ΔΗΜΙΟΥΡΓΙΑ ΒΑΣΕΩΝ ΚΑΝΟΝΩΝ ΑΠΟ ΔΕΔΟΜΕΝΑ
Μέρος Δ': Δημιουργία και Αξιολόγηση Βάσης Κανόνων

Διδάσκων:

Ι. ΧΑΤΖΗΛΥΓΕΡΟΥΔΗΣ

Πανεπιστήμιο Πατρών, Τμήμα Μηχ/κών Η/Υ και Πληροφορικής

ΔΗΜΙΟΥΡΓΙΑ ΒΑΣΗΣ ΚΑΝΟΝΩΝ

- Οι κανόνες που έχουν εξαχθεί μέσω της μεθόδου ΔΑ
  - ✓Μπορεί να έχουν μη ικανοποιητική απόδοση ή η απόδοσή τους να μπορεί να βελτιωθεί.
  - ✓Μπορεί κάποιοι από αυτούς να μην ανταποκρίνονται στην πραγματικότητα.
- Οπότε απαιτείται αναθεώρηση των κανόνων.

ΔΗΜΙΟΥΡΓΙΑ ΒΑΣΗΣ ΚΑΝΟΝΩΝ

- Διαδικασία αναθεώρησης κανόνων:
  1. Υλοποίηση των κανόνων μέσω ενός εργαλείου ανάπτυξης συστημάτων κανόνων.
  2. Αναθεώρηση των κανόνων με τη βοήθεια εμπειρογνώμονα του πεδίου και του συνόλου δεδομένων.
  3. (Επαν)αξιολόγηση των κανόνων.
  4. Αν τα αποτελέσματα είναι ικανοποιητικά, τότε σταμάτα. Αλλιώς, πήγαινε στο βήμα 2.

ΕΡΓΑΛΕΙΑ ΥΠΟΣΤΗΡΙΞΗΣ ΕΞΑΓΩΓΗΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗΣ ΚΑΝΟΝΩΝ

- **WEKA**
  - ✓Προσφέρει σύνοψη δεδομένων
  - ✓Οπτικοποίηση κατανομής παραδειγμάτων
  - ✓Χειρισμός ελλιπών τιμών
  - ✓Χειρισμός θορύβου δεδομένων
  - ✓Κάνει επιλογή χαρακτηριστικών
  - ✓Κάνει διακριτοποίηση
  - ✓Παράγει κανόνες

ΕΡΓΑΛΕΙΑ ΥΠΟΣΤΗΡΙΞΗΣ ΕΞΑΓΩΓΗΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗΣ ΚΑΝΟΝΩΝ

- CLIPS (JESS)
  - ✓Κέλυφος ανάπτυξης συστημάτων βασισμένων σε κανόνες.
  - ✓Υλοποίηση προτασιακών κανόνων και κανόνων πρώτης τάξεως.
  - ✓Ελεύθερο λογισμικό ανοικτού κώδικα.
  - ✓Διαθέτει εργαλεία εκσφαλμάτωσης.
  - ✓Επιλογή στρατηγικών συλλογισμού και επίλυσης συγκρούσεων.

ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ-ΠΡΟΒΛΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ

Μετρήσεις

- ✓ TP (true positives): πλήθος περιπτώσεων που ανήκουν σε μια κλάση C και ταξινομήθηκαν σ' αυτή .
- ✓ FN (false negatives): πλήθος περιπτώσεων που ανήκουν σε μια κλάση C και δεν ταξινομήθηκαν σ' αυτή .
- ✓ FP (false positives): πλήθος περιπτώσεων που δεν ανήκουν σε μια κλάση C και ταξινομήθηκαν σ' αυτή.
- ✓ TN (true negatives): πλήθος περιπτώσεων που δεν ανήκουν σε μια κλάση C και δεν ταξινομήθηκαν σ' αυτή .

ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ-ΠΡΟΒΛΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ

- **Μετρικές**
  - **Accuracy** = (TP+TN)/(TP+FN+FP+TN) (Ορθότητα)
  - **Sensitivity** = TP/(TP+FN) (Ευαισθησία)
  - **Specificity** = TN/(TN+FP) (Εξειδίκευση).
  - **Precision** = TP/(TP+FP) (Ακρίβεια)

  Εναλλακτικά της «ορθότητας» χρησιμοποιείται ο:

  Error rate = (FN+FP)/(TP+FN+FP+TN) (λόγος λάθους)

Σημαντικότερη η «ορθότητα», αφού εξασφαλίσουμε ότι «ευαισθησία» και «εξειδίκευση» είναι ισορροπημένες.

ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ-ΠΡΟΒΛΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ

- Οι παραπάνω μετρικές χρησιμοποιούνται ως έχουν σε περιπτώσεις δυαδικής εξόδου (ΝΑΙ-ΌΧΙ, ΑΛΗΘΗΣ-ΨΕΥΔΗΣ).
- Σε περιπτώσεις εξόδων πολλαπλών κλάσεων, η μέτρηση TN και οι μετρικές που την χρησιμοποιούν (ορθότητα, εξειδίκευση) χάνουν την αξιοπιστία τους. Τότε χρησιμοποιούνται οι μετρικές:
  - **Recall** = TP/(TP+FN) (Ανάκληση)
  - **Precision** = TP/(TP+FP) (Ακρίβεια)
  - **F_measure** = (2*precision*recall)/(precision+recall)

Για κάθε κλάση υπολογίζονται οι μετρικές και στη συνέχεια κάποιος Μ.Ο., ενδεχομένως με χρήση βαρών.

ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ-ΠΡΟΒΛΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ

- Για παράδειγμα αν έχουμε m κλάσεις και το σύνολο ελέγχου έχει k στιγμιότυπα με ki τα στιγμιότυπα της κλάσης i τότε μπορούμε να υπολογίσουμε τον μέσο όρο της Ανάκλησης ως εξής:

$$Weight\_Avg\_Recall = \sum_{i=1}^{m} \frac{k_i}{k} \times Recall_i$$

ΒΙΒΛΙΟΓΡΑΦΙΑ

- J. Han and M. Kamber, Data Mining-Cocepts and Technics, Morgan Kaufmann (Elsevier), 2nd Edition, 2006 (κεφ. 2 και 6).
- P.-N. Tan, M. Steibach, V. Kumar, Introduction to Data Mining, Addisson-Wesley, 2006 (κεφ. 4: www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf)
- I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157-1182.
- S. Kotsiantis, D. Kanellopoulos, Discretization Techniques: A recent survey, GESTS International Journal of Computer Science and Engineering, 32(1), 2006, 47-58.
- WEKA: www.cs.Waikato.ac.nz/ml/weka
- CLIPS: clipsrules.sourceforge.net