2ο Σύνολο Ασκήσεων
Ημερομηνία Παράδοσης: 18 Μαΐου 2011, στο μάθημα
Ενότητα: Κανόνες Συσχέτισης

Ποσοστό του τελικού βαθμού: 30 % του ως απαλλακτικές
15% αν δώσετε τελικό διαγώνισμα

Οι ασκήσεις είναι απαλλακτικές, με την έννοια ότι μπορεί να αναπληρώσουν το τελικό διαγώνισμα (δείτε και τη σελίδα του μαθήματος).

Άσκηση 1 [σε ομάδες έως 2 ατόμων] [35 μονάδες]
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 μόνο σε ordinal (κατηγορικά) 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 2 (Πίνακας 1), και στα σύνολα mushrooms, 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).

(i) Για κάθε ένα από τα σύνολα, εξηγείστε την επιλογή των τιμών που δώσατε στις παραμέτρους εισόδου. Ειδικά για το σύνολο της Άσκησης 2 θέστε τις παραμέτρου ώστε να πάρετε τα συχνά στοιχειοσύνολα και τους κανόνες που υπολογίσατε στα αντίστοιχα ερωτήματα της Άσκησης 2.

(ii) Διαλέξτε τρεις από τους κανόνες για το σύνολο weather-nominal. Κατατάξετε τους σε φθίνουσα διάταξη σύμφωνα με τα ακόλουθα κριτήρια: (1) υποστήριξη, (2) εμπιστοσύνη, (3) ενδιαφέρον (interest) και (4) IS.

Άσκηση 2 [ατομική] [45 μονάδες]
Θεωρείστε τις συναλλαγές του Πίνακα 1 και ελάχιστη υποστήριξη 3 (minsup = 30%). Στο Σχήμα 1 δίνετε το πλέγμα στοιχειοσυνόλων.

Πίνακας 1. Βάση Συναλλαγών για την Άσκηση 2

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

(α) Εφαρμόστε τον αλγόριθμο a-priori στις συναλλαγές του Πίνακα 1. Για τη δημιουργία των υποψήφιων συχνών στοιχειοσύνολων μεγέθους $k + 1$, $k > 1$, θεωρείστε τη μέθοδο που συνενώνει δύο συχνά στοιχειοσύνολα μεγέθους k. Χαρακτηρίστε κάθε κόμβο στο Σχήμα 1 με ένα από τα παρακάτω γράμματα:

- Με το γράμμα **Σ,** αν το αντίστοιχο στοιχειοσύνολο θεωρήθηκε συχνό.
- Με το γράμμα **Ο,** αν το αντίστοιχο στοιχειοσύνολο δε δημιουργήθηκε κατά τη φάση της δημιουργίας υποψηφίων στοιχειοσυνόλων,
- Με το γράμμα **Ψ,** αν το αντίστοιχο στοιχειοσύνολο δημιουργήθηκε κατά τη φάση της δημιουργίας υποψηφίων στοιχειοσυνόλων, αλλά ψαλιδίστηκε (pruned) χωρίς μέτρηση της υποστήριξής του (σημειώστε και γιατί που ψαλιδίστηκε).

- Με το γράμμα **N,** αν το αντίστοιχο στοιχειοσύνολο βρέθηκε μη συχνό μετά από μέτρηση της υποστήριξής του (δηλαδή, το στοιχειοσύνολο δημιουργήθηκε ως υποψήφιο, δε ψαλιδίστηκε, υπολογίστηκε η υποστήριξη του κοιτάζοντας τις συναλλαγές και η υποστήριξη αυτή βρέθηκε μικρότερη της ελάχιστης).

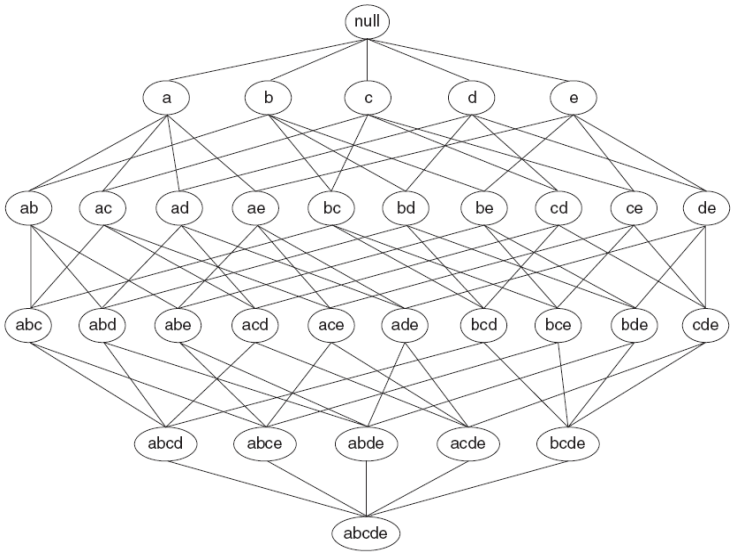
(β) Δώστε τα συχνά στοιχειοσύνολα (δηλαδή αυτά που έχουν χαρακτηριστεί με το γράμμα Σ) *με τη σειρά* που τα παράγει ο a-priori.

(γ) Εφαρμόστε τον αλγόριθμο FP-growth στις συναλλαγές του Πίνακα 1. Για τη δημιουργία του δέντρου των συναλλαγών, ταξινομείστε τα στοιχεία 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) Με βάση τα συχνά στοιχειοσύνολα που έχετε υπολογίσει, δώστε τους κανόνες με ελάχιστη εμπιστοσύνη 65%.

(ε) Κατασκευάστε τον πίνακα ενδεχομένων για τα {d} και {e} και υπολογίστε την εμπιστοσύνη, το lift, τη συσχέτιση και το ενδιαφέρον (interest) των δύο σχετικών κανόνων.

(στ) Σημειώστε στο πλέγμα του Σχήματος 1 ποια από τα συχνά στοιχειοσύνολα είναι (i) maximal (μέγιστα) και (ii) ποια είναι closed (κλειστά). Εξηγείστε την απάντησή σας.



Σχήμα 1: Πλέγμα στοιχειοσυνόλων για την Άσκηση 2

Άσκηση 3 [ατομική] [20 μονάδες]

Θέλετε να υπολογίσετε τα *σπάνια στοιχειοσύνολα*, όπου ένα στοιχειοσύνολο είναι σπάνιο, αν η υποστήριξη του είναι μικρότερη από μια τιμή υποστήριξης minsup.

(α) Δώστε μια apriori ιδιότητα για τα σπάνια στοιχειοσύνολα ανάλογη της apriori για τα συχνά στοιχειοσύνολα και περιγράψτε έναν αλγόριθμο για τον υπολογισμό των σπάνιων στοιχειοσυνόλων που να χρησιμοποιεί αυτήν την ιδιότητα.

(β) Ένας εναλλακτικός τρόπος υπολογισμού των σπάνιων στοιχειοσυνόλων θα ήταν να βρούμε τα συχνά (δηλαδή να πάρουμε αυτά με υποστήριξη τουλάχιστον ίση με minsup) και να πάρουμε το συμπλήρωμά τους. Συγκρίνετε αυτόν τον αλγόριθμο με αυτόν του ερωτήματος (α).

(γ) Δώστε αντίστοιχους ορισμούς για τα μέγιστα σπάνια στοιχειοσύνολα, και τα κλειστά σπάνια στοιχειοσύνολα. Για τα κλειστά σπάνια στοιχειοσύνολα εξηγείστε πως μπορείτε να τα χρησιμοποιείστε για τον υπολογισμό της υποστήριξης όλων των σπάνιων στοιχειοσυνόλων.