2ο Σύνολο Ασκήσεων

Καταληκτική Ημερομηνία Παράδοσης: 29 Μαΐου 2009, στις 17:00, στο Εργαστήριο Κατανεμημένων (B15)
Ενότητα: Κανόνες Συσχέτισης, Ταξινόμηση

Οι ασκήσεις με χαρακτηρισμό **Α** είναι ατομικές, ενώ οι ασκήσεις με χαρακτηρισμό **Δ** μπορεί να γίνουν σε ομάδες έως 2 ατόμων. Οι ασκήσεις με **(*)** είναι προαιρετικές με την παρακάτω έννοια: μπορείτε να τις παραδώσετε αντί τελικής εξέτασης. Θα υπάρχουν αντίστοιχες και στα άλλα σύνολα. Για αυτούς που θα επιλέξουν να ολοκληρώσουν όλες τις ασκήσεις (δηλαδή και τις προαιρετικές), ο τελικός βαθμός τους θα προκύψει από τον βαθμό τους στα σύνολα ασκήσεων. Για τους υπόλοιπους, οι ασκήσεις θα συμμετέχουν με ποσοστό 50% στον τελικό τους βαθμό.

Για τους αλγόριθμους, μπορείτε να χρησιμοποιήστε τα εργαλεία WEKA, MATLAB, δικό σας κώδικα, ή κάποιο άλλο εργαλείο. Πληροφορίες για τα εργαλεία WEKA και MATLAB υπάρχουν στην ιστοσελίδα του μαθήματος.

Ποσοστό επί του τελικού βαθμού: **45 %** για όσους ασχοληθούν με τις ασκήσεις (*)
23 % για τους υπόλοιπους

Άσκηση 1 [Α, (*)]

Θεωρείστε τον σύνολο δοσοληψιών του Πίνακα 1 και ελάχιστη υποστήριξη 3 (minsup = 30%). Στο Σχήμα 1 δίνετε το πλέγμα στοιχειοσυνόλων.

Πίνακας 1. Βάση Δοσοληψιών για την Άσκηση 1

TID	**Στοιχεία**
T10	{*a*, *b*, *d*, e}
T20	{*b*, *c*, *d*}
T30	{*a*, *b*, *d*, e}
T40	{*a*, c, *d*, *e*}
T50	{*b*, c, *d*, *e*}
T60	{*b*, *d*, *e*}
T70	{*c*, *d*}
T80	{*a*, *b*, *c*}
T90	{*a*, *d*, *e*}
T100	{*b*, *d*}

(α) Εφαρμόστε τον αλγόριθμο a-priori. Χαρακτηρίστε κάθε κόμβο στο Σχήμα 1 με ένα από τα παρακάτω γράμματα:

- Με το γράμμα **Ο,** αν δε θεωρήθηκε υποψήφιο στοιχειοσύνολο από τον αλγόριθμο a-priori. Ένα στοιχείο δεν θεωρείτε υποψήφιο για δύο λόγους: (1) δε δημιουργήθηκε κατά τη φάση της δημιουργίας υποψηφίων στοιχειοσυνόλων ή (2) δημιουργήθηκε αλλά ψαλιδίστηκε (pruned) επειδή διαπιστώθηκε ότι κάποιο από τα υποσύνολα του δεν ήταν συχνό
- Με το γράμμα **Σ,** αν θεωρήθηκε συχνό.
- Με το γράμμα **Ν,** αν βρέθηκε μη συχνό μετά από υπολογισμό της υποστήριξης του χρησιμοποιώντας τις δοσοληψίες (δηλαδή, δημιουργήθηκε ως υποψήφιο, δε ψαλιδίστηκε, υπολογίσαμε από τις δοσοληψίες την υποστήριξη του και βρέθηκε μικρότερη της ελάχιστης).

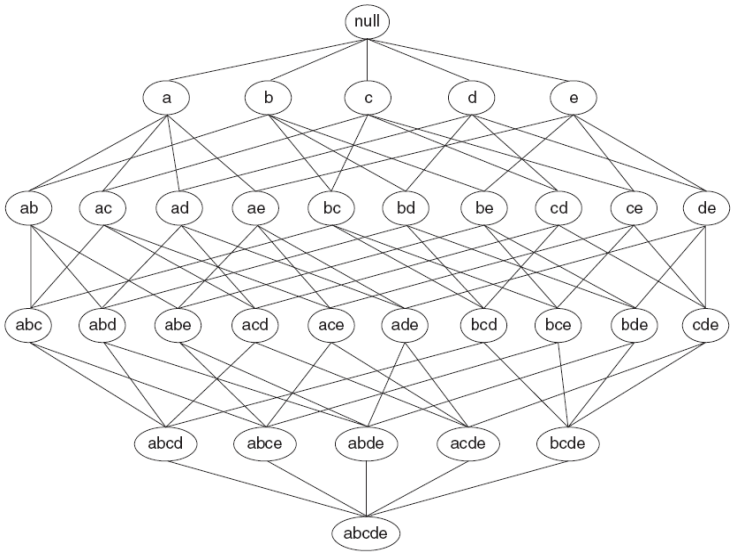
(β) Δώστε τα συχνά στοιχειοσύνολα (δηλαδή αυτά που έχουν χαρακτηριστεί με το γράμμα Σ) *με τη σειρά* που τα υπολογίζει ο a-priori.

(γ) Εφαρμόστε τον αλγόριθμο FP-growth. Θεωρείστε τα στοιχεία ταξινομημένα με βάση τη συχνότητα εμφάνισής τους. Δώστε: (i) το αρχικό FP-δέντρο, (ii) τα προθεματικά δέντρα για το πρώτο βήμα (δηλαδή, για την πρώτη κατάληξη που εξετάζει ο αλγόριθμος) και (iii) όλα τα συχνά στοιχειοσύνολα *με τη σειρά* που αυτά παράγονται.

(δ) Από τα συχνά στοιχειοσύνολα που έχετε υπολογίσει, δώστε τους κανόνες με ελάχιστη εμπιστοσύνη 80%.

(ε) Δώστε τον πίνακα συνάφειας για τα σύνολα {*a*} και {*d*} και σχολιάστε την στατιστική τους ανεξαρτησία καθώς και την εμπιστοσύνη, το lift και interest των σχετικών κανόνων.

(στ) Σημειώστε στο πλέγμα του Σχήματος 1 ποια από τα συχνά στοιχειοσύνολα είναι (i) maximal και (ii) ποια είναι closed (κλειστά).
(ζ) Είναι δυνατόν να βελτιώσουμε τον αλγόριθμο FP-growth για την περίπτωση που θέλουμε να υπολογίσουμε maximal ή κλειστά στοιχειοσύνολα; Αν όχι εξηγείστε γιατί, αν ναι περιγράψτε πως.



Σχήμα 1: Πλέγμα στοιχειοσυνόλων για την Άσκηση 1

Άσκηση 2 [Α, ()]

(α) Έστω ότι η υποστήριξη των κανόνων A → B και B → C είναι μεγαλύτερη από κάποια ελάχιστη τιμή *minconf.* Είναι δυνατόν ο κανόνας A → C να έχει εμπιστοσύνη μικρότερη του *minconf;*
(β) Δώστε ένα παράδειγμα που δύο κανόνες έχουν την ίδια εμπιστοσύνη και διαφορετική υποστήριξη και ένα παράδειγμα που έχουν την ίδια υποστήριξη και διαφορετική εμπιστοσύνη Σχολιάστε τι αυτό σημαίνει.

Άσκηση 3 [Δ, ()]

Σκοπός της άσκησης είναι η εξοικείωσή σας με ένα εργαλείο για εξόρυξη κανόνων συσχέτισης. Μπορείτε να χρησιμοποιείστε το εργαλείο WEKA που υλοποιεί τον αλγόριθμο apriori.
Το εργαλείο WEKA υποστηρίζει κανόνες μόνον σε ordinal γνωρίσματα, για αυτό, αριθμητικά δεδομένα θα χρειαστούν προ-επεξεργασία (χρησιμοποιείστε ένα κατάλληλο φίλτρο από αυτά που είναι διαθέσιμα στο `Filter`).
(α) Εξηγείστε τι σημαίνει κάθε παράμετρος εισόδου του αλγορίθμου. (πχ στην περίπτωση της WEKA, οι παράμετροι `MetricType`, `minMetric` κοκ)
(β) Τρέξτε τον αλγόριθμο κανόνων συσχέτισης στα σύνολα δεδομένων της Άσκησης 1 (Πίνακας 1), και στα σύνολα weather-nominal και weather (είναι τα ίδια δεδομένα όπως τα weather-nominal αλλά με αριθμητικές τιμές, για να τα χρησιμοποιείστε πρέπει να τα φιλτράρετε, χρησιμοποιείστε ένα κατάλληλο φίλτρο που να *μην* παράγει όμως ακριβώς τα ίδια δεδομένα με το weather-nominal). Τα σύνολα δεδομένων για την Άσκηση 1 θα πρέπει να τα γράψετε σε αρχείο arff με τη δομή που έχουν τα σύνολα weather-nominal και weather που θα τα βρείτε στη σελίδα του μαθήματος.
(i) Για κάθε μία περίπτωση, εξηγείστε την επιλογή των τιμών που δώσατε στις παραμέτρους εισόδου. Ειδικά για το σύνολο της Άσκησης 1 θέστε τις παραμέτρου ώστε να πάρετε τα συχνά στοιχειοσύνολα και τους κανόνες που υπολογίσατε στα αντίστοιχα ερωτήματα της Άσκησης 1.
(ii) Διαλέξτε έναν από τους κανόνες για το σύνολο weather-nominal. Εκτιμήστε το ενδιαφέρον τους με βάση κάποιες από τις μετρικές που συζητήσαμε στο μάθημα και τις σχετικές μετρικές του εργαλείου (πχ στην περίπτωση της WEKA, `lift`, `convinction`, κοκ)

Άσκηση 4 [Α, (*)]

Δείξτε ότι η εντροπία ενός κόμβου δεν αυξάνει αν διασπαστεί σε μικρότερους κόμβους.

Άσκηση 5 [Α, (*)]
Θεωρείστε τα δεδομένα εκπαίδευσης του Πίνακα 1.

(α) Δείξτε ποιος από τους δύο διαχωρισμός είναι καλύτερος, αυτός με βάση το Α1 ή αυτός με βάση το Α2 αν χρησιμοποιήσουμε:
(i) ευρετήριο Gini
(ii) λάθος ταξινόμησης
(iii) κέρδος πληροφορίας
(β) Κατασκευάστε τον Πίνακα Confusion για το δέντρο απόφασης που προκύπτει από την επιλογή του Ερωτήματος (α) για την περίπτωση (iii). Υπολογίστε την πιστότητα (accuracy), ανάκληση (recall) και ακρίβεια (precision).
(γ) Για το γνώρισμα Α3 που είναι συνεχές, υπολογίστε το κέρδος πληροφορίας σε κάθε σημείο διαχωρισμού.

Πίνακας 2. Σύνολο Δεδομένων για την Άσκηση 5

ID	A1	A2	A3	Κλάση
1	T	F	7.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	9.0	-
8	T	T	2.0	+
9	F	T	5.0	-

Άσκηση 6 [Δ, ()]
Σκοπός της άσκησης είναι η εξοικείωσή σας με ένα εργαλείο για ταξινόμηση με χρήση δέντρων απόφασης. Αν χρησιμοποιείστε WEKA, χρησιμοποιείστε το J48 (υλοποιεί τον C4.5).
(α) Τρέξτε τον αλγόριθμο για τα δεδομένα weather θεωρώντας ως γνώρισμα «κλάση» το play. Χρησιμοποιείστε 10 cross-validation. Δώστε το δέντρο που προκύπτει. Εξηγείστε τις τιμές των παραμέτρων.
(β) Επαναλάβετε το (α) τώρα χρησιμοποιώντας (i) 66% των δεδομένων ως δεδομένα εκπαίδευσης και 33% ως δεδομένου ελέγχου και (ii) 33% των δεδομένων ως δεδομένα εκπαίδευσης και 66% ως δεδομένου ελέγχου. Εξηγείστε τις τιμές των παραμέτρων.
(γ) Συγκρίνετε τα δέντρα που προκύπτουν για το (α) και β(i) και β(ii) χρησιμοποιώντας κάποια από το μέτρα που μελετήσαμε στο μάθημα.
(δ) Τρέξτε τον αλγόριθμο για τα δεδομένα zoo θεωρώντας ως γνώρισμα «κλάση» το type. Χρησιμοποιείστε 10 cross-validation. Δώστε το δέντρο που προκύπτει. Εξηγείστε τις τιμές των παραμέτρων.