ΔΗΜΙΟΥΡΓΙΑ ΒΑΣΕΩΝ ΚΑΝΟΝΩΝ ΑΠΟ ΔΕΔΟΜΕΝΑ
Μέρος Α: Προετοιμασία συνόλου δεδομένων

Διδάσκων:

Ι. ΧΑΤΖΗΛΥΓΕΡΟΥΔΗΣ

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

ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

- Το σύνολο δεδομένων D αναφέρεται στο γνωστικό πεδίο του προβλήματος που θέλουμε να επιλύσουμε.
- Αποτελείται από ένα (μεγάλο) αριθμό στοιχείων:

$$D = \{t_1, t_2, \ldots, t_n\}$$

που τα ονομάζουμε *παραδείγματα*.

- Κάθε παράδειγμα t_i αποτελείται από p τιμές, που αντιστοιχούν σε p *χαρακτηριστικά* (ή παραμέτρους):

$$t_i = <t_{i1}, t_{i2}, \ldots, t_{ip}>$$

ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

- Τα χαρακτηριστικά-παράμετροι σχετίζονται με την επίλυση του προβλήματος. Το τελευταίο χαρακτηριστικό συνήθως είναι το *χαρακτηριστικό-στόχος*.
- Τα χαρακτηριστικά παίρνουν τιμές διαφόρων τύπων: πραγματικές-ακέραιες, συνεχείς-διακριτές κλπ. Οι τιμές του χαρακτηριστικού-στόχου είναι συνήθως διακριτές, αποτελούν όλες τις δυνατές απαντήσεις στο πρόβλημα και συνιστούν *κλάσεις*.
- Επομένως, το σύνολο δεδομένων περιέχει αντιστοιχίες μεταξύ συνδυασμών τιμών των χαρακτηριστικών και των κλάσεων.

ΠΑΡΑΔΕΙΓΜΑ

Χαρακτηριστικά

Χαρακτηριστικό-Στόχος

No	Outlook	Temp.	Humid.	Wind	**PlayTennis**
1	Sunny	30°C	High	Weak	No
2	Sunny	29°C	High	Strong	No
3	Overcast	27°C	High	Weak	Yes
4	Rain	20°C	High	Weak	Yes
5	Rain	13°C	Normal	Weak	Yes
6	Rain	15°C	Normal	Strong	No
7	Overcast	12°C	Normal	Strong	Yes
8	Sunny	18°C	High	Weak	No
9	Sunny	10°C	Normal	Weak	Yes
10	Rain	22°C	Normal	Weak	Yes
...	...	...	...	...	...

ΠΟΙΟ ΕΙΝΑΙ ΤΟ ΖΗΤΟΥΜΕΝΟ;

- Η δημιουργία ενός συστήματος βασισμένου σε κανόνες που θα επιλύει ένα πρόβλημα, που αναφέρεται σ'ένα (σχετικά στενό) γνωστικό πεδίο (π.χ. το πρόβλημα της διάγνωσης ασθενειών των οστών).
- Οι κανόνες εξάγονται από ένα σύνολο πραγματικών δεδομένων σχετικών με το πρόβλημα.

ΔΙΑΔΙΚΑΣΙΑ ΔΗΜΙΟΥΡΓΙΑΣ ΒΑΣΗΣ ΚΑΝΟΝΩΝ

1. Προεπεξεργασία συνόλου δεδομένων.
2. Εφαρμογή μεθόδου εξαγωγής κανόνων.
3. Έλεγχος ακρίβειας κανόνων και πιθανές διορθώσεις.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

- ✓ Επιλογή Χαρακτηριστικών (Feature Selection)
- ✓ Διακριτοποίηση (Discretization)
- ✓ Διαχείριση Ελλιπών Τιμών (Missing Values Handling)

ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

- Ο προσδιορισμός των πιο σημαντικών χαρακτηριστικών για την δημιουργία του συνόλου δεδομένων για την εξαγωγή ενός συνόλου κανόνων παίζει καθοριστικό ρόλο στην απόδοση του παραγόμενου μοντέλου ταξινόμησης, ιδιαίτερα σε προβλήματα με πολύ μεγάλο αριθμό χαρακτηριστικών.

- Η επιλογή χαρακτηριστικών αναφέρεται στον εντοπισμό και στην απομάκρυνση άσχετων (irrelevant) και πλεοναζόντων (redundant) χαρακτηριστικών σε σχέση με το χαρακτηριστικό-στόχο.

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (ενδεικτικά)

- **Χρήση Μεθόδου Φίλτρου (Filter)**
  - ✓ Γίνεται κατάταξη των χαρακτηριστικών με βάση την ικανότητά τους να διαχωρίζουν τις κλάσεις του χαρακτηριστικού-στόχου. Η ικανότητα αυτή εκτιμάται με βάση κάποια μετρική όπως η συσχέτιση (correlation), η εντροπία (entropy) κ.ά.
  - ✓ Υπολογιστικά αποδοτική.
- **Χρήση Μεθόδου Περιτυλίγματος (Wrapper)**
  - ✓ Εκτιμάται η διαχωριστική ικανότητα υποσυνόλων χαρακτηριστικών με βάση κάποιο αλγόριθμο ταξινόμησης.
  - ✓ Υπολογιστικά απαιτητική.

ΔΙΑΚΡΙΤΟΠΟΙΗΣΗ

- Αρκετοί αλγόριθμοι εξαγωγής κανόνων απαιτούν τα χαρακτηριστικά να έχουν διακριτές τιμές. Το σύνολο δεδομένων όμως συνήθως περιέχει και χαρακτηριστικά με συνεχείς τιμές.
- Η διακριτοποίηση συνίσταται στη μετατροπή των τιμών ενός χαρακτηριστικού συνεχούς τιμής σε διακριτές τιμές.
- Κάποια ορολογία:
  - **Σημείο διάσπασης (split point)**: χωρίζει μια συνεχή περιοχή σε δύο (διακριτά) διαστήματα
  - **Τάξη διάσπασης (arity):** ο αριθμός των ζητούμενων διακριτών τιμών-περιοχών ενός συνεχούς χαρακτηριστικού

ΔΙΑΚΡΙΤΟΠΟΙΗΣΗ

- Μεγαλύτερη τάξη διάσπασης σημαίνει μεγαλύτερη ακρίβεια αναπαράστασης. Υπάρχει μια διελκυστίνδα μεταξύ τάξης διάσπασης (έστω k) και πολυπλοκότητας αναπαράστασης: όσο μεγαλύτερο το k τόσο μεγαλύτερη η ακρίβεια αναπαράστασης, τόσο μεγαλύτερη η πολυπλοκότητα αναπαράστασης, δηλ. τόσο μικρότερη η κατανοησιμότητά της και αντίστροφα.

ΓΕΝΙΚΗ ΔΙΑΔΙΚΑΣΙΑ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗΣ

1. Διάταξη των τιμών του προς διακριτοποίηση συνεχούς χαρακτηριστικού
2. Αξιολόγηση υποψήφιου σημείου διάσπασης ή γειτονικών περιοχών για συγχώνευση.
3. Διάσπαση περιοχής ή συγχώνευση περιοχών συνεχών τιμών με βάση κάποιο κριτήριο.
4. Τερματισμός με βάση το κριτήριο τερματισμού, αλλιώς πήγαινε στο 2.

(Παραλλαγή από Kotsiantis and Kanellopoulos, 2006)

ΜΕΘΟΔΟΙ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗΣ

- Δύο βασικοί μέθοδοι προσδιορισμού των (υποψήφιων) σημείων διακριτοποίησης είναι:
  - ✓ top-down: ξεκινά με ένα κενό σύνολο σημείων και διασπά διαστήματα.
  - ✓ bottom-up: ξεκινά με ένα σύνολο που έχει σαν σημεία όλες τις τιμές και συγχωνεύει διαστήματα.
- Στην προσπάθεια διακριτοποίησης απαιτείται ένας συμβιβασμός μεταξύ ποιότητας της πληροφορίας (ομοιογενή διαστήματα σε σχέση με το χαρακτηριστικό-στόχο) και στατιστικής ποιότητας (ικανό μέγεθος δείγματος παραδειγμάτων σε κάθε διάστημα για εξασφάλιση γενίκευσης).

ΜΕΘΟΔΟΙ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗΣ

- Δύο βασικά κριτήρια διάσπασης ή συγχώνευσης διαστημάτων είναι:
  - ✓ **Στατιστική ομοιότητα:** Π.χ. ένα διάστημα διασπάται αν ένα σημείο το χωρίζει σε δύο υποδιαστήματα που στατιστικά διαφέρουν σημαντικά ως προς την σχέση των τιμών με τις κλάσεις (υπάρχουν διάφορα κριτήρια-π.χ. χ^2). Αντίστοιχα, δύο διαστήματα συγχωνεύονται αν είναι στατιστικά όμοια.
  - ✓ **Ομοιογένεια πληροφορίας:** Η εντροπία της πληροφορίας κλάσης των διαστημάτων χρησιμοποιείται ως κριτήριο για τη συγχώνευσή τους ή τη διάσπασή τους.

ΜΕΘΟΔΟΙ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗΣ

Μια συνήθης κατηγοριοποίηση των μεθόδων διακριτοποίησης:

- Μη επιβλεπόμενες (Unsupervised)
  Δεν χρησιμοποιούν πληροφορία για τις κλάσεις.
  - Διακριτοποίηση ίσου εύρους (equal-width discretization)
  - Διακριτοποίηση ίσης συχνότητας (equal-frequency discretization)
- Επιβλεπόμενες (Supervised)
  Χρησιμοποιούν πληροφορία για τις κλάσεις
  - Μέθοδοι που χρησιμοποιούν στατιστική ομοιότητα ή ομοιογένεια πληροφορίας.

ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΟΙ ΜΕΘΟΔΟΙ ΔΙΑΚΡΙΤΟΠΟΙΗΣΗΣ

- Διακριτοποίηση ίσου εύρους
  1. Προσδιορισμός ελάχιστης και μέγιστης τιμής του χαρακτηριστικού
  2. Διαχωρισμός του προκύπτοντος διαστήματος σε τόσα ίσα υποδιαστήματα όσα ορίζονται από τον χρήστη
- Διακριτοποίηση ίσης συχνότητας
  1. Προσδιορισμός ελάχιστης και μέγιστης τιμής του χαρακτηριστικού
  2. Ταξινόμηση των τιμών του χαρακτηριστικού κατά αύξουσα σειρά
  3. Διαίρεση του προκύπτοντος διαστήματος σε τόσα διαστήματα όσα ορίζονται από τον χρήστη, ώστε κάθε διάστημα να περιέχει τον ίδιο αριθμό διατεταγμένων τιμών (πιθανόν και διπλότυπων)

ΔΙΑΧΕΙΡΙΣΗ ΕΛΛΙΠΩΝ ΤΙΜΩΝ

- Μπορεί να λείπουν τιμές σε ορισμένα παραδείγματα για κάποιο ή κάποια χαρακτηριστικά. Συνήθως εκτιμούμε τις τιμές που λείπουν με βάση τις υπάρχουσες σε άλλα παραδείγματα.
- Στρατηγικές
  - ✓ Διαγραφή των παραδειγμάτων με ελλιπείς τιμές.
  - ✓ Δίνουμε την τιμή που είναι πιο κοινή σε όλα τα παραδείγματα του συνόλου.
  - ✓ Δίνουμε την τιμή που είναι πιο κοινή στα παραδείγματα που έχουν τιμή χαρακτηριστικού-στόχου ίδια με το υπό εξέταση παράδειγμα.
  - ✓ Δίνουμε τη μέση τιμή των τιμών στα υπόλοιπα παραδείγματα του συνόλου.
  - ✓ Χρήση regression analysis για πρόβλεψη των ελλιπών τιμών.