Ανάλυση συσχετίσεων

Ιωάννης Μαδεμλής

Ανάλυση συσχετίσεων

- Μία μεγάλη κατηγορία προβλημάτων εξόρυξης γνώσης είναι τα προβλήματα **ανάλυσης συσχετίσεων** (association analysis).
  - Ανακύπτουν συνήθως κατά την ανάλυση μεγάλων βάσεων δεδομένων οι οποίες περιέχουν εγγραφές *συναλλαγών* (π.χ., αγορές καταναλωτών σε πολυκαταστήματα).

- Συνήθως κάθε συναλλαγή χαρακτηρίζεται από έναν μοναδικό ακέραιο αύξοντα αριθμό και περιέχει ένα σύνολο αγορασμένων προϊόντων.

- Στόχος είναι η ανακάλυψη *συσχετίσεων* μεταξύ προϊόντων: η αγορά ενός συγκεκριμένου είδους Α συνοδεύεται συνήθως από την αγορά και ενός άλλου είδους Β.
  - **Κανόνες συσχέτισης**.

Ανάλυση συσχετίσεων

Ανάλυση συσχετίσεων

Ανάλυση συσχετίσεων

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Πηγή: Tan et al., *Introduction to Data Mining, 2006.*

Ανάλυση συσχετίσεων

- Δεδομένα της ανωτέρω μορφής μπορούν επίσης να αναπαρασταθούν με *ασύμμετρες δυαδικές μεταβλητές*:
  - Αν η i-οστή τιμή της j-οστής εγγραφής είναι 0, τότε η j-οστή συναλλαγή δεν περιλαμβάνει αγορά του i-οστού είδους/προϊόντος.
  - Αν είναι 1, τότε το περιλαμβάνει.

Ανάλυση συσχετίσεων

- Άρα, αν I είναι το πλήθος των δυνατών ειδών και T το ολικό πλήθος συναλλαγών/εγγραφών, τότε το σύνολο δεδομένων αναπαρίσταται ως ένας αραιός δυαδικός πίνακας $\mathbf{A} \in \{0,1\}^{T \times I}$.

Ανάλυση συσχετίσεων

- Κάθε δυνατό υποσύνολο των διαθέσιμων προϊόντων λέγεται **σύνολο ειδών** (item set).
- Κάθε σύνολο ειδών χαρακτηρίζεται από το πλήθος k των προϊόντων τα οποία περιέχει.
  - Αν η i-οστή γραμμή του πίνακα $\mathbf{A}$ έχει k άσσους και I-k μηδενικά, τότε εκφράζει μία συναλλαγή/σύνολο ειδών με πλήθος προϊόντων k.
- **Μετρητής υποστήριξης** $\sigma(\mathcal{P})$ ενός συγκεκριμένου συνόλου ειδών $\mathcal{P}$ με πλήθος προϊόντων k, είναι το πλήθος των συναλλαγών (διαφορετικών γραμμών του $\mathbf{A}$) οι οποίες περιέχουν και τα k προϊόντα του $\mathcal{P}$.
  - Μπορεί να περιέχουν και άλλα προϊόντα (εκτός $\mathcal{P}$).

Ανάλυση συσχετίσεων

- Παράδειγμα: στο παρακάτω σύνολο συναλλακτικών δεδομένων, το σύνολο ειδών $\mathcal{P}$ = {Beer, Diapers, Milk} (μεγέθους 3) έχει μετρητή υποστήριξης 2 ($\sigma(\mathcal{P}) = 2$).

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Πηγή: Tan et al., *Introduction to Data Mining, 2006.*

Ανάλυση συσχετίσεων

Ανάλυση συσχετίσεων

- Κανόνας συσχέτισης είναι ένας κανόνας συνεπαγωγής της μορφής $X \rightarrow Y$, όπου X και Y ξένα μεταξύ τους σύνολα ειδών.
- Δύο μετρικές ορίζονται για την αξιολόγηση ενός κανόνα συσχέτισης:
  - **Υποστήριξη** $s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{T}$.
  - **Εμπιστοσύνη** $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$.
- Η υποστήριξη ποσοτικοποιεί τη συχνότητα επαλήθευσης του κανόνα στο ολικό σύνολο δεδομένων $\mathbf{A}$.
  - Κανόνας μικρής υποστήριξης μπορεί να προκύπτει από απλή τύχη και δεν έχει ενδιαφέρον.
- Η εμπιστοσύνη ποσοτικοποιεί τη συχνότητα επαλήθευσης του κανόνα στο σύνολο των εμφανίσεων του X μες στο $\mathbf{A}$.
  - Κανόνας μικρής εμπιστοσύνης δεν είναι αξιόπιστος.
  - Μεγαλύτερη εμπιστοσύνη σημαίνει μεγαλύτερη πιθανότητα εμφάνισης του Y σε μία συναλλαγή, δοθέντος ότι εμφανίζεται το X.

Ανάλυση συσχετίσεων

Ανάλυση συσχετίσεων

- Προσοχή: Οι κανόνες συσχέτισης υποδηλώνουν *τάση συνεμφάνισης και όχι αιτιότητα*.
  - ΝΑΙ: Τα X και Y τείνουν να εμφανίζονται από κοινού στην ίδια συναλλαγή.
  - ΟΧΙ: Η εμφάνιση του X προκαλεί την εμφάνιση και του Y.

- Υπάρχει ένας τετριμμένος τρόπος εύρεσης όλων των ισχυόντων κανόνων συσχέτισης σε ένα σύνολο δεδομένων, με τιμές υποστήριξης και εμπιστοσύνης μεγαλύτερες από κάποιο επιθυμητό κατώφλι.
  - Εξαντλητική απαρίθμηση όλων των δυνατών κανόνων.
  - Υπολογισμός των αντίστοιχων τιμών υποστήριξης και εμπιστοσύνης τους.
  - Απαλοιφή όσων κανόνων έχουν ανεπαρκή τιμή υποστήριξης ή εμπιστοσύνης.

Ανάλυση συσχετίσεων

Ανάλυση συσχετίσεων

- Ο τετριμμένος αλγόριθμος έχει τεράστια υπολογιστική πολυπλοκότητα.

- Οι πρακτικοί αλγόριθμοι ανάλυσης συσχετίσεων προσπαθούν να μειώσουν αυτή την πολυπλοκότητα.

- Συνήθως διασπούν το πρόβλημα σε δύο διαφορετικά υποπροβλήματα τα οποία επιλύονται διαδοχικά:
  - **Παραγωγή συχνών συνόλων ειδών**.
    - Εύρεση όλων των συνόλων ειδών με επαρκή μετρητή υποστήριξης (*συχνά σύνολα ειδών*).
  - **Παραγωγή κανόνων**.
    - Εξαγωγή όλων των κανόνων με επαρκή τιμή εμπιστοσύνης από τα συχνά σύνολα ειδών (*ισχυροί κανόνες*).

Παραγωγή συχνών συνόλων ειδών

Αλγόριθμος Apriori

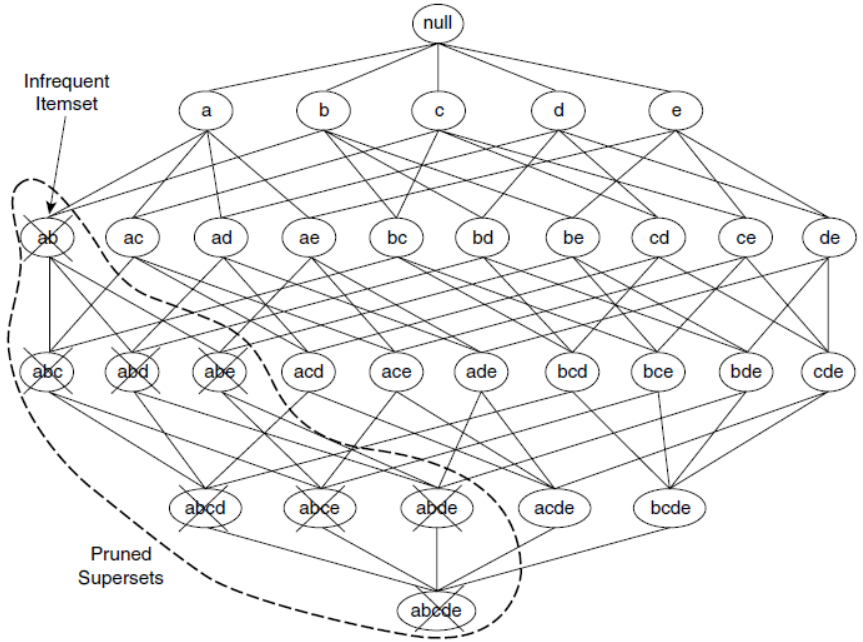
- Αφελής προσέγγιση στην παραγωγή συχνών συνόλων ειδών:
  - Διαδοχική απαρίθμηση όλων των δυνατών υποσυνόλων προϊόντων και υπολογισμός του πλήθους εμφανίσεών τους στο σύνολο δεδομένων (μετρητές υποστήριξης).

- Η **αρχή Apriori** μπορεί να αξιοποιηθεί για να απορρίψουμε εκ των προτέρων ένα σημαντικό ποσοστό των δυνατών συνόλων ειδών και να μην υπολογίσουμε καθόλου τους μετρητές υποστήριξής τους.

Παραγωγή συχνών συνόλων ειδών

- Αρχή Apriori: αν ένα σύνολο ειδών είναι *συχνό*, τότε όλα τα υποσύνολά του επίσης είναι *συχνά*.
- Άρα, αν ένα σύνολο ειδών είναι *σπάνιο*, τότε όλα τα υπερσύνολά του επίσης είναι *σπάνια*.
  - Μόλις εντοπίσουμε ένα σύνολο ειδών με μικρό μετρητή υποστήριξης, κατευθείαν κλαδεύουμε/αγνοούμε όλα τα υπερσύνολά του.

Αλγόριθμος Apriori



Δένδρο συνόλων ειδών σε ένα σύνολο συναλλαγών με τα προϊόντα a, b, c, d και e.

Πηγή: Tan et al., *Introduction to Data Mining, 2006.*

Αλγόριθμος Apriori

Παραγωγή συχνών συνόλων ειδών

- Ο αλγόριθμος Apriori (1994) στηρίζεται σε αυτή την ιδέα για να κλαδέψει επανειλημμένα ένα δένδρο απαρίθμησης συνόλων ειδών.
  - Κάθε κόμβος του δένδρου περιέχει ένα από όλα τα δυνατά σύνολα ειδών.
  - Το δένδρο διατρέχεται κατά πλάτος (BFS) εκκινώντας από μεμονωμένα προϊόντα, δηλαδή από σύνολα ειδών μεγέθους 1.
  - Το $(i+1)$-οστό επίπεδο του δένδρου περιέχει σύνολα ειδών μεγέθους $i+1$.
  - Όταν ο αλγόριθμος επισκέπτεται έναν κόμβο υπολογίζει τον μετρητή υποστήριξής του στο δοθέν σύνολο δεδομένων.
  - Αν ο τρέχων μετρητής υποστήριξης είναι μικρότερος από ένα κατώφλι-υπερπαράμετρο, τότε κατευθείαν κλαδεύεται όλο το κλαδί του δένδρου το οποίο εκκινεί από τον τρέχοντα κόμβο.

Αλγόριθμος Apriori

Παραγωγή συχνών συνόλων ειδών

- Ο αλγόριθμος επιφέρει σημαντική επιτάχυνση σε σχέση με την αφελή λύση.
  - Κατά τη διάσχιση και αξιολόγηση των κόμβων του επιπέδου $i+1$ δεν χρειάζεται να ληφθούν υπόψη τα σύνολα ειδών με υποσύνολα τα οποία είχαν προηγουμένως αναγνωριστεί ως σπάνια στο επίπεδο i.
  - Τα σπάνια σύνολα ειδών αγνοούνται και δεν χρησιμοποιούνται για την παραγωγή νέων (μεγαλύτερου μεγέθους) συνόλων ειδών προς διάσχιση και αξιολόγηση.

Αλγόριθμος Apriori

Παραγωγή συχνών συνόλων ειδών

- Πώς γίνεται η παραγωγή των υποψήφιων συνόλων ειδών μεγέθους $i+1$;
  - Τα **συχνά** σύνολα ειδών μεγέθους i επεκτείνονται με την προσθήκη των **συχνών** μεμονωμένων προϊόντων (μεγέθους 1).
- Όμως κάποιο από τα υποψήφια σύνολα ειδών μεγέθους $i+1$ τα οποία παράγονται έτσι ενδέχεται να περιέχει σπάνια υποσύνολα μεγέθους i.
  - Άρα αυτό το υποψήφιο σύνολο ειδών είναι **σπάνιο**, ως υπερσύνολο σπάνιου υποσυνόλου (αρχή Apriori).
  - **Θα πρέπει να υπολογιστεί ο μετρητής υποστήριξής του** για να αξιολογηθεί ως μικρής υποστήριξης.
    - Τότε θα κλαδευτεί όλο το κλαδί του δένδρου το οποίο εκκινεί από αυτό.
  - Ειδικές δομές δεδομένων χρησιμοποιούνται για γρήγορο υπολογισμό των μετρητών υποστήριξης.

Παραγωγή συχνών συνόλων ειδών

- Παράδειγμα: έστω τα εξής συχνά σύνολα ειδών μεγέθους 2 και μεγέθους 1:

{Beer, Diapers}
{Bread, Diapers}
{Bread, Milk}
{Diapers, Milk}

Beer
Bread
Diapers
Milk

Αλγόριθμος Apriori

- Κατά την παραγωγή των συνόλων ειδών μεγέθους 3, θα παραχθεί το {Beer, Diapers, Milk}.
  - Όμως αυτό περιέχει το σπάνιο υποσύνολο {Beer, Milk}, άρα με βάση την αρχή Apriori είναι και το ίδιο σπάνιο.
  - Θα σημειωθεί ως σπάνιο με υπολογισμό του μετρητή υποστήριξής του και σύγκρισή του με το δοθέν κατώφλι υποστήριξης.

Αλγόριθμος Apriori

Παραγωγή κανόνων

- Αφού έχουν εξαχθεί όλα τα συχνά σύνολα ειδών και οι αντίστοιχοι μετρητές υποστήριξης, πρέπει να εξαχθούν από αυτά οι κανόνες με τιμή εμπιστοσύνης η οποία υπερβαίνει κάποιο δοθέν κατώφλι (υπερπαράμετρος).
  - Ο υπολογισμός της εμπιστοσύνης απαιτεί μόνο γνώση των ήδη αποθηκευμένων μετρητών υποστήριξης.
- Παράδειγμα: από το ίδιο συχνό σύνολο ειδών {Bread, Diapers, Milk} (μεγέθους 3), μπορούν να προκύψουν πολλαπλοί διαφορετικοί κανόνες.
  - Π.χ. {Bread, Diapers} → {Milk}, {Diapers, Milk} → {Bread}, , {Diapers} → {Bread, Milk}, κλπ.
  - Όσο μεγαλύτερο μέγεθος έχει το υποσύνολο ειδών του δεύτερου σκέλους, τόσο μικρότερο μέγεθος έχει το υποσύνολο του πρώτου σκέλους, και αντιστρόφως.

Παραγωγή κανόνων

Αλγόριθμος Apriori

- Αφελής λύση:
  - Εξαντλητική απαρίθμηση όλων των δυνατών κανόνων από κάθε συχνό σύνολο ειδών.
  - Για κάθε δυνατό κανόνα, υπολογισμός της τιμής εμπιστοσύνης του.
  - Εξάλειψη κάθε υποψήφιου κανόνα με τιμή εμπιστοσύνης μικρότερη του δοθέντος κατωφλίου.
- Στο τέλος θα απομείνουν μόνον οι ζητούμενοι ισχυροί κανόνες.
- Εναλλακτικά, μπορούμε να εφαρμόσουμε μία παραλλαγή της αρχής Apriori και να κατασκευάσουμε ένα δένδρο κανόνων.
  - Ξεχωριστά για κάθε συχνό σύνολο ειδών.

Παραγωγή κανόνων

Αλγόριθμος Apriori

- Αλγόριθμος Apriori:
  - Παράγεται/διασχίζεται ξανά μία δενδρική δομή, εκκινώντας από τα φύλλα.
  - Σε κάθε επίπεδο αυξάνεται κατά ένα το μέγεθος του υποσυνόλου ειδών στο δεύτερο σκέλος της συνεπαγωγής.
  - Οι κόμβοι του επιπέδου $i+1$ παράγονται συνενώνοντας τα δεύτερα σκέλη των κόμβων/κανόνων **υψηλής εμπιστοσύνης** του προηγούμενου επιπέδου i.
  - Οι κόμβοι μικρής εμπιστοσύνης του επιπέδου i αγνοούνται και κλαδεύεται εκ των προτέρων το τμήμα του δένδρου το οποίο εκκινεί από αυτούς.
    - Άρα εξοικονομούμε υπολογισμούς εμπιστοσύνης για τους κλαδεμένους κόμβους.

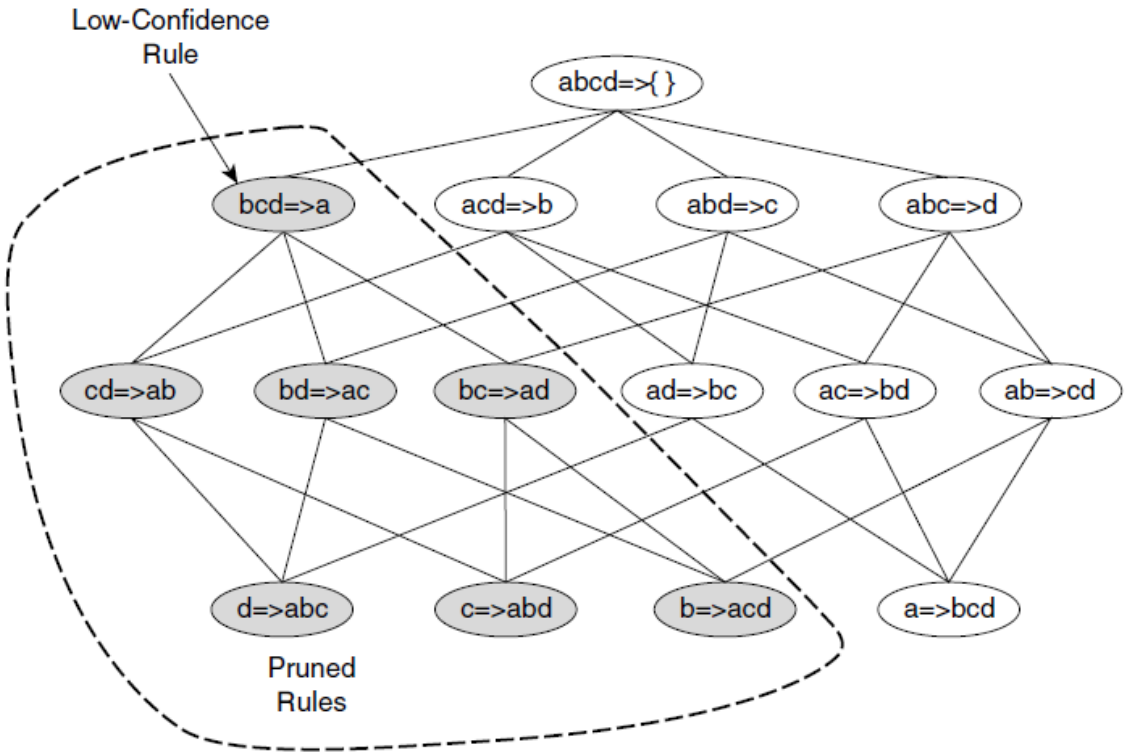
Αλγόριθμος Apriori

Παραγωγή κανόνων

- Βάση του αλγορίθμου είναι μία μαθηματικά αποδεδειγμένη παραλλαγή της αρχής Apriori:
  - Έστω συχνό σύνολο ειδών Y και ένα υποσύνολό του X.
  - Αν ένας κανόνας $X \rightarrow Y - X$ είναι μικρής εμπιστοσύνης, τότε κάθε κανόνας $X' \rightarrow Y - X'$, όπου X' είναι υποσύνολο του X, επίσης είναι μικρής εμπιστοσύνης.
    - Προφανώς, όσο μεγαλύτερου μεγέθους το υποσύνολο ειδών X, τόσο μικρότερου μεγέθους το υποσύνολο ειδών $Y - X$, και αντιστρόφως.

Παραγωγή κανόνων

Αλγόριθμος Apriori



Δένδρο κανόνων από το συχνό σύνολο ειδών {a, b, c, d}.

Πηγή: Tan et al., *Introduction to Data Mining, 2006.*

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr