Παρουσίαση Διπλωματικής Εργασίας

# Εξόρυξη Γνώσης από Βιολογικά Δεδομένα

**Καρυπίδης Γεώργιος (Μ27/03)**

**Επιβλέπων Καθηγητής: Ιωάννης Βλαχάβας**

**MIS – Πανεπιστήμιο Μακεδονίας**

**Φεβρουάριος 2005**

# Εξόρυξη Γνώσης από Βιολογικά Δεδομένα

**ΠΕΡΙΕΧΟΜΕΝΑ ΕΡΓΑΣΙΑΣ**

- Μοριακή Βιολογία & Βιοπληροφορική
- Ανακάλυψη Γνώσης σε Β.Δ.
(Knowledge Discovery in Databases)
- Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία

# Ενότητα 1: Μοριακή Βιολογία & Πληροφορική

- Περιγραφή μερικών βασικών εννοιών στη Βιολογία:
  - Κύτταρο & Πρωτεΐνες
  - Γενετικό Υλικό
  - Γονιδίωμα & Γονιδιωματική
  - Γονιδιακή Έκφραση
  - Βιοπληροφορική

# Μοριακή Βιολογία & Πληροφορική(1/6)

- Κύτταρο & Πρωτεΐνες

- Το κύτταρο αποτελεί το θεμελιώδες δομικό στοιχείο όλων των οργανισμών
- Κοινά χαρακτηριστικά όλων των κυττάρων είναι η πλασματική (κυτταρική) μεμβράνη, κυτταρόπλασμα, ριβοσώματα, DNA ή RNA
- Πρωτεΐνες αποτελούν τη θεμελιώδη δομική και ενεργειακή μονάδα του κυττάρου και κατ' επέκταση του οργανισμού
- Οι πρωτεΐνες είναι αλυσίδες αμινοξέων
- Υπάρχουν 20 διαφορετικά είδη αμινοξέων
- Η παραγωγή της πρωτεΐνης γίνεται με μεταβολισμό
- Η πληροφορία που χρειάζεται για την παραγωγή της πρωτεΐνης βρίσκεται στο γενετικό υλικό (DNA)

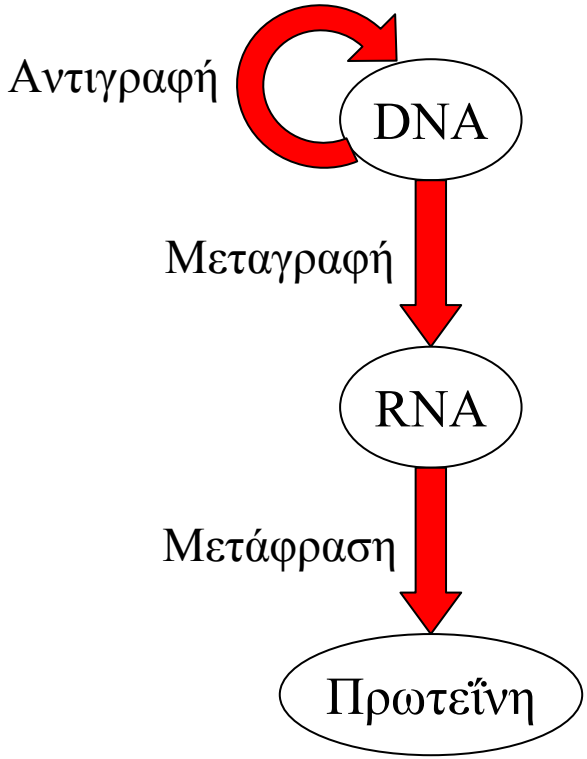
# Μοριακή Βιολογία & Πληροφορική(2/6)

- Γενετικό Υλικό (DNA, RNA)

- Το γενετικό υλικό είναι αλυσίδες νουκλεοτίδιων που καλούνται DNA και RNA (νουκλεϊκά οξέα)
- Τα νουκλεοτίδια είναι τεσσάρων ειδών
- Ο συνδυασμός του μη-καθορισμένου μήκους της αλυσίδας με τα 4 διαφορετικά είδη νουκλεοτιδίων καθιστούν το DNA & RNA ιδανικό για την κωδικοποίηση πληροφορίας
- Τα 3 είδη RNA (mRNA, rRNA, tRNA) αποτελούν τα βασικά συστατικά στη σύνθεση της πρωτεΐνης

# Μοριακή Βιολογία & Πληροφορική(3/6)

- Κεντρικό Δόγμα της Μοριακής Βιολογίας




mRNA
αποκαλύπτει τα γονίδια και την έκφρασή τους

# Μοριακή Βιολογία & Πληροφορική(4/6)

- Γονιδίωμα & Γονιδιωματική

- Το «γονιδίωμα» είναι το σύνολο του νουκλεϊκού οξέος
  - Οργανώνεται σε χρωμοσώματα και γονίδια
- Η Γονιδιωματική είναι ο τομέας της γενετικής που ασχολείται με την μελέτη και έρευνα του γονιδιώματος
- Σκοπός της γονιδιωματικής η χαρτογράφηση της δομής και της λειτουργίας των γονιδίων (δομική & λειτουργική γονιδιωματική)
- Το ανθρώπινο γονιδίωμα χαρτογραφήθηκε με το “Human Genome Project”

# Μοριακή Βιολογία & Πληροφορική(5/6)

- Γονιδιακή Έκφραση

- Γονιδιακή έκφραση είναι η ποσοτικοποίηση της έκφρασης ενός γονιδίου για την παραγωγή πρωτεΐνης
- Πρόσφατη εξέλιξη η ανακάλυψη τεχνικών απεικόνισης της έκφρασης χιλιάδων γονίδιων σε 1 μόνο πείραμα
- Συσχετίζει τα γονίδια με φάρμακα & ασθένειες – μεταλλάξεις
- Οι δύο κυριότερες τεχνικές:
  - SAGE (Serial Analysis of Gene Expression)
  - Μικροσυστοιχίες DNA (microarrays)

# Μοριακή Βιολογία & Πληροφορική(6/6)

- **Βιοπληροφορική**

- Βάσεις Δεδομένων
  - Ακολουθιών - Μοριακές (GenBank, …)
  - Γονιδιωμάτων (GOLD, …)
- Μέθοδοι – Αλγόριθμοι:
  - Αναγνώριση γονιδίων (GeneID, …)
  - Αναζήτησης ομοιότητας (BLAST, …)
  - Ποσοτικοποίηση & Ανάλυση δεδομένων γονιδιακής έκφρασης (Sage, GeneChip – κατηγοριοποίηση, ομαδοποίηση, ...)

# Ενότητα 2: Ανακάλυψη Γνώσης σε ΒΔ

- Η Εξόρυξη σε Δεδομένα εργαλείο Υποστήριξης Λήψης Αποφάσεων (Decision Support Systems) σε πολλούς τομείς όπως στις Επιχειρήσεις και στις Επιστήμες
- Αξιοποιεί μεθόδους και αλγόριθμους από άλλες επιστήμες: των Βάσεων Δεδομένων, της Στατιστικής, της Μηχανικής Μάθησης
- Περιγραφή
  - Της διαδικασίας Ανακάλυψης Γνώσης σε ΒΔ
  - Μεθόδων & Αλγόριθμων Εξόρυξης Γνώσης

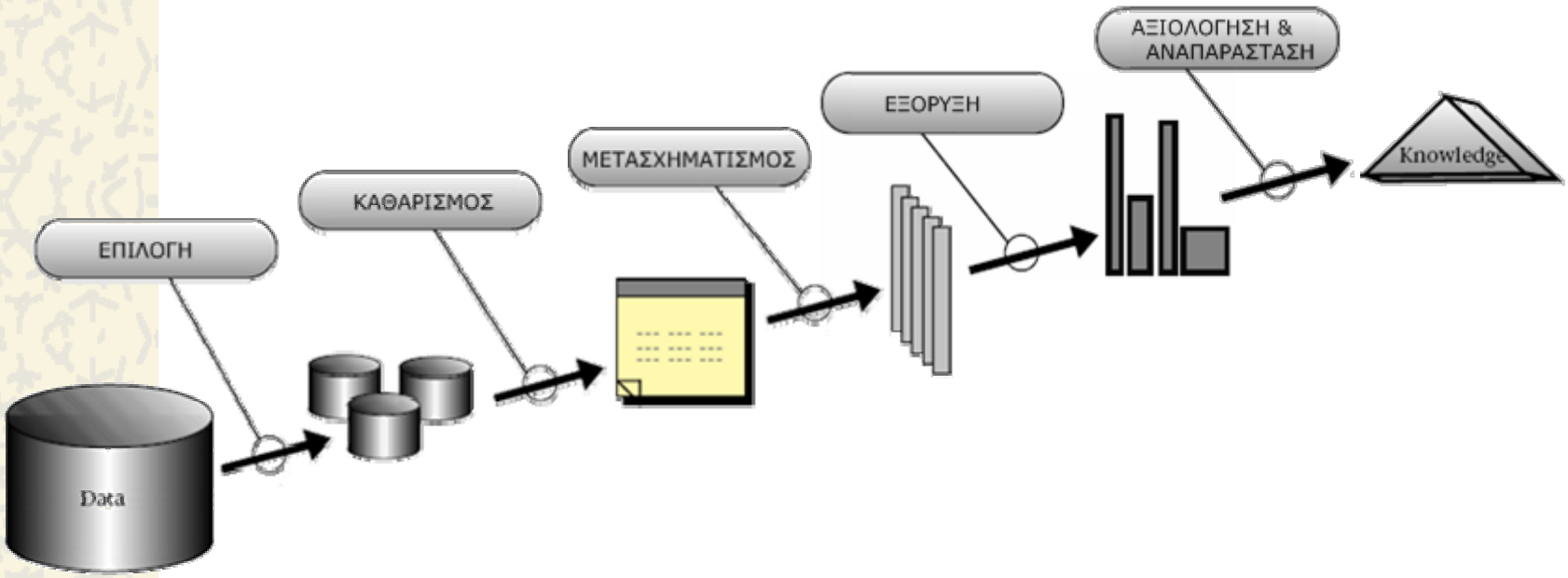
# Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(1/7)

- Βήματα

- Προεπεξεργασία Δεδομένων
  - Επιλογή
  - Καθαρισμός
  - Μετασχηματισμός
- Εφαρμογή Μεθόδων Εξόρυξης Γνώσης
  - κατηγοριοποίηση
  - εμπειρική συσχέτιση μεταβλητών
  - ομαδοποίηση
  - κανόνες συσχέτισης
- Αξιολόγηση & Αναπαράσταση Γνώσης

# Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(2/7)

- Σειρά των φάσεων εξόρυξης



# Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(3/7)

- Εξόρυξη Γνώσης

ΜΟΝΤΕΛΑ - ΠΡΟΤΥΠΑ

| Πρόβλεψης<br>με επίβλεψη | Πληροφόρησης<br>χωρίς επίβλεψη |
|---|---|

ΜΕΘΟΔΟΙ

| | |
|---|---|
| Κατηγοριοποίηση | Ομαδοποίηση |
| Εμπειρική Συσχέτιση Μεταβλητών | Κανόνες Συσχέτισης |

# Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(4/7)

- Κατηγοριοποίηση

- Η κατηγοριοποίηση αφορά την κατασκευή μοντέλου για την πρόβλεψη της κατηγορίας της κάθε κατάστασης (εγγραφής)
- Η κατασκευή γίνεται με τα υπάρχοντα δεδομένα
- Το μοντέλο ανάλογα με τον αλγόριθμο αναπαρίσταται με κανόνες, δένδρα απόφασης, μαθηματική συνάρτηση
- Δύο βασικές τεχνικές: Επαγωγική Μάθηση με Δένδρα Απόφασης (ID3) & Απλοί Κατηγοριοποιητές Bayes (Naive Bayes)
- Άλλοι Αλγόριθμοι: K-nearest neighbors, Support Vector Machines, Voted Classification, Weighted gene voting

# Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(5/7)

## • Εμπειρική Σχέση Μεταβλητών

- Η εμπειρική σχέση μεταβλητών αφορά την κατασκευή μοντέλου για την πρόβλεψη αριθμητικής τιμής
- Η κατασκευή γίνεται με τα υπάρχοντα δεδομένα
- Κυριότερες Μέθοδοι:
  - Γραμμική / Μη-Γραμμική Παρεμβολή &
  - Νευρωνικά Δίκτυα

# Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(6/7)

- Ομαδοποίηση & Κανόνες Συσχέτισης (1/2)

- Ανακάλυψη προτύπων στα δεδομένα
- Ανακάλυψη με παρατήρηση και όχι με παραδείγματα
- Στην ομαδοποίηση 2 οι κυριότερες κατηγορίες μεθόδων: διαχωρισμού, ιεραρχικές
- Διαχωρισμού: k ομάδες με κριτήριο την απόσταση (Ευκλείδεια, ...)
- Ιεραρχικές: Συγχωνευτικές & Διαιρετικές (κάθε δεδομένο μια ομάδα ↔ όλα τα δεδομένα μια ομάδα) με κριτήριο την απόσταση

# Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων(7/7)

- Ομαδοποίηση & Κανόνες Συσχέτισης (2/2)

- Η ομαδοποίηση δεν αποκαλύπτει τη σχέση μεταξύ των ομαδοποιημένων δεδομένων => κανόνες συσχέτισης
- Σκοπός η αναζήτηση συχνών συνόλων αντικειμένων (itemsets) => αντιπροσωπεύουν κανόνες συσχέτισης μεταξύ των δεδομένων
- Κριτήριο: υποστήριξη (χρησιμότητα) & εμπιστοσύνη (βεβαιότητα)
- Μια εφαρμογή των μεθόδων «Μάθησης Χωρίς Επίβλεψη» είναι η ανακάλυψη συνεκφρασμένων ομάδων γονιδίων (synexpression groups)

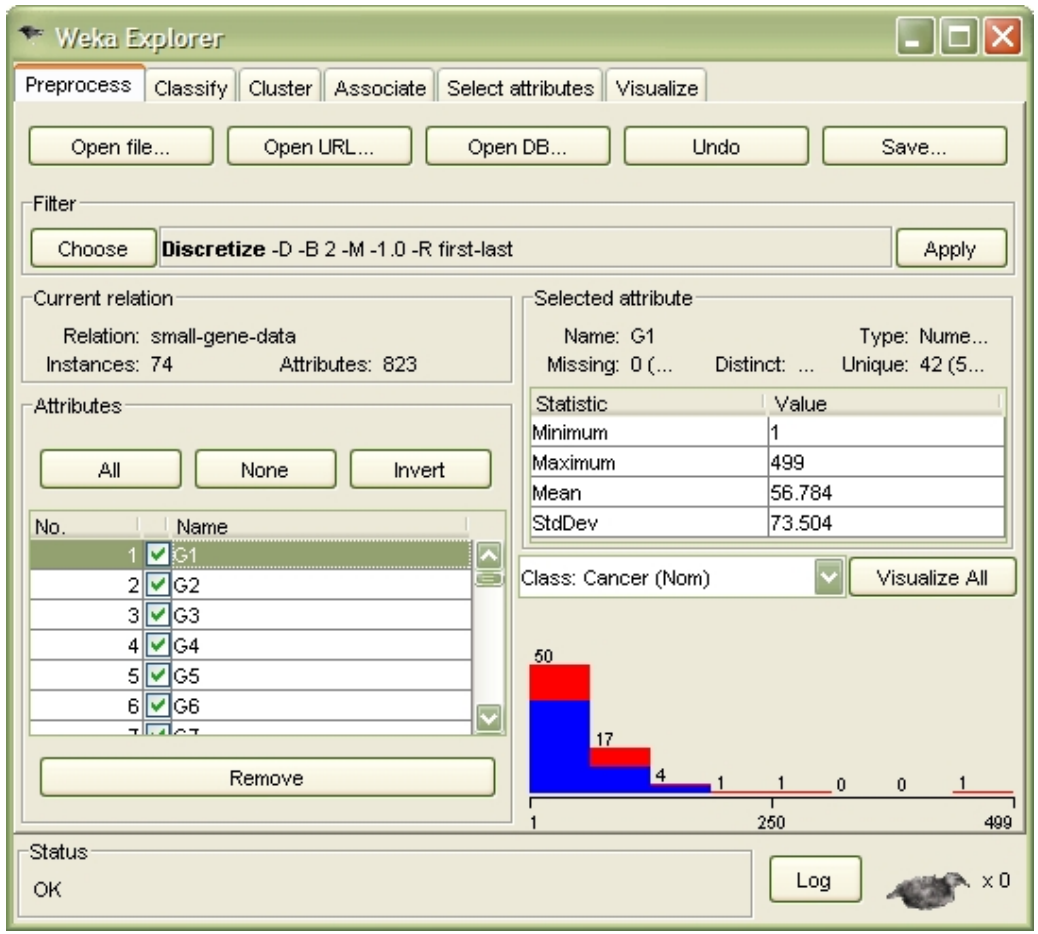
# Ενότητα 3: Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία

- Παρουσίαση εφαρμογής αλγορίθμων με σκοπό την πρόβλεψη καρκινικών οργανισμών
- Εφαρμογή
- Αξιολόγηση
- Σύγκριση

# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(1/9)

- Περιβάλλον εργασίας WEKA

- Γραφικό Περιβάλλον
- Βασισμένο στην Java
- Open Source
- Διαθέσιμοι αρκετοί αλγόριθμοι από όλες τις μεθόδους
- Έλλειψη documentation



# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(2/9)

- Δεδομένα

- Προέλευση δεδομένων από το διαγωνισμό Discovery Challenge 2004
- Τα δεδομένα παράχθηκαν με πειράματα SAGE
- Περιγραφή σε 3 αρχεία
  - Ποσοτικοποιημένη έκφραση των γονίδιων
  - Περιγραφή των βιολογικών καταστάσεων
  - Περιγραφή των γονιδίων
- 822 γονίδια σε 74 βιολογικές καταστάσεις

# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(3/9)

## • Προεπεξεργασία (1/2)

- Επιλογή

  για την προσθήκη του χαρακτηριστικού για το οποίο θα φτιαχτεί το μοντέλο πρόβλεψης – "tissue type" → "cancer"

- Καθαρισμός & Μετατροπή

  για την μετατροπή των τιμών σε {"yes", "no"} & συμπλήρωση των κενών τιμών στο χαρακτηριστικό "cancer"

- Μορφοποίηση του αρχείου δεδομένων

  WEKA → ARFF ("@relation … @attributes … @data …")

- Μετασχηματισμός

  -διακριτοποίηση (ομαδοποίηση) των δεδομένων σε 2 κατηγορίες {εκφρασμένα, μη-εκφρασμένα} υπολογιστικά με το WEKA

  -πριν την εξόρυξη υπολογιστικά επιλέγουμε τα γονίδια που δείχνουν μεγαλύτερη συσχέτιση με το χαρακτηριστικό πρόβλεψης - Αλγόριθμοι: Ranker με Information Gain & Επιλογή των 100 καλύτερων

# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(4/9)

- Προεπεξεργασία (2/2)

- Τα δεδομένα πριν τη διακριτοποίηση

```
  6,10,0,4,1,7, ... ,no
 23,13,1,0,6,2, ... ,no
 37, 2,0,1,1,0, ... ,yes
 19, 7,0,0,0,0, ... ,yes
 48, 4,0,0,1,0, ... ,yes
145,10,2,0,0,2, ... ,yes
 28,31,0,5,4,0, ... ,yes
```

- Τα δεδομένα μετά την διακριτοποίηση

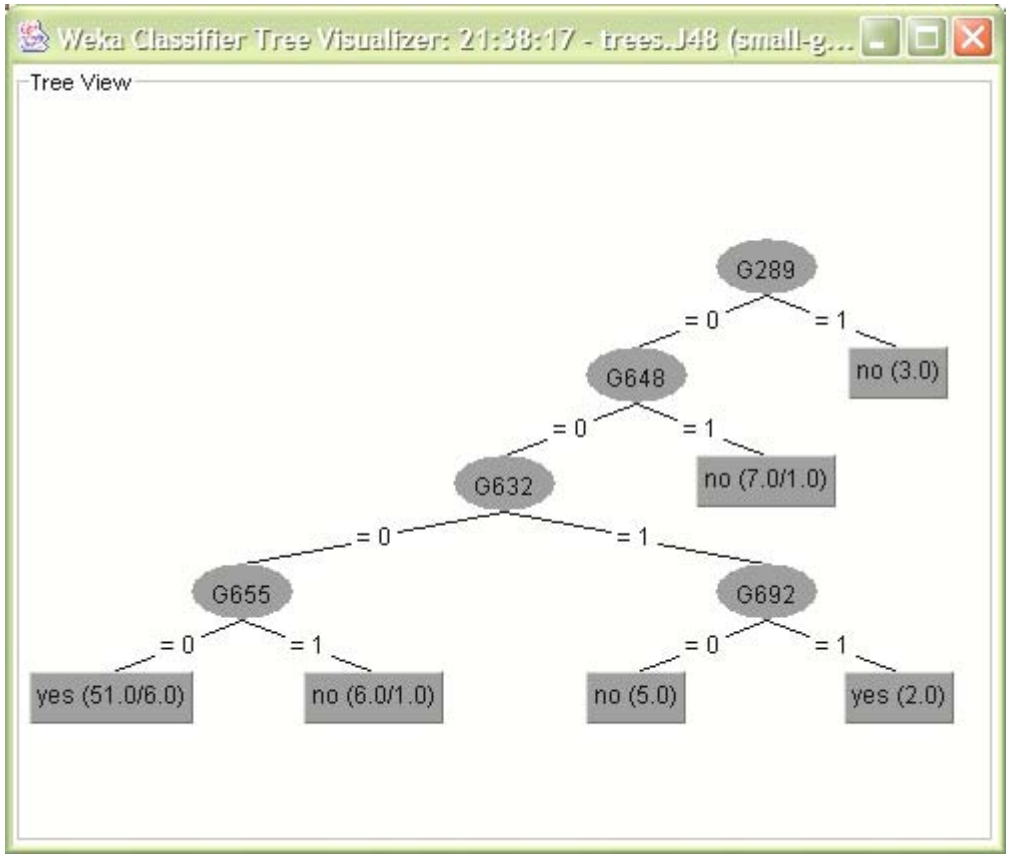
```
(-inf-250],(-inf-30],(-inf-1.5],(3.5-inf), (-inf-5.5],(-inf-8], ... ,no
(-inf-250],(-inf-30],(-inf-1.5],(-inf-3.5], (5.5-inf),(-inf-8], ... ,no
(-inf-250],(-inf-30],(-inf-1.5],(-inf-3.5],(-inf-5.5],(-inf-8], ... ,yes
(-inf-250],(-inf-30],(-inf-1.5],(-inf-3.5],(-inf-5.5],(-inf-8], ... ,yes
(-inf-250],(-inf-30],(-inf-1.5],(-inf-3.5],(-inf-5.5],(-inf-8], ... ,yes
(-inf-250],(-inf-30], (1.5-inf),(-inf-3.5],(-inf-5.5],(-inf-8], ... ,yes
(-inf-250], (30-inf),(-inf-1.5], (3.5-inf),(-inf-5.5],(-inf-8], ... ,yes
```

- Πρόβλεψη → Κατηγοριοποίηση
  - ID3, C4.5
  - Naive Bayes
  - Ripper
  - K-Nearest Neighbors

# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(6/9)

- Εξόρυξη Γνώσης με Δένδρο Απόφασης

Το δένδρο που κατασκεύασε ο C4.5 (J48)



# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(7/9)

- Αξιολόγηση (1/2)

- Έλεγχος διασταύρωσης 10-επαναλήψεων (10-fold cross validation)
  - 10 ίσου μεγέθους τμήματα διαφορετικά μεταξύ τους
  - Σε κάθε επανάληψη χρησιμοποιείται ένα τμήμα για έλεγχο και τα υπόλοιπα για εκπαίδευση
  - Η εκτίμηση του σφάλματος είναι ο μέσος όρος των επιμέρους σφαλμάτων
  - Θεωρείται η καλύτερη μέθοδος αξιολόγησης
- Τα αποτελέσματα αξιολογούνται με κριτήριο:
  - τις σωστές κατηγοριοποιήσεις
  - τον πίνακα “confusion matrix”

# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(8/9)

- Αξιολόγηση (2/2)

Αποτελέσματα της αξιολόγησης από την εφαρμογή του Naive Bayes:

```
Correctly Classified Instances          64               86.4865 %
Incorrectly Classified Instances        10               13.5135 %

...

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.857     0.12       0.933     0.857     0.894    yes
  0.88      0.143      0.759     0.88      0.815    no

=== Confusion Matrix ===

  a  b   <-- classified as
 42  7 |  a = yes
  3 22 |  b = no
```

64 σωστές κατηγοριοποιήσεις – 42 στο yes & 22 στο no =>
"**yes**": True Positive Rate (42 σωστές από τις 49 που έκανε) = 85,7%
& Precision (42 σωστές από τις 45 που έπρεπε να κάνει) = 93,3%

# Εφαρμογή Μεθόδων Εξόρυξης Γνώσης στη Μοριακή Βιολογία(9/9)

- Σύγκριση

| | ID3 | C4.5 | Naive Bayes | Ripper | 10-NN |
|---|---|---|---|---|---|
| Σύνολο | 74 | 74 | 74 | 74 | 74 |
| **Σωστές κατηγοριοποιήσεις** | **53** | **58** | **64** | **51** | **57** |
| Λανθασμένες κατηγοριοποιήσεις | 21 | 16 | 10 | 23 | 17 |
| **% (σωστές κατηγοριοποιήσεις)** | **72%** | **78%** | **86%** | **69%** | **77%** |
| "yes" (TP) | 35 | 40 | 42 | 37 | 39 |
| % | 71% | 82% | 86% | 76% | 80% |
| "no" (TP) | 18 | 18 | 22 | 14 | 18 |
| % | 72% | 72% | 88% | 56% | 72% |

# Συμπεράσματα

- Naive Bayes
  - Καλύτερα αποτελέσματα τόσο στις σωστές κατηγοριοποιήσεις συνολικά όσο και για την κάθε κατηγορία
- Αδυναμία: η έλλειψη ειδικής γνώσης (βιολογική)
  - Καλύτερη προεπεξεργασία, εξόρυξη & αξιολόγηση
- Στο μέλλον απαραίτητη η συμμετοχή μοριακού βιολόγου για ρεαλιστικά αποτελέσματα
- Βελτιώσεις στους αλγόριθμους εξόρυξης γνώσης

Ευχαριστώ