ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

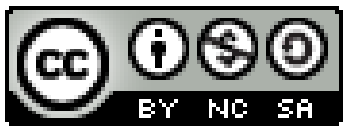
ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΙΚΑ
ΜΑΘΗΜΑΤΑ

Συστήματα Γνώσης

Θεωρητικό Κομμάτι Μαθήματος

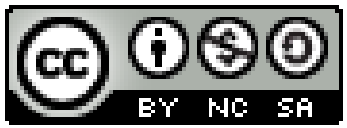
Ενότητα 11: Εφαρμογές Συστημάτων Γνώσης - Κατηγοριοποίηση

Νίκος Βασιλειάδης, Αναπλ. Καθηγητής
Τμήμα Πληροφορικής



Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΙΚΑ
ΜΑΘΗΜΑΤΑ

Εφαρμογές Συστημάτων Γνώσης

Κατηγοριοποίηση

Εφαρμογές Συστημάτων Γνώσης

- Τρεις σημαντικές κατηγορίες εφαρμογών των συστημάτων γνώσης
  - Κατηγοριοποίηση
  - Διαμόρφωση
  - Διάγνωση
- Γενική ανάλυση των χαρακτηριστικών των προβλημάτων, των μοντέλων και των μεθοδολογιών της κάθε κατηγορίας
- Μελέτη περιπτώσεων: DENDRAL, MYCIN, INTERNIST, PROSPECTOR, XCON, κτλ.

Κατηγοριοποίηση

- Ο προσδιορισμός της κατηγορίας στην οποία ανήκει ένα αντικείμενο.
- Είσοδος: ένα σύνολο δεδομένων που περιγράφουν το αντικείμενο.
- Έξοδος: η κατηγορία στην οποία αυτό ανήκει.
- Κυριότερο χαρακτηριστικό: Η επιλογή της κατηγορίας γίνεται από ένα προκαθορισμένο σύνολο κατηγοριών.
  - Κάθε αντικείμενο δεν ανήκει σε μια μοναδική κατηγορία.
  - Κάθε αντικείμενο δεν ανήκει σίγουρα σε κάποια κατηγορία.

Κατηγοριοποίηση

- Τα μέλη μιας κατηγορίας έχουν αρκετά κοινά χαρακτηριστικά μεταξύ τους.
- Συνήθως οι κατηγορίες είναι οργανωμένες σε ιεραρχίες, ώστε:
  - Οι υποκατηγορίες να έχουν τις ιδιότητες των υπερκατηγοριών.
  - Οι κατηγορίες του ίδιου επίπεδου να έχουν αλληλοαναιρούμενες ιδιότητες.
- Εφαρμογές:
  - Π.χ. διάγνωση, διαμόρφωση, επιδιόρθωση βλαβών, κλπ.
  - Η λογική της κατάταξης ενός προβλήματος σε κάποια κατηγορία της οποίας η λύση είναι γνωστή ταιριάζει με τον καθημερινό τρόπο επίλυσης προβλημάτων των ανθρώπων.

«Εξαντλητική» Κατηγοριοποίηση

- Στις απλές περιπτώσεις, αρκεί η απλή σύγκριση μερικών "επιφανειακών" χαρακτηριστικών του αντικειμένου.
- Όταν υπάρχουν πολλές ιδιότητες και πολύπλοκη ιεραρχία κατηγοριών:
  - Τα επιφανειακά χαρακτηριστικά δεν επαρκούν για την κατάταξη σε κάποιο κλαδί και επίπεδο της ιεραρχίας.
  - Η εξαντλητική σύγκριση όλων των χαρακτηριστικών δεν είναι πρακτικά εφαρμόσιμη.

Ευριστική Κατηγοριοποίηση
Heuristic Classification

- Προσπαθεί να κατατάξει τα αντικείμενα σε κατηγορίες που βρίσκονται στα φύλλα της ιεραρχίας χωρίς να διέλθει από όλα τα επίπεδα και να κάνει όλες τις συγκρίσεις.
  - Η κατάταξη γίνεται πιο γρήγορα, αλλά με μικρότερη ακρίβεια.
- Χρησιμοποιεί εμπειρική γνώση για τα αντικείμενα, τις κατηγορίες και τις συσχετίσεις τους, που προέρχεται από ανθρώπους-ειδικούς.

Φάσεις Ευριστικής Κατηγοριοποίησης

- Αφαίρεση / γενίκευση των δεδομένων (data abstraction)
- Ευριστική ταυτοποίηση (heuristic match)
- Επιλογή λύσης (solution refinement)

Γενίκευση δεδομένων

- Επικέντρωση μόνο στα σημαντικά χαρακτηριστικά ενός δεδομένου.

ΕΑΝ ένα βακτήριο ζει σε περιβάλλον στο οποίο δεν υπάρχει ελεύθερο οξυγόνο

ΤΟΤΕ πρόκειται για αναερόβιο βακτήριο

- Απλοποίηση ποσοτικών δεδομένων.

ΕΑΝ ο ασθενής είναι ενήλικος, ΚΑΙ

ο αριθμός των λευκών αιμοσφαιρίων είναι < 2500/cm3

ΤΟΤΕ ο αριθμός των λευκών αιμοσφαιρίων είναι μικρός

- Ιεραρχική οργάνωση των εννοιών.

ΕΑΝ το άτομο είναι πατέρας

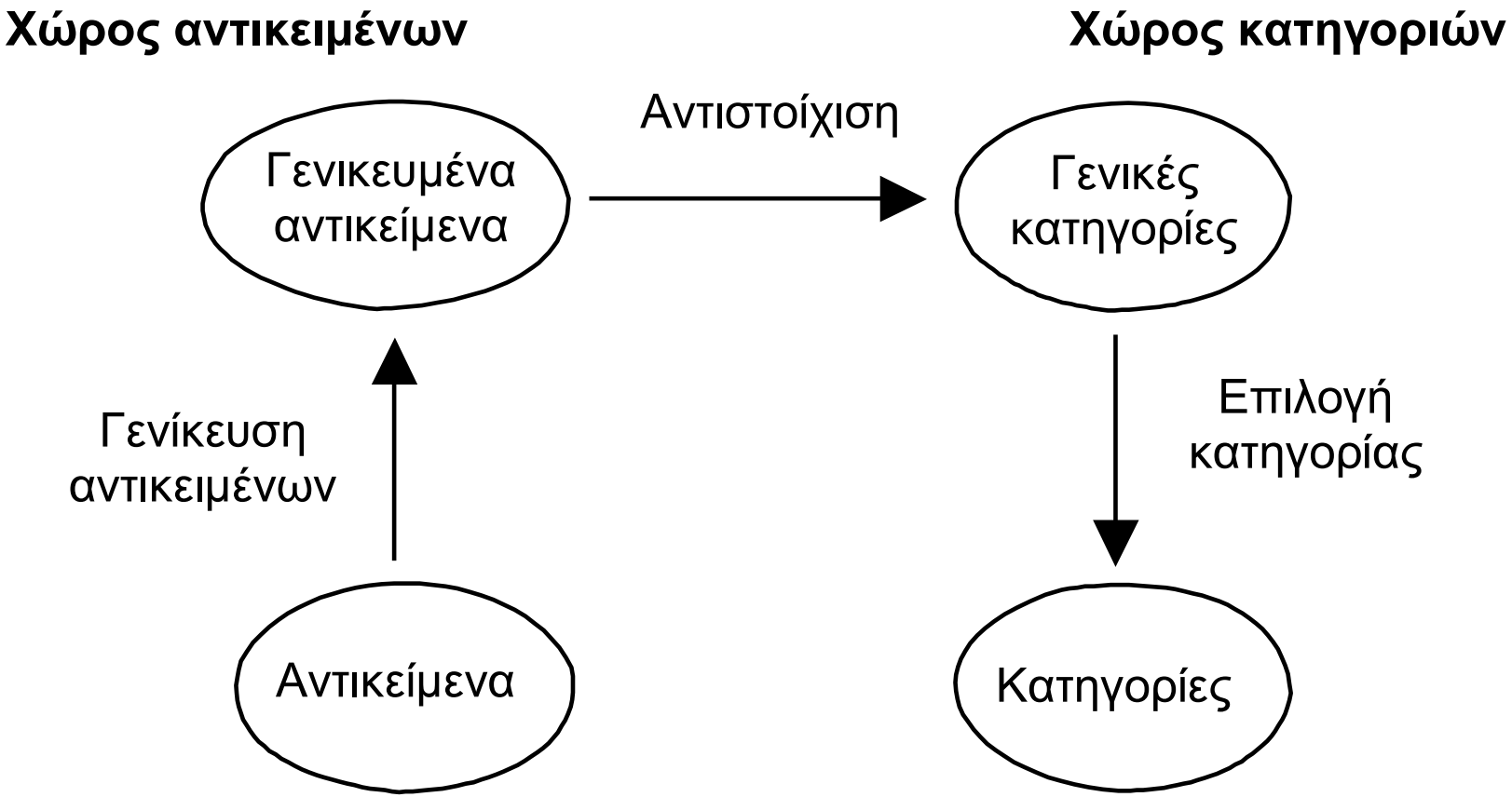
ΤΟΤΕ το άτομο είναι άντρας

Λοιπές Φάσεις

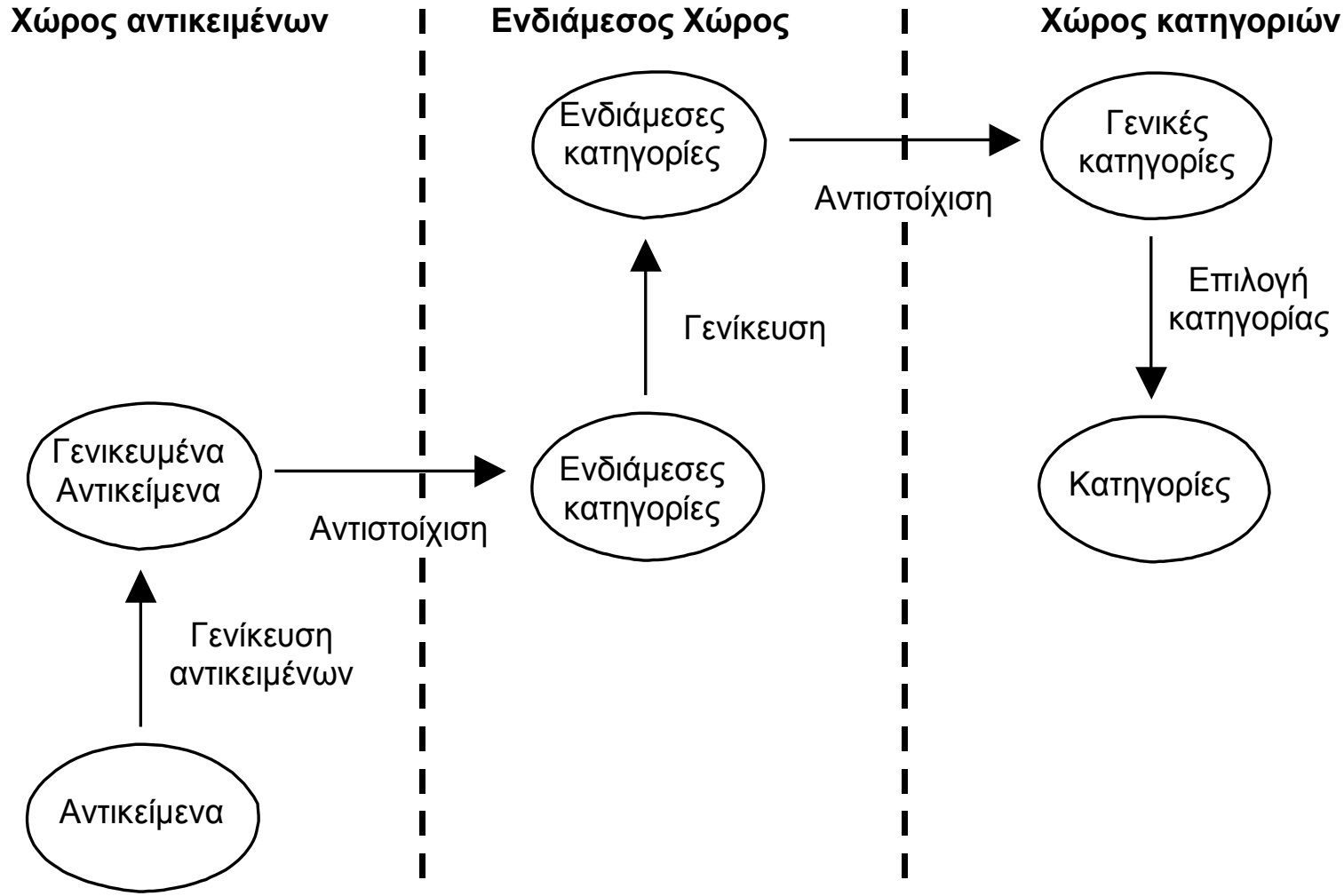
- **Ευριστική ταυτοποίηση (heuristic match)** των γενικευμένων αντικειμένων σε μια γενικότερη περιγραφή ενός συνόλου κατηγοριών.
  - Π.χ., ο πυρετός (γενίκευση υψηλής θερμοκρασίας) **μπορεί** να είναι ένδειξη μόλυνσης, η οποία εξειδικεύεται σε πολλές διαφορετικές μορφές.
- **Επιλογή** μιας συγκεκριμένης κατηγορίας-λύσης από το γενικό σύνολο κατηγοριών (**solution refinement**).
  - Π.χ., το είδος της μόλυνσης πρέπει να διαγνωστεί με ακρίβεια ώστε να δοθεί η σωστή θεραπεία.

Συστήματα Ευριστικής Κατηγοριοποίησης

Φάσεις Λειτουργίας



Πολυβάθμια Συστήματα Κατηγοριοποίησης



Συζευκτικό Μοντέλο Κατηγοριοποίησης

- **Χώρος δεδομένων D**: Πεπερασμένο σύνολο χαρακτηριστικών $\{\mathbf{D_i}\}$ του αντικειμένου που πρέπει να καταταχθεί σε μια κατηγορία.
- **Χώρος κατηγοριών-λύσεων S**: Πεπερασμένο σύνολο λύσεων $\{\mathbf{S_j}\}$.
- Τα $\mathbf{D_i}$ παίρνουν τιμές 0 και 1 ή το σύμβολο "?" που υποδηλώνει άγνωστη τιμή.
- Το σύνολο τιμών των $\mathbf{D_i}$ ονομάζεται διάνυσμα τιμών.

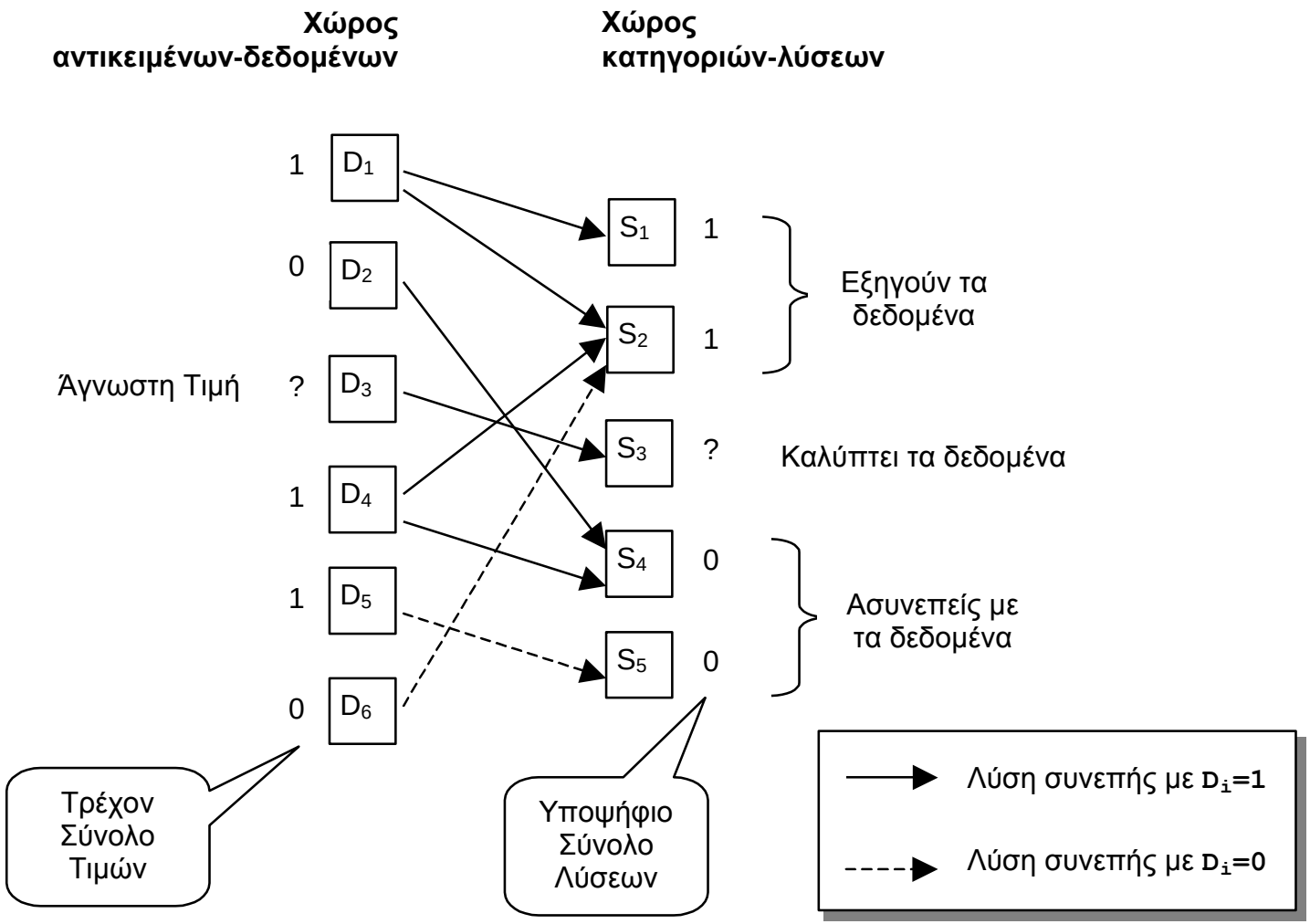
Συζευκτικό Μοντέλο Κατηγοριοποίησης

- Για κάθε υποψήφια λύση $\mathbf{S_j}$ υπάρχει ένα πρότυπο που προσδιορίζει τις απαραίτητες συνθήκες **συνέπειας** μεταξύ λύσεων και δεδομένων.
- $\mathbf{C(S_j,D_i)=1}$
  - Το $\mathbf{S_j}$ είναι συνεπές με $\mathbf{D_i=1}$, δηλαδή δεν μπορεί να είναι λύση αν $\mathbf{D_i=0}$.
- $\mathbf{C(S_j,D_i)=0}$
  - Το $\mathbf{S_j}$ είναι συνεπές με $\mathbf{D_i=0}$, δηλαδή δεν μπορεί να είναι λύση αν $\mathbf{D_i=1}$.
- $\mathbf{C(S_j,D_i)=?}$
  - Οι τιμές του $\mathbf{D_i}$ δεν έχουν σχέση με την συνέπεια των λύσεων $\mathbf{S_j}$

Συζευκτικό Μοντέλο Κατηγοριοποίησης

- Μια υποψήφια λύση (κατηγορία) $\mathbf{S_j}$ είναι **ασυνεπής** με ένα διάνυσμα τιμών και απορρίπτεται, αν τουλάχιστον ένα από τα δεδομένα είναι ασυνεπές με αυτήν.
- Μια λύση είναι **συνεπής** (**consistent**) αν δεν υπάρχει τιμή που να είναι ασυνεπής.
- Η λύση $\mathbf{S_j}$ **καλύπτει** το δεδομένο $\mathbf{D_i}$ αν:
  - $\mathbf{D_i=1}$ και $\mathbf{C(S_j,D_i)=1}$, ή
  - $\mathbf{D_i=0}$ και $\mathbf{C(S_j,D_i)=0}$.
- Μια λύση $\mathbf{S_j}$ **ταιριάζει** ή **εξηγεί** τα δεδομένα, αν όλα τα δεδομένα τα σχετικά με την $\mathbf{S_j}$ είναι γνωστά και όλες οι τιμές τους είναι συνεπείς με αυτήν.

Παράδειγμα


Χώρος αντικειμένων-δεδομένων
Χώρος κατηγοριών-λύσεων
1
D1
0
D2
Άγνωστη Τιμή
?
D3
1
D4
1
D5
0
D6
S1
1
S2
1
S3
?
S4
0
S5
0
Εξηγούν τα δεδομένα
Καλύπτει τα δεδομένα
Ασυνεπείς με τα δεδομένα
Τρέχον Σύνολο Τιμών
Υποψήφιο Σύνολο Λύσεων
Λύση συνεπής με Di=1
Λύση συνεπής με Di=0

Συζευκτικό Μοντέλο Κατηγοριοποίησης

- Ο προσδιορισμός της συνέπειας μιας λύσης γίνεται με τη διάδοση των τιμών από τα δεδομένα στις υποψήφιες λύσεις.
- Με τις **συνεχείς γραμμές**, οι τιμές διαδίδονται όπως είναι.
- Στις **διακεκομμένες γραμμές** οι τιμές αναστρέφονται (δηλ, το 1 γίνεται 0 και αντίστροφα).
- Το ? διαδίδεται πάντα ως έχει.

Συζευκτικό Μοντέλο Κατηγοριοποίησης

- Η κατάσταση μιας λύσης $\mathbf{S_j}$ προσδιορίζεται ως εξής:
  - Αν όλες οι τιμές στο $\mathbf{S_j}$ είναι 1, τότε το $\mathbf{S_j}$ ταιριάζει ή εξηγεί τα δεδομένα.
  - Αν κάποια τιμή στο $\mathbf{S_j}$ είναι 0, τότε το $\mathbf{S_j}$ είναι ασυνεπές και απορρίπτεται.
  - Αν όλες οι τιμές στο $\mathbf{S_j}$ είναι 1 και ?, τότε το $\mathbf{S_j}$ είναι συνεπές ή καλύπτει τα δεδομένα.
- Υπάρχει πιθανότητα κάποια διανύσματα τιμών να μην ταιριάζουν με καμία κατηγορία ή να ταιριάζουν με περισσότερες από μια κατηγορίες.

Μέθοδοι Κατηγοριοποίησης

- **Βασικός στόχος**: Ο αποκλεισμός των εναλλακτικών μονοπατιών στο χώρο αναζήτησης που αποτελείται από υποψήφιες κατηγορίες των δεδομένων εισόδου.
- Μέθοδοι:
  - Κ1: Παραγωγή και Δοκιμή
  - Κ2: Από τα Δεδομένα σε Πιθανές Λύσεις
  - Κ3: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τις Λύσεις
  - Κ4: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τα Δεδομένα

Κ1: Παραγωγή και Δοκιμή

1. Θέσε την κενή λίστα ως λίστα των κατηγοριών-λύσεων
2. Πάρε τα δεδομένα εισόδου και γενίκευσέ τα
3. Για κάθε υποψήφια γενική κατηγορία
   i. Έλεγξε αν τα δεδομένα εισόδου ανήκουν στην υποψήφια γενική κατηγορία
   ii. Εάν ναι, πρόσθεσε την υποψήφια γενική κατηγορία στη λίστα των κατηγοριών
4. Ανέφερε τις λύσεις από τη λίστα των κατηγοριών-λύσεων

Κ1: Παραγωγή και Δοκιμή

- Οι προϋποθέσεις για τη χρήση ενός τέτοιου αλγόριθμου είναι οι εξής:
  - Το σύνολο των κατηγοριών (λύσεων) είναι αρκετά μικρό, έτσι ώστε η εξαντλητική σύγκριση να είναι πρακτικά εφικτή.
  - Όλα τα απαραίτητα δεδομένα μπορούν να αποκτηθούν στην αρχή της διαδικασίας, έτσι ώστε να θεωρηθεί ότι η απουσία κάποιων δεδομένων σημαίνει πως αυτά δεν υφίστανται.
- Ουσιαστικά, ελάχιστα συστήματα χρησιμοποιούν αυτή τη μέθοδο, η οποία παρατίθεται ως βάση για την παρουσίαση των υπολοίπων μεθόδων.

Κ2: Από τα Δεδομένα σε Πιθανές Λύσεις

- Μειώνει τον υπολογιστικό χρόνο όταν υπάρχει μεγάλος αριθμός λύσεων.
  - Επικεντρώνεται μόνο σε εκείνες που *πιθανώς* ικανοποιούν τα δεδομένα.
  - Ασχολείται με κάθε υποψήφια λύση μόνο μια φορά επειδή καταγράφει ποιες λύσεις έχει ήδη δοκιμάσει.
- Χρησιμοποιεί τις ακόλουθες ειδικές διαδικασίες:
  - *Διαδικασία γενίκευσης δεδομένων*
  - *Διαδικασία ανάκλησης υποψήφιων λύσεων*
  - *Διαδικασία ελέγχου λύσεων*

Διαδικασία γενίκευσης δεδομένων
Data abstractor

- Χρησιμοποιεί τις τεχνικές γενίκευσης που αναφέρθηκαν και είναι από μόνη της ένα μικρό σύστημα κατηγοριοποίησης.
- Για κάθε σύνολο δεδομένων μπορεί να υπάρχουν περισσότερες από 1 δυνατότητες γενίκευσης.

Διαδικασία ανάκλησης υποψήφιων λύσεων

Candidate retriever

- Μπορούν εν δυνάμει να **εξηγήσουν** ένα δεδομένο.
- Σύμφωνα με το συζευκτικό μοντέλο κατηγοριοποίησης αυτή η διαδικασία μπορεί να υλοποιηθεί επιστρέφοντας όλες τις λύσεις που είναι **συνεπείς** ή **καλύπτουν** τα δεδομένα.
- Η διαδικασία δεν πρέπει να επιστρέφει πολύ λίγες λύσεις, γιατί τότε θα χαθούν κάποιες, ούτε πάρα πολλές, γιατί τότε δε θα υπάρχει βελτίωση στην απόδοση σε σχέση με την Κ1.

Διαδικασία ελέγχου λύσεων
Solution tester

- Βαθμολογεί τις υποψήφιες κατηγορίες-λύσεις βασιζόμενος σε διάφορα κριτήρια και επιλέγει στο τέλος μία ή περισσότερες.
- Π.χ. προϋπάρχουσες πιθανότητες

Αλγόριθμος K2

1. Θέσε την κενή λίστα ως λίστα των κατηγοριών-λύσεων.
2. Πάρε τα δεδομένα εισόδου.
3. Γενίκευσε τα δεδομένα με τη διαδικασία data_abstractor.
4. Βρες τις υποψήφιες γενικές κατηγορίες με τη διαδικασία candidate_retriever.
5. Για κάθε υποψήφια γενική κατηγορία:
   i. Έλεγξε αν τα δεδομένα εισόδου ανήκουν στην υποψήφια γενική κατηγορία (solution_tester)
   ii. Εάν ναι, πρόσθεσε την υποψήφια γενική κατηγορία στη λίστα των κατηγοριών.
6. Ανέφερε τις λύσεις από τη λίστα των κατηγοριών-λύσεων.

Κ3: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τις Λύσεις

- Αποτελεί μια ιεραρχική τεχνική παραγωγής και δοκιμής
- Υποθέτει ότι οι πιθανές κατηγορίες-λύσεις είναι ιεραρχικά διατεταγμένες
  - Διασχίζει την ιεραρχία από πάνω προς τα κάτω και πρώτα σε πλάτος.
  - Οι τελικές κατηγορίες-λύσεις βρίσκονται στα φύλλα του δένδρου.
  - Οι λύσεις των ενδιάμεσων επιπέδων είναι μερικές ή γενικές.

Κ3: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τις Λύσεις

- Σε κάθε επίπεδο, η μέθοδος συγκρίνει τις υποψήφιες λύσεις με τα δεδομένα και απορρίπτει κλαδιά του δένδρου.
  - Μπορεί να ζητήσει πρόσθετα δεδομένα για να επιτύχει λεπτομερέστερη διάκριση μεταξύ των υποψήφιων λύσεων.
- Στη συνέχεια η μέθοδος προχωρά στον έλεγχο λύσεων σε κατώτερο επίπεδο του δένδρου, ακολουθώντας μόνο τα κλαδιά που δεν έχουν απορριφθεί.
  - Αν μια κατηγορία απορριφθεί, είναι προφανές ότι απορρίπτονται και όλες οι υποκατηγορίες που περιλαμβάνει.

Αλγόριθμος K3 (1/2)

1. Θέσε την κενή λίστα ως λίστα των κατηγοριών-λύσεων.
2. Για κάθε επίπεδο της ιεραρχίας των κατηγοριών, επανέλαβε:
   i. Πάρε τα δεδομένα εισόδου που χρειάζονται για να γίνει διάκριση των υποψήφιων λύσεων σε αυτό το επίπεδο.
   ii. Γενίκευσε τα δεδομένα με τη διαδικασία data_abstractor.
   iii. Για κάθε υποψήφια γενική κατηγορία αυτού του επιπέδου:

   …

Αλγόριθμος Κ3 (2/2)

2. …
   iii. Για κάθε υποψήφια γενική κατηγορία αυτού του επιπέδου:
      a. Έλεγξε αν τα δεδομένα εισόδου ανήκουν στην υποψήφια γενική κατηγορία με τη διαδικασία solution_tester.
      b. Εάν όχι, τότε απέρριψε την υποψήφια γενική κατηγορία.
      c. Εάν ναι, τότε έλεγξε αν η υποψήφια γενική κατηγορία είναι τερματική στην ιεραρχία
         1) Εάν όχι, τότε αφαίρεσε την υποψήφια γενική κατηγορία από τη λίστα και πρόσθεσε τις κατηγορίες των κόμβων-παιδιών του δένδρου.
         2) Εάν ναι, τότε πρόσθεσε την υποψήφια γενική κατηγορία στη λίστα
3. Βαθμολόγησε τις λύσεις από τη λίστα και επέστρεψέ τες.

Προϋποθέσεις Εφαρμογής Κ3

- Οι κατηγορίες πρέπει να μπορούν να διαταχθούν σε ιεραρχία
- Σε *κάθε* επίπεδο πρέπει να υπάρχει ένα υποσύνολο των δεδομένων που μπορεί να κάνει διάκριση μεταξύ των γενικών κατηγοριών του συγκεκριμένου επιπέδου.
- Η επιλογή του υποσυνόλου πρέπει να είναι τέτοια ώστε το δένδρο να είναι όσο πιο **ισορροπημένο** γίνεται
  - Ο διαχωρισμός των πληθυσμών να είναι όσο το δυνατόν μεγαλύτερος
  - Σε κάθε βήμα να είναι δυνατός ο αποκλεισμός ενός σημαντικού αριθμού λύσεων (αποδοτικότητα).

Προϋποθέσεις Εφαρμογής Κ3

- Ομοιότητα με τα ***δένδρα κατηγοριοποίησης*** ή ***απόφασης***.
- Ο στόχος είναι να βρεθούν οι πιο ειδικές κατηγορίες που είναι *συνεπείς* με τα δεδομένα
  - Οι τερματικοί κόμβοι (φύλλα) του δένδρου
- Ο αλγόριθμος σταματά όταν φθάσει στους τερματικούς κόμβους
  - Ή αν δεν υπάρχουν άλλα κλαδιά προς εξερεύνηση

Κ4: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τα Δεδομένα

- Η Κ3 εξακολουθεί να είναι απλή
  - Ανεξάρτητα από την περίπτωση ξεκινά ζητώντας πάντα τα ίδια δεδομένα.
  - Σε προβλήματα π.χ. διάγνωσης βλαβών ή ασθενειών αυτό θα ήταν εξαιρετικά χρονοβόρο αφού σχεδόν πάντα κάποιος ξεκινά τη διάγνωση από τα δεδομένα-συμπτώματα που έχει.
- Η μέθοδος Κ4:
  - Χρησιμοποιεί και αυτή έναν ιεραρχικό χώρο αναζήτησης
  - Αλλά αποφεύγει να ζητήσει από το χρήστη δεδομένα που δεν έχουν σχέση με την περίπτωση που αντιμετωπίζεται.

Κ4: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τα Δεδομένα

- Η Κ4 υποθέτει ότι όλα τα σχετικά δεδομένα έχουν δοθεί στην αρχή
  - Όπως η Κ2
- Η Κ4 προχωρά στην επιλογή των υποψήφιων κατηγοριών
  - Αποκλείει από την ιεραρχία όλα εκείνα τα κλαδιά που δεν είναι συμβατά με τα δεδομένα που δόθηκαν
  - Όπως η Κ3
- Στη συνέχεια αποκλείονται κάποιες από τις αρχικές υποψήφιες λύσεις, που δεν επιβεβαιώνουν τα δεδομένα

Αλγόριθμος K4 (1/2)

1. Θέσε την κενή λίστα ως λίστα των κατηγοριών-λύσεων.
2. Πάρε τα δεδομένα εισόδου.
3. Γενίκευσε τα δεδομένα με τη διαδικασία data_abstractor.
4. Βρες τις υποψήφιες γενικές κατηγορίες με τη διαδικασία candidate_retriever.
5. Για κάθε επίπεδο της ιεραρχίας των κατηγοριών που ξεκινάει από τις υποψήφιες γενικές κατηγορίες (**και όχι από την κορυφή της ιεραρχίας**), επανέλαβε:
   i. Πάρε τα δεδομένα εισόδου που χρειάζονται για να γίνει διάκριση των υποψήφιων λύσεων σε αυτό το επίπεδο.
   ii. Γενίκευσε τα δεδομένα με τη διαδικασία data_abstractor.
   iii. Πρόσθεσε νέες υποψήφιες γενικές κατηγορίες λόγω των καινούριων δεδομένων.

Αλγόριθμος K4 (2/2)

iv. Για κάθε υποψήφια γενική κατηγορία αυτού του επιπέδου:

   a. Έλεγξε αν τα δεδομένα εισόδου ανήκουν στην υποψήφια γενική κατηγορία με τη διαδικασία solution_tester.

   b. Εάν όχι, τότε απέρριψε την υποψήφια γενική κατηγορία.

   c. Εάν ναι, τότε έλεγξε αν η υποψήφια γενική κατηγορία είναι τερματική στην ιεραρχία

      1) Εάν όχι, τότε αφαίρεσε την υποψήφια γενική κατηγορία από τη λίστα και πρόσθεσε τις κατηγορίες των κόμβων-παιδιών του δένδρου.

      2) Εάν ναι, τότε πρόσθεσε την υποψήφια γενική κατηγορία στη λίστα των λύσεων.

6. Βαθμολόγησε τις λύσεις από τη λίστα και επέστρεψέ τες.

Πλεονεκτήματα Κ4

- Η μέθοδος Κ4 βελτιώνει την Κ3
  - Θεωρεί ότι όσα δεδομένα δόθηκαν στην αρχή είναι τα μόνα σχετικά με την περίπτωση
  - Δεν εξετάζει καθόλου τις γενικές κατηγορίες οι οποίες δεν εξηγούν τα αρχικά δεδομένα
  - Όταν φτάσει στη γενίκευση των δεδομένων, **η Κ4 λειτουργεί αντίστροφα, από τις λύσεις προς τα δεδομένα**, όπως και η Κ3, και ρωτά για επιπλέον δεδομένα τα οποία χρειάζεται για να αποκλείσει κάποιες υποψήφιες λύσεις
    - Στη φάση αυτή, κάποιες από τις γενικές κατηγορίες που είχαν αρχικά αποκλειστεί μπορεί με τα καινούρια δεδομένα να θεωρηθούν εκ νέου υποψήφιες.

Πλεονεκτήματα Κ4

- Η μέθοδος κινείται σε **κάθε** κύκλο:
  - προς τα πάνω, γενικεύοντας τα νέα δεδομένα που αποκτά
  - προς τα κάτω, αποκλείοντας κάποια κλαδιά του δένδρου λόγω των νέων δεδομένων
- Η Κ4 παρουσιάζεται ευέλικτη και πιο αποδοτική, γιατί επικεντρώνεται στην εξέταση μόνο των λύσεων που είναι απολύτως σχετικές με τα δεδομένα.

Μελέτη Περίπτωσης
Το Σύστημα DENDRAL

- Ενδιάμεσος σταθμός μεταξύ προγραμμάτων "έξυπνης αναζήτησης" και εμπείρων συστημάτων (απ' ευθείας καταγραφή ειδικής γνώσης).
- Η αναζήτηση λύσεων περιορίζεται με πληροφορίες του χρήστη (καθοδηγεί το πρόγραμμα στην αναζήτηση εναλλακτικών επιλογών).
- Αναπτύχθηκε στο πανεπιστήμιο STANFORD από το 1965.

Αντικείμενο DENDRAL

- Ο καθορισμός της μοριακής δομής (**στερεοχημικό τύπος**) αγνώστων οργανικών ουσιών
  - Από την ανάλυση των αποτελεσμάτων του **φασματογράφου μάζας**
  - Με τη χρήση τροποποιημένης μεθόδου **δημιουργίας και ελέγχου (generate-and-test)** εναλλακτικών λύσεων.

Λίγη Φυσική-Χημεία!

- Με το φασματογράφο μάζας η ένωση διασπάται σε τμήματα (όχι πάντα με τον ίδιο τρόπο), τα οποία διαχωρίζονται και καταγράφεται η έντασή τους.
  - Η παρουσία, η σχετική ένταση ή η απουσία τμημάτων από το φάσμα δίνει αφορμή για υποθέσεις σχετικά με τη δομή ή για περιορισμούς που πρέπει να ικανοποιεί.
- **Στερεοχημικός τύπος**: δομή των ατόμων που αποτελούν το μόριο στο χώρο, καθώς και το είδος των μεταξύ τους χημικών δεσμών
  - Δεδομένος ο μοριακός τύπος (π.χ. $C_6H_{13}NO_2$)
  - Τα ισομερή που αντιστοιχούν σε ένα μοριακό τύπο είναι συνήθως πάρα πολλές (π.χ. 10000)

Βάση γνώσης DENDRAL

- Περιέχει **περιορισμούς** που πρέπει να ικανοποιεί η ένωση
  - Βασίζονται στην παρουσία-απουσία τμημάτων από το φάσμα
  - Χρησιμοποιούνται για να περιορίσουν το μεγάλο αριθμό εναλλακτικών δομών.
- **Απαιτούμενοι**: Οι υποψήφιες ενώσεις πρέπει να ικανοποιούν τα στοιχεία που παρατηρήθηκαν.
- **Απαγορευτικοί**: Οι πιθανές ενώσεις πρέπει να είναι χημικά σταθερές
- Μπορούν να δοθούν με διάφορους τρόπους:
  - Δομικός σκελετός πάνω στον οποίο πρέπει να τοποθετηθούν τα άτομα
  - Κανόνες πιθανών τρόπων διάσπασης μιας ένωσης

Λειτουργία DENDRAL

- Συσχετίζει τις κορυφές εντάσεων του φάσματος με τμήματα της διασπασμένης ένωσης
- Συνδυάζει τα τμήματα ώστε να ταιριάζουν στο δομικό σκελετό
- Δίνει στην έξοδο μια λίστα ημιτελών υποθέσεων σχετικά με τη δομή της άγνωστης ένωσης

Έλεγχος εκτέλεσης DENDRAL

- **Δημιουργία και έλεγχος υποθέσεων** (**hypothesize-and-test**) (μέθοδος K1).
- Αρχικά τα δεδομένα συνιστούν ένα μεγάλο σύνολο υποψήφιων λύσεων (υποθέσεις).
- Κάθε υπόθεση μπορεί να ελεγχθεί με την ύπαρξη ή την απουσία σχετικών δεδομένων και μπορεί είτε να γίνει περισσότερο συγκεκριμένη ή να αποκλειστεί.
- Η διαδικασία επαναλαμβάνεται, προσθέτοντας περισσότερους περιορισμούς που μειώνουν ακόμα περισσότερο το σύνολο των υποθέσεων.

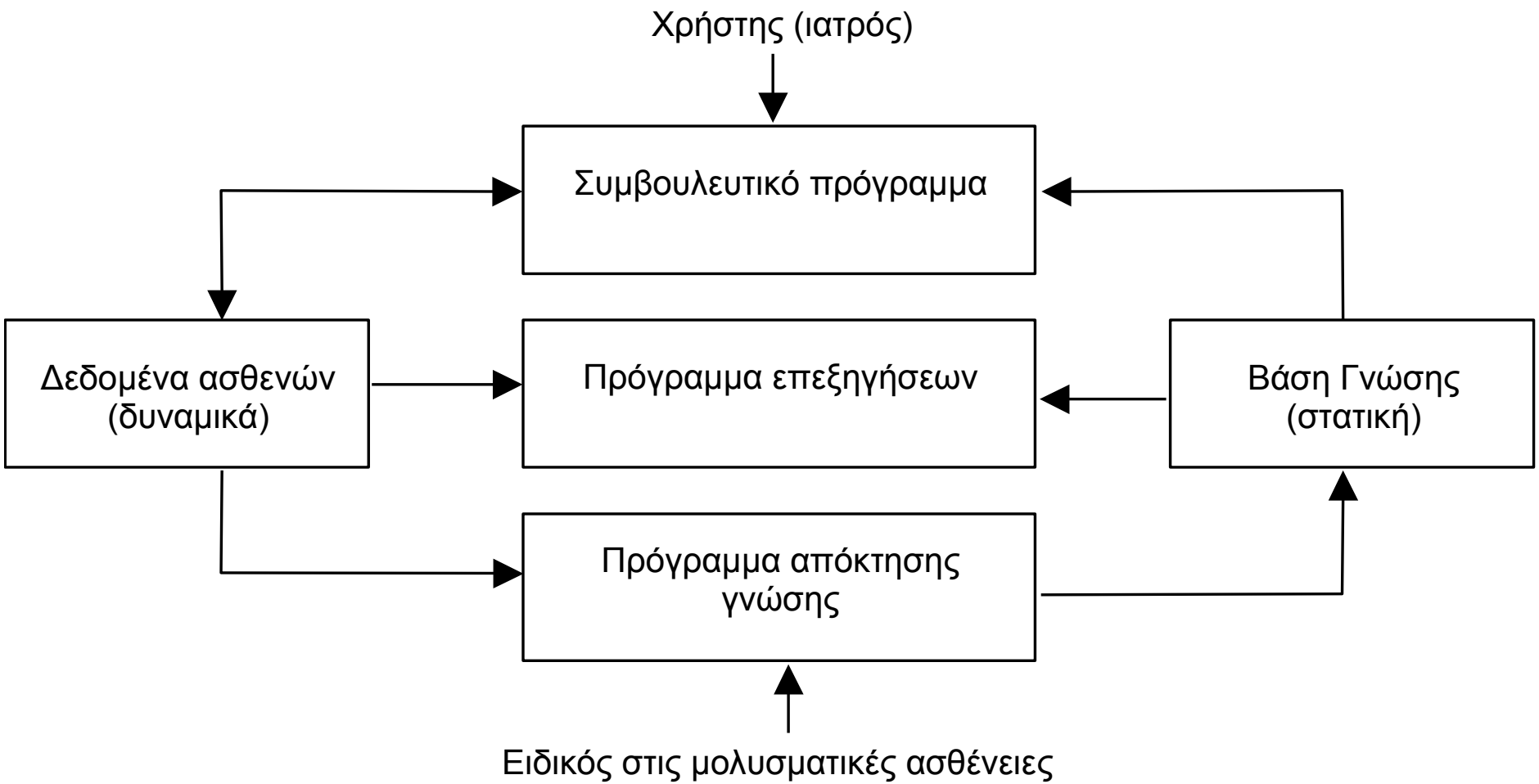
Μελέτη Περίπτωσης
Το Σύστημα MYCIN

- Αντιμετώπιση μολύνσεων του αίματος από βακτήρια-μικρόβια
  - Έγκαιρη διάγνωση πιθανών μικροοργανισμών που προκαλούν τη μόλυνση.
  - Πρόταση αντιβιοτικών(-ου) για την αντιμετώπιση
- Λειτουργεί παρόμοια με την **απόδειξη θεωρημάτων**
  - Η επίτευξη ενός στόχου αναλύεται σε επίτευξη υποστόχων
  - Ερευνά το μεγαλύτερο μέρος των εναλλακτικών διαδρομών προς την επίλυση ενός προβλήματος
  - Αξιολογεί τις εναλλακτικές διαδρομές βάσει κριτηρίων

Βήματα Λειτουργία MYCIN

- Λήψη απόφασης για το αν ο ασθενής έχει κάποια σοβαρή μόλυνση.
- Καθορισμός των πιθανών μικροοργανισμών (μικροβίων) που εμπλέκονται.
- Επιλογή του συνόλου των κατάλληλων φαρμάκων.
- Επιλογή του καταλληλότερου φαρμάκου ή συνδυασμού φαρμάκων.

Βασική Δομή MYCIN



Βάση Γνώσης Mycin

- Κανόνες της μορφής:

IF **condition$_1$ AND … AND condition$_m$**

THEN **conclusion$_1$ AND … AND conclusion$_n$**

- Παράδειγμα κανόνα (κατηγορία μικροοργανισμού):

IF **The stain of the organism is Gram negative, and**

The morphology of the organism is rod, and

The aerobicity of the organism is aerobic

THEN **There is strongly suggestive evidence (0.8) that**

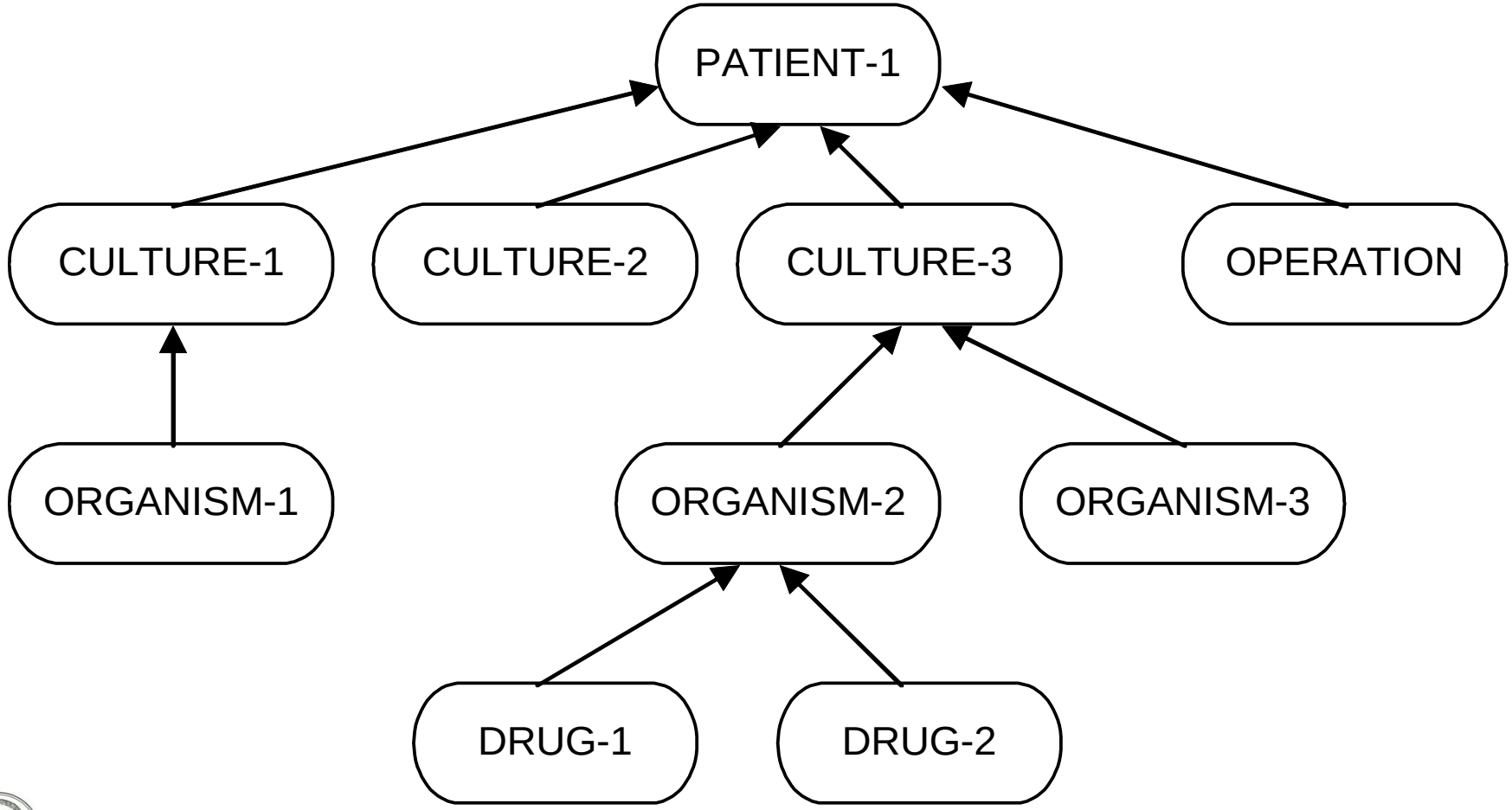
the class of the organism is Enterobacteriaceae

  - Ο αριθμός **0.8** είναι η **βεβαιότητα** του κανόνα
    - Καθορίζει πόσο σίγουρο είναι το συμπέρασμά, με την προϋπόθεση ότι ικανοποιούνται οι συνθήκες

Βάση Γνώσης MYCIN

- Η βάση γνώσης περιέχει επίσης γεγονότα και ορισμούς διαφόρων μορφών:
  - **Απλές λίστες**, π.χ. η λίστα όλων των μικροοργανισμών που γνωρίζει το σύστημα.
  - **Πίνακες γνώσης**: Περιέχουν εγγραφές κλινικών παραμέτρων και τις τιμές που παίρνουν.
  - Ένα σύστημα **ταξινόμησης** των **κλινικών παραμέτρων** ανάλογα με την κατηγορία στην οποία ανήκουν, π.χ. αν είναι χαρακτηριστικά ασθενών ή μικροοργανισμών.
- Οι πληροφορίες για τον ασθενή αποθηκεύονται στο **context tree** (**δένδρο περιβάλλοντος**).

Δένδρο Περιβάλλοντος Ασθενή



Έλεγχος Εκτέλεσης στο MYCIN

- Οι κανόνες εκτελούνται ανάστροφα από κάποιον αρχικό στόχο-σκοπό (σύσταση κατάλληλης θεραπείας).
- Ο αρχικός στόχος σταδιακά αναλύεται σε απλούστερους υποστόχους
  - Πρέπει να επιτευχθούν για να επιτευχθεί και ο αρχικός στόχος
- Οι υποστόχοι περιλαμβάνουν:
  - τον καθορισμό των εμπλεκομένων μικροοργανισμών και
  - τη διαπίστωση της σοβαρότητάς τους ως παράγοντες της μολυσματικής ασθένειας
- Οι περισσότεροι υποστόχοι έχουν δικούς τους υποστόχους,
  - Π.χ. καθορισμός των ιδιοτήτων της χρώσης και της μορφολογίας του μικροοργανισμού.

Έλεγχος Εκτέλεσης στο MYCIN

- Οι απλούστεροι στόχοι είναι η ανάκληση γεγονότων από τη βάση δεδομένων ή το χρήστη.
  - Εργαστηριακά δεδομένα που δεν μπορούν να εξαχθούν με λογικούς συμπερασμούς.
- **Βασικός στόχος**:

IF **υπάρχει κάποιος μικροοργανισμός που χρειάζεται αντιμετώπιση ΚΑΙ**
όλοι οι υπόλοιποι μικροοργανισμοί έχουν αντιμετωπισθεί

THEN **φτιάξε μία λίστα με πιθανές θεραπείες-φάρμακα ΚΑΙ**
εξακρίβωσε την καλύτερη από αυτές.

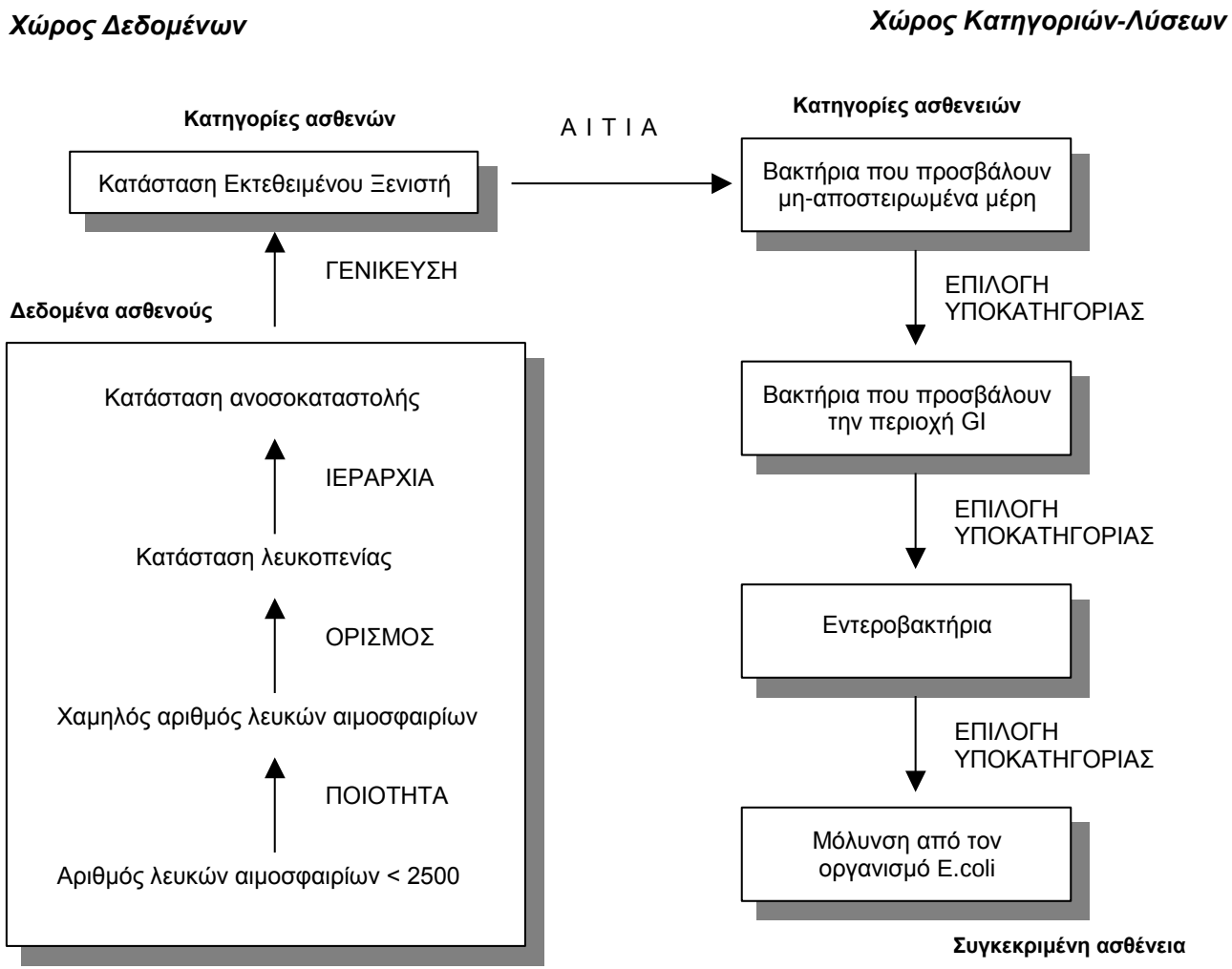
Ανάστροφη Ακολουθία Εκτέλεσης στο MYCIN

- Οι κανόνες που περιέχουν τις **κύριες παραμέτρους** εκτελούνται πριν από τους υπόλοιπους.
- Πολλοί από τους συμπερασμούς στο MYCIN είναι **αβέβαιοι**.
  - Συλλέγονται πληροφορίες από **όλους** τους σχετικούς κανόνες.
  - Η **βεβαιότητά** τους **συνδυάζεται** για να εξαχθεί η τελική βεβαιότητα κάποιου συμπεράσματος.
  - Αν κάποιος κανόνας έχει **βεβαιότητα 1.0** τότε χρησιμοποιείται **μόνο** αυτός για την εξαγωγή συμπεράσματος.
  - Αν κάποιο συμπέρασμα εξαχθεί με βεβαιότητα **μεταξύ –0.2 και +0.2**, τότε θεωρείται ότι δεν εξήχθη καθόλου (αυθαίρετο).
  - Αν κάποια συνθήκη κανόνα είναι από την αρχή **σίγουρα ψευδής (βεβαιότητα -1.0)**, τότε ο κανόνας δεν εξετάζεται καθόλου
  - Αν κάποια παράμετρος εξαχθεί με **απόλυτη βεβαιότητα (1.0)**, τότε προηγούνται οι κανόνες που την έχουν στη συνθήκη τους

Κατηγοριοποίηση στο MYCIN

- Χρησιμοποιεί μέθοδο κατηγοριοποίησης που είναι πιο κοντά στην Κ3.
  - Το σύστημα οδηγείται από τις λύσεις προς τα δεδομένα.
  - Χρησιμοποιεί δενδροειδή ιεράρχηση των υπο-στόχων.
  - Το MYCIN χρησιμοποιεί πρώτα σε βάθος αναζήτηση (ενώ η Κ3 πρώτα σε πλάτος).
- Βασικά στάδια κατηγοριοποίησης:
  - Χρησιμοποιούνται οι σχέσεις **γενίκευσης**, για να βρεθεί μία γενική κατηγορία ασθενούς για την οποία μπορεί να βρει μία πιθανή **αιτία** (ασθένεια).
  - Η γενική κατηγορία γίνεται ολοένα και πιο συγκεκριμένη, **επιλέγοντας υποκατηγορίες** της αρχικής ασθένειας από την ιεραρχία των μικροβίων.

Βασικά Στάδια Κατηγοριοποίησης



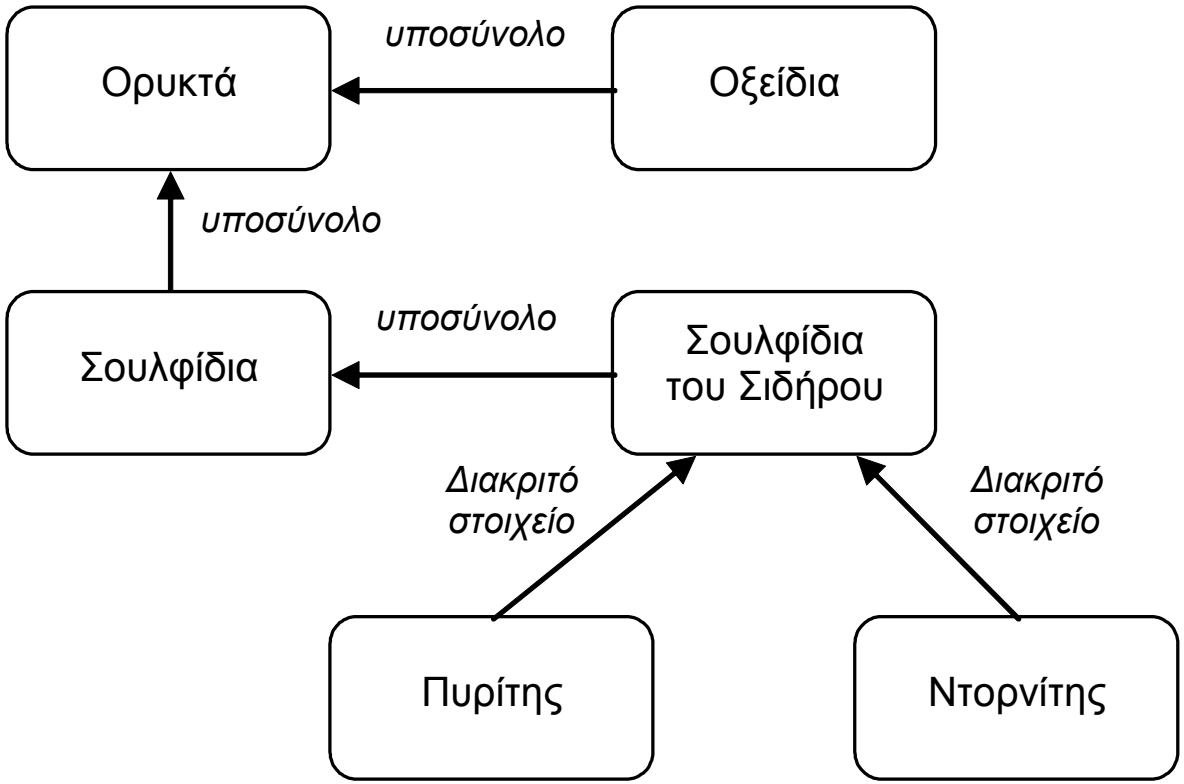
Μελέτη Περίπτωσης
Το Σύστημα PROSPECTOR

- Έμπειρο σύστημα για την αξιολόγηση γεωλογικών δεδομένων και τον καθορισμό της πιθανότητα ύπαρξης αξιόλογων ορυκτών κοιτασμάτων σε μια περιοχή.
- Αναπτύχθηκε τη δεκαετία του 1970 στο Stanford Research Institute.
- Χρησιμοποιήθηκε πειραματικά και προέβλεψε την ύπαρξη ορυκτού κοιτάσματος μολυβδαίνιου (επιβεβαιώθηκε με γεωτρήσεις).

Μελέτη Περίπτωσης
Το Σύστημα PROSPECTOR

- Λειτουργεί σε δύο φάσεις
  - Δέχεται παρατηρήσεις (δεδομένα) και κάνει ερωτήσεις προκειμένου να καθορίσει την πιθανή ύπαρξη κάποιου χρήσιμου αποθέματος.
  - Αν η πιθανότητα είναι καλή, εμφανίζει γραφική απεικόνιση των πιθανών σημείων γεώτρησης.
- Η γνώση αναπαριστάται με δύο είδη δικτύων.
  - **Σημασιολογικά δίκτυα:** Αναπαριστούν γεωλογικές γνώσεις.
  - **Δίκτυα Συμπερασμού**: Αναπαριστούν τους κανόνες.

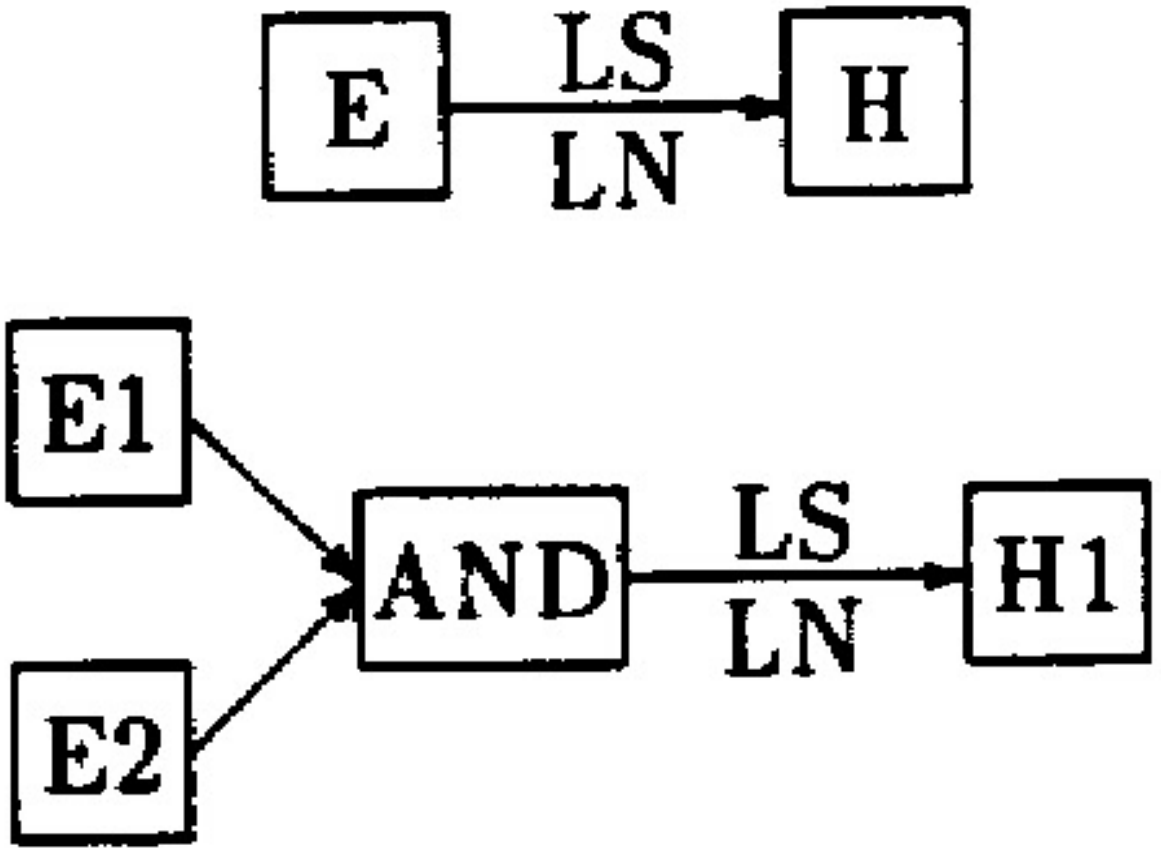
Σημασιολογικά Δίκτυα



Σημασιολογικά Δίκτυα

- Αναπαριστούν γεωλογικές γνώσεις.
- Αξιοποιούνται για **εξαγωγή συμπερασμάτων**
  - Π.χ., αν υπάρχουν πυρίτες στην περιοχή, τότε συμπεραίνεται ότι υπάρχει θειούχος σίδηρος και γενικά θειούχες ενώσεις.
- Αξιοποιούνται για έλεγχο πληροφοριών που έδωσε ο χρήστης
  - Π.χ., αν δεν υπάρχουν θειούχα κοιτάσματα στην περιοχή δεν μπορούν να υπάρχουν και πυρίτες.

Δίκτυα Συμπερασμού



Δίκτυα Συμπερασμού

- Αναπαριστούν τους κανόνες.
- Υπάρχει ένα δίκτυο για κάθε ορυκτό.
- Τα τόξα αναπαριστούν τη σχέση της συνεπαγωγής ενώ οι κόμβοι τις λογικές πράξεις.
- Κάθε κόμβος έχει μια προϋπάρχουσα πιθανότητα, δηλαδή πιθανότητα να υπάρχει μία παρατήρηση Ε στην περιοχή χωρίς την ύπαρξη άλλων αποδεικτικών στοιχείων.

Δίκτυα Συμπερασμού

- Με τη χρήση των μεγεθών της **λογικής επάρκειας** (LS) και **αναγκαιότητας** (LN) η πιθανότητα κάθε κόμβου του δικτύου μεταβάλλεται, λόγω της παρουσίας άλλων πληροφοριών στο δίκτυο.
- Οι πιθανότητες και τα μεγέθη LS, LN παρέχονται από τους ειδικούς-γεωλόγους.
  - Υποκειμενικότητα - Πιθανότητα σφάλματος.
  - Ο τρόπος δόμησης της γνώσης και της αβεβαιότητας είναι κατάλληλος μόνο για μικρό αριθμό δικτύων (15 δίκτυα, 150 κανόνες)

Λογική Επάρκεια
Logical Sufficiency – LS

- Η λογική επάρκεια εκφράζει πόσο πιθανότερο είναι να συνδεθεί ένα γεγονός *E* με την αλήθεια ενός υποθετικού συμπεράσματος *H*, παρά με την άρνηση του *H* (συμβολίζεται ¬*H*)
- $LS = \frac{P(E|H)}{P(E|\neg H)}$

Λογική Αναγκαιότητα
Logical Necessity – LN

- Η λογική αναγκαιότητα εκφράζει το πόσο πιθανότερο είναι να συνδεθεί η απουσία ενός γεγονότος Ε με την αλήθεια του υποθετικού συμπεράσματος Η παρά με την άρνηση του Η
- $$LN = \frac{P(\neg E|H)}{P(\neg E|\neg H)}$$

Έλεγχος Εκτέλεσης στο PROSPECTOR

- Εισαγωγή Δεδομένων
- Προώθηση Πιθανοτήτων
- Επιβεβαίωση Υπόθεσης

Εισαγωγή Δεδομένων

- Ο χρήστης παραθέτει ένα σύνολο παρατηρήσεων.
  - Κάθε πληροφορία συνοδεύεται από μία τιμή από –5 ως 5
  - Υπολογίζεται η πιθανότητα ύπαρξης της παρατήρησης **E**, βάσει των παρατηρήσεων **E'** του χρήστη.
- –5 $\rightarrow$ **P(E|E')=0**: το **E δεν υπάρχει** βάσει των παρατηρήσεων του χρήστη.
- 0 $\rightarrow$ **P(E|E')=P(E)**: Οι παρατηρήσεις δε μεταβάλλουν τις προϋπάρχουσες πιθανότητες.
- +5 $\rightarrow$ **P(E|E')=1**: Το **E υπάρχει** βάσει των παρατηρήσεων

Προώθηση Πιθανοτήτων

- Για κάθε πληροφορία, το σύστημα προωθεί στο δίκτυο τις μεταβολές των πιθανοτήτων με **ορθή ακολουθία εκτέλεσης**
- Συνεχίζεται έως ότου προκύψει η ύπαρξη ή όχι κάποιου ορυκτού στην περιοχή.
- Το πιο πιθανό από τα δίκτυα επιλέγεται ως υποψήφιο για την επόμενη φάση.

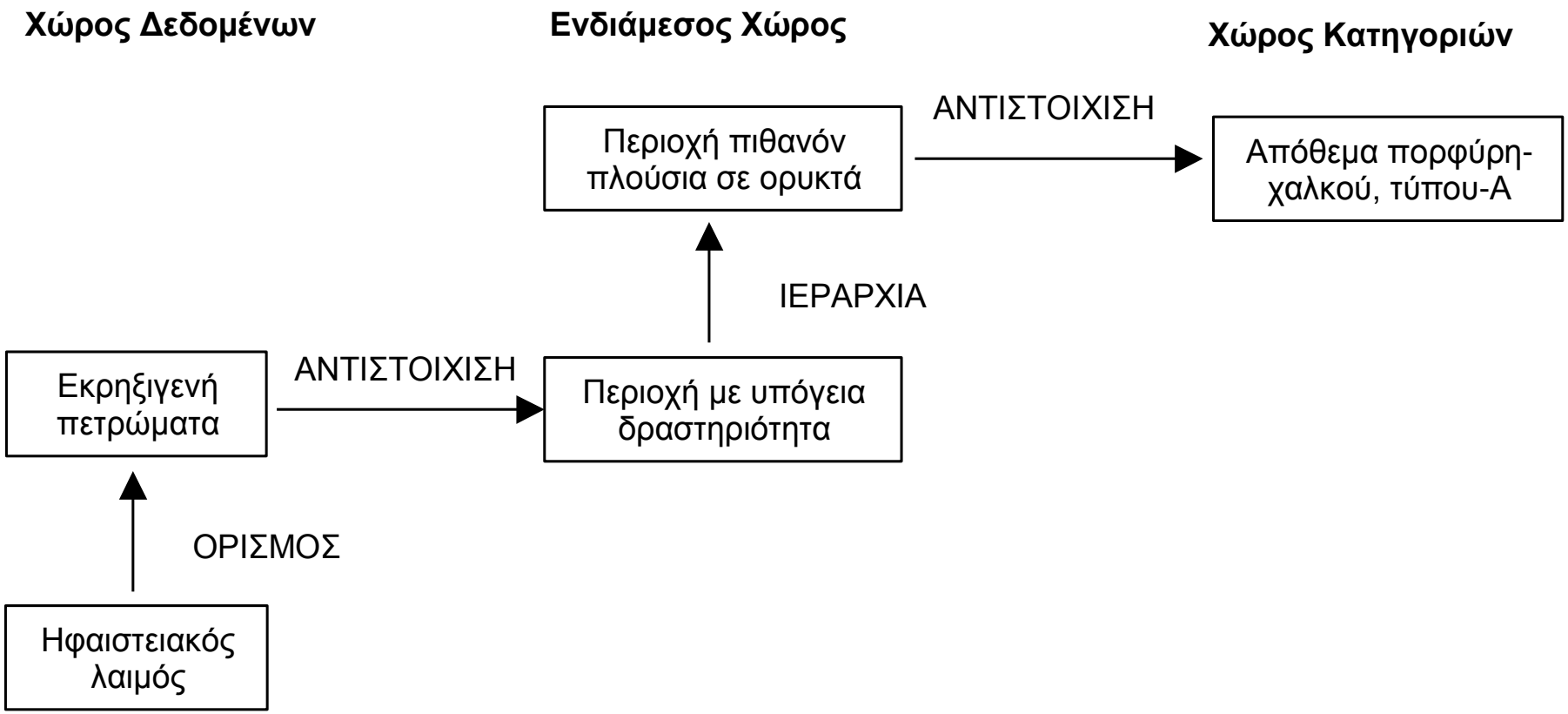
Επιβεβαίωση Υπόθεσης

- Το σύστημα δουλεύει από το συμπέρασμα προς τις παρατηρήσεις (ανάστροφη ακολουθία) ώστε να επιβεβαιώσει την επιλεχθείσα υπόθεση.
- Η διαδικασία συνεχίζεται έως ότου το σύστημα φθάσει σε τερματικούς κόμβους.
  - Επανάληψη της **εισαγωγής δεδομένων**.
  - Εκ νέου **προώθηση πιθανοτήτων** προς τον τελικό κόμβο.
  - Αν ο τελικός κόμβος εξακολουθεί να είναι το πιο πιθανό συμπέρασμα, τότε ο κύκλος συνεχίζεται για το ίδιο δίκτυο με την **επιβεβαίωση της υπόθεσης**.
  - Αλλιώς επιλέγεται το πιο πιθανό από τα υπόλοιπα δίκτυα.

Κατηγοριοποίηση στο PROSPECTOR

- Χρησιμοποιεί τη μέθοδο κατηγοριοποίησης K4.
  - Αρχικά οδηγείται από τα δεδομένα του χρήστη σε πιθανές λύσεις.
  - Στη συνέχεια προσπαθεί να φτάσει από τις πιθανές υποψήφιες λύσεις στα δεδομένα που τις στηρίζουν.
- Χρησιμοποιεί ενδιάμεσες υποθέσεις στην κατηγοριοποίηση.
  - Χρήσιμα ενδιάμεσα συμπεράσματα σχετικά με την περιοχή που βρίσκεται κάποιο πέτρωμα, την ηλικία και τον τρόπο σχηματισμού του, κλπ.

Χώροι αναζήτησης στο Prospe



ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΙΚΑ
ΜΑΘΗΜΑΤΑ

Τέλος Ενότητας

Επεξεργασία: Εμμανουήλ Ρήγας

Θεσσαλονίκη, 17/3/2014

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης