη μία κλάση καταλήγουν σε κάποιο χώρο ενώ τα στιγμιότυπα που ανήκουν στην άλλη κλάση καταλήγουν σε άλλο χώρο διαφορετικό από τον προηγούμενο, όπως φαίνεται παρακάτω:



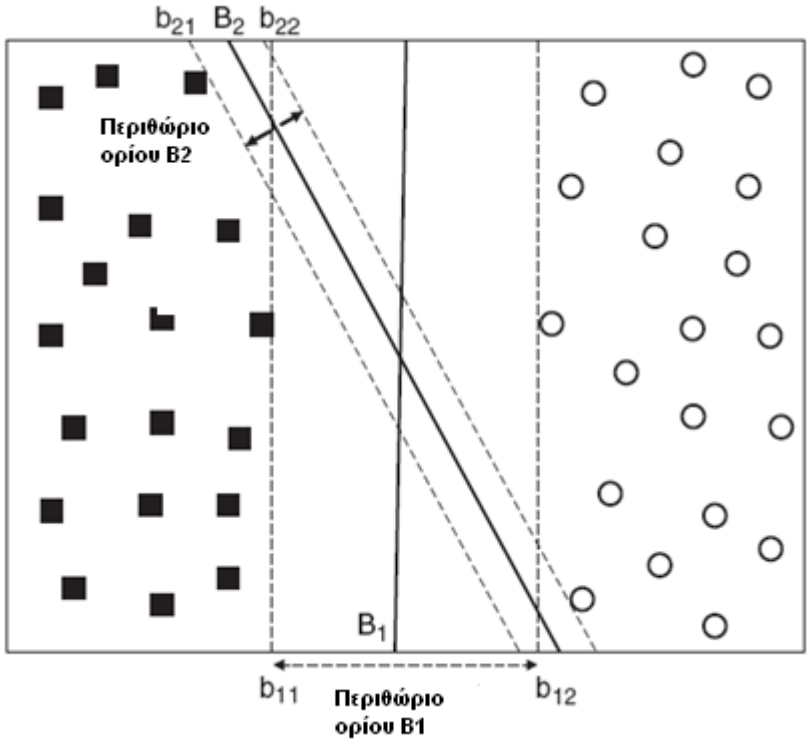
Με τετράγωνο συμβολίζονται τα στιγμιότυπα της μίας κλάσης και με κύκλο τα στιγμιότυπα της άλλης.

Η εκπαίδευση του ταξινομητή γίνεται μέσω της εύρεσης ενός γεωμετρικού ορίου το οποίο να διαχωρίζει τα στιγμιότυπα της μίας κλάσης από αυτά της άλλης.



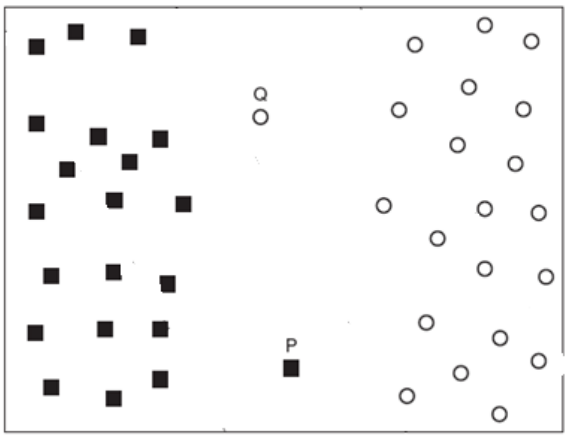
Το όριο αυτό αποτελεί ένα επίπεδο και ονομάζεται **όριο απόφασης**. Ένα νέο στιγμιότυπο ανήκει σε μία από τις δύο κλάσεις ανάλογα με τον ημιχώρο στον οποίο ανήκει. Όπως βλέπουμε, υπάρχουν πολλά δυνατά όρια απόφασης που μπορούν να βρεθούν για το δεδομένο σύνολο εκπαίδευσης. Πρέπει να επιλέξουμε το καταλληλότερο για την εκπαίδευση του ταξινομητή ώστε να ελαχιστοποιηθούν τα σφάλματα ταξινόμησης του συνόλου εξέτασης στη συνέχεια.

Κάθε όριο απόφασης συνοδεύεται από ένα ζεύγος υπερεπιπέδων. Τα υπερεπίπεδα είναι παράλληλα και συμμετρικά ως προς το επίπεδο απόφασης σε τέτοια απόσταση από αυτό ώστε οριακά να μην συμπεριλαμβάνουν κάποιο στιγμιότυπο.

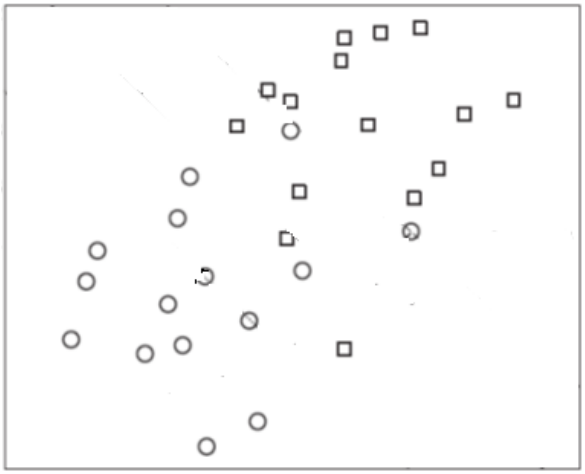


Με αυτόν τον τρόπο ορίζεται το περιθώριο του κάθε ορίου και εκτείνεται μέχρι να συναντήσει το πρώτο στιγμιότυπο εκπαίδευσης. Έχει διαπιστωθεί ότι τα όρια με τα μεγαλύτερα περιθώρια έχουν και τα καλύτερα αποτελέσματα ταξινόμησης. Ένα νέο στιγμιότυπο είναι πιθανότερο να ταξινομηθεί σωστά αν επιλεγεί όριο απόφασης με μεγάλο περιθώριο. Σε αντίθετη περίπτωση, ένα όριο με μικρό περιθώριο δεν θα έχει καλά αποτελέσματα και το εξαγόμενο μοντέλο θα πάσχει όπως λέμε από υπερπροσαρμογή. Εμείς θέλουμε το όριο να απέχει όσο το δυνατόν περισσότερο από τα στιγμιότυπα εκπαίδευσης. Συνεπώς το πρόβλημά μας ανάγεται στην εύρεση του ορίου με τα **υπερεπίπεδα μεγίστου περιθωρίου.**

Βέβαια, έως τώρα έχουμε υποθέσει ότι τα στιγμιότυπα εκπαίδευσης είναι σαφώς διαχωρίσιμα και ότι δεν υπερκαλύπτει η μία περιοχή την άλλη. Στην πράξη τα στιγμιότυπα μπορεί να έχουν την παρακάτω εικόνα



όπου τα Q, P αποτελούν θόρυβο και αναγκαστικά θα γίνει διαχωρισμός με μικρό περιθώριο ή ακόμα χειρότερα



όπου υπάρχει αρκετά μεγάλος βαθμός αβεβαιότητας και δεν είναι δυνατός ο διαχωρισμός χωρίς σφάλματα. Αρχικά θα αναφερθούμε στη διαδικασία εύρεσης του ορίου με το μέγιστο περιθώριο στη διαχωρίσιμη περίπτωση και στη συνέχεια με μερικές τροποποιήσεις θα μεταβούμε και στη μη διαχωρίσιμη περίπτωση.

1.6.2 Γραμμική μηχανή διανυσμάτων υποστήριξης: Διαχωρίσιμη περίπτωση

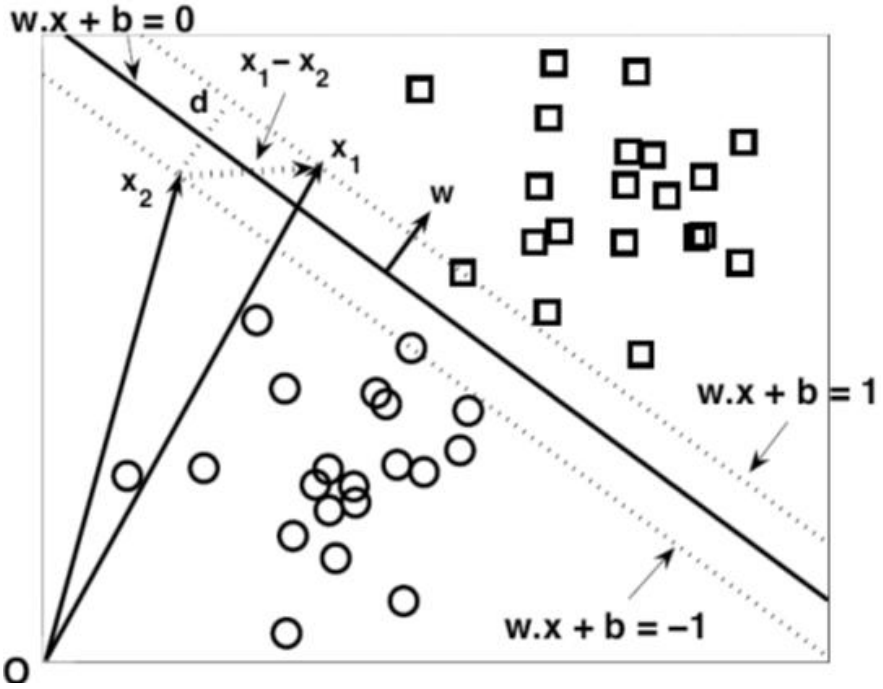
Η γραμμική μηχανή διανυσμάτων υποστήριξης (**linear SVM**) είναι ένας ταξινομητής ο οποίος αναζητά το όριο απόφασης μεγίστου περιθωρίου, γι'αυτό και είναι γνωστός ως ταξινομητής μεγίστου περιθωρίου.

Γραμμικό όριο απόφασης: Έστω το διάνυσμα χαρακτηριστικών $X=\{x_1, x_2, \dots, x_n\}$ και το σύνολο των δύο κλάσεων $y=\{1, -1\}$. Κάθε στιγμιότυπο του X ανήκει στην κλάση 1 ή στην κλάση -1. Το γραμμικό όριο απόφασης μπορεί να γραφεί στην ακόλουθη μορφή

$$\mathbf{w}X + b = 0 \quad (1),$$

όπου **w** και b είναι οι παράμετροι του μοντέλου.

Στο ακόλουθο σχήμα φαίνεται ένα σύνολο εκπαίδευσης δύο διαστάσεων.



Κάθε στιγμιότυπο που βρίσκεται πάνω στο όριο απόφασης πρέπει να ικανοποιεί την ανωτέρω εξίσωση. Για κάθε τετράγωνο $\mathbf{x_s}$ το οποίο είναι πάνω από το όριο απόφασης ισχύει ότι

$$\mathbf{w}\cdot\mathbf{x}_s + b = k \quad (2),$$

όπου k > 0. Ομοίως, για κάθε κύκλο $\mathbf{x_c}$ κάτω από το όριο θα ισχύει ότι

$$\mathbf{w}\cdot\mathbf{x}_c + b = k' \quad (3),$$

όπου k’ < 0. Αν δώσουμε στα τετράγωνα την κλάση 1 και στους κύκλους την κλάση -1, τότε για κάθε στιγμιότυπο εξέτασης **z** η κλάση μπορεί να υπολογιστεί ως ακολούθως:

$$y = \begin{cases} 1, & \text{αν } \mathbf{w}\cdot\mathbf{z} + b > 0 \\ -1, & \text{αν } \mathbf{w}\cdot\mathbf{z} + b < 0 \end{cases} \quad (4)$$

Περιθώριο ενός γραμμικού ταξινομητή: Ένας κύκλος λοιπόν που βρίσκεται κάτω από το όριο απόφασης ικανοποιεί την εξίσωση (2) στην οποία έχουμε k > 0 ενώ ένα τετράγωνο πάνω από το όριο την εξίσωση (3) όπου k’ < 0. Μπορούμε κανονικοποιήσουμε τις παραμέτρους **w** και b ώστε τα δύο παράλληλα υπερεπίπεδα b_1 και b_2 να εκφράζονται ως ακολούθως:

$$\beta_1: \quad \mathbf{w}\cdot\mathbf{x} + b = 1 \quad (5)$$

$$\beta_2: \quad \mathbf{w}\cdot\mathbf{x} + b = -1 \quad (6)$$

Το περιθώριο του ορίου απόφασης δίνεται από την απόσταση των δύο υπερεπιπέδων. Αν $\mathbf{x_1}$ σημείο του β_1 και $\mathbf{x_2}$ σημείο του β_2, αφαιρώντας την (6) από την (5) έχουμε

$$\mathbf{w}(\mathbf{x}_1 - \mathbf{x}_2) = 2,$$

παίρνουμε τα μέτρα,

$$\|\mathbf{w}\| d = 2$$

$$d = \frac{2}{\|\mathbf{w}\|} \quad (7)$$

Μετά και την απαίτηση της κανονικοποίησης των παραμέτρων w, b ώστε να πάρουμε τις ανωτέρω εκφράσεις των υπερεπιπέδων, θα πρέπει να ισχύουν οι ακόλουθες συνθήκες για κάποιο i στιγμιότυπο του συνόλου εκπαίδευσης:

$$\mathbf{w}\mathbf{x}_i + b \geq 1, \text{ an } y_i = 1 \quad (7)$$

$$\mathbf{w}\mathbf{x}_i + b \leq -1, \text{ an } y_i = -1 \quad (8),$$

δηλαδή τα στιγμιότυπα που ανήκουν στην κλάση 1 θα πρέπει να βρίσκονται στο υπερεπίπεδο b_1 ή πάνω από αυτό και τα στιγμιότυπα της -1 στο υπερεπίπεδο b_2 ή κάτω από αυτό. Οι δύο αυτές ανισότητες συμπυκνώνονται στη μορφή

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N$$

Τελικά το πρόβλημά μας ανάγεται στο ακόλουθο **πρόβλημα βελτιστοποίησης με περιορισμούς:**

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2}$$

ώστε $y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N$

Θα το επιλύσουμε με τη βοήθεια των **πολλαπλασιαστών Lagrange**. Έστω η αντικειμενική συνάρτηση $L_p(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \lambda_i (y_i(\mathbf{w}\mathbf{x}_i + b) - 1)$ (9), η οποία πρέπει να ελαχιστοποιηθεί. Οι πολλαπλασιαστές Lagrange είναι οι λ_i. Θέλουμε

$$\frac{\partial L_p}{\partial w_k} = 0 \text{ ή } w_k = \sum_{i=1}^{N} \lambda_i y_i x_i \quad (10)$$

$$\frac{\partial L_p}{\partial b} = 0 \text{ ή } \sum_{i=1}^{N} \lambda_i y_i = 0 \quad (11)$$

Παρ'όλα αυτά, λόγω των ανισοτικών περιορισμών, οι εξισώσεις δεν επαρκούν για την επίλυση του συστήματος. Αν αντί για ανισότητες είχαμε ισότητες, θα είχαμε N εξισώσεις περιορισμών και τις δύο ανωτέρω σχέσεις, αρκετά για την επίλυση του συστήματος. Γι'αυτό το λόγο οι περιορισμοί ανισοτήτων μετασχηματίζονται ως εξής:

$\lambda_i \geq 0$ (12)

$\lambda_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0$ (13) *Συνθήκες Karush-Kuhn-Tucker(KKT)*

Ουσιαστικά οι παραπάνω συνθήκες μας λένε ότι ένα στιγμιότυπο εκπαίδευσης βρίσκεται πάνω σε ένα από τα δύο υπερεπίπεδα αν $\lambda_i > 0$ ή εκτός των δύο υπερεπιπέδων αν $\lambda_i = 0$. Τα στιγμιότυπα εκπαίδευσης με $\lambda_i > 0$ ονομάζονται **διανύσματα υποστήριξης**. Τελικά, οι συνθήκες που πρέπει να ισχύουν για να ελαχιστοποιείται η (9) είναι οι (10), (11), (12), (13).

Απ'την (13) έχουμε $\sum_{i=1}^{N} \lambda_i \cdot (y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1) = 0$ ή $\sum_{i=1}^{N} \lambda_i \cdot y_i \cdot \mathbf{x}_i \cdot \mathbf{w} = 0$. Από (10) έχουμε $\mathbf{w} \cdot \mathbf{w} = 0$, δηλαδή $\|\mathbf{w}\|^2 = 0$.

Στη συνέχεια αντικαθιστούμε τις σχέσεις (10), (11) στην (9) και μπορούμε να πάρουμε μια έκφραση της L_P περιέχουσα μόνο τους πολλαπλασιαστές Lagrange. Έτσι θα έχουμε

$$L_D(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \sum_{ij} \lambda_i \lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j \quad (14)$$

Τα x_i και x_j είναι τα δεδομένα εκπαίδευσης. Πλέον επειδή ο δεύτερος όρος είναι αρνητικός το πρόβλημα Lagrange μετατρέπεται σε ένα πρόβλημα μεγιστοποίησης, το δυαδικό της ελαχιστοποίησης. Όταν τελικά βρεθούν τα λ_i χρησιμοποιούνται οι (10) και (13) για να βρεθούν οι επιτρεπτές λύσεις των w,b.

1.6.3 Γραμμική μηχανή διανυσμάτων υποστήριξης: Μη Διαχωρίσιμη περίπτωση

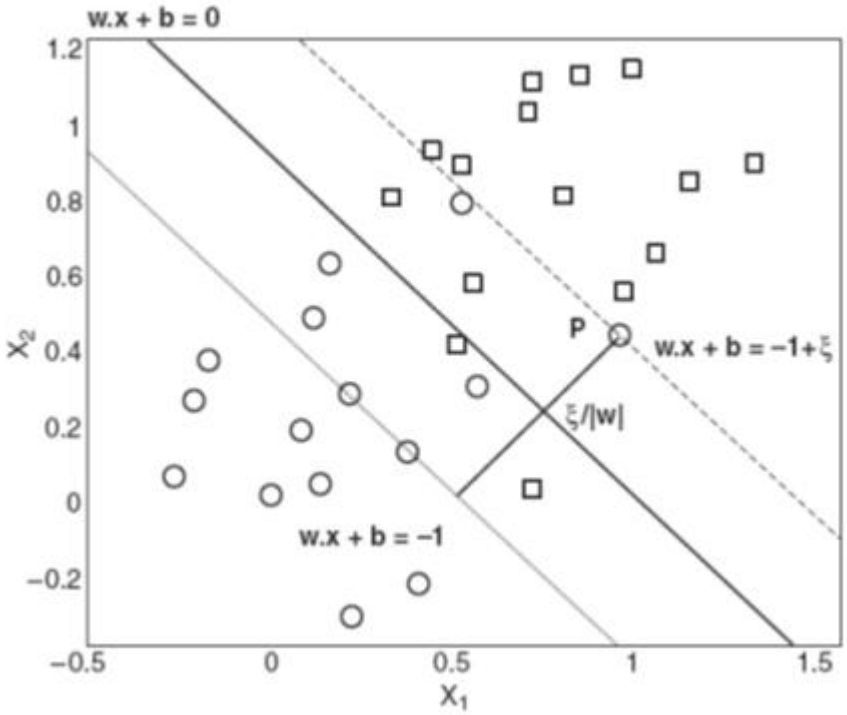
Στις περισσότερες περιπτώσεις τα δεδομένα εκπαίδευσης δεν είναι διαχωρίσιμα. Οι δύο ομάδες των στιγμιοτύπων δε είναι ξένες αλλά υπερκαλύπτουν η μία την άλλη. Πρέπει να εκφράσουμε λοιπόν σε λιγότερο αυστηρή μορφή τους ανισοτικούς περιορισμούς και να δείξουμε κάποια ανοχή στα σφάλματα διαχωρισμού επειδή δεν μπορούμε να κάνουμε αλλιώς. Οι περιορισμοί ανισότητας γίνονται

$\mathbf{w} \cdot \mathbf{x}_i + b \geq 1 - \xi_i$,αν $y_i = 1$ (15)

$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 + \xi_i$, αν $y_i = -1$ (16)

όπου για κάθε i: $\xi_i > 0$.

Οι ξ_i ονομάζονται **χαλαρές μεταβλητές**. Ένα παράδειγμα χαλαρής μεταβλητής φαίνεται παρακάτω:



Προφανώς η μεταβλητή ξ δεν μπορεί να είναι αυθαίρετα μεγάλη, εφόσον θα έχουμε μεγάλο αριθμό σφαλμάτων διαχωρισμού. Θέλουμε λοιπόν να ορίσουμε κάποιο μέτρο ποινής το οποίο να αυξάνει όσο το ξ είναι μεγαλύτερο από όσο πρέπει. Ένας απλός τρόπος είναι να τροποποποιήσουμε την αντικειμενική συνάρτηση $f(w)= \|w\|^2 / 2$ καταλλήλως. Η νέα συνάρτηση f(w) θα είναι

$$f(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + C\left(\sum_{i=1}^{N} x_i\right)^k,$$

όπου τα C, k καθορίζονται από το χρήστη, δηλαδή ο χρήστης επιλέγει πόση ποινή θα επιβάλλεται σε κάθε ευρισκόμενη ξ_i από την επίλυση του συστήματος στη συνέχεια.

Η συνάρτηση Lagrange γράφεται λοιπόν για k=1

$$L_p(\mathbf{w}, b, \boldsymbol{\lambda}, \mathbf{x}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \lambda_i [y_i(\mathbf{w}\mathbf{x}_i + b) - 1 + x_i] - \sum_{i=1}^{N} \mu_i x_i \quad (17)$$

Ο τελευταίος όρος εκφράζει το ότι απαγορεύεται να είναι αρνητικά τα ξ ($\mu_i >= 0$). Οι συνθήκες Karush - Kuhn – Tucker γίνονται

$$x_i \geq 0,\ \lambda_i \geq 0,\ \mu_i \geq 0, \quad (18)$$

$$\lambda_i [y_i (w \cdot x_i + b) - 1 + x_i] = 0, \quad (19)$$

$$\mu_i x_i = 0 \quad (20)$$

Επίσης,

$$\frac{\partial L_p}{\partial w_j} = 0 \Rightarrow w_j = \sum_{i=1}^{N} \lambda_i y_i x_{ij} \quad (21)$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow -\sum_{i=1}^{N} \lambda_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0 \quad (22)$$

$$\frac{\partial L_p}{\partial x_i} = C - \lambda_i - \mu_i = 0 \Rightarrow \lambda_i + \mu_i = C \quad (23)$$

Αντικαθιστούμε τις (21), (22), (23) στην Lagrange και έχουμε την ακόλυθη δϋική συνάρτηση

$$\begin{aligned} L_D &= \frac{1}{2} \sum_{ij} \lambda_i \lambda_j y_i y_j x_i x_j + C \sum_i x_i \\ &\quad - \sum_{i=1} \lambda_i y_i \left(\sum_j \lambda_j y_j x_i x_j + b \right) - 1 + x_i] \\ &\quad - \sum_i (C - \lambda_i) x_i \\ &= \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i x_j \end{aligned} \quad (24)$$

Από την 23 βλέπουμε ότι λ<=C, άρα 0<= λ <=C στα μη διαχωρίσιμα δεδομένα. Τέλος, χρησιμοποιούνται οι σχέσεις (19), (21) για να βρεθούν τα επιτρεπτά w, b.

ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΕΦΑΛΑΙΟΥ 1

[1] Ιωάννης Βλαχάβας, Πέτρος Κεφάλας, Νικόλαος Βασιλειάδης, Φώτης Κόκκορας, Ηλίας Σακελλαρίου. Τεχνητή Νοημοσύνη, 3η έκδοση.

[2] Ian Witten, Eibe Frank, Mark Hall. Data Mining 3rd edition

[3] D. Koller and N. Friedman (2009). Probabilistic Graphical Models: Principles and Techniques. edited by . MIT Press.

[4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Εισαγωγή στην εξόρυξη δεδομένων.

[5] Aji, S. M. and R. J. McEliece (2000). The generalized distributive law. *IEEE Trans. Information Theory 46,* 325-343.

[6] H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI),* pp. 21-30-*σελίδες 4,5*
[7] Barash, Y. and N. Friedman (2002). Context-specific Bayesian clustering for gene expression data, Journal of Computational Biology 9,*169-191.*
[8] Bernardo, J. and A. Smith (1994). *Bayesian Theory.* New York: John Wiley and Sons.
[9] Becker, A. and D. Geiger (1994). The loop cutset problem. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI),* pp. 60-68
[10] Becker, A., R. Bar-Yehuda, and D. Geiger (1999). Random algorithms for the loop cutset problem. *In* Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI), *pp. 49-56.*
[11] Bidyuk, B. and R. Dechter (2007). Cutset sampling for bayesian networks. *Journal of Artificial Intelligence Research 28,*1-48.
[12] *New Complexity Results for MAP in Bayesian Networks, Cassio P. de Campos*
[13] Beinlich, L., H. Suermondt, R. Chavez, and G. Cooper (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine,* pp. 247-256. Springer Verlag.
[14] Besag, J. (1977b). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36,*192-236.
[15] Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Information Science and Statistics (M. Jordan, J. Kleinberg, and B. Schokopf, editors). New York: Springer-Verlag.
[16] Bouckaert, R. (1993). Probabilistic network construction using the minimum description length principle. In *Proc. European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty,* pp. 41-48.
[17] Casella, G. and R. Berger (1990). *Statistical Inference.* Wadsworth.
[18]Chan, H. and A. Darwiche (2002). When do numbers really matter? *Journal of Artificial Intelligence Research 17,* 265-287.
[19] Cheeseman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman (1988). Autoclass: a Bayesian classification system. In *Proc. 5th International Conference on Machine Learning (ICML)*
[20]Chickering, D., D. Geiger, and D. Heckerman (1995, January). Learning Bayesian networks: Search methods and experimental results. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics,* pp. 112-128.
[21] hickering, D. M., D. Heckerman, and C. Meek (1997). A Bayesian approach to learning Bayesian networks with local structure. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI),* pp. 80-89.
[22]Cooper, G. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning 9,* 309-347.
[23] Cowell, R. (2005). Local propagation in conditional gaussian Bayesian networks. *Journal of Machine Learning Research 6,*1517-1550.

[24] Dawid, A. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B 41,*1-31.
[25] Dean, T. and K. Kanazawa (1989). A model for reasoning about persistence and causation. *Computational Intelligence* 5(3), 142-150.
[26] Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4), 309–347

ΚΕΦΑΛΑΙΟ 2: ΕΞΟΡΥΞΗ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΚΕΙΜΕΝΑ

Η πληθώρα ψηφιακών υπολογιστικών συσκευών και η χρήση τους στις επικοινωνίες έχει ως αποτέλεσμα την αυξανόμενη ζήτηση σε συστήματα και αλγορίθμους ικανούς για εξόρυξη πληροφορίας από κείμενα. Επομένως, η ανάπτυξη τεχνικών για εξόρυξη μπηροφορίας από μη δομημένα, ημιδομημένα και πλήρως δομημένα κείμενα έχει γίνει πολύ σημαντική στην ακαδημαϊκή κοινότητα και στη βιομηχανία. Σκοπός είναι η αναζήτηση προτύπων και συσχετίσεων, δηλαδή η ταξινόμηση των κειμένων ανάλογα με το θέμα τους σε πεπερασμένο αριθμό κατηγοριών ή κλάσεων (**μάθηση με επίβλεψη**), είτε η ομαδοποίηση των κειμένων ανάλογα με το θέμα τους χωρίς να έχουμε ορίσει αριθμό κλάσεων (**μάθηση χωρίς επίβλεψη**).

Ένα σύνολο δεδομένων αποτελούμενο από διαφορετικά κείμενα και τις ταξινομήσεις τους εισάγεται σε έναν αλγόριθμο μάθησης χωρίς επίβλεψη όπως στην ανωτέρω περίπτωση των πέντε πελατών της τράπεζας, με σκοπό την αναγνώριση προτύπου. Ποια μορφή όμως έχουν τα κείμενα; Σίγουρα όχι την αρχική μορφή τους, αλλά μία αφαιρετική αναπαράστασή της η οποία αποτελείται από **λέξεις-κλειδιά (keywords)**. Μέσω των λέξεων-κλειδιών, οι οποίες βρίσκονται αυτόματα[6], μπορούμε να συγκεντρώσουμε τα ουσιώδη σημεία ενός κειμένου που χρειαζόμαστε για την ταξινόμηση και να απορρίψουμε τις άχρηστες λεπτομέρειες. Ταυτόχρονα πετυχαίνουμε και μια πολύ πιο συνεπτυγμένη αναπαράσταση του κειμένου. Η εξαγωγή λέξεων-κλειδιών από ένα κείμενο μπορεί να γίνει αυτόματα.

2.1 Αυτόματη εξαγωγή λέξεων-κλειδιών

Έστω το παρακάτω κείμενο:

Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.

Λέξεις-κλειδιά που εξάγονται από κάποιον αναγνώστη

linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets

Ας δούμε τώρα τον τρόπο αυτόματης εξαγωγής λέξεων κλειδιών

Υποψήφιες λέξεις-κλειδιά

Αρχικά, το κείμενο μετατρέπεται σε ένα διάνυσμα λέξεων και φράσεων το οποίο περιέχει τα ουσιαστικά και τους επιθετικούς προσδιορισμούς ή μετοχές του

κειμένου. Τα υπόλοιπα μέρη του λόγου (προθέσεις, άρθρα, ρήματα) απορρίπτονται επειδή δεν περιέχουν την ουσία της πληροφορίας που μας ενδιαφέρει [3]. Έτσι έχουμε το παρακάτω διάνυσμα:

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

Στη συνέχεια δημιουργείται ένας τετραγωνικός πίνακας (ΕΙΚΟΝΑ 1) κάθε γραμμή και στήλη του οποίου αντιπροσωπεύει μία λέξη του κειμένου. Κάθε μη διαγώνιο στοιχείο αντιπροσωπεύει μία φράση και πόσες φορές αυτή εμφανίζεται. Τα διαγώνια στοιχεία αντιπροσωπεύουν πόσες φορές εμφανίζεται η κάθε λέξη μεμονωμένα. Όπου δεν υπάρχει τιμή σημαίνει ότι ο συγκεκριμένος συνδυασμός λέξεων δεν σχηματίζει φράση που να υπάρχει στο κείμενο. Ορίζουμε τα ακόλουθα μεγέθη:

α)συχνότητα λέξης (freq(w))
β) βαθμός λέξης (deg(w))
γ)lλόγος βαθμού προς συχνότητας(deg(w)/freq(w))

Παρακάτω φαίνονται τα μεγέθη για κάθε μία από τις λέξεις. Γενικά , ο deg(w) έχει μεγαλύτερη τιμή για τις λέξεις που παρατηρούνται συχνά μόνες τους αλλά και μέσα σε φράσεις[8]. Πρόκειται για το άθροισμα των τιμών κάθε γραμμής. Για παράδειγμα, deg(minimal)>deg(systems). Λέξεις που παρατηρούνται συχνά χωρίς να λαμβάνονται υπ’όψιν οι φράσεις στις οποίες συμμετέχουν έχουν μεγαλύτερο freq(w): freq(systems)>freq(minimal). Πρόκειται για τις τιμές των διαγωνίων στοιχείων. Επομένως, λέξεις οι οποίες εμφανίζονται περισσότερο σε φράσεις παρά μόνες τους έχουν μεγάλο λόγο deg(w)/freq(w):deg(diophantine)/freq(diophantine)>deg(linear)/freq(linear).

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
algorithms	2						1																		
bounds		1																							1
compatibility			2																						
components				1																					
constraints					1								1												
constructing						1																			
corresponding	1						1																		
criteria								2																	
diophantine									1	1			1												
equations									1	1			1												
generating											1			1				1							
inequations												2			1						1				
linear					1				1	1			2												
minimal											1			3				2	1			1			
natural															1		1								
nonstrict												1				1									
numbers															1		1								
set														2				3				1			
sets											1			1					1						
solving																				1					
strict												1									1				
supporting														1				1				1			
system																							1		
systems																								4	
upper		1																							1

ΕΙΚΟΝΑ 1: Πίνακας με τις υποψήφιες λέξεις- κλειδιά και τις συχνότητές τους

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

ΕΙΚΟΝΑ 2: Η συχνότητα, ο βαθμός και ο λόγος των δύο μεγεθών για κάθε λέξη χωριστά.

Τελική επιλογή λέξεων-κλειδιών[6]

Μετά τον υπολογισμό των μεγεθών freq(w), deg(w) και deg(w)/freq(w) για κάθε λέξη, επιλέγονται οι λέξεις με τις μεγαλύτερες T τιμές ως λέξεις κλειδιά του κειμένου. Έχει καθοριστεί (Mihalcea, Tarau, 2004) ο T να ισούται με το 1/3 του αριθμού των υποψηφίων λέξεων-κλειδιών. Στην περίπτωσή μας T=28/3 δηλαδή 9. Παρακάτω (ΕΙΚΟΝΑ 3) φαίνεται η σύκριση του «χειροκίνητου» τρόπου εύρεσης

λέξεων-κλειδιών με τον αυτόματο. Από τις εννιά λέξεις που βγήκαν με τον αλγόριθμο έξι ταιριάζουν με αυτές που βγάλαμε με το χέρι. Αν και η φράση *natural numbers* είναι παρόμοια με την *set of natural numbers*, εκλαμβάνεται ως διαφορετική. Άρα έχουμε επιτυχία 6/9 δηλαδή 67%. Αν θεωρήσουμε ως σύνολο μόνο τις λέξεις που βγάλαμε με το χέρι τότε πετυχαίνουμε ένα ποσοστό 6/7 δηλαδή 86%.

Αυτόματη εξαγωγή	*Εξαγωγή με το χέρι*
minimal generating sets	minimal generating sets
linear diophantine equations	linear Diophantine equations
minimal supporting set	
minimal set	
linear constraints	linear constraints
natural numbers	
strict inequations	strict inequations
nonstrict inequations	nonstrict inequations
upper bounds	upper bounds
	set of natural numbers

ΕΙΚΟΝΑ 3: Σύγκριση των δύο τρόπων εξαγωγής λέξεων-κλειδιών.

2.2 Τελική αναπαράσταση κειμένου

Μετά την τελική επιλογή των όρων κάθε κείμενο κωδικοποιείται ως ένα αριθμητικό διάνυσμα του οποίου τα στοιχεία αντιστοιχούν στους ανακτηθέντες όρους. Κάθε όρος σχετίζεται με ένα τοπικό και συνολικό βάρος το οποίο αντιπροσωπεύει τη σημαντικότητα του όρου στο κείμενο και στον κορμό αντιστοίχως. Δοθέντος ενός όρου t και ενός κειμένου d, το βάρος του όρου ορίζεται ως εξής

$$w_{t,d}=\log(1+tf_{t,d})\log(|D|/df_t)$$

όπου $tf_{t,d}$ είναι η συχνότητα όρων (tf) του t sto d, df, t είναι η συχνότητα κειμένων (df) του t, ή ο αριθμός των κειμένων σε μια συλλογή D που περιέχει τον t, και |D| είναι το μέγεθος της συλλογής. Ο δεύτερος όρος του γινομένου είναι η αντίστροφη συχνότητα κειμένων (idf-inverse document frequency) του t. Ο συγκεκριμένος τρόπος απόδοσης βάρους παράγει καλά αποτελέσματα ταξινόμησης.

2.3 Λανθάνουσα σημασιολογική δεικτοδότηση

Πριν αναφέρουμε τεχνικές μάθησης χωρίς επίβλεψη οι οποίες εφαρμόζονται σε ένα σύνολο εκπαίδευσης, ακολουθείται μία διαδικασία. Έστω ο παρακάτω πίνακας, ο οποίος αναπαριστά τις λέξεις έξι κειμένων. Η αναπαράσταση γίνεται σε δυαδική μορφή (1 αν η λέξη ανήκει σε κάποιο κείμενο, 0 αν δεν ανήκει) και όχι με την

τεχνική των βαρών που περιγράφηκε προηγουμένως. D={d1, d2, d3, d4, d5, d6} το σύνολο των κειμένων.

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

Ορίζουμε ως κριτήριο ομοιότητας δύο κειμένων το **συνημίτονο της γωνίας** των διανυσμάτων τους, υπολογιζόμενο με τη βοήθεια του εσωτερικού γινομένου. Για παράδειγμα, έστω d1=[1 0 1 1 0] και d2=[0 1 1 0 0]. Έχουμε $\vec{d1}\cdot\vec{d2} = d1 * d2^T = 1$ και $\cos(\angle(\vec{d1},\vec{d2})) = 1/\sqrt{6}$. Όσο μικρότερη είναι η γωνία δύο διανυσμάτων, τόσο περισσότερο "μοιάζουν" μεταξύ τους.

Παρατηρούμε ότι ενώ τα διανύσματα d2 και d3 έχουν παρόμοια θεματολογία εφόσον περιέχουν τις συνώνυμες λέξεις ship και boat, η γωνία των δύο διανυσμάτων είναι 90^o. Άρα τα d2 και d3 κείμενα εκλαμβάνονται ως ανόμοια. Υπάρχει άλλος τρόπος να αναγνωριστεί μαθηματικά η συσχέτισή τους;

Στο σύνολο των έξι κείμενων εφαρμόζεται μια τεχνική ονομαζόμενη **Singular Value Decomposition (SVD)**, με τη βοήθεια της οποίας επιτυγχάνεται η ανακάλυψη λανθανουσών συσχετίσεων μεταξύ κειμένων **(λανθάνουσα σημασιολογική δεικτοδότηση, Latent Semantic Indexing-LSI)** [9], μη εντοπιζόμενες με κριτήριο τις μεταξύ των διανυσμάτων γωνίες. Ακολουθεί η περιγραφή της SVD και εν συνεχεία ο τρόπος που τη χρησιμοποιούμε για να κάνουμε LSI.

Singular Value Decomposition (SVD)

Ένας οποιοσδήποτε πίνακας A διαστάσεων M΄ N (M<>N) γράφεται στη μορφή

$$A = U S V^T$$

U: πίνακας με στήλες τα ιδιοδιανύσματα του $\mathbf{AA^T}$.

V: πίνακας με στήλες τα ιδιοδιανύσματα του $\mathbf{A^TA}$.

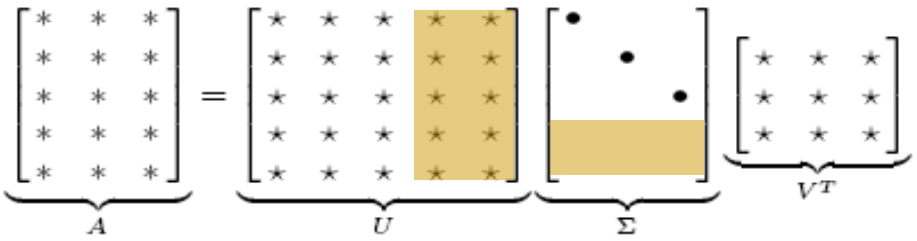
Σ: πίνακας με την τετραγωνική ρίζα των κοινών ιδιοτιμών των AA^T και A^TA. Αν M>N (εικόνα 3) οι ιδιοτιμές του A^TA είναι υποσύνολο των ιδιοτιμών του AA^T και το αντίστροφο αν M<N (εικόνα 4). Το πλήθος των κοινών ιδιοτιμών (λαμβάνοντας και την πολλαπλότητα υπ'όψιν) ισούται με *min(M,N)*. Υπενθυμίζουμε ότι *rank (A) <= min(M,N)*

Ο Σ αποτελείται από τον διαγώνιο πίνακα με τιμές $\sigma_i : s_i = \sqrt{l_i}, \text{diag}(s_i)$ όπως προαναφέρθηκε και επαυξάνεται από γραμμές ή στήλες

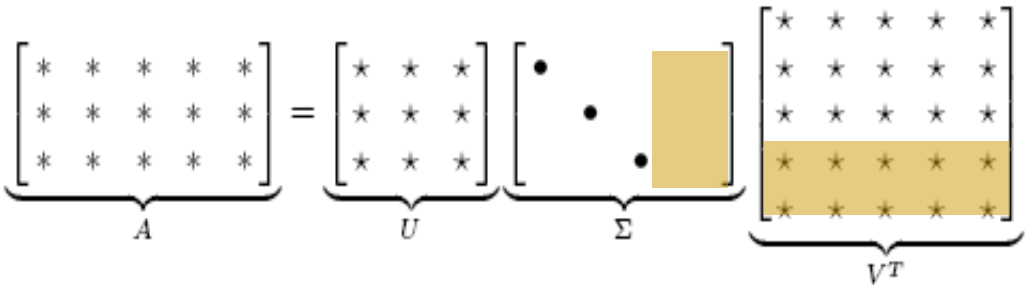
με μηδενικά, ανάλογα με το αν M>N ή M<N αντίστοιχα (σκιασμένα τμήματα του Σ στις παρακάτω εικόνες). Ο Σ έχει διαστάσεις MxN

Στην εικόνα 3 (M>N) τα σκιασμένα ιδιοδιανύσματα αντιστοιχούν στις μη κοινές ιδιοτιμές των πινάκων AA^T και A^TA, αντιπροσωπευόμενες από τις δύο σκιασμένες μηδενικές γραμμές του Σ.

Στην εικόνα 4 (M<N) τα σκιασμένα ιδιοδιανύσματα αντιστοιχούν στις μη κοινές ιδιοτιμές των πινάκων AA^T και A^TA, αντιπροσωπευόμενες από τις δύο σκιασμένες μηδενικές στήλες του Σ.



ΕΙΚΟΝΑ 3



ΕΙΚΟΝΑ 4

Η απόδειξη παραλεἰ πεται.

Ακολουθεί παράδειγμα. Έστω

$$A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Παρατηρούμε ότι M=3, N=2. Η SVD του είναι

$$\begin{bmatrix} 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

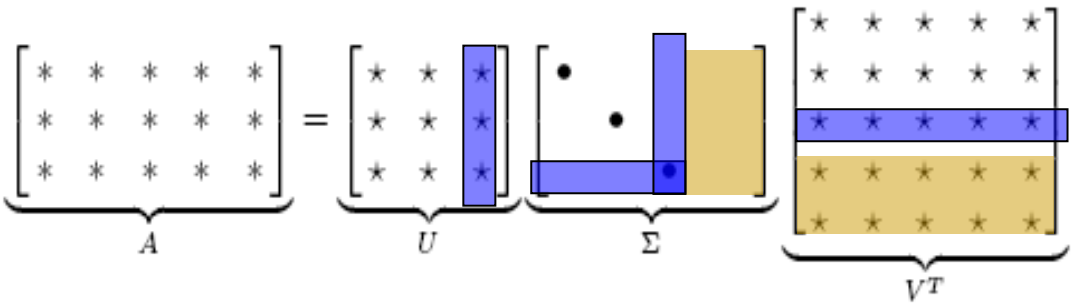
Τα σκιασμένα τμήματα της εικόνας 3 μπορούν να παραληφθούν και η παραγοντοποίηση του A να πάρει τη μορφή

$$A = \begin{bmatrix} 0 & 2/\sqrt{6} \\ 1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{3} \end{bmatrix} * \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Προσέγγιση ελαττωμένου βαθμού (Low-rank approximation)

Θα δούμε τώρα πώς αξιοποιούμε την SVD με σκοπό την LSI. Αυτό που κάνουμε είναι να πάρουμε την SVD μορφή του A και να παράγουμε έναν πίνακα Ακ βαθμού κ (συνήθως επιλέγουμε κ<<r, όπου r ο βαθμός του πίνακα A) ο οποίος θα έχει μηδενισμένες τις r-κ μικρότερες ιδιοτιμές του Σ και τα αντίστοιχα ιδιοδιανύσματα. Στο παρακάτω σχήμα λόγου χάριν μηδενίστηκε η μικρότερη από τις τρεις με τα αντίστοιχα ιδιοδιανύσματα.

$$A_k = U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_k,0,\ldots,0)\,V^T$$



Παράδειγμα:

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

Η SVD του C αποτελείται από τους παρακάτω πίνακες:

U	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
wood	−0.70	0.35	0.15	−0.58	0.16
tree	−0.26	0.65	−0.41	0.58	−0.09

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0.00	0.00	0.58	0.00	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

Επιλέγουμε να κρατήσουμε τις κ=2 μεγαλύτερες ιδιοτιμές. Ο πίνακας C2 αποτελείται από το γινόμενο των παρακάτω πινάκων

U	1	2	3	4	5
ship	−0.44	−0.30	0.00	0.00	0.00
boat	−0.13	−0.33	0.00	0.00	0.00
ocean	−0.48	−0.51	0.00	0.00	0.00
wood	−0.70	0.35	0.00	0.00	0.00
tree	−0.26	0.65	0.00	0.00	0.00

Σ_2	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00

V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00

Στη συνέχεια φαίνεται η σύγκριση του C με τον C2

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

C_2	d_1	d_2	d_3	d_4	d_5	d_6
ship	0.85	0.52	0.28	0.13	0.21	−0.08
boat	0.36	0.36	0.16	−0.20	−0.02	−0.18
ocean	1.01	0.72	0.36	−0.04	0.16	−0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	−0.39	−0.08	0.90	0.41	0.49

Ενώ λοιπόν στον C η γωνία d2 και d3 ήταν 90^o ,στον C2 είναι κάτω από 90^o,άρα αναγνωρίστηκε σε έναν βαθμό τουλάχιστον η λανθάνουσα συσχέτιση των δύο κειμένων. Συνολικά, αν πάρουμε έναν οποιονδήποτε συνδυασμό διανυσμάτων θα παρατηρήσουμε ότι συγκλίνουν περισσότερο από τα αντίστοιχα στον C. Αν μηδενίζαμε μία παραπάνω ιδιοτιμή σίγουρα θα αναδεικνυόταν περισσότερο η συσχέτιση των d2 και d3 μέσω του πίνακα C1, πλην όμως ο C1 θα απέκλεινε ακόμα περισσότερο από τον C, κάτι μη επιθυμητό.

Γενικά, όσο περισσότερες κ μεγαλύτερες ιδιοτιμές διατηρούνται, τόσο ο Ακ προσεγγίζει τον Α αλλά και τόσο μικρότερη είναι η λανθάνουσα δεικτοδότηση. Η κατάλληλη επιλογή του κ, η οποία συμβιβάζει τη διατήρηση της αρχικής

πληροφορίας του Α χωρίς μεγάλο σφάλμα με την λανθάνουσα δεικτοδότηση αποτελεί ένα ανοικτό πρόβλημα στη διανυσματική ανάλυση κειμένων.

2.4 Εξόρυξη πληροφορίας από κείμενα στη βιοϊατρική

2.4.1 Αναγνώριση ονοματικών οντοτήτων

Η αναγνώριση ονοματικών βιοϊατρικών οντοτήτων (Named Entity Recognition, NER) αναφέρεται στην εργασία του αυτόματου εντοπισμού των εμφανίσεων βιολογικών ή ιατρικών όρων σε αδόμητο κείμενο. Στις συνήθεις οντότητες ενδιαφέροντος περιλαμβάνονται τα ονόματα γονιδίων και πρωτεϊνών, τα ιατρικά προβλήματα και οι θεραπείες, τα ονόματα φαρμάκων και οι δοσολογίες τους, καθώς και άλλα καλώς καθορισμένα δεδομένα από σημασιολογικής άποψης, τα οποία κατατάσσονται στον βιοϊατρικό τομέα. Παρόλο που συχνά περιγράφεται ως μεμονωμένη εργασία, η NER είναι τυπικά μια διαδικασία τριών βημάτων που περιλαμβάνει τον προσδιορισμό των ορίων της υπο-συμβολοσειράς μιας οντότητας εντός του κειμένου, την αντιστοίχιση της οντότητας σε μια προκαθορισμένη κλάση και την επιλογή του συνήθους ονόματος (preferred name) ή του μοναδικού αναγνωριστικού (unique identifier) της έννοιας την οποία ονομάζει η οντότητα. Αυτή η τελευταία υπο-εργασία, η κανονικοποίηση της οντότητας, συχνά αντιμετωπίζεται ως ξεχωριστό πρόβλημα από τη NER, αλλά στην παρούσα εργασία αναλύεται εν συντομία στα πλαίσια της περιγραφής των πολλών ζητημάτων που καθιστούν τη NER μια απαιτητική διαδικασία στον βιοϊατρικό τομέα.

Η NER είναι μια ιδιαίτερα απαιτητική εργασία στην περίπτωση των βιοϊατρικών κειμένων για μια σειρά από λόγους. Το βασικότερο εμπόδιο οφείλεται στη δυναμική φύση της επιστημονικής ανακάλυψης. Στον βιοϊατρικό τομέα συναντάται ένας τεράστιος αριθμός σημασιολογικά σχετικών οντοτήτων, ο οποίος αυξάνεται διαρκώς και με μεγάλη ταχύτητα καθώς προστίθενται νέες επιστημονικές ανακαλύψεις.

Αυτός ο συνεχώς αναπτυσσόμενος κατάλογος σχετικών όρων αποτελεί πρόκληση για τα συστήματα NER που βασίζονται σε λεξικά γνωστών όρων ή άλλες συντηρούμενες πηγές για τον εντοπισμό ονοματικών οντοτήτων, καθώς οι εν λόγω πηγές δεν μπορούν ποτέ να είναι πλήρεις όσο η επιστημονική εξέλιξη συνεχίζεται.

Μια άλλη πρόκληση για τη βιοϊατρική NER είναι η συνωνυμία. Στη βιοϊατρική βιβλιογραφία παρατηρείται το φαινόμενο η ίδια έννοια να εκφράζεται χρησιμοποιώντας διαφορετικές λέξεις. Για παράδειγμα, η "καρδιακή προσβολή" και το "έμφραγμα του μυοκαρδίου" αναφέρονται στο ίδιο ιατρικό πρόβλημα και, συνεπώς, ένα σύστημα NER θα πρέπει να αναγνωρίζει αυτούς τους όρους ως στιγμιότυπα της ίδιας έννοιας, παρόλο που εκφράζονται με διαφορετικό τρόπο. Στις περιπτώσεις όπου χρησιμοποιούνται πολλά συνώνυμα για μια συγκεκριμένη έννοια, η ενοποίηση της γνώσης από πολλαπλές πηγές αποδεικνύεται δύσκολη χωρίς μια περιεκτική πηγή συνωνύμων όπως το UMLS Metathesaurus ή το Gene Ontology. Ωστόσο, δεδομένου του ταχέος αυξανόμενου αριθμού των βιοϊατρικών

οντοτήτων, οι πηγές αυτές είναι απίθανο να καταστούν πλήρεις σε μια δεδομένη στιγμή, με αποτέλεσμα την ύπαρξη συνωνυμικών σχέσεων που δεν εντοπίζονται.

Τέλος, η συχνή χρήση ακρωνυμίων και συντομογραφιών στη βιοϊατρική βιβλιογραφία καθιστά δύσκολο τον αυτόματο εντοπισμό των εννοιών στις οποίες αναφέρονται οι εν λόγω όροι. Συχνά, η επιτυχής ανάλυση ακρωνυμίων και συντομογραφιών εξαρτάται σε μεγάλο βαθμό από τα συμφραζόμενα του κειμένου στο οποίο εμφανίζονται οι όροι, καθώς ο ίδιος όρος μπορεί να σχετίζεται με διαφορετικές έννοιες. Για παράδειγμα, η συντομογραφία *RA* μπορεί να αναφέρεται σε ένα από τα "right atrium" (δεξιά κοιλία), "rheumatoid arthritis" (ρευματοειδής αρθρίτιδα), "refractory anemia" (ανθεκτική αναιμία), "renal artery" (νεφρική αρτηρία) ή σε μία από πολλές άλλες έννοιες . Για την αντιμετώπιση των δυσκολιών που σχετίζονται με τα ακρωνύμια, τις συντομογραφίες και τη συνωνυμία, τα συστήματα NER περιλαμβάνουν συνήθως την εκτέλεση κάποιας μορφής κανονικοποίησης οντοτήτων.

Για τα συστήματα NER που αναλύουν μεγάλο αριθμό βιοϊατρικών κειμένων, είναι σημαντικό να εξεταστεί το ζήτημα της ποιότητας που μπορεί να αναμένεται από τη χρησιμοποιούμενη μέθοδο. Συνήθως, η απόδοση των συστημάτων NER μετράται με όρους ακρίβειας (**precision**), ανάκλησης (**recall**) και βαθμολογίας-F (**F-score**).

Πρόσφατες αξιολογήσεις με ευρεία συμμετοχή σε επίπεδο επιστημονικής κοινότητας, υποδεικνύουν ότι τα συστήματα NER έχουν τυπικά τη δυνατότητα επίτευξης ευνοϊκών αποτελεσμάτων. Για παράδειγμα, τα συστήματα με την καλύτερη απόδοση πέτυχαν βαθμολογίες-F 0,83 και 0,87 για την πρώτη[10] και τη δεύτερη εργασία[11] αναγνώρισης αναφορών γονιδίων BioCreAtIve, 0,85 για την εργασία εξαγωγής εννοιών i2b2 [12] και 0,73 για την εργασία αναγνώρισης βιο-οντοτήτων JNLPBA [13]. Παρόλο που τα συστήματα NER δύνανται να προσαρμοστούν ειδικά για συγκεκριμένες εργασίες εξαγωγής πληροφοριών, οι βασικές μέθοδοί τους μπορούν να ομαδοποιηθούν γενικά ανάλογα με τις βασικές προσεγγίσεις που ακολουθούν, οι οποίες περιγράφονται παρακάτω.

2.4.2 Χρήση μεθόδου βασισμένης σε λεξικό

Οι μέθοδοι που βασίζονται σε λεξικά -μία εκ των βασικότερων προσεγγίσεων στη βιοϊατρική NER- χρησιμοποιούν περιεκτικούς καταλόγους βιοϊατρικών όρων για τον εντοπισμό των εμφανίσεων οντοτήτων στο κείμενο. Αυτά τα συστήματα εξακριβώνουν εάν μια λέξη ή ομάδα λέξεων που έχει επιλεγεί από το κείμενο, ταιριάζει επακριβώς με έναν όρο από μια βιοϊατρική πηγή.

Οι προσεγγίσεις που βασίζονται σε λεξικά, παρουσιάζουν γενικά σχετικά υψηλή ακρίβεια, αλλά έχουν το μειονέκτημα της ανεπαρκούς ανάκλησης εξαιτίας της ύπαρξης ορθογραφικών λαθών και μορφολογικών παραλλαγών[14]. Ωστόσο, υπάρχει επίσης η πιθανότητα χαμηλών επιπέδων ακρίβειας λόγω της ομωνυμίας. Για παράδειγμα, πολλά ονόματα και συντομογραφίες γονιδίων (π.χ. "an," "by," και "can") έχουν τις ίδιες λεξικές αναπαραστάσεις με ευρέως χρησιμοποιούμενες λέξεις της Αγγλικής γλώσσας[15]. Για το λόγο αυτό, συχνά χρησιμοποιείται κάποια μορφή μη ακριβούς αντιστοίχισης για τη βελτίωση της ακρίβειας και της ανάκλησης των

προσεγγίσεων που βασίζονται σε λεξικά. Ορισμένες μέθοδοι βελτιώνουν την απόδοσή τους δημιουργώντας αρχικά διαφορετικές παραλλαγές συλλαβισμού για τους όρους μιας βιοϊατρικής πηγής και, στη συνέχεια, ενημερώνοντας κατάλληλα τους καταλόγους με τις εναλλακτικές αυτές μορφές των όρων[16,17]. Με αυτόν τον τρόπο οι μέθοδοι έχουν τη δυνατότητα αντιστοίχισης χρησιμοποιώντας την ενισχυμένη πηγή. Άλλες μέθοδοι χρησιμοποιούν αλγορίθμους όπως οι BLAST® για την εκτέλεση κατά προσέγγισης αντιστοίχισης συμβολοσειρών, αντί για απόλυτη αντιστοίχιση. Παρά τις βελτιώσεις αυτές, οι μέθοδοι που βασίζονται σε λεξικά χρησιμοποιούνται συχνότερα σε συνδυασμό με περισσότερο προηγμένες μεθόδους NER.

2.4.3 Χρήση μεθόδου βασισμένης σε κανόνες σύνθεσης βιοϊατρικών όρων

Μια διαφορετική προσέγγιση στη NER συνίσταται στον ορισμό κανόνων που περιγράφουν τα συνθετικά μοτίβα των ονοματικών βιοϊατρικών οντοτήτων και του συγκειμένου τους. Ορισμένα παραδείγματα προσεγγίσεων που βασίζονται σε κανόνες είναι τα συστήματα EMPathIE και PASTA, τα οποία χρησιμοποιούν γραμματικές χωρίς συμφραζόμενα που αναγνωρίζουν ενζυματικές αλληλεπιδράσεις και πρωτεϊνικές δομές. Άλλα συστήματα χρησιμοποιούν κανόνες που βασίζονται σε μοτίβα, οι οποίοι χρησιμοποιούν τα ορθογραφικά και λεξικά χαρακτηριστικά στοχευμένων κατηγοριών οντοτήτων για την αναγνώριση ονομάτων πρωτεϊνών και χημικών ουσιών [18]. Αυτές οι απλούστερες μέθοδοι δύνανται να βελτιωθούν μέσω της επιπρόσθετης χρήσης πληροφοριών από τα συμφραζόμενα (contextual information) [19] και των αποτελεσμάτων της συντακτικής ανάλυσης (syntactic parsing) για τον προσδιορισμό των ορίων της οντότητας [20]. Ωστόσο, παρόλο που οι προσεγγίσεις που βασίζονται σε κανόνες τυπικά επιτυγχάνουν καλύτερες αποδόσεις σε σχέση με αυτές που βασίζονται σε λεξικά, η χειροκίνητη δημιουργία των απαιτούμενων κανόνων είναι μια χρονοβόρα διαδικασία και οι εν λόγω κανόνες είναι δύσκολο να επεκταθούν σε άλλες κλάσεις οντοτήτων δεδομένου ότι συνήθως χαρακτηρίζονται από μεγάλο βαθμό ειδικότητας για τους σκοπούς της επίτευξης υψηλής ακρίβειας.

2.4.4 Χρήση τεχνικών μηχανικής μάθησης με επίβλεψη

Ολοένα και συχνότερα συναντάται το φαινόμενο οι προσεγγίσεις NER να βασίζονται σε στατιστικές μεθόδους αντί για, ή σε συνδυασμό με, προσεγγίσεις που βασίζονται σε κανόνες ή λεξικά. Σε αντίθεση με τις προαναφερθείσες προσεγγίσεις, οι στατιστικές μέθοδοι βασίζονται εξ αρχής σε κάποια μορφή αλγορίθμου μηχανικής μάθησης για τον εντοπισμό των βιοϊατρικών οντοτήτων. Παρόλο που οι επιτηρούμενες προσεγγίσεις μηχανικής μάθησης πρέπει να "εκπαιδεύονται" με παρατηρήσεις που λαμβάνονται από μεγάλα επισημειωμένα σώματα κειμένων, πρόσφατες εργασίες έχουν διερευνήσει την αυτόματη παραγωγή εκπαιδευτικών

δεδομένων για την εργασία NER μέσω της χρήσης της μεθόδου Bootstrap και άλλων ημι-επιτηρούμενων στατιστικών τεχνικών [21, 22, 23]. Οι συνήθεις στατιστικές μέθοδοι που χρησιμοποιούνται στη NER μπορούν να ομαδοποιηθούν ως προσεγγίσεις που βασίζονται σε ταξινόμηση ή ακολουθία.

Οι προσεγγίσεις που βασίζονται σε ταξινόμηση μετασχηματίζουν την εργασία NER σε ένα πρόβλημα ταξινόμησης, το οποίο μπορεί να εφαρμοστεί σε μεμονωμένες λέξεις ή ομάδες λέξεων. Οι συνήθεις ταξινομητές που χρησιμοποιούνται στη βιοϊατρική NER περιλαμβάνουν τους Naive Bayes [24] και Support Vector Machine (SVM) [25, 26, 27, 28]. Παρόλο που η ταξινόμηση φράσεων που αποτελούνται από πολλές λέξεις είναι εφικτή, μία από τις δημοφιλείς προσεγγίσεις ακολουθεί το πρόγραμμα μορφολογικής ανάλυσης BIO [29], όπου οι μεμονωμένες λεκτικές μονάδες ταξινομούνται με βάση τη θέση τους στην αρχή (B) μιας οντότητας, εντός (I) των ορίων μιας οντότητας και εκτός (O) των ορίων μιας οντότητας. Ωστόσο, παρά την επιτυχία του, το συγκεκριμένο πρόγραμμα μορφολογικής ανάλυσης μπορεί να αποδειχθεί προβληματικό σε περίπτωση αλληλεπικάλυψης των ορίων των οντοτήτων και αρκετοί συγγραφείς έχουν ασχοληθεί με το πρόβλημα της αναγνώρισης ένθετων βιοϊατρικών οντοτήτων [30, 31]. Η απόδοση των προσεγγίσεων που βασίζονται σε ταξινόμηση εξαρτάται σε μεγάλο βαθμό από την επιλογή των χαρακτηριστικών που χρησιμοποιούνται για την εκπαίδευση και πολλοί συγγραφείς έχουν διερευνήσει διάφορους συνδυασμούς τέτοιων χαρακτηριστικών. Για παράδειγμα, ο Kazama και οι συνεργάτες του [86] και ο Mit-sumori και οι συνεργάτες του [32], εξέτασαν τις μορφοσυντατικές ιδιότητες των ονοματικών οντοτήτων, ενώ οι Takeuchi και Collier [33] τα ορθογραφικά χαρακτηριστικά και το κυρίαρχο ουσιαστικό. Ο Yamamoto και οι συνεργάτες του [34] έχουν επίσης διερευνήσει πληθώρα χαρακτηριστικών μεταξύ των οποίων τα όρια, οι μορφολεξικές και συντακτικές ιδιότητες, καθώς και ένα χαρακτηριστικό που βασίζεται σε λεξικό, το οποίο υποδεικνύει εάν μια λέξη εμφανίζεται σε μια βιοϊατρική πηγή. Λόγω της ευαισθησίας των προσεγγίσεων που βασίζονται σε ταξινόμηση όσον αφορά την επιλογή των χαρακτηριστικών, η αυτόματη επιλογή χαρακτηριστικών είναι ένα ζήτημα ιδιαίτερης σημασίας. Ο Hakenberg και οι συνεργάτες του [35] έχουν εκτελέσει μια συστηματική αξιολόγηση των συνηθέστερων χαρακτηριστικών και έχουν αναλύσει την επιρροή τους στην ποιότητα της πρόβλεψης των συστημάτων NER που βασίζονται σε ταξινόμηση.

Σε αντίθεση με τις προσεγγίσεις που βασίζονται σε ταξινόμηση, τα συστήματα NER που βασίζονται σε ακολουθία χρησιμοποιούν ολόκληρες ακολουθίες λέξεων αντί για μεμονωμένες λέξεις ή φράσεις. Εκπαιδεύονται μέσω σωμάτων κειμένων με γλωσσικές ετικέτες (tagged corpora) και στόχος τους είναι η πρόβλεψη των πιθανότερων ετικετών για μια δεδομένη ακολουθία παρατηρήσεων. Ένα σύνηθες στατιστικό πλαίσιο που χρησιμοποιείται για τη βιοϊατρική NER είναι το Hidden Markov Model (HMM) [36]. Οι μέθοδοι που βασίζονται στο Μοντέλο Μέγιστης Εντροπίας του Markov (Maximum Entropy Markov Model) επίσης χρησιμοποιούνται συχνά [37]. Ωστόσο, τα Υπό συνθήκη τυχαία πεδία (Conditional Random Fields, CRF) [38] συχνά εμφανίζονται ως ανώτερα στατιστικά πλαίσια για τη βιοϊατρική NER. Για

παράδειγμα, τα CRF χρησιμοποιήθηκαν από το σύστημα με την καλύτερη απόδοση στην εργασία εξαγωγής ιατρικών εννοιών i2b2 [39] και από συστήματα με υψηλή βαθμολογία στις εργασίες αναγνώρισης αναφορών γονιδίων BioCreAtIve [40] και αναγνώρισης βιο-οντοτήτων JNLPBA. Όπως ισχύει και με άλλες στατιστικές μεθόδους, οι προσεγγίσεις που βασίζονται σε ακολουθία μπορούν να "εκπαιδευτούν" σε μια σειρά από χαρακτηριστικά, συμπεριλαμβανομένων των ορθογραφικών χαρακτηριστικών [36], των πληροφοριών προθήματος και επιθήματος και των συνόλων ετικετών μέρους του λόγου (part-of-speech tags) ενισχυμένων ώστε να περιλαμβάνουν ετικέτες για κατηγορίες οντοτήτων [37].

Πολλές προσεγγίσεις δεν χρησιμοποιούν απλά μια μεμονωμένη μέθοδο για την εκτέλεση της βιοϊατρικής NER, αλλά αντιθέτως βασίζονται σε πολλαπλές τεχνικές και διάφορες πηγές. Αυτές οι υβριδικές προσεγγίσεις συχνά έχουν αρκετά επιτυχημένα αποτελέσματα στο συνδυασμό των προσεγγίσεων που βασίζονται σε λεξικά ή κανόνες με στατιστικές μεθόδους. Σε μια απόπειρα απόδειξης των πλεονεκτημάτων των υβριδικών προσεγγίσεων, ο Abacha και οι συνεργάτες του [41] συνέκριναν την απόδοση των συνήθων προσεγγίσεων που βασίζονται σε κανόνες και των στατιστικών προσεγγίσεων στην αναγνώριση ιατρικών οντοτήτων και κατέληξαν στο συμπέρασμα ότι οι υβριδικές προσεγγίσεις που χρησιμοποιούν μηχανική μάθηση και γνώση πεδίου έχουν την καλύτερη απόδοση.

Υπάρχουν πολυάριθμα υβριδικά συστήματα βιοϊατρικής NER. Για παράδειγμα, ο Sasaki και οι συνεργάτες του χρησιμοποίησαν μια προσέγγιση που βασίζεται σε λεξικά για τον εντοπισμό γνωστών ονομάτων πρωτεϊνών παράλληλα με επισημείωση μερών του λόγου (part-of speech tagging). Στη συνέχεια χρησιμοποίησαν μια προσέγγιση που βασίζεται σε CRF για να μειώσουν τον αριθμό των λανθασμένων θετικών και των λανθασμένων αρνητικών στην προκύπτουσα επισημασμένη ακολουθία. Άλλες μέθοδοι δημιουργούν "meta-learners" από πολλαπλές στατιστικές μεθόδους. Για παράδειγμα, ο Zhou και οι συνεργάτες του [42] χρησιμοποίησαν ένα "meta-learner" που αποτελείται από δύο HMM εκπαιδευμένα σε διαφορετικά σώματα κειμένων, οι παράγωγοι των οποίων συνδυάζονται με ένα SVM για την αναγνώριση ονομάτων πρωτεϊνών και γονιδίων. Παρομοίως, οι Mika και Rost [43] συνέθεσαν ένα "meta-learner" για την αναγνώριση ονομάτων πρωτεϊνών από τρία SVM εκπαιδευμένα σε διαφορετικά σώματα κειμένων και σύνολα χαρακτηριστικών, οι παράγωγοι των οποίων συνδυάζονται στη συνέχεια με ένα τέταρτο SVM. Τέλος, οι Cai και Cheng παρουσίασαν μια προσέγγιση για τη βιοϊατρική NER στην οποία χρησιμοποιούνται τρεις διαφορετικοί ταξινομητές για τη βελτίωση της ικανότητας γενίκευσης του συστήματος.

2.4.5 Εξαγωγή σχέσεων

Οι περισσότερες εργασίες εξαγωγής πληροφοριών στον βιοϊατρικό τομέα εκτείνονται πέρα από την απλή αναγνώριση ονοματικών οντοτήτων και περιλαμβάνουν επίσης τον προσδιορισμό των σχέσεων ανάμεσα σε αυτές τις οντότητες. Στην απλούστερή τους μορφή, οι συσχετίσεις ανάμεσα στις βιοϊατρικές

οντότητες είναι δυαδικές και αφορούν μόνο τις σχέσεις ζεύγους μεταξύ δύο οντοτήτων. Ωστόσο, οι βιοϊατρικές σχέσεις μπορεί να περιλαμβάνουν περισσότερες από δύο οντότητες και αυτές οι πολύπλοκες συσχετίσεις αναλύονται παρακάτω με την εργασία εξαγωγής συμβάντων. Σκοπός της εργασίας εξαγωγής σχέσεων, επομένως, είναι ο εντοπισμός των εμφανίσεων συγκεκριμένων τύπων σχέσεων ανάμεσα σε ζεύγη δεδομένων οντοτήτων. Παρόλο που οι συνήθεις κατηγορίες οντοτήτων (π.χ. γονίδια ή φάρμακα) είναι γενικά αρκετά εξειδικευμένες, οι τύποι των καθορισμένων σχέσεων μπορεί να είναι ευρείς και να περιλαμβάνουν οποιονδήποτε τύπο βιοϊατρικής συσχέτισης ή να είναι ειδικοί, για παράδειγμα, χαρακτηρίζοντας μόνο τις συσχετίσεις ρύθμισης των γονιδίων.

Μια σειρά από βιοϊατρικές σχέσεις έχουν αποτελέσει αντικείμενο των εργασιών εξαγωγής πληροφοριών της βιβλιογραφίας. Στη σημερινή γενωμική εποχή, μεγάλο μέρος αυτής της προσπάθειας έχει επικεντρωθεί σε αλληλεπιδράσεις αυτόματης εξαγωγής ανάμεσα σε γονίδια και πρωτεΐνες. Συγκεκριμένα, εξαιτίας του εξαιρετικά σημαντικού ρόλου της στην κατανόηση βιολογικών διεργασιών, οι Αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης (PPI) είναι ένα από τα θέματα που συγκεντρώνουν το μεγαλύτερο όγκο έρευνας στον τομέα της εξαγωγής βιοϊατρικών πληροφοριών.

Άλλες συσχετίσεις ενδιαφέροντος περιλαμβάνουν τις αλληλεπιδράσεις ανάμεσα σε πρωτεΐνες και σημεία μεταλλάξεων , πρωτεΐνες και τις περιοχές πρόσδεσής τους , γονίδια και νόσους, και γονίδια και φαινοτυπικό υπόβαθρο . Στον κλινικό τομέα, οι σχέσεις ανάμεσα στα προβλήματα που παρουσιάζουν οι ασθενείς και τις εξετάσεις ή θεραπείες στις οποίες υποβάλλονται είναι ένας ολοένα και σημαντικότερος τύπος συσχέτισης, ειδικά λαμβάνοντας υπόψη την αυξανόμενη παρουσία των ηλεκτρονικών συστημάτων ιατρικών αρχείων.

Η εξαγωγή βιοϊατρικών σχέσεων αντιμετωπίζει πολλές από τις ίδιες προκλήσεις με την NER, συμπεριλαμβανομένης της δημιουργίας επισημειωμένων σωμάτων κειμένων υψηλής ποιότητας για την εκπαίδευση και αξιολόγηση των συστημάτων εξαγωγής σχέσεων. Σε σύγκριση με την επισημείωση ονοματικών οντοτήτων, η επισημείωση σχέσεων είναι μια αρκετά πιο πολύπλοκη διαδικασία, καθώς οι σχέσεις γενικά εκφράζονται ως ασυνεχή εύρη κειμένου και οι τύποι των σχέσεων που λαμβάνονται υπόψη είναι συνήθως ειδικοί για κάθε εφαρμογή. Επιπλέον, δεδομένου ότι συχνά δεν υπάρχει ευρεία συναίνεση σχετικά με τη βέλτιστη μέθοδο επισημείωσης των δεδομένων τύπων σχέσεων, οι προκύπτουσες πηγές είναι σε μεγάλο βαθμό ασύμβατες και, κατά συνέπεια, η ποιότητα των μεθόδων που χρησιμοποιούν αυτές τις πηγές είναι δύσκολο να αξιολογηθεί. Για παράδειγμα, ο Pyysalo και οι συνεργάτες του εκτέλεσαν συγκριτική ανάλυση πέντε σωμάτων κειμένων PPI και βρήκαν ότι η απόδοση των συστημάτων εξαγωγής PPI τελευταίας τεχνολογίας, μετρούμενη με βάση τη βαθμολογία-F, εμφάνιζε διακυμάνσεις της τάξης των 19 ποσοστιαίων μονάδων κατά μέσο όρο, η οποία έφτανε και τις 30 ποσοστιαίες μονάδες στα σώματα κειμένων που αξιολογήθηκαν. Η συμμετοχή σε αξιολογήσεις με ευρεία συμμετοχή της επιστημονικής κοινότητας, οι οποίες έχουν ως αποκλειστικό αντικείμενο την εργασία εξαγωγής σχέσεων, είναι ζωτικής σημασίας όσον αφορά τη λήψη επισημειωμένων σωμάτων εκπαίδευσης.

Οι εργασίες εξαγωγής σχέσεων έχουν αποτελέσει αντικείμενο πολλών πρόσφατων forum αξιολόγησης και οι εν λόγω εργασίες περιλαμβάνουν την πρόκληση γενετικής αλληλεπίδρασης LLL , την εργασία εξαγωγής PPI BioCreAtIve και την εργασία εξαγωγής σχέσεων i2b2 . Σκοπός της πρόκλησης LLL ήταν η εξαγωγή σχέσεων πρωτεϊνών και γονιδίων από περιλήψεις που περιέχονται στο MEDLINE και το σύστημα με την καλύτερη απόδοση πέτυχε βαθμολογία-*F* 0,54 στον προσδιορισμό αυτών των σχέσεων. Η εργασία BioCreAtIve αποτελούνταν από τέσσερις υπο-εργασίες που σχετίζονται με την εξαγωγή PPI. Αυτές οι προκλήσεις περιελάμβαναν την ταξινόμηση περιλήψεων του PubMed με βάση τη σχετικότητά τους ως προς την επισημείωση PPI, τον προσδιορισμό δυαδικών αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης από πλήρη άρθρα, την εξαγωγή μεθόδων αλληλεπίδρασης πρωτεϊνών και την ανάκτηση κειμενικών ενδείξεων που περιγράφουν τις αλληλεπιδράσεις. Το σύστημα με την καλύτερη απόδοση πέτυχε βαθμολογία ακρίβειας της τάξης του 0,37 σε ανάκληση 0,33 για την εξαγωγή δυαδικών σχέσεων PPI. Τέλος, σκοπός της πρόκλησης εξαγωγής σχέσεων i2b2 ήταν ο προσδιορισμός των σχέσεων ιατρικού προβλήματος-θεραπείας, προβλήματος-εξέτασης και προβλήματος-προβλήματος σε κλινικές σημειώσεις. Οι συμμετέχοντες κλήθηκαν, για παράδειγμα, να προσδιορίσουν εάν δύο συνυπάρχουσες έννοιες προβλήματος και θεραπείας σχετίζονταν και, εάν όντως σχετίζονταν, κατά πόσον η θεραπεία του ασθενούς βελτιώθηκε, επιδεινώθηκε ή προκάλεσε το ιατρικό πρόβλημα.

Το σύστημα με την καλύτερη απόδοση στην εξαγωγή σχέσεων i2b2 που δοκιμάστηκε, πέτυχε βαθμολογία *F* 0,74. Όπως ισχύει και στα forum με αντικείμενο την αξιολόγηση της εργασίας NER, οι αξιολογήσεις με ευρεία συμμετοχή της επιστημονικής κοινότητας όπως αυτές, έχουν παίξει ζωτικό ρόλο στην ανάπτυξη και την εξέλιξη των προσεγγίσεων εξαγωγής σχέσεων.

Οι προσεγγίσεις εξαγωγής σχέσεων σήμαναν την εξέλιξη από τα απλά συστήματα που βασίζονται μόνο σε στατιστικά στοιχεία συνύπαρξης στα πολύπλοκα συστήματα που χρησιμοποιούν συντακτική ανάλυση και ανάλυση εξάρτησης (dependency parsing). Παρακάτω περιγράφονται ορισμένες πρόσφατες προσεγγίσεις στην εργασία εξαγωγής σχέσεων. Επεξηγήσεις πρόσθετων μεθόδων βρίσκονται σε άλλες έρευνες εξόρυξης βιοϊατρικών κειμένων με αντικείμενο την εργασία εξαγωγής σχέσεων.

Η απλούστερη μέθοδος προσδιορισμού των σχέσεων ανάμεσα σε βιοϊατρικές οντότητες είναι η συλλογή στιγμιοτύπων συνύπαρξης των οντοτήτων. Εάν οι οντότητες αναφέρονται επανειλημμένα μαζί, τότε υπάρχει μεγαλύτερη πιθανότητα να σχετίζονται με κάποιον τρόπο, αλλά ο τύπος και η κατεύθυνση αυτής της σχέσης συνήθως δεν είναι δυνατόν να προσδιοριστεί μόνο με τα στατιστικά στοιχεία συνύπαρξης. Για παράδειγμα, ο Chen και οι συνεργάτες του χρησιμοποίησαν στατιστικά στοιχεία συνύπαρξης για τον υπολογισμό του βαθμού συσχέτισης ανάμεσα σε νόσους και φάρμακα που εξήχθησαν από κλινικά αρχεία και τη βιοϊατρική βιβλιογραφία. Οι προσεγγίσεις συνύπαρξης παρουσιάζουν συχνά υψηλή ανάκληση και χαμηλή ακρίβεια.

Οι προσεγγίσεις που βασίζονται σε κανόνες περιγράφουν τα γλωσσικά μοτίβα που παρουσιάζουν συγκεκριμένες σχέσεις. Σε αντίθεση με τα συστήματα που βασίζονται στη συνύπαρξη όρων, οι προσεγγίσεις που βασίζονται σε κανόνες παρουσιάζουν υψηλή ακρίβεια και χαμηλή ανάκληση. Οι κανόνες που χρησιμοποιούνται για την εξαγωγή σχέσεων μπορούν να οριστούν χειροκίνητα από ειδικούς του εκάστοτε τομέα ή να εξαχθούν από επισημειωμένα σώματα κειμένων με τη βοήθεια αλγορίθμων μηχανικής μάθησης .

Οι προσεγγίσεις που βασίζονται σε ταξινόμηση χρησιμοποιούνται επίσης συχνά για τον προσδιορισμό σχέσεων, ειδικά αυτών που περιλαμβάνουν ιατρικές οντότητες. Ο Roberts και οι συνεργάτες του έχουν περιγράψει ένα επιτηρούμενο σύστημα μηχανικής μάθησης, το οποίο εκπαιδεύεται με επιφανειακά χαρακτηριστικά που εξάγονται από ογκολογικές εκθέσεις και ανιχνεύει διάφορες κλινικές σχέσεις στις περιγραφές των ασθενών. Παρομοίως, ο Rink και οι συνεργάτες του έχουν περιγράψει ένα σύστημα που ανακαλύπτει σχέσεις ανάμεσα σε ιατρικά προβλήματα, θεραπείες και εξετάσεις που αναφέρονται σε ηλεκτρονικά ιατρικά αρχεία. Το σύστημα βασίζεται στην επιτηρούμενη μηχανική μάθηση, καθώς και σε λεξικά, συντακτικά χαρακτηριστικά και σημασιολογικά χαρακτηριστικά συγκειμένου. Ο Bundschus και οι συνεργάτες του χρησιμοποίησαν CRF για τον προσδιορισμό και την ταξινόμηση των σχέσεων ανάμεσα σε νόσους και θεραπείες που έχουν εξαχθεί από περιλήψεις του PubMed και τις σχέσεις ανάμεσα σε γονίδια και νόσους από τη βάση δεδομένων GeneRIF για τον άνθρωπο. Τέλος, οι Abach και Zweigenbaum έχουν περιγράψει μια υβριδική προσέγγιση που χρησιμοποιεί μοτίβα που έχουν αναπτυχθεί από ειδικούς του τομέα καθώς και ταξινόμηση SVM για την εξαγωγή σχέσεων ανάμεσα σε νόσους και θεραπείες σε ιατρικά κείμενα.

Ένα σημαντικό βήμα προόδου στις μεθόδους εξαγωγής σχέσεων είναι η χρήση συντακτικών δομών. Συγκεκριμένα, η ανάλυση της εξάρτησης (dependency parsing) δίνει τη δυνατότητα δημιουργίας διδακτικών συντακτικών περιγραφών βιοϊατρικών κειμένων με τη μορφή δένδρων ή γραφημάτων εξάρτησης, τα οποία κωδικοποιούν τις γραμματικές σχέσεις ανάμεσα σε φράσεις ή λέξεις. Ο Fundel και οι συνεργάτες του προχώρησαν στη δημιουργία δένδρων εξάρτησης από περιλήψεις του MEDLINE. Το σύστημά τους εν συνεχεία εφαρμόζει τρεις κανόνες εξαγωγής σχέσεων στις συντακτικές δομές με σκοπό τον προσδιορισμό των συσχετίσεων ανάμεσα σε γονίδια και πρωτεΐνες. Παρομοίως, ο Rinaldi και οι συνεργάτες του συνδύασαν συντακτικά μοτίβα που έχουν ληφθεί από δενδρικές δομές εξάρτησης με σκοπό την υποστήριξη της δημιουργίας ερωτημάτων (querying) στη βιοϊατρική βιβλιογραφία ως προς τις αλληλεπιδράσεις ανάμεσα σε γονίδια και πρωτεΐνες. Ο Miyao και οι συνεργάτες του εκτέλεσαν βαθιά συντακτική ανάλυση για την επισημείωση δομών κατηγορήματος-ορισμάτων (predicate-argument) σε περιλήψεις του MEDLINE. Το σύστημά τους βασίζεται εν συνεχεία στη δομική αντιστοίχιση των σημασιολογικών επισημειώσεων για τον προσδιορισμό και την ανάκτηση σχεσιακών εννοιών. Σε μια άλλη εργασία, ο Miyao και οι συνεργάτες του αξιολόγησαν διάφορους αναλυτές και τις εξαγόμενες αναπαραστάσεις τους ως προς την ικανότητά τους να

βελτιώσουν την ακρίβεια όταν χρησιμοποιούνται ως μέρος ενός συστήματος εξαγωγής PPI.

Δεδομένης της αυξανόμενης διαθεσιμότητας μεγάλων σωμάτων κειμένων που περιέχουν σχεσιακές επισημειώσεις, πολλές προσεγγίσεις χρησιμοποιούν αλγορίθμους μηχανικής μάθησης για την εξαγωγή χρήσιμων πληροφοριών από συντακτικές δομές αντί για την απλή χειροκίνητη εφαρμογή παρεπόμενων μοτίβων. Όσον αφορά τη μηχανική μάθηση που βασίζεται σε kernel, αρκετοί συγγραφείς έχουν προτείνει kernel με δυνατότητα μέτρησης της ομοιότητας ανάμεσα σε δένδρα ή γραφήματα συντακτικής ανάλυσης. Ο Airola και οι συνεργάτες του έχουν περιγράψει ένα kernel γραφήματος συνόλου διαδρομών (all-paths) για τον υπολογισμό της ομοιότητας ανάμεσα σε γραφήματα εξάρτησης. Στη συνέχεια χρησιμοποιείται η συνάρτηση kernel για την εκπαίδευση ενός Συστήματος διανυσμάτων υποστήριξης (Support Vector Machine) ελαχίστων τετραγώνων με σκοπό τον προσδιορισμό των αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης. Ο Kim και οι συνεργάτες του έχουν υποδείξει τέσσερα kernel εξαγωγής γενετικών σχέσεων που ορίζονται στη συντομότερη διαδρομή συντακτικής εξάρτησης ανάμεσα σε δύο ονοματικές οντότητες. Τέλος, ο Miwa και οι συνεργάτες του έχουν περιγράψει ένα πλαίσιο για το συνδυασμό των παραγώγων από πολλαπλά kernel και συντακτικούς αναλυτές για την εξαγωγή αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης.

Η συντακτική ανάλυση συχνά συνοδεύεται από την επισήμανση σημασιολογικού ρόλου (semantic role labeling), μια τεχνική επεξεργασίας φυσικής γλώσσας που προσδιορίζει τους σημασιολογικούς ρόλους λέξεων και φράσεων σε προτάσεις και τις εκφράζει ως δομές κατηγορήματος-ορισμάτων. Ο Tsai και οι συνεργάτες του έχουν κατασκευάσει ένα σύστημα επισήμανσης ρόλων που χρησιμοποιεί ένα μοντέλο μηχανικής μάθησης μέγιστης εντροπίας για την εξαγωγή βιοϊατρικών σχέσεων από ένα προκαθορισμένο τμήμα του σώματος κειμένων GENIA. Όπως αναλύεται παρακάτω, η επισημείωση των σημασιολογικών ρόλων για τις ονοματικές βιοϊατρικές οντότητες έχει καταστήσει εφικτή την εξαγωγή μια σειράς από περίπλοκες συσχετίσεις οντοτήτων.

2.4.6 Εξαγωγή συμβάντων

Πρόσφατα έχει παρατηρηθεί μια μετατόπιση στην εξαγωγή βιοϊατρικών πληροφοριών από την αναγνώριση δυαδικών σχέσεων σε μια πιο φιλόδοξη τεχνική, αυτή του προσδιορισμού περίπλοκων, ένθετων δομών συμβάντων. Τα συμβάντα συνήθως χαρακτηρίζονται από ρήματα ή ουσιαστικοποιημένα ρήματα. Για παράδειγμα, στην πρόταση *"το glnAP2* μπορεί να ενεργοποιηθεί από το *NifA,"* το ρήμα *ενεργοποιηθεί* προσδιορίζει το συμβάν και τα *glnAP2* και *NifA* αποτελούν τα ορίσματα του συμβάντος. Σε αντίθεση με τις απλές δυαδικές σχέσεις, στο συμβάν και τα ορίσματά του εκχωρούνται τόσο ετικέτες εννοιών όσο και σημασιολογικοί ρόλοι. Σε αυτό το παράδειγμα, το ρήμα *ενεργοποιηθεί* υποδεικνύει ένα συμβάν

θετικής ρύθμισης, στο οποίο αναμένεται η πρωτεΐνη *(NifA)* να δράσει ως η αιτία του συμβάντος και το γονίδιο *(glnAP2*) ως το θέμα του συμβάντος .

Μια άλλη σημαντική διάκριση ανάμεσα στην εξαγωγή δυαδικών σχέσεων και περίπλοκων συμβάντων είναι ότι τα συμβάντα μπορούν να είναι ένθετα, με ένα συμβάν να δρα ως "συμμετέχων" σε ένα άλλο συμβάν. Για παράδειγμα, στην πρόταση "η RFLAT-1 ενεργοποιεί την έκφραση του γονιδίου RANTES", υπάρχουν δύο συμβάντα . Ένα συμβάν υποδεικνύεται από το ουσιαστικοποιημένο ρήμα *έκφραση,* θέμα του οποίου είναι το γονίδιο *RANTES* και το άλλο συμβάν υποδεικνύεται από το ρήμα *ενεργοποιεί* , αιτία του οποίου είναι η πρωτεΐνη *RFLAT-1* και θέμα του οποίου είναι αυτή καθαυτή η έκφραση του γονιδίου. Συνεπώς, οι αναπαραστάσεις των συμβάντων, σε αντίθεση με τις δυαδικές σχέσεις, έχουν τη δυνατότητα να εντοπίσουν πολλούς διαφορετικούς τύπους συσχετίσεων με έναν αυθαίρετο αριθμό οντοτήτων και συμβάντων που σχετίζεται μέσω μιας σειράς από σημασιολογικούς ρόλους.

Εξαιτίας της πολυπλοκότητας των βιοϊατρικών συμβάντων, η αποτελεσματική εξαγωγή σχέσεων απαιτεί συνήθως διεξοδική ανάλυση της δομής της πρότασης. Η εξαγωγή συμβάντων υποβοηθείται ιδιαίτερα από τη χρήση τεχνικών σημασιολογικής επεξεργασίας και βαθιάς συντακτικής ανάλυσης, οι οποίες έχουν τη δυνατότητα ανάλυσης τόσο της συντακτικής όσο και της σημασιολογικής δομής των βιοϊατρικών κειμένων. Η ανάλυση εξάρτησης (depedency parsing) είναι μια ιδιαίτερα χρήσιμη τεχνική για τον εντοπισμό σημασιολογικών πτυχών όπως οι σχέσεις κατηγορήματος-ορισμάτων, οι οποίες έχει καταδειχθεί ότι αποτελούν μια αποτελεσματική αναπαράσταση για την εξαγωγή συμβάντων. Παρά την πολυπλοκότητα της εργασίας, η εξαγωγή συμβάντων παρουσιάζει ένα ευρύ φάσμα δυνατοτήτων στον βιοϊατρικό τομέα και χρησιμοποιείται ολοένα και περισσότερο για την επισημείωση βιοϊατρικών μονοπατιών, την επισημείωση της Γονιδιακής Οντολογίας και τον εμπλουτισμό των βιολογικών βάσεων δεδομένων.

Το εντεινόμενο ενδιαφέρον γύρω από την εξαγωγή συμβάντων ωθείται κατά κύριο λόγο από τη διαθεσιμότητα, κυρίως στον τομέα της βιολογίας συστημάτων, σωμάτων κειμένων που περιέχουν τις απαραίτητες επισημειώσεις για την εκπαίδευση και αξιολόγηση των στατιστικών μεθόδων εξαγωγής συμβάντων. Το σώμα κειμένων BioInfer ήταν το πρώτο δημόσια διαθέσιμο σώμα κειμένων στον βιοϊατρικό τομέα που ενσωμάτωσε επισημειώσεις συμβάντων. Άλλα σώματα κειμένων με επισημειώσεις συμβάντων είναι τα GENIA Event Corpus και Gene Regulation Event Corpus. Αξίζει να σημειωθεί ότι το σώμα κειμένων GENIA παραμένει μία από τις ευρύτερα χρησιμοποιούμενες πηγές στην εξόρυξη από βιοϊατρικά κείμενα και τα δεδομένα για την κοινή εργασία εξαγωγής σχέσεων BioNLP προετοιμάστηκαν με βάση την πηγή αυτή.

Η κοινή εργασία BioNLP '09 ήταν η πρώτη του είδους αξιολόγηση μεθόδων εξαγωγής σχέσεων με ευρεία συμμετοχή της επιστημονικής κοινότητας. Η κύρια πρόκληση συνίστατο στην εξαγωγή των τύπων των συμβάντων που σχετίζονται με τη βιολογία των πρωτεϊνών από περιλήψεις του MEDLINE. Οι τύποι συμβάντων που αποτελούσαν αντικείμενο της πρόκλησης περιελάμβαναν, μεταξύ άλλων, την

έκφραση γονιδίων, την αντιγραφή, τον εντοπισμό, την πρόσδεση και τη ρύθμιση. Ο τύπος συμβάντων πρόσδεσης ήταν περισσότερο περίπλοκος σε σχέση με τους υπόλοιπους, καθώς απαιτούσε την ανίχνευση ενός αυθαίρετου αριθμού ορισμάτων, ενώ οι τύποι συμβάντων ρύθμισης ήταν αξιοσημείωτοι ως προς την παροχή δυνατότητας στα άλλα συμβάντα να δράσουν ως αιτία ή θέμα τους. Το σύστημα με την καλύτερη απόδοση πέτυχε βαθμολογία-F 0,52 στην πρωταρχική εργασία εξαγωγής συμβάντων. Στην κοινή εργασία BioNLP '11 επαναλήφθηκε η αξιολόγηση από την προηγούμενη συνάντηση, αλλά περιελήφθησαν επίσης εργασίες με στόχο τον εντοπισμό τύπων συμβάντων σε άλλους δευτερεύοντες τομείς της βιολογίας. Σε μια υπο-εργασία συγκρίσιμη με αυτήν της πρώτης συνάντησης, το σύστημα με καλύτερη απόδοση πέτυχε βαθμολογία-*F* 0,57, το οποίο συνιστά σημαντική βελτίωση για την επιστημονική κοινότητα. Τα επιτυχημένα συστήματα στις συναντήσεις κοινών εργασιών BioNLP βασίζονταν σε μια σειρά τεχνικών συμπεριλαμβανομένης της μηχανικής μάθησης, των λογικών δικτύων Markov και της ανάλυσης εξάρτησης. Παρακάτω περιγράφονται μια σειρά από προσεγγίσεις στην εξαγωγή βιοϊατρικών συμβάντων.

Τα περισσότερα συστήματα εξαγωγής συμβάντων ακολουθούν προσεγγίσεις ταχείας επεξεργασίας που διαιρούν την εργασία σε μια αλληλουχία τριών σταδίων. Πρώτον, τα συστήματα προβλέπουν ένα υποψήφιο σύνολο λέξεων εναύσματος (trigger words) συμβάντων. Οι λέξεις εναύσματος είναι συχνά τα ρήματα ή ουσιαστικοποιημένα ρήματα που υποδεικνύουν έναν συγκεκριμένο τύπο συμβάντος, όπως "φωσφορυλίωση", "ενεργοποιεί" ή "αναστέλλει". Τα συστήματα εν συνεχεία καθορίζουν εάν οι αναγνωρισμένες ονοματικές οντότητες ή λέξεις ενεργοποίησης αποτελούν συγκεκριμενοποίηση των ορισμάτων των συμβάντων. Το τελευταίο στάδιο της διαδικασίας είναι ένα βήμα σημασιολογικής μετα-επεξεργασίας που προσαρτά ορίσματα σε εναύσματα συμβάντων μετά από επιβολή περιορισμών στον τύπο και τον αριθμό των ορισμάτων που επιτρέπονται για έναν δεδομένο τύπο συμβάντων.

Αυτή η βασική αρχιτεκτονική είναι μια συχνά χρησιμοποιούμενη προσέγγιση στις εργασίες εξαγωγής συμβάντων. Ο Bjorne και οι συνεργάτες του έχουν περιγράψει το σύστημα με την καλύτερη απόδοση στην εργασία εξαγωγής συμβάντων BioNLP '09. Η μέθοδός τους προβλέπει την εκπαίδευση SVM πολλαπλής κατηγοριοποίησης για την ανίχνευση ορισμάτων και εναυσμάτων συμβάντων χρησιμοποιώντας ένα εκτεταμένο σύνολο χαρακτηριστικών, ειδικά αυτών που απορρέουν από τα γραφήματα συντακτικής ανάλυσης σχέσεων εξάρτησης (dependency parse graphs). Το σύστημά τους χρησιμοποιεί εν συνεχεία μια προσέγγιση που βασίζεται σε κανόνες για την προσάρτηση ορισμάτων στα αντίστοιχα συμβάντα τους. Αυτή η προσέγγιση έχει συνδυαστεί με το BANNER για την εκτέλεση της εξαγωγής συμβάντων σε μη επισημασμένο υποσύνολο παραπομπών από το PubMed. Ο Miwa και οι συνεργάτες του έχουν περιγράψει μια προσέγγιση εξαγωγής συμβάντων παρόμοια με αυτήν του Bjorne και των συνεργατών του, η οποία όμως αντί να επιτυγχάνει τη βελτίωση χρησιμοποιώντας έναν ταξινομητή και πρόσθετα χαρακτηριστικά για το συγκεκριμένο βήμα αυτό, αντί να χρησιμοποιεί μια

προσέγγιση που βασίζεται σε κανόνες για την προσάρτηση συμμετεχόντων συμβάντων σε λέξεις εναύσματος.

Ο Buyko και οι συνεργάτες του έχουν περιγράψει ένα σύστημα που χρησιμοποιεί μια προσέγγιση που βασίζεται σε λεξικά για τον εντοπισμό εναυσμάτων συμβάντων και ένα σύνολο ταξινομητών που βασίζονται σε χαρακτηριστικά και kernel, τα οποία έχουν εκπαιδευτεί χρησιμοποιώντας "περικομμένα" γραφήματα εξάρτησης για τον εντοπισμό συμμετεχόντων συμβάντων. Οι Kilicoglu και Bergler έχουν επίσης χρησιμοποιήσει μια προσέγγιση που βασίζεται σε λεξικά για τον εντοπισμό εναυσμάτων συμβάντων, αλλά ανέπτυξαν κανόνες με βάση τις διαδρομές συντακτικής εξάρτησης για την ανίχνευση συμμετεχόντων συμβάντων. Τέλος, ο Cohen και οι συνεργάτες του έχουν περιγράψει μια προσέγγιση που βασίζεται σε μοτίβα για την εξαγωγή συμβάντων, η οποία χρησιμοποιεί το σύστημα OpenDMAP για τον καθορισμό οντοτήτων και τύπων συμβάντων καθώς και για την επιβολή περιορισμών στα ορίσματα συμβάντων.

Πρόσφατα προτάθηκαν κοινές προσεγγίσεις πρόβλεψης, οι οποίες έχουν ως σκοπό την υπέρβαση του προβλήματος των αλληλοδιαδοχικών αστοχιών που εμφανίζονται σε ορισμένες από τις παραπάνω προσεγγίσεις. Για παράδειγμα, μέσω του διαχωρισμού της ανίχνευσης εναυσμάτων συμβάντων και ορισμάτων, το σύστημα πιθανώς να μην εξαγάγει σωστά ένα συμβάν εάν αποτύχει να ανιχνεύσει μια λέξη-έναυσμα στο πρώτο στάδιο της διαδικασίας. Οι Poon και Vanderwende έχουν προτείνει μια μέθοδο που βασίζεται στα λογικά δίκτυα του Markov, η οποία προβλέπει από κοινού συμβάντα και ορίσματα. Για κάθε λέξη, το σύστημα προβλέπει εάν πρόκειται για λέξη εναύσματος συμβάντος και για κάθε ακμή συντακτικής εξάρτησης, το σύστημα προβλέπει εάν πρόκειται για διαδρομή ορίσματος που οδηγεί σε θέμα ή αιτία συμβάντος. Επιπλέον, οι Riedel και Mc-Callum έχουν προτείνει μια ομάδα τριών μοντέλων από κοινού πρόβλεψης που βασίζονται στη λογική του Markov και χαρακτηρίζονται από μικρότερη υπολογιστική πολυπλοκότητα σε σχέση με προηγούμενες εργασίες ενώ οδηγούν και σε καλύτερα αποτελέσματα εξαγωγής συμβάντων.

ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΕΦΑΛΑΙΟΥ 2

[1] Michael W.Berry, Jacob Kogan. Text Mining, Applications and Theory

[2] Andrade M and Valencia A 1998 Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14(7), 600–607.

[3] Fox C 1989 A stop list for general text. *ACM SIGIR Forum*, vol. 24, pp. 19–21.ACM,New York, USA.

[4] Hulth A 2003 Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, vol. 10, pp. 216–223 Association for Computational Linguistics, Morristown, NJ, USA.

[5] Hulth A 2004 *Combining machine learning and natural language processing for automatic keyword extraction*. Stockholm University, Faculty of Social Sciences, Department of Computer and Systems Sciences (together with KTH).

[6] Jones K 1972 A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1), 11–21.

[7] Jones S and Paynter G 2002 Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. *Journal of the American Society for Information Science and Technology*.

[8] Matsuo Y and Ishizuka M 2004 Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* **13**(1), 157–169.

[9] G, Wong A and Yang C 1975 A vector space model for automatic indexing. *Communications of the ACM* **18**(11), 613–620.

[10] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2, 2005.

[11] L. Smith, L. Tanabe, R. Johnson nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. Struble, R. Povinelli, A. Vlachos, W. Baumgartner, L. Hunter, B. Carpenter, R. Tzong-Han Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Mana-Lopez, J. Mata, and W. Wilbur. Overview of BioCreAtIve II: Gene mention recognition. *Genome Biology*, 9(Suppl 2):S2, 2008.

[12] O. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[13] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 70–75, 2004.

[14] O. Tuason, L. Chen, L. H., and C. Friedman. Biological nomenclatures: A source of lexical knowledge and ambiguity. In *Pacific Symposium on Biocomputing*, pages 238–249, 2004.
[15] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, 2004.
[16] Y. Tsuruoka and J. Tsujii. Probabilistic term variant generator for biomedical terms. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 167–173, 2003.
[17] Y. Tsuruoka and J. Tsujii. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 41–48, 2003.
[18] M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. A biological named entity recognizer. In *Pacific Symposium on Biocomputing*, pages 427–438, 2003.
[19] W.-J. Hou and H.-H. Chen. Enhancing performance of protein name recognizers using collocation. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 25–32, 2003.
[20] K. Franz´en, G. Eriksson, F. Olsson, L. Asker, P. Lid´en, and J. C¨oster. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3):49–61, 2002.
[21] A. Vlachos and C. Gasperin. Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 138–145, 2006
[22] A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe. Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, 37(6):396–410, 2004.
[23] Y. Usami, H.-C. Cho, N. Okazaki, and J. Tsujii. Automatic acquisition of huge training data for bio-medical named entity recognition. In *Proceedings of BioNLP 2011 Workshop*, pages 65–73, 2011.
[24] C. Nobata, N. Collier, and J.-i. Tsujii. Automatic term identification and classification in biology texts. In *Proceedings of the Natural Language Pacific Rim Symposium*, pages 369–374, 1999.
[25] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 Workshop on Natural Language Processing*

in the Biomedical Domain - Volume 3, pages 1–8, 2002
[26] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6(Suppl 1):S8, 2005.
[27] K. Takeuchi and N. Collier. Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, 33(2):125–137, 2005.
[28] K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto. Protein name tagging for biomedical annotation in text. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 65–72, 2003.
[29] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In *3rd ACL SIGDAT Workshop on Very Large Corpora*, pages 82–94, 1995.
[30] B. Gu. Recognizing nested named entities in GENIA corpus. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 112–113, 2006.
[31] B. Alex, B. Haddow, and C. Grover. Recognising nested named entities in biomedical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 65–72, 2007.
[32] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6(Suppl 1):S8, 2005.
[33] K. Takeuchi and N. Collier. Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, 33(2):125–137, 2005.
[34] K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto. Protein name tagging for biomedical annotation in text. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, pages 65–72, 2003.
[35] J. Hakenberg, S. Bickel, C. Plake, U. Brefeld, H. Zahn, L. Faulstich, U. Leser, and T. Scheffer. Systematic feature evaluation for gene name recognition. *BMC Bioinformatics*, 6(Suppl 1):S9, 2005.
[36] N. Collier, C. Nobata, and J.-i. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, pages 201–207, 2000.
[37] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair. Exploiting context for biomedical entity recognition:

From syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 88–91, 2004.
[38] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107, 2004
[39] O. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.
[40] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2, 2005.
[41] A. B. Abacha and P. Zweigenbaum. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, pages 56–64, 2011.
[42] G. Zhou, D. Shen, J. Zhang, J. Su, and S. Tan. Recognition of protein/ gene names from text using an ensemble of classifiers. *BMC Bioinformatics*, 6(Suppl 1):S7, 2005.
[43] S. Mika and B. Rost. Protein names precisely peeled off free text. *Bioinformatics*, 20(suppl 1):i241–i247, 2004.

ΚΕΦΑΛΑΙΟ 3: ΤΟ ΛΟΓΙΣΜΙΚΟ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ WEKA

Το πρόγραμμα Weka αποτελεί μία συλλογή από αλγορίθμους μηχανικής μάθησης και εργαλεία προεπεξεργασίας (**pre-processing tools**). Περιλαμβάνει τους βασικότερους αλγορίθμους εξόρυξης πληροφορίας και παρέχει εκτενή υποστήριξη στον χρήστη που πειραματίζεται με την εξόρυξη πληροφορίας, περιλαμβάνοντας την προετοιμασία των δεδομένων εισόδου, στατιστική αποτίμηση των εξαγομένων από τους αλγορίθμους μοντέλων καθώς και οπτικοποίηση των αποτελεσμάτων. Πέρα από τους αλγορίθμους το πρόγραμμα διαθέτει και μεγάλη ποικιλία εργαλείων προεπεξεργασίας.

3.1 Προετοιμασία δεδομένων εισόδου

Σε κάθε πρόβλημα εξόρυξης δεδομένων, πρώτα είναι απαραίτητο να συγκεντρώσουμε όλα τα δεδομένα σε ένα σύνολο στιγμιοτύπων. Η εργασία αυτή δεν είναι τόσο εύκολη όσο δείχνει επειδή τα δεδομένα προέρχονται συνήθως από διαφορετικές πηγές και είναι ασύμβατα. Για παράδειγμα, έστω ότι στα πλαίσια μιας πραγματικής επιχειρηματικής εφαρμογής επιθυμούμε να κάνουμε έρευνα αγοράς. Προκύπτει ότι είναι αναγκαία η άντληση πληροφοριών από διαφορετικά τμήματα της επιχείρησης. Χρειάζονται δεδομένα από το τμήμα πωλήσεων, το τμήμα τιμολόγησης πελατών και το τμήμα παροχής υπηρεσιών προς τους πελάτες.

Η προσπάθεια αυτή **ενοποίησης δεδομένων (data integration)** από διαφορετικές πηγές συνήθως παρουσιάζει πολλές προκλήσεις κυρίως πρακτικής φύσεως. Διαφορετικά τμήματα χρησιμοποιούν διαφορετικούς τρόπους διατήρησης ιστορικών, διαφορετικές συμβάσεις, διαφορετικές χρονικές περιόδους και γενικά διαφορετικές αναπαραστάσεις. Τα δεδομένα πρέπει να συλλεχθούν, να ολοκληρωθούν και να "καθαριστούν". Η δημιουργία βάσης δεδομένων με ολοκληρωμένα και κανονικοποιημένα δεδομένα από όλα τα τμήματα της επιχείρησης έχει μεγάλη στρατηγική αξία και αποτελεί απαραίτητο βήμα πριν την εξόρυξη πληροφορίας.

3.1.1 Μορφή εισόδου ARFF (ARFF format) και CSV (CSV format)

Όταν έχει συγκεντρωθεί ένα κανονικοποιημένο και ολοκληρωμένο σύνολο δεδομένων, ένας συνήθης τρόπος αναπαράστασής του είναι ένα **ARFF αρχείο (ARFF file)**. Παρακάτω φαίνεται ένα ARFF αρχείοπου περιλαμβάνει τα δεδομένα του παραδείγματος (αριθμητικά χαρακτηριστικά) με τον καιρό από το πρώτο κεφάλαιο.

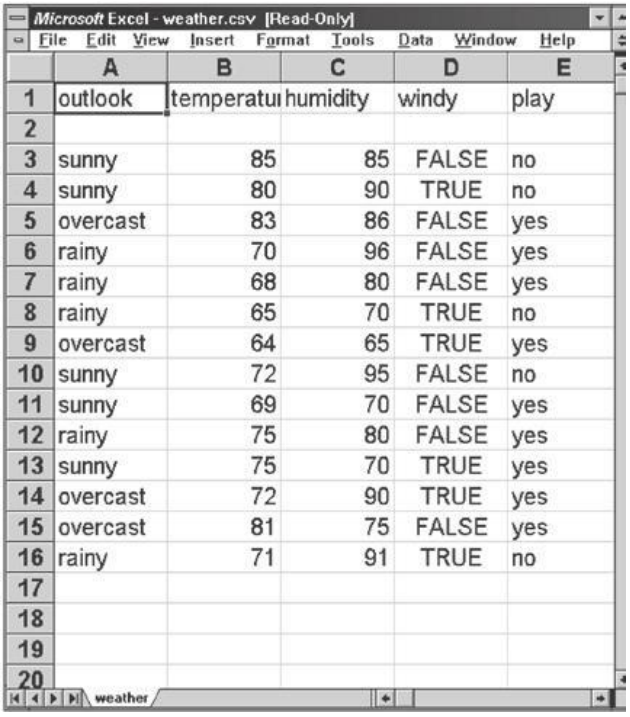
```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }
```

```
@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false,
yes rainy, 70, 96, false,
yes rainy, 68, 80, false,
yes rainy, 65, 70, true,
no overcast, 64, 65, true,
yes sunny, 72, 95, false,
no sunny, 69, 70, false,
yes rainy, 75, 80, false,
yes sunny, 75, 70, true,
yes overcast, 72, 90,
true, yes overcast, 81,
75, false, yes rainy, 71,
91, true, no
```

Στη γραμμή `@relation` καθορίζεται το όνομα της σχέσης στην οποία ανήκουν τα χαρακτηριστικά του συνόλου δεδομένων και είναι απαραίτητη. Ακολουθεί το όνομα του κάθε χαρακτηριστικού μετά τη δεσμευμένη λέξη `@attribute` καθώς και οι δυνατές τιμές του. Το ερωτηματικό στο play τοποθετείται απλά για να το ξεχωρίσουμε ως μεταβλητή ταξινόμησης, μπορεί και να παραληφθεί. Η επιλογή του attribute για μεταβλητή ταξινόμησης θα γίνει αργότερα μέσα στο πρόγραμμα αφού φορτωθούν τα δεδομένα. Αν στα `@data` έχουμε άγνωστες τιμές τοποθετούμε το σύμβολο ?. Για σχόλια χρησιμοποιούμε το σύμβολο `%`. Στο τέλος το αρχείο σώζεται με επέκταση .arff. Χρειάζεται κωδικοποίηση ANSI για να διαβαστεί από το Weka και να μην υπάρχει σφάλμα. Γι'αυτό το λόγο προτιμώνται προγράμματα σύνταξης κειμένου όπως το Notepad++, στα οποία γίνεται επιλογή της κωδικοποίησης.

Το format CSV είναι για βάσεις δεδομένων. Τα περισσότερα προγράμματα βάσεων δεδομένων επιτρέπουν την εξαγωγή δεδομένων σε μια κωδικοποίηση τιμών χωρισμένων με κόμματα(**comma- separated value format CSV**), δηλαδή μεταξύ των εγγραφών παρεμβάλλονται κόμματα.



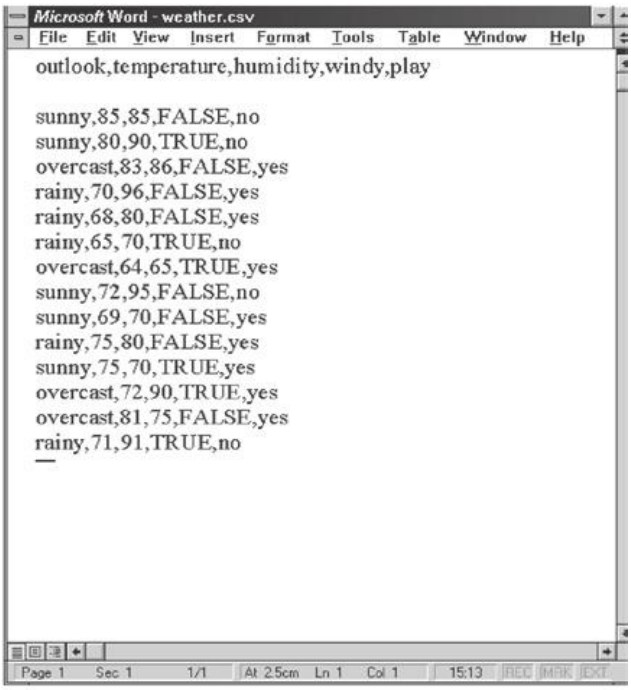
Microsoft Excel - weather.csv [Read-Only]

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E
1	outlook	temperatu	humidity	windy	play
2					
3	sunny	85	85	FALSE	no
4	sunny	80	90	TRUE	no
5	overcast	83	86	FALSE	yes
6	rainy	70	96	FALSE	yes
7	rainy	68	80	FALSE	yes
8	rainy	65	70	TRUE	no
9	overcast	64	65	TRUE	yes
10	sunny	72	95	FALSE	no
11	sunny	69	70	FALSE	yes
12	rainy	75	80	FALSE	yes
13	sunny	75	70	TRUE	yes
14	overcast	72	90	TRUE	yes
15	overcast	81	75	FALSE	yes
16	rainy	71	91	TRUE	no
17					
18					
19					
20					

weather

Αρχική μορφή δεδομένων



Microsoft Word - weather.csv

File Edit View Insert Format Tools Table Window Help

```
outlook,temperature,humidity,windy,play

sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Page 1 Sec 1 1/1 At 2.5cm Ln 1 Col 1 15:13 REC MARK EXT

CSV μορφή

Τα CSV αρχεία αντιμετωπίζονται ακριβώς όπως τα ARFF.

3.1.2 Αραιά δεδομένα

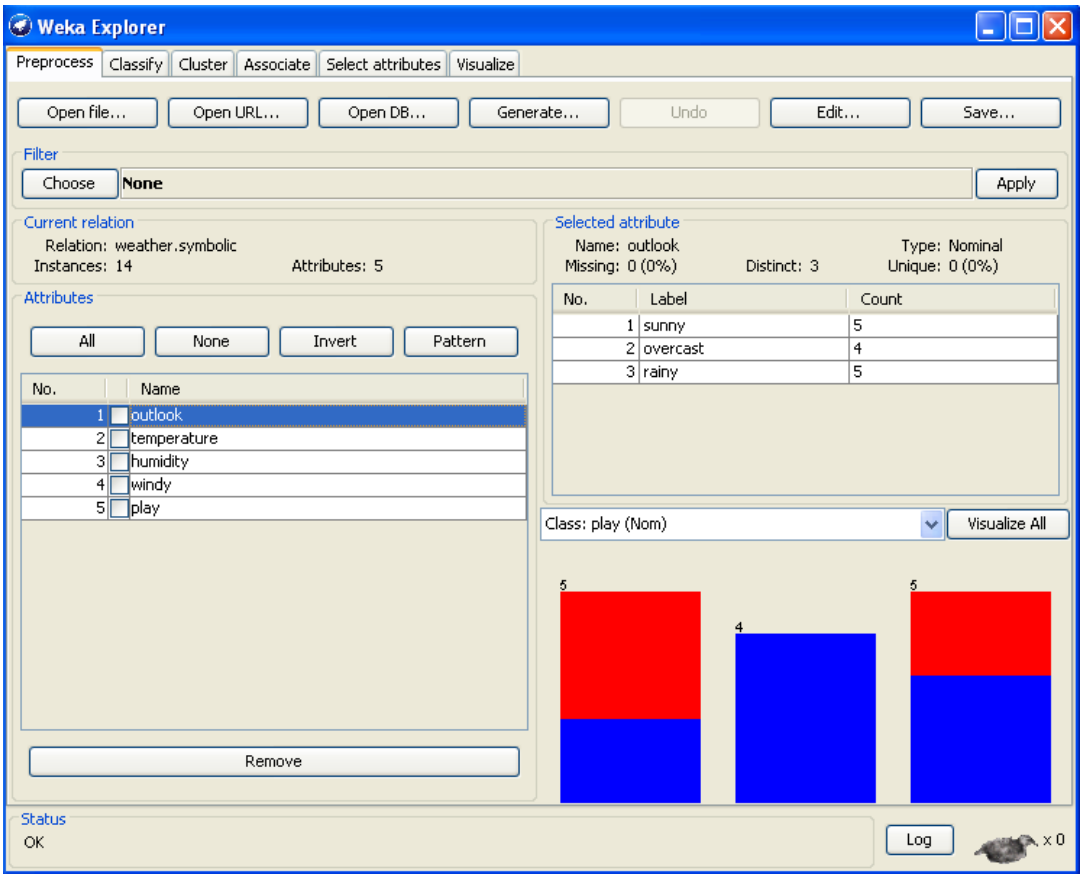
Υπάρχουν περιπτώσεις στις οποίες τα δεδομένα μοιάζουν με αραιούς πίνακες όπως στην παρακάτω περίπτωση

```
0, X, 0, 0, 0, 0, Y, 0, 0, 0, "class A"
0, 0, 0, W, 0, 0, 0, 0, 0, 0, "class B"
```

όπου τα περισσότερα χαρακτηριστικά έχουν μηδενική τιμή. Στην πράξη, αν ένα σούπερμαρκετ χωρίσει σε δύο κατηγορίες τους πελάτες του και καταγράψει την ποσότητα που αγοράζει κάθε κατηγορία από ένα προϊόν (στήλες τα προϊόντα, γραμμές οι δύο κατηγορίες πελατών), είναι λογικό να υπάρχουν πολλά μηδενικά από τη στιγμή που κάθε φορά αγοράζει κάποιος λίγα από τα συνολικά είδη. Παρόμοιο θέμα θα συναντήσουμε παρακάτω και στη μετατροπή ενός συνόλου κειμένων σε μήτρα αριθμητικών διανυσμάτων με σκοπό την εφαρμογή αλγορίθμων ταξινόμησης. Οι στήλες θα αντιπροσωπεύουν τις κοινές λέξεις των διαφορετικών κειμένων και οι γραμμές το εκάστοτε κείμενο. Τοποθετούμε την τιμή 1 αν το κείμενο περιέχει μια λέξη, 0 αν δεν την περιέχει. Πάλι τα δεδομένα θα είναι πολύ αραιά.

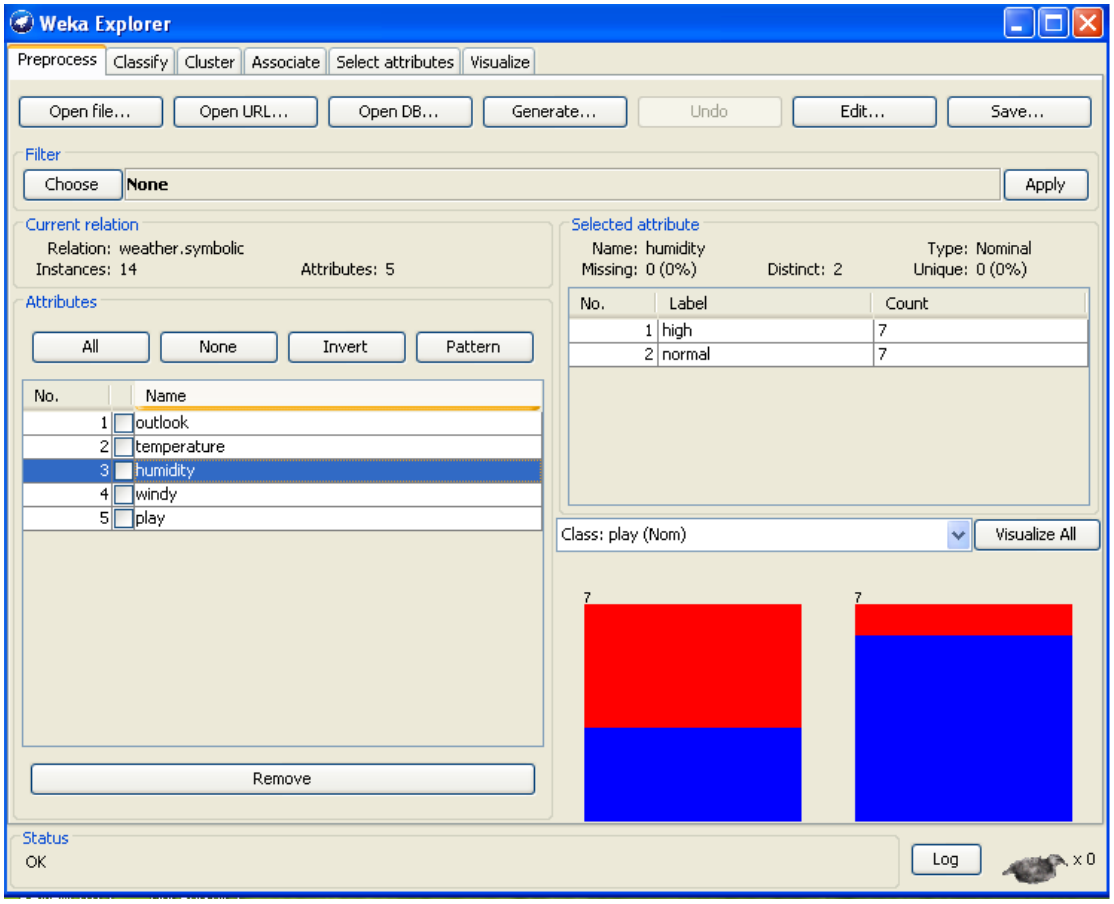
3.2 Το περιβάλλον του explorer

3.2.1 Προεπεξεργασία



Αρχικά πρέπει να φορτώσουμε τα δεδομένα μας. Πηγαίνουμε στην επιλογή Open file και επιλέγουμε το είδος του αρχείου που μας ενδιαφέρει. Αν είναι σε .csv μορφή, κάνουμε την αντίστοιχη επιλογή. Στο φάκελο data των εγκατεστημένων αρχείων του προγράμματος υπάρχει πληθώρα συνόλων δεδομένων arff μορφής για πειραματισμό. Για τις ανάγκες της επίδειξης διαλέγουμε το weather.nominal αρχείο. Στη συνέχεια πατάμε open και εμφανίζεται η παραπάνω οθόνη (αν είχαμε τα δεδομένα σε μορφή .csv πάλι η ίδια οθόνη θα εμφανιζόταν).

Παρατηρούμε ότι κάτω από τα attributes εμφανίζεται η επιλογή remove. Μας δίνει τη δυνατότητα να διαγράψουμε όσα attributes δε χρειαζόμαστε. Δεξιά, φαίνονται τα χαρακτηριστικά του εκάστοτε επιλεγμένου attribute: το missing αναφέρεται σε τυχόν άγνωστες τιμές σημειωθείσες στο arff αρχείο με ? και το unique πόσες από τις τιμές εμφανίζονται μόνο μια φορά στο σύνολο δεδομένων. Ανάλογα με την επιλογή μεταβλητής ταξινόμησης (class) και το επιλεγμένο αριστερά attribute αναπαρίσταται γραφικά το πλήθος εμφανίσεων της τιμής του attribute μεμονωμένα αλλά και σε σχέση με τις τιμές της μεταβλητής ταξινόμησης. Όπως φαίνεται παραπάνω, αν είναι επιλεγμένο το outlook, αναπαρίσταται το πλήθος εμφανίσεων κάθε τιμής του και πόσες φορές αυτή συνδέεται με τις τιμές yes , no της μεταβλητής ταξινόμησης play (yes=μπλε, no=κόκκινο). Αν είναι επιλεγμένο το humidity έχουμε την παρακάτω εικόνα



Προς το παρών δε μας χρειάζεται η επιλογή Filter. Περιέχει αλγορίθμους οι οποίοι μετατρέπουν συγκεκριμένους τύπους δεδομένων εισόδου σε άλλους κατάλληλους για εκπαίδευση ταξινομητών (όπως η μετατροπή ενός κειμένου σε αριθμητικό διάνυσμα). Θα την αναλύσουμε παρακάτω όταν μας χρειαστεί.

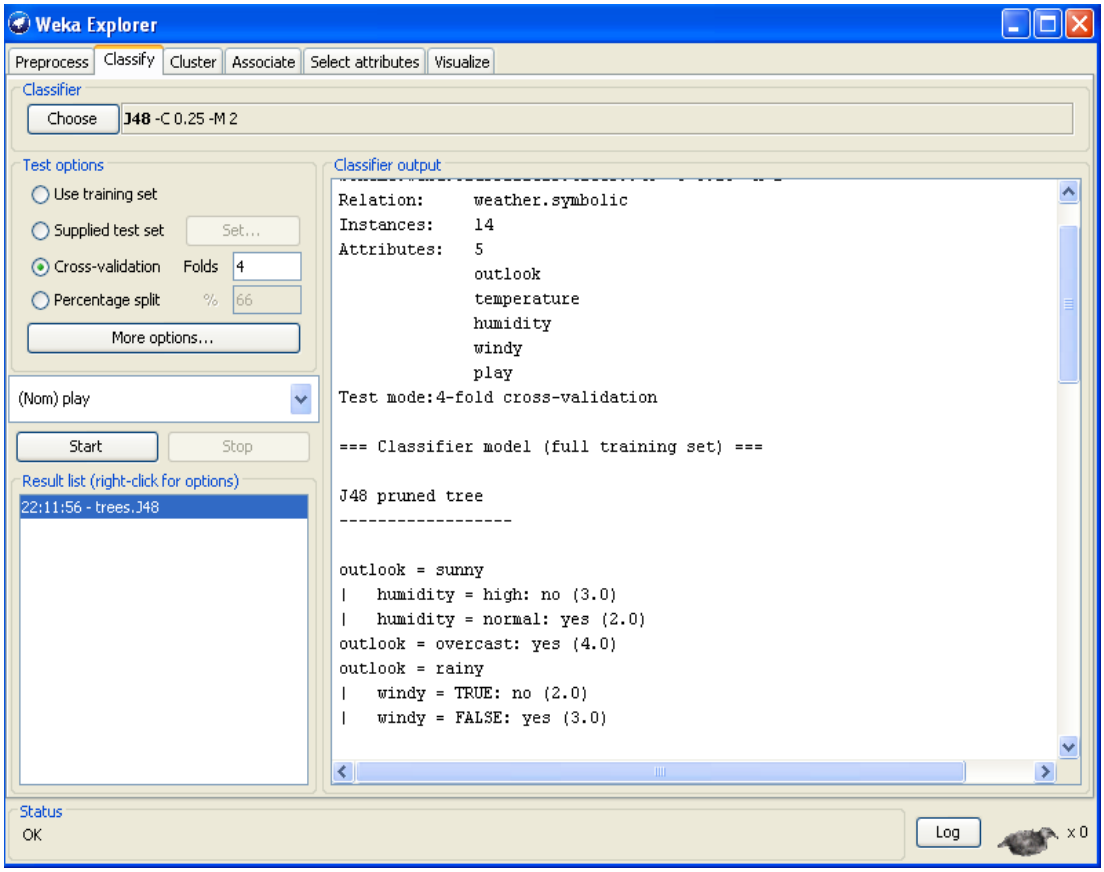
3.2.2 Εκπαίδευση και αποτίμηση ταξινομητών

Τώρα είμαστε έτοιμοι να εκπαιδεύσουμε κάποιον ταξινομητή της αρεσκείας μας με σκοπό την εξαγωγή μοντέλου. Βρισκόμαστε στην επιλογή Classify. Επιλέγουμε να κατασκευάσουμε δέντρο ταξινόμησης. Ο αλγόριθμος για το δέντρο ταξινόμησης επιλέγεται στο κουμπί classifier από τη λίστα trees και είναι ο J48.

Το μοντέλο J48 εξάγεται με είσοδο στιγμιότυπα εκπαίδευσης. Για να κρίνουμε την απόδοση του μοντέλου επιλέγουμε ένα νέο σύνολο στιγμιοτύπων και του ζητάμε να τα ταξινομήσει. Το νέο αυτό σύνολο ονομάζεται **σύνολο εξέτασης (test set)**. Στο πλαίσιο Test options δίνεται η δυνατότητα επιλογής του test set με τέσσερις διαφορετικούς τρόπους:

- Το test set να είναι το ίδιο με το training set.
- Τα στιγμιότυπα εξέτασης να επιλέγονται και να εισάγονται από το χρήστη. Το test set πρέπει να έχει ακριβώς τα ίδια attributes και όνομα στη σχέση με το training ώστε να είναι συμβατά μεταξύ τους. Στις κλάσεις των νέων στιγμιοτύπων στην περιοχή @data τοποθετούμε ? εφόσον είναι άγνωστες και πρέπει να προβλεφθούν.
- μέσω της **διασταυρωτής επαλήθευσης (cross validation)**. Στη γενική μέθοδο τα στιγμιότυπα εισόδου χωρίζονται τυχαία σε κ ομάδες στιγμιοτύπων. Στο πρόγραμμα, όπως θα δούμε παρακάτω, αν κάθε ομάδα αποτελείται από x στιγμιότυπα η πρώτη ομάδα αποτελείται από τα πρώτα x στιγμιότυπα, η δεύτερη από τα επόμενα x και ούτω καθ΄εξής. Κάθε ομάδα αποτελεί και ένα test set. Σε κάθε ομάδα αντιστοιχεί ένα training set αποτελούμενο από τα υπόλοιπα στιγμιότυπα εισόδου που δεν ανήκουν στην ομάδα. Άρα έχουμε συνολικά κ ζεύγη training- test sets. Η πρόβλεψη για κάθε test set γίνεται μέσω του προκύπτοντος μοντέλου από το αντίστοιχό του training set. Ουσιαστικά πραγματοποιείται μια διαδικασία επαναλαμβανόμενης εκπαίδευσης για την πρόβλεψη κάθε ομάδας στιγμιοτύπων. Στο περιβάλλον του weka το πλήθος κ των ομάδων ορίζεται από το χρήστη.
- Διαχωρισμός τον δεδομένων εισόδου σε δύο μέρη, ένα σύνολο εκπαίδευσης και ένα σύνολο εξέτασης. Ο χρήστης ορίζει το ποσοστό του δεδομένων εισόδου που θα χρησιμοποιηθεί για εκπαίδευση (*percentage split)* και τα στιγμιότυπα επιλέγονται τυχαία.

Πριν την εκτέλεση του αλγορίθμου στο κουμπί More options επιλέγουμε output predictions ώστε να φανεί στην έξοδο πώς ταξινομεί το μοντέλο το σύνολο εξέτασης που του παρέχουμε. Στην έξοδο φαίνεται το ακόλουθο αποτέλεσμα:



Για τη γραφική αναπαράσταση του δέντρου πατάμε δεξί κλικ στο όνομα του αλγορίθμου στο result list και επιλέγουμε visualize tree. Παρακάτω φαίνεται η συνολική έξοδος:

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
```

```
              windy
              play
Test mode:4-fold cross-validation

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

J48 pruned tree
------------------

outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :        5

Size of the tree :  8


Time taken to build model: 0.02 seconds

=== Predictions on test data ===

inst#,    actual, predicted, error, probability distribution
     1       2:no      1:yes      +  *1      0
     2       2:no       2:no          0     *1
     3      1:yes      1:yes         *1      0
     4      1:yes       2:no      +   0     *1
     1       2:no      1:yes      +  *0.667  0.333
     2      1:yes       2:no      +   0.25  *0.75
     3      1:yes      1:yes         *0.667  0.333
     4      1:yes      1:yes         *1      0
     1       2:no       2:no          0.333 *0.667
```

```
     2      1:yes      1:yes          *1      0
     3      1:yes       2:no      +   0.333 *0.667
     1       2:no      1:yes      +  *0.8    0.2
     2      1:yes       2:no      +   0     *1
     3      1:yes       2:no      +   0     *1

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

Correctly Classified Instances           6               42.8571 %
Incorrectly Classified Instances         8               57.1429 %
Kappa statistic                         -0.1429
Mean absolute error                      0.5393
Root mean squared error                  0.6715
Relative absolute error                114.3495 %
Root relative squared error            137.8741 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
Class
                 0.444     0.6         0.571     0.444     0.5         0.444
yes
                 0.4       0.556       0.286     0.4       0.333       0.444
no
Weighted Avg.    0.429     0.584       0.469     0.429     0.44        0.444

=== Confusion Matrix ===

 a b   <-- classified as
 4 5 | a = yes
 3 2 | b = no
```

Οι εσφαλμένες προβλέψεις δίνονται με +. Επιπλέον παρατηρούμε ότι οι προβλέψεις δίνονται σε δύο μορφές: στις **απόλυτες**(στήλη predicted) και στις

πιθανοτικές(στήλη probability distribution). Παρά το γεγονός ότι ο αλγόριθμος J48 φύσει δεν κάνει πιθανοτικές προβλέψεις αλλά απόλυτες, υπολογίζεται και η πιθανότητα του στιγμιοτύπου να ανήκει σε μία από τις δύο κλάσεις όπως γίνεται στο Naïve Bayes. Άλλωστε είναι γενικά προτιμότερο να γνωρίζουμε ότι κάποιο στιγμιότυπο ανήκει στην κλάση x με πιθανότητα 56%, παρά ότι απλά ανήκει στην κλάση x.

Συνάρτηση απώλειας τετραγώνων (quadric loss function): Εκτός από την απλή παράθεση των προβλέψεων του ταξινομητή ενδιαφερόμαστε να ορίσουμε και κάποια μεγέθη αξιολόγησής του. Αν κ είναι το πλήθος των κλάσεων, σε κάθε στιγμιότυπο αντιστοιχίζεται ένα διάνυσμα $p=(p_1, p_2, \ldots, p_\kappa)$ όπου p_j, $1 \leq j \leq \kappa$, η πιθανότητα πρόβλεψης το στιγμιότυπο να ανήκει στην κλάση j και το διάνυσμα $a=(a_1, a_2, \ldots, a_\kappa)$ όπου $a_j=1$ ή 0 ανάλογα με το αν η πραγματική κλάση του στιγμιοτύπου είναι η j ή όχι. Η συνάρτηση απώλειας τετραγώνων ορίζεται ως η ποσότητα $\sum_{j=1}^{k} (p_j - a_j)^2$ και εκφράζει το τετραγωνικό σφάλμα για κάθε στιγμιότυπο. Αν θέλουμε να υπολογίσουμε το συνολικό σφάλμα του ταξινομητή σε όλα τα στιγμιότυπα τότε απλά αθροίζουμε τα επί μέρους σφάλματα, δηλαδή υπολογίζουμε την ποσότητα $\sum_{i=1}^{n} \sum_{j=1}^{k} (p_{j,i} - a_{j,i})^2$.Φυσικά όλα τα παραπάνω ισχύουν για ονομαστικές μεταβλητές ταξινόμησης. Στην περίπτωση των **αριθμητικών μεταβλητών ταξινόμησης** σε κάθε στιγμιότυπο i αντιστοιχεί ο αριθμός πρόβλεψης p_i (μπορεί να είναι κάποιο φυσικό μέγεθος, για παράδειγμα θερμοκρασία) και ο πραγματικός αριθμός a_i με τη συνάρτηση απώλειας τετραγώνων να ορίζεται ανάλογα.

Για την αποτίμηση της απόδοσης του ταξινομητή ορίζονται τα ακόλουθα μεγέθη, τα οποία βασίζονται στη συνάρτηση απώλειας τετραγώνων και χρησιμοποιούνται στην πράξη:

<u>*Μέσο απόλυτο σφάλμα(Mean absolute error)*</u>: Το μέσο απόλυτο σφάλμα ορίζεται ως $\frac{\sum_{i=1}^{n} \sum_{j=1}^{k} |p_{j,i} - a_{j,i}|}{n}$.

<u>*Μέσο τετραγωνικό σφάλμα (Mean squared error)*</u> : $\frac{\sum_{i=1}^{n} \sum_{j=1}^{k} (p_{j,i} - a_{j,i})^2}{n}$

<u>*Τυπική απόκλιση σφάλματος(Root mean squared error):*</u> $\sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{k} (p_{j,i} - a_{j,i})^2}{n}}$

Τι μας λέει για παράδειγμα η γνώση μιας τιμής απόλυτου σφάλματος 0,2; Μας πληροφορεί ότι το μέσο σφάλμα του ταξινομητή είναι 0,2, δε μας πληροφορεί όμως

για το σφάλμα σε σχέση με το μέγεθος του συνόλου στιγμιοτύπων. Το 0,2 θα μπορούσε να είναι 1/5, δηλαδή να αναφέρεται σε πλήθος 5 στιγμιοτύπων, ή 2/10, δηλαδή να αναφέρεται σε πλήθος 10 στιγμιοτύπων. Άλλο είναι να έχουμε σφάλμα μεγέθους 1 σε 5 στιγμιότυπα και άλλο να έχουμε σφάλμα μεγέθους 2 σε 10 στιγμιότυπα. Το ίδιο ισχύει και για την τιμή 0,5393 του παραδείγματος. Συνεπώς χρειαζόμαστε άλλα κριτήρια περισσότερο αντιπροσωπευτικά της απόδοσης ενός ταξινομητή.

Σχετικό απόλυτο σφάλμα (Relative mean absolute error): Η λογική του είναι διαφορετική από αυτή των προηγουμένων κριτηρίων. Θα προσπαθήσουμε να δώσουμε εξηγήσεις θεωρώντας ότι η μεταβλητή ταξινόμησης είναι αριθμητική. Ο τύπος είναι $\frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=1}^{n} |a_i - \bar{a}|}$ με $\bar{a} = \frac{\sum_{i=1}^{n} p_i}{n}$. Ο παρονομαστής εκφράζει την συνολική απόκλιση των πραγματικών τιμών από τη μέση τιμή τους. Όσο αυτή είναι μεγαλύτερη, τόσο συγχωρείται η απόκλιση των προβλεπόμενων τιμών από τις πραγματικές (αριθμητής) και το σχετικό σφάλμα είναι μικρό. Αν όμως οι πραγματικές τιμές έχουν μικρή απόκλιση από το μέσο όρο, δηλαδή συγκλίνουν περισσότερο σε κάποια τιμή, τότε ο ταξινομητής δε δικαιολογείται να δώσει μεγάλο αριθμητή και αν το κάνει το σχετικό σφάλμα θα είναι μεγάλο. Ανάλογη είναι και η φυσική σημασία για την περίπτωση της ονομαστικής μεταβλητής ταξινόμησης, στην περίπτωση της οποίας ο παρονομαστής είναι το σφάλμα που προκύπτει από τον υπολογισμό της πιθανότητας εμφάνισης κάθε κλάσης (P(yes), P(no)), δηλαδή την απόκλιση(d_{error}) της κατανομής p_{yes}-p_{no} από την 0-1 κατανομή.

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{k} |p_{j,i} - a_{j,i}|}{d_{error}}$$

Σχετικό τετραγωνικό σφάλμα: $$\frac{\sum_{i=1}^{n}\sum_{j=1}^{k} (p_{j,i} - a_{j,i})^2}{d_{error}}$$

Τυπική απόκλιση σχετικού σφάλματος: $$\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{k} (p_{j,i} - a_{j,i})^2}{d_{error}^{\,2}}}$$

Συντελεστής Συσχέτισης: Έστω οι αριθμητικές μεταβλητές ταξινόμησης A,P, με τιμές τις πραγματικές και τις προβλεπόμενες αντιστοίχως. Ορίζεται ως $\frac{S_{ap}}{S_a * S_p}$ όπου σ_{ap} η τυπική απόκλιση του γινομένου μεταβλητών A*P, σ_a η τυπική απόκλιση της A, σ_p η

τυπική απόκλιση της P. Το μέγεθος εκφράζει τη συσχέτιση των μεταβλητών P και A, το κατά πόσο οι προβλέψεις p_i προσεγγίζουν τις a_i.

Από το πρόγραμμα επίσης καταγράφεται πόσες ήταν οι σωστές και οι λάθος προβλέψεις για κάθε κλάση σε ένα πίνακα που ονομάζεται **confusion matrix**.

Actual/predicted	yes	no
Yes	4	5
no	3	2

Το άθροισμα κάθε γραμμής είναι το πλήθος των πραγματικών στιγμιοτύπων κάθε κλάσης. Τα διαγώνια στοιχεία κάθε γραμμής αποτελούν τις σωστές προβλέψεις του ταξινομητή για την αντίστοιχη κλάση. Τα μη διαγώνια κάθε γραμμής είναι τα στιγμιότυπα της αντίστοιχης κλάσης τα οποία λανθασμένα έχουν ταξινομηθεί στις υπόλοιπες κλάσεις. 'Αρα κάθε στήλη μας δείχνει τα σωστά και λάθος ταξινομηθέντα στιγμιότυπα στην κλάση.

Για κάθε κλάση τα σωστά ταξινομημένα στιγμιότυπα ονομάζονται **αληθώς θετικά (TP)**, ενώ τα στιγμιότυπα που λανθασμένα ταξινομούνται στις υπόλοιπες κλάσεις **ψευδώς αρνητικά (FN).** Τα στιγμιότυπα που λανθασμένα ταξινομούνται σε μία κλάση ονομάζονται **ψευδώς θετικά (FP)** και αυτά που ορθώς ταξινομούνται στις υπόλοιπες κλάσεις **αληθώς αρνητικά (TN)**. Τα μεγέθη TP, FN, FP, TN ορίζονται για κάθε κλάση χωριστά. Βάσει αυτών ορίζονται και οι παρακάτω ποσότητες:

TP rate: Ο λόγος $\frac{TP}{TP + FN} 100\,\%$

FP rate: Ο λόγος $\frac{FP}{FP + TN} 100\,\%$

Recall: Ίδιο με το TP rate.

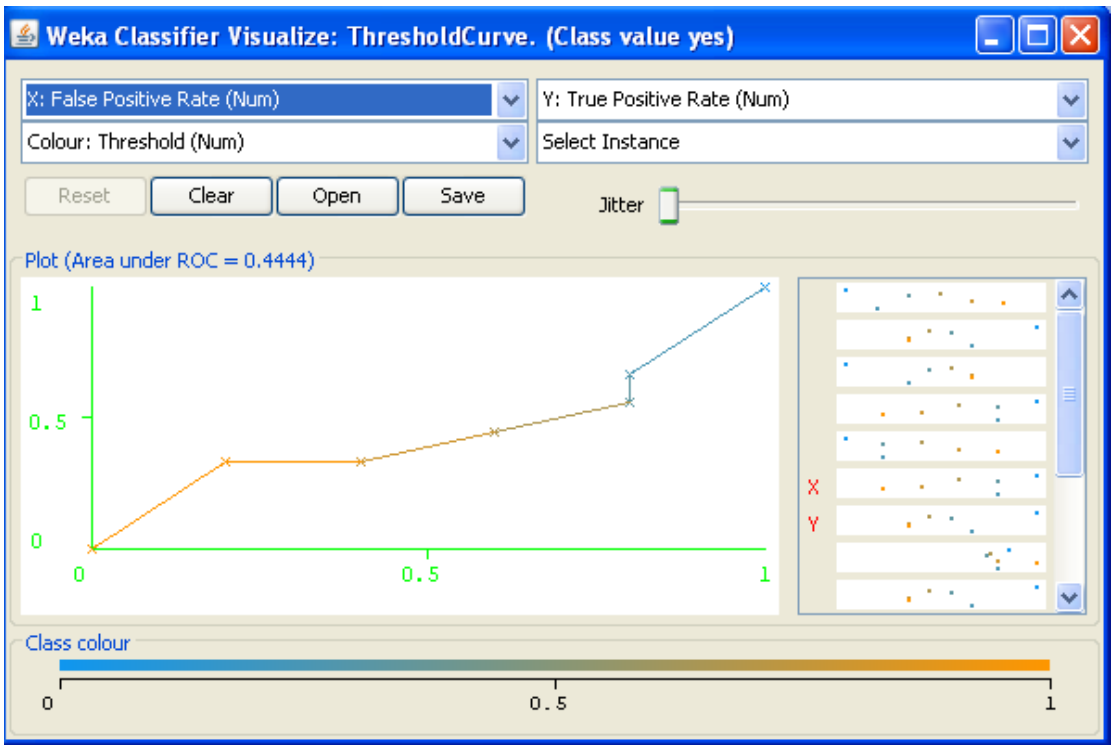
Precision: Ο λόγος $\frac{TP}{TP + FP} 100\,\%$

F-measure: $\frac{2 * recall * precision}{recall + precision} = \frac{2 * TP}{2 * TP + FP + FN} 100\,\%$

Καμπύλες ROC (Receiver Operating Characteristic): Πρόκειται για τις χαρακτηριστικές καμπύλες του ταξινομητή, μία για κάθε κλάση. Αναπαριστούν γραφικά το TP ποσοστό σε σχέση με το FP. Το κριτήριο για να αποφασίσει το μοντέλο αν ένα στιγμιότυπο ανήκει στη θετική κλάση είναι η κατανομή πιθανότητας πρόβλεψης. Αν μια πιθανότητα είναι μεγαλύτερη του 0,5 για τη θετική κλάση, σε αυτήν κατατάσσεται το στιγμιότυπο. Με αυτόν τον τρόπο δημιουργείται μια αναλογία TP-FP στο σύνολο των θετικών προβλέψεων. Αυτή η ποσότητα από μόνη της όμως δεν είναι και τόσο αντιπροσωπευτική της απόδοσης TP-FP του μοντέλου. Και αυτό γιατί το μοντέλο προβλέπει και αρνητικά στιγμιότυπα, με αποτέλεσμα οι

θετικές προβλέψεις να μην είναι αρκετές για να βγάλουμε ένα αντιπροσωπευτικό ποσοστό TP-FP. Αν όμως υποθέταμε ότι το μοντέλο κατατάσσει θετικό οποιοδήποτε στιγμιότυπο με πιθανότητα θετικής κλάσης πάνω από 0,2 για παράδειγμα; Δε θα είχαμε περισσότερες θετικές προβλέψεις για να βρούμε την αναλογία TP-FP; Γι'αυτό το λόγο επινοήθηκε η καμπύλη ROC, η οποία για διαφορετικά κατώφλια θετικών αποφάσεων αναπαριστά την αναλογία TP-FP. Τα κατώφλια παίρνουν τιμές από 0 μέχρι 1.

Πώς λαμβάνονται; Τοποθετούμε σε αύξουσα σειρά τις πιθανότητες πρόβλεψης, ξεκινώντας από τις πιθανότητες χειρότερης πρόβλεψης που είναι οι μικρότερες και καταλήγωντας στις καλύτερες προβλέψεις που είναι οι μεγαλύτερες. Κάθε μία από αυτές τις πιθανότητες την θεωρούμε κατώφλι θετικής απόφασης για το οποίο υπολογίζεται η αναλογία TP-FP στο σύνολο εκπαίδευσης. Έτσι, με τη βοήθεια χαρακτηριστικών σημείων δημιουργείται η καμπύλη. Όσο πιο πάνω αυτή η καμπύλη βρίσκεται από την ευθεία αναφοράς y=x (ίσο ποσοστό TP, FP για οποιοδήποτε κατώφλι) τόσο καλύτερο είναι το ποσοστό TP σε σχέση με το FP για διαφορετικά κατώφλια. Φυσικά, τμήματα της χαρακτηριστικής μπορεί να βρίσκονται πάνω ή κάτω από τη ευθεία αναφοράς. Η ιδανικότερη χαρακτηριστική είναι η οριζόντια TP=1.



Στο πρόγραμμα υπολογίζεται το εμβαδόν της επιφάνειας κάτω από την καμπύλη ROC *(ROC area)*. Το μέγιστο εμβαδόν που μπορεί να προκύψει και το οποίο αντιστοιχεί στην ιδανική χαρακτηριστική ισούται με μονάδα.

Οι καμπύλες ROC χρησιμοποιούνται συχνά για την αποτίμηση ταξινομητών[3].

ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΕΦΑΛΑΙΟΥ 3

[1] Ian Witten, Eibe Frank, Mark Hall. Data Mining 3rd edition

[2] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Εισαγωγή στην εξόρυξη δεδομένων.

[3] Beck, J. R., & Schultz, E. K. (1986). The use of ROC curves in test performance evaluation. *Archives of Pathology and Laboratory Medicine*, 110, 13–20.

ΚΕΦΑΛΑΙΟ 4: ΕΠΑΓΡΥΠΝΗΣΗ ΙΑΤΡΟΤΕΧΝΟΛΟΓΙΚΟΥ ΕΞΟΠΛΙΣΜΟΥ

Το σύστημα υγείας διεθνώς στηρίζεται σε μεγάλο βαθμό στη χρήση υψηλής τεχνολογίας ιατροτεχνολογικών προϊόντων. Υπάρχουν ορισμένες περιπτώσεις στις οποίες τα εν λόγω μηχανήματα αποτυγχάνουν να εκτελέσουν τη λειτουργία τους λόγω διαφόρων σφαλμάτων που μπορούν να υπάρξουν, με αποτέλεσμα να τίθεται σε κίνδυνο η ζωή των ασθενών. Εν συνεχεία συντάσσονται αναφορές για τις **αστοχίες** αυτές των προϊόντων, στις οποίες περιγράφεται η αστοχία, οι πληροφορίες του προϊόντος (είδος, κατασκευαστής), και αποστέλλονται στη διοίκηση του νοσοκομείου και στις αρμόδιες αρχές για τη λήψη των απαραίτητων μέτρων.

4.1 Κατηγοριοποίηση των αστοχιών

Οι αστοχίες του ιατρικού εξοπλισμού μπορούν να ταξινομηθούν με τρεις διαφορετικούς τρόπους:

i)ως προς τη βασική λειτουργία των ιατρικών μηχανημάτων

ii)ως προς την αιτία αστοχίας

iii)ως προς αν φταίει το υλικό (hardware) ή το λογισμικό (software)

4.1.1 Κατηγοριοποίηση των αστοχιών ως προς τη λειτουργία του ιατροτεχνολογικού εξοπλισμού

Οι γενικευμένες ομάδες (κλάσεις) ιατρικών προϊόντων όπως αυτές ορίζονται από τη διεθνώς αποδεκτή και αναγνωρισμένη ονοματολογία της GMDN (Global Medical Device Nomenclature) είναι οι εξής:

01 Active implantable devices - Ενεργά εμφυτεύσιμα προϊόντα
02 Anaesthetic and respiratory devices - Συσκευές αναισθησίας και αναπνευστικές συσκευές
03 Dental devices - Οδοντιατρικές συσκευές
04 Electro mechanical medical devices - Ηλεκτρομηχανικές συσκευές
05 Hospital hardware - Νοσοκομειακός εξοπλισμός
06 In vitro diagnostic devices - Διαγνωστικές συσκευές in vitro
07 Non-active implantable devices - Μη ενεργά εμφυτεύσιμα προϊόντα
08 Ophthalmic and optical devices - Οφθαλμολογικές συσκευές και συσκευές οπτικής
09 Reusable devices - Επαναχρησιμοποιήσιμα εργαλεία
10 Single use devices - Συσκευές μίας χρήσης
11 Assistive products for persons with disability - Τεχνική βοήθεια ατόμων με ειδικές ανάγκες
12 Diagnostic and therapeutic radiation devices - Διαγνωστικές και θεραπευτικές συσκευές ακτινοβολίας
13 Complementary therapy devices - Συμπληρωματικές συσκευές θεραπείας
14 Biological-derived devices - Βιολογικά παράγωγα μηχανήματα

15 Healthcare facility products and adaptations - Προϊόντα εγκαταστάσεων υγειονομικής περίθαλψης και προσαρμογές
16 Laboratory equipment - Εργαστηριακός εξοπλισμός

4.1.2 Κατηγοριοποίηση των αστοχιών ως προς την αιτία πρόκλησής τους

Η κατηγοριοποίηση των αστοχιών ως προς την αιτία πρόκλησής τους γίνεται με διάφορους τρόπους. Αφορά σε προβλήματα που προκύπτουν σχετικά με τον τρόπο λειτουργίας του συστήματος, με χειρισμούς δεδομένων εισόδου καθώς και συσκευές εξόδου και οπτικοποίησης:

- **Behaviour / Συμπεριφορά**- το σύστημα εκτελεί μια φυσική κίνηση η οποία οφείλεται σε κάποια έξοδο μιας συνάρτησης
- **Response / Απάντηση** - για μια λειτουργία που δεν θα έπρεπε να συμβεί
- **Data / Δεδομένα** - όταν υπάρχει μια αλλαγή ή απώλεια δεδομένων
- **Display / Προβολή** - αυτά τα συμπτώματα σχετίζονται με οπτική απεικόνιση των κειμένων, των αριθμών, ή εικόνες σε διάφορα μέσα, όπως είναι οι οθόνες και εκτυπωτές
- **Function / Λειτουργία** - λάθη στους υπολογισμούς και σε ορισμένες λειτουργίες του κώδικα
- **General / Γενικά** - για τις περιπτώσεις που δεν έχει επαρκείς πληροφορίες
- **Input / Είσοδος** - συμπτώματα που σχετίζονται με την είσοδο κάποιας λειτουργίας ή μονάδας, π.χ. δεδομένα που εισέρχονται από τον χρήστη, ή να διαβάζονται από ένα σκληρό δίσκο
- **Output / Έξοδος** - συνήθως το αποτέλεσμα μιας συνάρτησης ή μιας μονάδας
- **Service / Υπηρεσία** - δυσλειτουργίες του συστήματος με τρόπο που εμπλέκονται περισσότερα από ένα υποσυστήματα
- **System / Σύστημα** - ένα συνολικό πρόβλημα του συστήματος
- **User instructions / Οδηγίες χρήσης** - αναφέρεται στην έγγραφη τεκμηρίωση που δόθηκε στον χρήστη.

4.1.3 Κατηγοριοποίηση των αστοχιών ως προς το λογισμικό ή υλικό του ιατροτεχνολογικού εξοπλισμού

Το λογισμικό αποτελεί αναπόσπαστο κομμάτι ενός εύρους συσκευών οι οποίες πραγματοποιούν κρίσιμες για τη ζωή του ασθενούς λειτουργίες καθώς και απλούστερες οι οποίες χρειάζονται σε καθημερινή βάση. Καθώς οι ασθενείς καθίστανται όλο και περισσότερο εξαρτημένοι από ψηφιακές συσκευές, το λογισμικό πλέον συνδέεται στενά με πιθανά δυσμενή περιστατικά που μπορούν να συμβούν. Η ανάγκη για αντιμετώπιση των αδυναμιών του λογισμικού είναι ιδιαίτερα επιτακτική στις εμφυτεύσιμες συσκευές, τις χρησιμοποιούμενες από

εκατομμύρια ασθενείς για τη θεραπεία χρονίων καρδιακών παθήσεων, επιληψίας, διαβήτη, παχυσαρκίας ακόμη και κατάθλιψης.

Γενικά ένα πρόβλημα λογισμικού δεν αντιμετωπίζεται τόσο εύκολα όσο το hardware. Αρκετά συχνά είναι δύσκολος ο εντοπισμός και η επιδιόρθωση σφαλμάτων με τόσες πολλές και περίπλοκες λειτουργίες από τις οποίες αποτελείται. Το γεγονός αυτό μπορεί να οδηγήσει ακόμα και σε θάνατο στη χειρότερη περίπτωση όπως συνέβη σε τουλάχιστον 212 ασθενείς με εμφυτευμένες καρδιακές συσκευές την περίοδο 1997-2003 από τους 450.000 παγκοσμίως που επίσης είχαν χρησιμοποιήσει συσκευές του ιδίου είδους[1].

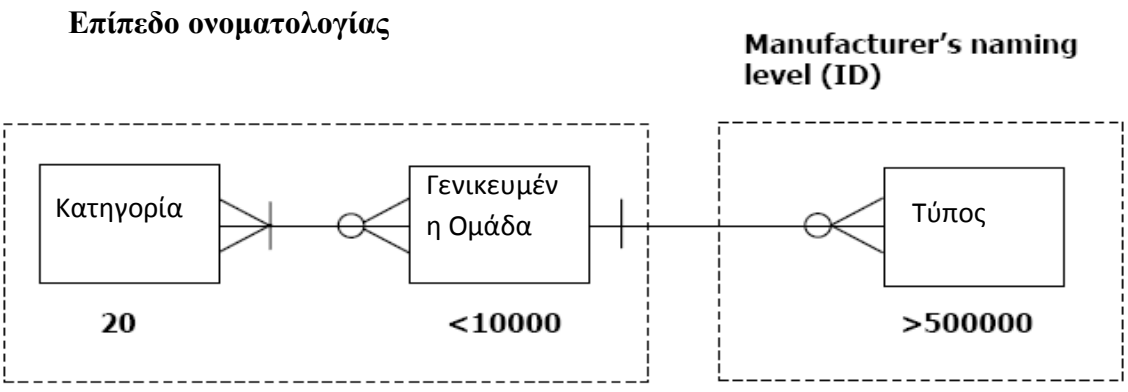
4.2 Ονοματολογία και κωδικοποίηση ιατρικών μηχανημάτων

Διαχρονικά έχουν γίνει διάφορες προσπάθειες για την κατάρτιση μιας κοινής ονοματολογίας και κωδικοποίησης ιατρικού εξοπλισμού διεθνούς αποδοχής. Η πρώτη προσπάθεια έγινε από το μη κερδοσκοπικό αμερικανικό οργανισμό με αντικείμενο τη βϊοιατρική τεχνολογία ECRI το 1970 αλλά δεν είχε αρχικά επιτυχία εφόσον δεν έγινε αποδεκτό το πρότυπο σε χώρες πέραν της Αμερικής. Ανάλογες προσπάθειες από τον FDA είχαν αντίστοιχη κατάληξη. Τελικά, ένα πιο ολοκληρωμένο σύστημα, το UMDNS του ECRI έγινε αποδεκτό από περίπου 40 χώρες, ενώ το πρότυπο που υιοθετήθηκε στα μέσα της δεκαετίας του ‘90 από τα κράτη μέλη της Ευρωπαϊκής Ένωσης και ονομάζεται GMDN [2] (Global Medical Device Nomenclature) βασίζεται στο UMDNS.

4.2.1 Γενική Δομή GMDN

Η γενική δομή της GMDN ρυθμίζεται από τις απαιτήσεις του προτύπου ISO 15225:2000, *Nomenclature – Specification for a nomenclature system for medical devices for the purpose of regulatory data exchange* και της διόρθωσής του ISO 15225:2000/Amd 1:200X ή των ευρωπαϊκών ισοδύναμων EN προτύπων.

Η γενική μορφή της ονοματολογίας φαίνεται στο παρακάτω σχήμα



Όπως βλέπουμε, αποτελείται από τρία επίπεδα. Το πρώτο και γενικότερο επίπεδο περιλαμβάνει τις 20 κατηγοριοποιήσεις του ιατρικού εξοπλισμού. Στο δεύτερο

επίπεδο ανήκουν οι ονομασίες των προϊόντων τα οποία κατατάσσονται στις κατηγορίες του πρώτου επιπέδου. Το πλήθος των γενικευμένων ομάδων αγγίζει τις 10000. Για κάθε είδος προϊόντος υπάρχουν μάρκες από διαφορετικούς κατασκευαστές οι οποίες ξεπερνούν σε αριθμό τις 500000. Η τελευταία ανισότητα αντιπροσωπεύει ουσιαστικά το πλήθος των διαφορετικών ιατρικών προϊόντων που υπάρχουν μέχρι στιγμής παγκοσμίως.

4.2.2 Γενικευμένες Κατηγορίες Ιατροτεχνολογικών Προϊόντων

Το πρότυπο EN ISO 15225 ορίζει κωδικούς για 20 πιθανές κατηγορίες. Μόνο οι 16 από τις 20 κατηγορίες είναι ενεργοποιημένες. Οι Κατηγορίες 17-20 είναι κενές και αναμένεται να συμπληρωθούν με την εξέλιξη της τεχνολογίας.

Πίνακας 1 : **Γενικευμένες Κατηγορίες Ιατροτεχνολογικών Προϊόντων**

ΚΩΔΙΚΟΣ	**ΓΕΝΙΚΕΥΜΕΝΗ ΚΑΤΗΓΟΡΙΑ ΙΑΤΡΟΤΕΧΝΟΛΟΓΙΚΩΝ ΠΡΟΪΟΝΤΩΝ**	**DEVICE CATEGORY**
1	**Ενεργά εμφυτεύσιμα προϊόντα**	**Active implantable devices**
2	**Αναισθησιολογικά και αναπνευστικά προϊόντα**	**Anaesthetic and respiratory devices**
3	**Οδοντιατρικά προϊόντα**	**Dental devices**
4	**Ηλεκτρομηχανικά ιατροτεχνολογικά προϊόντα**	**Electro mechanical medical devices**
5	**Εξοπλισμός εγκαταστάσεων**	**Hospital hardware**
6	**In vitro διαγνωστικά προϊόντα**	**In vitro diagnostic devices, IVD**
7	**Μη ενεργά εμφυτεύσιμα προϊόντα**	**Non-active implantable devices**
8	**Οφθαλμολογικά προϊόντα και προϊόντα οπτικής**	**Ophthalmic and optical devices**
9	**Επαναχρησιμοποιήσιμα εργαλεία**	**Reusable devices**
10	**Προϊόντα μίας χρήσης**	**Single use devices**
11	**Τεχνικά βοηθήματα για άτομα με ειδικές ανάγκες**	**Assistive products for persons with disability**
12	**Προϊόντα διαγνωστικής και θεραπευτικής ακτινοβολίας**	**Diagnostic and therapeutic radiation devices**
13	**Προϊόντα συμπληρωματικής θεραπείας**	**Complementary therapy devices**
14	**Προϊόντα Βιολογικής προέλευσης**	**Biological-derived devices**

15	**Προϊόντα νοσοκομειακών εγκαταστάσεων και προσαρμογής**	**Healthcare facility products and adaptations**
16	**Εργαστηριακός Εξοπλισμός**	**Laboratory equipment**
17	**Κενή, αναμένεται να συμπληρωθεί**	**Reserved**
18	**Κενή, αναμένεται να συμπληρωθεί**	**Reserved**
19	**Κενή, αναμένεται να συμπληρωθεί**	**Reserved**
20	**Κενή, αναμένεται να συμπληρωθεί**	**Reserved**

4.3 Συστήματα κωδικοποίησης ιατροτεχνολογικού εξοπλισμού

Εκτός από τα πρότυπα ονοματολογίας έχουν αναπτυχθεί και συστήματα κωδικοποίησης των προϊόντων με σκοπό την αντιστοίχιση ενός κωδικού σε οποιοδήποτε προϊόν. Σκοπός της κωδικοποίησης είναι η ευκολότερη ανταλλαγή πληροφοριών και γι'αυτό είναι απαραίτητη συνοδεύουσα την ονοματολογία. Υπάρχει το διεθνώς αποδεκτό σύστημα της UMDNS αλλά και κάθε χώρα διατηρεί το δικό της σύστημα κωδικοποίησης. Μερικά από αυτά φαίνονται παρακάτω. Γενικά, στα περισσότερα εθνικά συστήματα κωδικοποίησης οι κωδικοί κλίνουν στη λογική της ονοματολογίας εκτός από της Νορβηγίας το οποίο λειτουργεί διαφορετικά και ενέπνευσε κιόλας το UMDNS.

FDA Standard Product Nomenclature

Ο συγκεκριμένος οργανισμός έχει υπό την εποπτεία του οποιοδήποτε είδος τροφίμου, φαρμακευτική ουσία ή ιατρική συσκευή κυκλοφορεί στην αγορά παγκοσμίως.

Κάθε συσκευή ή υλικό λαμβάνει ένα πενταψήφιο κωδικό. Οι δύο πρώτες θέσεις του κωδικού είναι ψηφία και αφορούν μία γενικευμένη κατηγορία. Οι τρεις τελευταίες θέσεις δείχνουν τη γενικευμένη ομάδα προϊόντων και είναι κεφαλαίοι λατινικοί χαρακτήρες.

Ιταλία

Τρία πεδία χρησιμοποιούνται για την ταξινόμηση του εκάστοτε προϊόντος, η γενικευμένη κατηγορία, ο κατασκευαστής και η ομάδα. Ο κωδικός κάθε προϊόντος είναι οκταψήφιος. Το πρώτο πεδίο περιέχει τρία γράμματα και αφορά στην κατηγορία, το δεύτερο τρία γράμματα για τον κατασκευαστή και το τρίτο δύο ψηφία για την ομάδα.

Νορβηγία

Σε αυτό το πρότυπο δεν ακολουθείται η λογική των δύο προηγουμένων, αυτή της ονοματολογίας, αλλά ο κωδικός είναι μοναδικός αριθμός μεγαλύτερος από 9999 χωρίς αυτός να κωδικοποιεί με κάποιον τρόπο την κατηγορία και την ομάδα του προϊόντος. Η απόδοση του αριθμού γίνεται τυχαία, ο αριθμός δεν έχει κάποιο νόημα και κάθε προϊόν έχει τον δικό του αριθμό. Οι κωδικοί οι μικρότεροι του 9999 (<9,999) αφιερώνονται για την εισαγωγή νέων προϊόντων από τους χρήστες. Για παράδειγμα, μπορεί ένα νοσοκομείο να χρησιμοποιεί κάποιες συσκευές οι οποίες δεν υπάρχουν σε άλλα. Έτσι, εισάγει τον δικό του κωδικό για τη συσκευή χωρίς να εγκυμονεί κίνδυνος χρήσης ενός κωδικού μεγαλύτερου από 9999, ο οποίος είναι κατειλημμένος και χρησιμοποιείται από το εθνικό σύστημα υγείας.

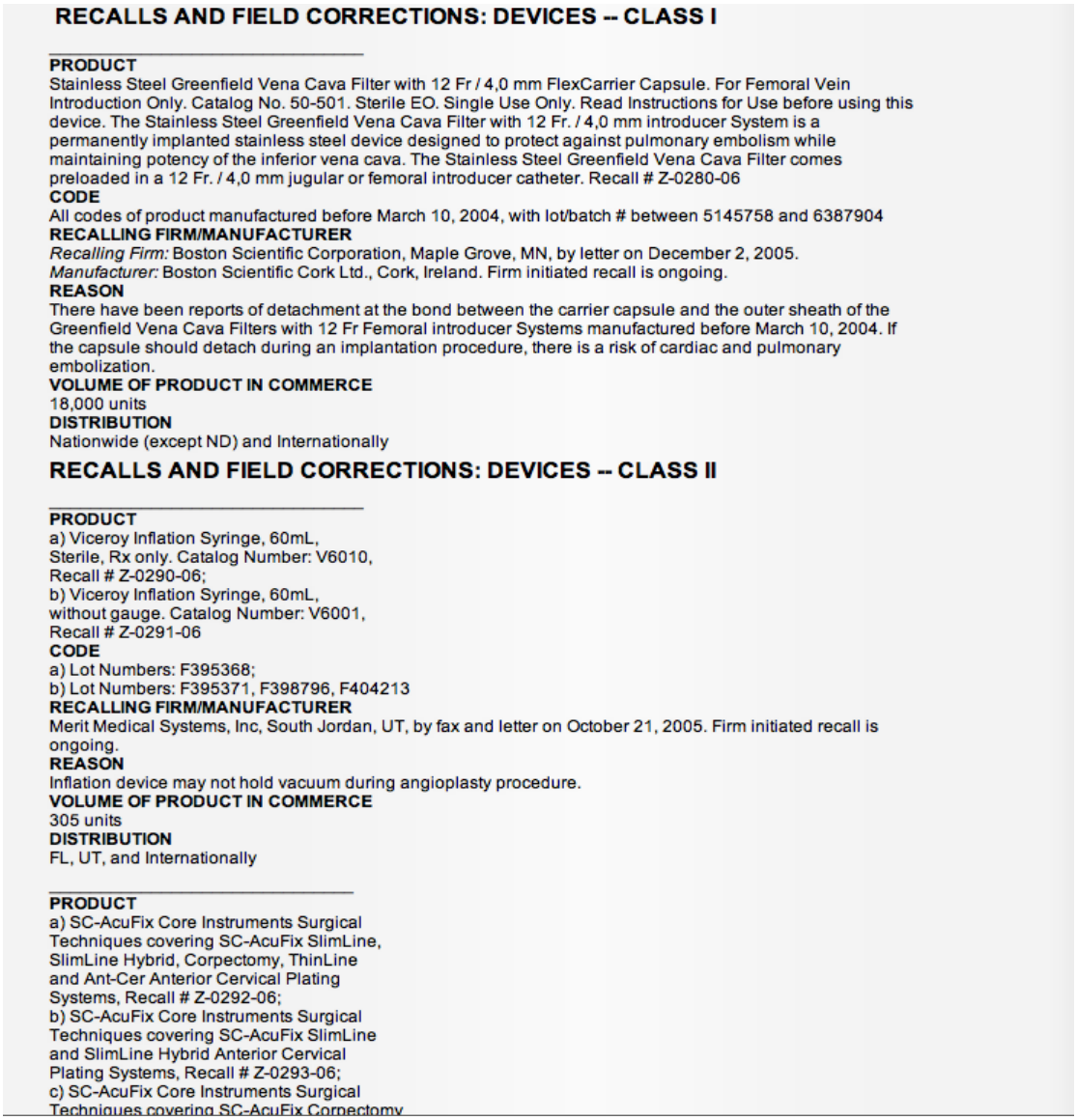
UMDNS / ECRI

Ο κωδικός κάθε προϊόντος είναι μεγαλύτερος ή ίσος του 10000 και αποδίδεται με τυχαίο τρόπο. Είναι εμφανείς οι ομοιότητες του διεθνούς με το νορβηγικό πρότυπο, στο οποίο άλλωστε βασίζεται.

4.4 Αναφορές αστοχιών σε διεθνείς βάσεις δεδομένων

Τα κράτη παγκοσμίως έχουν συμφωνήσει στην κατάρτιση του προτύπου κοινής αποδοχής GMDN για την κατηγοριοποίηση του ιατροτεχνολογικού εξοπλισμού. Οποιοδήποτε εθνικό σύστημα υγείας περιλαμβάνει ιατρικά προϊόντα τα οποία, ανεξάρτητα από τη μάρκα και τον κατασκευαστή τους, μπορεί να είναι οφθαλμολογικές, οδοντιατρικές συσκευές, ενεργά εμφυτεύσιμα προϊόντα ή επαναχρησιμοποιήσιμα εργαλεία ή οποιαδήποτε άλλη από τις 16 καθορισμένες κατηγορίες.

Ένα από τα πλεονεκτήματα της μοντελοποίησης της ονοματολογίας του εξοπλισμού, μοντέλο το οποίο αντιπροσωπεύει τον εξοπλισμό οποιουδήποτε εθνικού συστήματος υγείας, είναι ότι οι αναφορές δυσμενών περιστατικών παύουν πλέον να έχουν μόνο τοπικό χαρακτήρα. Η περιγραφή της αστοχίας, λόγου χάριν, μιας διαγνωστικής ακτινολογικής συσκευής στη νοσοκομειακή μονάδα μιας χώρας μπορεί να κινητοποιήσει κάποιο άλλο νοσοκομείο, ακόμα και σε άλλη χώρα, για τη λήψη κατάλληλων μέτρων πάνω στις δικές του ακτινοδιαγνωστικές συσκευές. Ειδικά στις μέρες μας, με τη ραγδαία εξέλιξη των επικοινωνιών και του διαδικτύου, υπάρχει η δυνατότητα διατήρησης ολόκληρων βάσεων δεδομένων αποτελούμενες από αναφορές αστοχιών με συχνές και διαρκείς ενημερώσεις. Τέτοιες βάσεις δεδομένων συντηρούνται από ιατρικούς οργανισμούς όπως είναι οι FDA (U.S Food and Drug Administration), EUDAMED (EUropean DAtabase on MEdical Devices), ECRI (Emergency Care Research Institute), MDA (Medical Devices Agency), IMB (Irish Medicine Board) και προσπελαύνονται μέσω αντίστοιχων ιστοσελίδων. Παραδείγματα της ιστοσελίδας FDA φαίνονται παρακάτω



RECALLS AND FIELD CORRECTIONS: DEVICES -- CLASS I

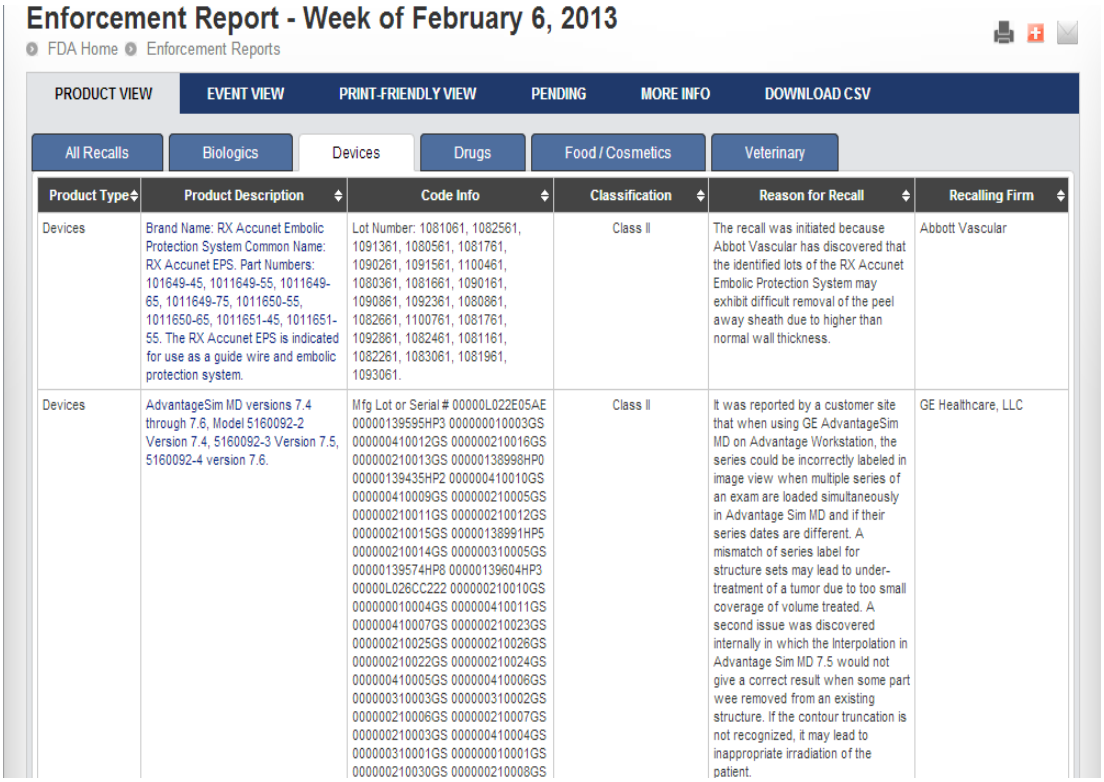
PRODUCT
Stainless Steel Greenfield Vena Cava Filter with 12 Fr / 4,0 mm FlexCarrier Capsule. For Femoral Vein Introduction Only. Catalog No. 50-501. Sterile EO. Single Use Only. Read Instructions for Use before using this device. The Stainless Steel Greenfield Vena Cava Filter with 12 Fr. / 4,0 mm introducer System is a permanently implanted stainless steel device designed to protect against pulmonary embolism while maintaining potency of the inferior vena cava. The Stainless Steel Greenfield Vena Cava Filter comes preloaded in a 12 Fr. / 4,0 mm jugular or femoral introducer catheter. Recall # Z-0280-06
CODE
All codes of product manufactured before March 10, 2004, with lot/batch # between 5145758 and 6387904
RECALLING FIRM/MANUFACTURER
Recalling Firm: Boston Scientific Corporation, Maple Grove, MN, by letter on December 2, 2005.
Manufacturer: Boston Scientific Cork Ltd., Cork, Ireland. Firm initiated recall is ongoing.
REASON
There have been reports of detachment at the bond between the carrier capsule and the outer sheath of the Greenfield Vena Cava Filters with 12 Fr Femoral introducer Systems manufactured before March 10, 2004. If the capsule should detach during an implantation procedure, there is a risk of cardiac and pulmonary embolization.
VOLUME OF PRODUCT IN COMMERCE
18,000 units
DISTRIBUTION
Nationwide (except ND) and Internationally

RECALLS AND FIELD CORRECTIONS: DEVICES -- CLASS II

PRODUCT
a) Viceroy Inflation Syringe, 60mL,
Sterile, Rx only. Catalog Number: V6010,
Recall # Z-0290-06;
b) Viceroy Inflation Syringe, 60mL,
without gauge. Catalog Number: V6001,
Recall # Z-0291-06
CODE
a) Lot Numbers: F395368;
b) Lot Numbers: F395371, F398796, F404213
RECALLING FIRM/MANUFACTURER
Merit Medical Systems, Inc, South Jordan, UT, by fax and letter on October 21, 2005. Firm initiated recall is ongoing.
REASON
Inflation device may not hold vacuum during angioplasty procedure.
VOLUME OF PRODUCT IN COMMERCE
305 units
DISTRIBUTION
FL, UT, and Internationally

PRODUCT
a) SC-AcuFix Core Instruments Surgical
Techniques covering SC-AcuFix SlimLine,
SlimLine Hybrid, Corpectomy, ThinLine
and Ant-Cer Anterior Cervical Plating
Systems, Recall # Z-0292-06;
b) SC-AcuFix Core Instruments Surgical
Techniques covering SC-AcuFix SlimLine
and SlimLine Hybrid Anterior Cervical
Plating Systems, Recall # Z-0293-06;
c) SC-AcuFix Core Instruments Surgical
Techniques covering SC-AcuFix Corpectomy

Εικόνα 1: *Μορφή FDA μέχρι και 13 Ιουνίου 2012*



Enforcement Report - Week of February 6, 2013

FDA Home ⊙ Enforcement Reports

PRODUCT VIEW | EVENT VIEW | PRINT-FRIENDLY VIEW | PENDING | MORE INFO | DOWNLOAD CSV

All Recalls | Biologics | Devices | Drugs | Food / Cosmetics | Veterinary

Product Type	Product Description	Code Info	Classification	Reason for Recall	Recalling Firm
Devices	Brand Name: RX Accunet Embolic Protection System Common Name: RX Accunet EPS. Part Numbers: 101649-45, 1011649-55, 1011649-65, 1011649-75, 1011650-55, 1011650-65, 1011651-45, 1011651-55. The RX Accunet EPS is indicated for use as a guide wire and embolic protection system.	Lot Number: 1081061, 1082561, 1091361, 1080561, 1081761, 1090261, 1091561, 1100461, 1080361, 1081661, 1090161, 1090861, 1092361, 1080861, 1082661, 1100761, 1081761, 1092861, 1082461, 1081161, 1082261, 1083061, 1081961, 1093061.	Class II	The recall was initiated because Abbot Vascular has discovered that the identified lots of the RX Accunet Embolic Protection System may exhibit difficult removal of the peel away sheath due to higher than normal wall thickness.	Abbott Vascular
Devices	AdvantageSim MD versions 7.4 through 7.6, Model 5160092-2 Version 7.4, 5160092-3 Version 7.5, 5160092-4 version 7.6.	Mfg Lot or Serial # 00000L022E05AE 00000139595HP3 000000010003GS 000000410012GS 000000210016GS 000000210013GS 00000138998HP0 00000139435HP2 000000410010GS 000000410009GS 000000210005GS 000000210011GS 000000210012GS 000000210015GS 00000138991HP5 000000210014GS 000000310005GS 00000139574HP8 00000139604HP3 00000L026CC222 000000210010GS 000000010004GS 000000410011GS 000000410007GS 000000210023GS 000000210025GS 000000210026GS 000000210022GS 000000210024GS 000000410005GS 000000410006GS 000000310003GS 000000310002GS 000000210006GS 000000210007GS 000000210003GS 000000410004GS 000000310001GS 000000010001GS 000000210030GS 000000210008GS	Class II	It was reported by a customer site that when using GE AdvantageSim MD on Advantage Workstation, the series could be incorrectly labeled in image view when multiple series of an exam are loaded simultaneously in Advantage Sim MD and if their series dates are different. A mismatch of series label for structure sets may lead to under-treatment of a tumor due to too small coverage of volume treated. A second issue was discovered internally in which the Interpolation in Advantage Sim MD 7.5 would not give a correct result when some part wee removed from an existing structure. If the contour truncation is not recognized, it may lead to inappropriate irradiation of the patient.	GE Healthcare, LLC

Εικόνα 2: *Σημερινή μορφή FDA (από 20 Ιουνίου 2012)*

Όπως παρατηρούμε για κάθε προϊόν μας δίνεται πληροφορία για το όνομά του την περιγραφή, τον κωδικό και την περιγραφή αστοχίας (στήλη **Reason for Recall** δηλαδή η αστοχία αποτελεί και την αιτία ανάκλησης του ελαττωματικού προϊόντος) όχι όμως για την κλάση του. Και όταν μιλάμε για κλάση δεν εννούμε τη στήλη classification που φαίνεται παραπάνω, αλλά τους τρεις προαναφερθέντες τρόπους ταξινόμησης αστοχιών. Η ταξινόμηση κάθε αστοχίας (προς αποφυγήν συγχύσεων επαναλαμβάνουμε ότι δε φαίνεται στην ιστοσελίδα) πραγματοποιείται από ειδικευμένους επιστήμονες βιοϊατρικής [4]. Ουσιαστικά αντλείται η πληροφορία από την ιστοσελίδα και καταρτίζεται ένα νέο αρχείο με μια επιπλέον στήλη, αυτή της μεταβλητής ταξινόμησης. Η μεταβλητή ταξινόμησης μπορεί να είναι η λειτουργία, η αιτία αστοχίας ή το αν η αστοχία οφείλεται σε λογισμικό (software) ή υλικό (hardware). Στην παρούσα πτυχιακή θα ασχοληθούμε μόνο με τη μεταβλητή ταξινόμησης τιμών {hardware, software}.

Τα καταρτηθέντα αρχεία αποτελούν δεδομένα εκπαίδευσης για ταξινομητές με στιγμιότυπα εκπαίδευσης τις περιγραφές των αστοχιών. Με αυτόν τον τρόπο χτίζεται ένα μοντέλο ικανό για αυτόματη αναγνώριση της κλάσης μιας νέας αστοχίας χωρίς την ανάγκη χειροκίνητης αναγνώρισης από κάποιον ειδικό. Ένα άλλο πλεονέκτημα της δημιουργίας μοντέλου είναι η επίτευξη αφαιρετικής, συμπυκνωμένης και κυρίως περισσότερο δομημένης αναπαράστασης της ογκώδους πληροφορίας της βάσης FDA.

Πηγές δυσμενών περιστατικών όμως υπάρχουν και άλλες όπως προαναφέρθηκε. Στην περίπτωση που τα δεδομένα εκπαίδευσης προέρχονται από διαφορετικές

πηγές, αυτά είναι ετερογενή και μη συμβατά. Πριν χρησιμοποιηθούν περαιτέρω είναι αναγκαία μια διαδικασία **τυποποίησης**, στην υπηρεσία της οποίας επιστρατεύονται αλγόριθμοι ταξινόμησης όπως νευρωνικά δίκτυα, μπεϋζιανά δίκτυα, δέντρα αποφάσεων και μοντέλα Markov.

.

4.5 Εξαγωγή μοντέλων ταξινόμησης αστοχιών

Στη διάθεσή μας έχουμε ένα σύνολο εκπαίδευσης αποτελούμενο από δύο στήλες, η μία περιέχει την περιγραφή της αστοχίας και η άλλη την ταξινόμησή της (hardware, software, other αν δεν είναι ξεκάθαρη η φύση της αστοχίας). Συνολικά έχουμε 746 στιγμιότυπα, επεξεργασμένα από το Εργαστήριο Ιατρικής Φυσικής του Πανεπιστημίου Πατρών. Πριν χρησιμοποιηθούν τα στιγμιότυπα για την εξαγωγή του μοντέλου θα πρέπει να μετατραπούν σε κατάλληλη μορφή στην οποία να έχουν καθορισμένα και κοινά χαρακτηριστικά όπως τα παραδείγματα του κεφαλαίου 1. Γι'αυτόν το σκοπό χρησιμοποιείται ένα φίλτρο δεδομένων το οποίο μετατρέπει ένα σύνολο κειμένων σε μία μήτρα της οποίας οι γραμμές αντιπροσωπεύουν το εκάστοτε κείμενο και οι στήλες κάθε μία από τις διαφορετικές λέξεις που συναντώνται στα κείμενα. Αν δηλαδή έχουμε n κείμενα m διαφορετικών λέξεων η μήτρα θα έχει διαστάσεις n x m. Τα στοιχεία της μήτρας είναι 1 αν μία λέξη ανήκει σε κάποιο κείμενο ή 0 αν δεν ανήκει. Το φίλτρο αυτό ονομάζεται **StringToWordVector** και βρίσκεται στην τοποθεσία *Filters -> Unsupervised -> attribute*.

Να σημειωθεί σε αυτό το σημείο ότι για να λειτουργήσει το φίλτρο πρέπει το αρχείο .arff των δεδομένων εισόδου να διαθέτει δύο χαρακτηριστικά, ένα χαρακτηριστικό τύπου *string* το οποίο θα αντιπροσωπεύει το σύνολο των κειμένων και άλλο ένα που θα χρησιμοποιηθεί για μεταβλητή κλάσης. Κάθε στιγμιότυπο στην περιοχή @data αποτελεί και ένα κείμενο το οποίο βρίσκεται ανάμεσα σε quotes(‘ ‘). Κατά τη διάρκεια του φιλτραρίσματος κάθε μία από τις m λέξεις μετατρέπεται σε *αριθμητικό χαρακτηριστικό* το οποίο παίρνει μία από τις τιμές 0, 1. Η έξοδος θα είναι ένα αρχείο με τα m αυτά χαρακτηριστικά και κάθε στιγμιότυπο θα αποτελείται από m τιμές 0 ή 1. Αυτό το αρχείο χρησιμοποιείται στη συνέχεια για την εκπαίδευση του ταξινομητή. Στο GUI του WEKA μετά την εφαρμογή του φίλτρου μπορούμε να προχωρήσουμε αμέσως στο classification. Η n x m μήτρα φαίνεται στην επιλογή Edit.

Στη συνέχεια εφαρμόζουμε διάφορους αλγορίθμους ταξινόμησης με σκοπό την εξαγωγή μοντέλων και τη σύγκριση της απόδοσής τους. Σε καθέναν από τους αλγορίθμους τα σύνολα εξέτασης επελέχθησαν δια της μεθόδου της διασταυρωτής επαλήθευσης 10 ομάδων. Αρχικά επιλέγουμε τον Naïve Bayes. Η αποτίμηση του μοντέλου φαίνεται παρακάτω:

```
Time taken to build model: 3.09 seconds

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

Correctly Classified Instances         619               82.9759 %
Incorrectly Classified Instances       127               17.0241 %
Kappa statistic                          0.674
Mean absolute error                      0.118
Root mean squared error                  0.3122
Relative absolute error                 33.4708 %
Root relative squared error             74.4149 %
Total Number of Instances              746

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.911     0.236      0.783     0.911     0.843      0.93     software
                 0.766     0.094      0.885     0.766     0.821      0.929    hardware
                 0.545     0          1         0.545     0.706      0.905    other
Weighted Avg.    0.83      0.16       0.839     0.83      0.828      0.929

=== Confusion Matrix ===

   a   b   c   <-- classified as
 329  32   0 |   a = software
  85 278   0 |   b = hardware
   6   4  12 |   c = other
```

NAÏVE BAYES

Στη συνέχεια δοκιμάζουμε τον BayesNet:

```
Time taken to build model: 4.03 seconds

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

Correctly Classified Instances         551               73.8606 %
Incorrectly Classified Instances       195               26.1394 %
Kappa statistic                          0.4973
Mean absolute error                      0.1737
Root mean squared error                  0.3342
Relative absolute error                 49.2849 %
Root relative squared error             79.6502 %
Total Number of Instances              746

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.903     0.397      0.681     0.903     0.776      0.895    software
                 0.598     0.11       0.838     0.598     0.698      0.887    hardware
                 0.364     0          1         0.364     0.533      0.81     other
Weighted Avg.    0.739     0.246      0.767     0.739     0.731      0.889

=== Confusion Matrix ===

   a   b   c   <-- classified as
 326  35   0 |   a = software
 146 217   0 |   b = hardware
   7   7   8 |   c = other
```

BAYESNET

Παρατηρούμε ένα περίεργο εκ πρώτης όψεως φαινόμενο, και αυτό είναι η καλύτερη απόδοση του Naïve Bayes από τον BayesNet. Γενικά το φαινόμενο είναι τυχαίο και στα περισσότερα σύνολα εκπαίδευσης αριθμητικών χαρακτηριστικών δε συμβαίνει. Είναι αποτέλεσμα της προσαρμογής κανονικής κατανομής στα αριθμητικά χαρακ τηριστικά. Αν θέλουμε να έχουμε βέλτιστη απόδοση του αλγορίθμου μπεϋζιανών δικτύων ο μόνος τρόπος είναι να διακριτοποιήσουμε τα χαρακτηριστικά (ειδικά unsupervised φίλτρα για το σκοπό αυτό).

Ο αλγόριθμος δέντρων ταξινόμησης :

```
Time taken to build model: 31.36 seconds

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

Correctly Classified Instances         643               86.193  %
Incorrectly Classified Instances       103               13.807  %
Kappa statistic                          0.7367
Mean absolute error                      0.1082
Root mean squared error                  0.2834
Relative absolute error                 30.705  %
Root relative squared error             67.5282 %
Total Number of Instances              746

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.845     0.106      0.882     0.845     0.863      0.906    software
                 0.904     0.146      0.854     0.904     0.878      0.91     hardware
                 0.455     0.008      0.625     0.455     0.526      0.786    other
Weighted Avg.    0.862     0.123      0.861     0.862     0.86       0.904

=== Confusion Matrix ===

   a   b   c   <-- classified as
 305  50   6 |   a = software
  35 328   0 |   b = hardware
   6   6  10 |   c = other
```

ΔΕΝΤΡΟ ΤΑΞΙΝΟΜΗΣΗΣ

Τέλος, ο αλγόριθμος διανυσμάτων υποστήριξης:

```
Time taken to build model: 1.36 seconds

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

Correctly Classified Instances         681               91.2869 %
Incorrectly Classified Instances        65                8.7131 %
Kappa statistic                          0.8335
Mean absolute error                      0.2446
Root mean squared error                  0.3105
Relative absolute error                 69.3724 %
Root relative squared error             74.0003 %
Total Number of Instances              746

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.934     0.099      0.899     0.934     0.916      0.918    software
                 0.912     0.068      0.927     0.912     0.919      0.922    hardware
                 0.591     0.001      0.929     0.591     0.722      0.794    other
Weighted Avg.    0.913     0.081      0.913     0.913     0.912      0.916

=== Confusion Matrix ===

   a   b   c   <-- classified as
 337  23   1 |   a = software
  32 331   0 |   b = hardware
   6   3  13 |   c = other
```

ΔΙΑΝΥΣΜΑΤΑ ΥΠΟΣΤΗΡΙΞΗΣ

4.6 Συμπεράσματα

Η μήτρα σύγχυσης (confusion matrix) μας δίνει μία πρώτη και άμεση εικόνα για την απόδοση του εκάστοτε μοντέλου. Όπως αναφέρθηκε και στο κεφάλαιο 3 στην αποτίμηση ταξινομητών, οι στήλες της μήτρας μας πληροφορούν για το πλήθος των ορθώς ταξινομηθέντων στην αντίστοιχη κλάση στιγμιοτύπων καθώς και για τα λάθος ταξινομηθέντα σε αυτήν, προερχόμενα από διαφορετικές πραγματικές κλάσεις. Με άλλα λόγια τα μη διαγώνια στοιχεία αντιπροσωπεύουν τις διαρροές στιγμιοτύπων των υπολοίπων κλάσεων προς την κλάση της στήλης. Σε καθέναν από τους τέσσερις ταξινομητές παρατηρούμε ότι η κλάση με το μεγαλύτερο διαγώνιο στοιχείο, με τα περισσότερα αληθώς θετικά στιγμιότυπα, τραβάει και τον μεγαλύτερο αριθμό στιγμιοτύπων άλλων πραγματικών κλάσεων. Έτσι, βλέπουμε ότι στο μοντέλο Naïve Bayes τα ορθά software είναι 329 και τα ψευδή 85 + 6 = 91, περισσότερα από τα ψευδή hardware και τα ψευδή other. Στο δέντρο ταξινόμησης έχουμε 328 ορθές hardware κλάσεις και 50 + 6 = 56 λάθος. Εδώ οι διαρροές προς software (41) πλησιάζουν περισσότερο τις διαρροές προς hardware, λογικό εφόσον το πλήθος σωστών software είναι 305, λίγο μικρότερο από το αντίστοιχο hardware.

Από τα τρία μοντέλα**,** την καλύτερη απόδοση έχει η μηχανή διανυσμάτων υποστήριξης (91% σωστά, 9% λάθος). Πρόκειται πράγματι για έναν πολύ αξιόπιστο

αλγόριθμο ταξινόμησης ο οποίος εκμεταλλεύεται στο έπακρο την όποια τάση των στιγμιοτύπων να διαχωρίζονται σε ομάδες ως προς τις κλάσεις ενώ έχει και πολύ καλή προοπτική να ταξινομήσει σωστά νεοεμφανιζόμενα στιγμιότυπα μέσω της εύρεσης επιπέδου διαχωρισμού μεγίστου περιθωρίου. Συνολικά το σύνολο εκπαίδευσης περιλαμβάνει 361 αστοχίες τύπου software, 363 τύπου hardware και 22 other. Το μοντέλο προβλέπει σωστά 337 software, 331 hardware και 13 other. Εκτός από το υψηλό ποσοστό επιτυχίας δηλαδή, προσεγγίζει και την πραγματική κατανομή των στιγμιοτύπων στις τρεις κλάσεις χωρίς να ευνοεί κάποια από αυτές παραπάνω από ότι πρέπει. Γενικά πάντως, όλοι οι αλγόριθμοι είχαν πολύ καλή απόδοση σε ένα ογκώδες σύνολο εκπαίδευσης με πολλά διαφορετικά χαρακτηριστικά και μάλιστα με εξέταση διασταυρωτής επαλήθευσης, η οποία αποτελεί την πιο ρεαλιστική και αξιόπιστη μέθοδο λήψης συνόλων εξέτασης για ταξινομητές.

Η εξαγωγή των ανωτέρω μοντέλων αλλά και άλλων με χρήση διαφορετικών αλγορίθμων αποτελεί ένα πολύ μικρό κομμάτι ενός γενικότερου συστήματος επαγρύπνησης [3]. Αποτελεί έναν μη αιτιοκρατικό, εναλλακτικό του συμβατικού ελέγχου και συντήρησης τρόπο για την πρόληψη αστοχιών ιατροτεχνολογικών προϊόντων οι οποίες μπορεί να προκαλέσουν δυσμενή για τους ασθενείς περιστατικά. Σίγουρα δεν αντικαθιστά τον έλεγχο και τη συντήρηση, αλλά τους βοηθάει και τους συμπληρώνει.

ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΕΦΑΛΑΙΟΥ 4

[1] Killed by Code: Software Transparency in Implantable Medical Devices, Karen Sandler,Lysandra Ohrstrom, Laura Moy,Robert McVay

[2] Anand K, Saini SK, Singh BK, Veermaram C. Global medical device nomenclature: The concept for reducing device-related medical errors. Journal of Young Pharmacists, 2(4), 403-409 (2010).

[3] Bliznakov, Z; Malataras, P..; Pallikarakis, N..(2007) “Medical Equipment Inventorying and Installation of a Web-based Management System – Pilot Application in the Periphery of Crete, Greece” 11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007 (2007) 16: 1092-1095

[4] Z. Bliznakov, G. Mitalas, N. Pallikarakis, Analysis and Classification of Medical Device Recalls, World Congress on Medical Physics and Biomedical Engineering 2006, 27 Aug – 1 Sept 2006, Seoul, Korea.

[5] G. Pappous, Z. Bliznakov, G. Mitalas, N. Pallikarakis, Medical Device Software and Patient Safety. 3rd International Conference on Information Communication Technologies in Health, 7-9 July 2005, Samos Island, Greece.

[6] Ian Witten, Eibe Frank, Mark Hall. Data Mining 3rd edition

[7] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Εισαγωγή στην εξόρυξη δεδομένων.