Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Κανόνες Συσχετίσεων

Κανόνες Συσχετίσεων

Οι Κανόνες Συσχέτισης αφορούν την ανακάλυψη και διατύπωση σχέσεων που υπάρχουν στα δεδομένα. Οι σχέσεις αυτές προκύπτουν από τη συχνή εμφάνιση ζεύγους τιμών.

Για παράδειγμα μελετώντας τις συνήθειες των καταναλωτών σε ένα ηλεκτρονικό κατάστημα, μπορούμε να εξάγουμε το συμπέρασμα ότι όταν ένας καταναλωτής αγοράσει το προϊόν Α τότε είναι πολύ πιθανό να αγοράσει και το Β.

Κανόνες Συσχετίσεων - Ορισμοί

- Ένας κανόνας ορίζεται ως έξης

  $X \rightarrow Y$, όπου τα X και Y είναι αντικείμενα του συνόλου

π.χ. {Μπλοκ ζωγραφικής}→{Μαρκαδόροι}. Τα άτομα που αγόρασαν μπλοκ ζωγραφικής αγόρασαν επίσης και μαρκαδόρους.

Κανόνες Συσχετίσεων - Ορισμοί

Δύο ποσοτικά μεγέθη καθορίζουν πόσο ισχυρός είναι ο κανόνας X → Y

1. Support (s). Το ποσοστό των συναλλαγών (επί του συνόλου των συναλλαγών) που περιέχουν και το X και το Y.
2. Confidence (c). Είναι η πιθανότητα εμφάνισης του Y, όταν εμφανίζεται το X.

Κανόνες Συσχετίσεων - Ορισμοί

Παράδειγμα

Συναλλαγή	Προϊόντα
1	Α-Β-Γ
2	Γ-Δ
3	Α-Β
4	Α-Β-Δ
5	Α-Δ
6	Β-Γ

Κανόνας :Α->Β

- Το ζεύγος Α-Β εμφανίζεται 3 φορές.
- Οι συνολικές συναλλαγές είναι 6.
- Το Α εμφανίζεται 4 φορές ενώ το Β 3.
- Η υποστήριξη του κανόνα είναι 3/6 , 50%
- Η εμπιστοσύνη του κανόνα είναι 3/4,

Κανόνες Συσχετίσεων

Προκειμένου να βρούμε τους κανόνες εκτελούμε τις παρακάτω ενέργειες

1. Υπολογίζουμε όλους τους κανόνες.
2. Υπολογίζουμε τις τιμές support και confidence για κάθε κανόνα.
3. Κρατάμε μόνο αυτούς τους κανόνες που έχουν τιμές μικρότερες από αυτές που ορίσαμε για support και confidence

Κανόνες Συσχετίσεων

Προβλήματα παρατηρούνται με αυτή την τεχνική όταν έχουμε μεγάλα σύνολα δεδομένων όπου θα πρέπει να υπολογιστούν όλοι οι δυνατοί συνδυασμοί.

Σε τέτοιες περιπτώσεις ο έλεγχος όλων των κανόνων είναι πρακτικά αδύνατος καθώς απαιτείται πολύς χρόνος και πόροι του συστήματος.

Κανόνες Συσχετίσεων – Αλγόριθμος Apriori

Ένας από του πιο γνωστούς αλγορίθμους είναι ο Apriori ο οποίος είναι Brute-force και αποτελείται από τα παρακάτω βήματα

1. Υπολογίζουμε όλους τους κανόνες
2. Υπολογίζουμε τις τιμές support και confidence για κάθε κανόνα.
3. Κρατάμε τους κανόνες με τιμές μικρότερες από αυτές που ορίσαμε για support και confidence

Αλγόριθμος Apriori – Παράδειγμα #1

Έστω έχουμε το παρακάτω σύνολο

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}
{1,5}
{1,3,5}

θέτουμε σαν κριτήριο Support>=4

Αλγόριθμος Apriori – Παράδειγμα #1

Βήμα 1.
Συχνότητα εμφάνισης κάθε αντικειμένου

Στοιχείο (item)	Συχνότητα (Support)
{1}	5
{2}	6
{3}	5
{4}	5
{5}	3

Βήμα 2.
Συχνότητα εμφάνισης διπλών συνδυασμών

Στοιχείο (item)	Συχνότητα (Support)
{1,2}	3
{1,3}	2
{2,3}	1
{2,4}	1

Βήμα 3.
Συχνότητα εμφάνισης 3πλών συνδυασμών

Στοιχείο (item)	Συχνότητα (Support)
{1,2,3}	1

Αλγόριθμος Apriori – Παράδειγμα #2

Λίστα συνδυασμών

Α/Α	Λίστα Αντικειμένων
1	ΓΑΛΑ,ΨΩΜΙ,ΤΥΡΙ
2	ΨΩΜΙ,ΑΥΓΑ
3	ΨΩΜΙ,ΚΑΡΟΤΑ
4	ΓΑΛΑ,ΨΩΜΙ,ΑΥΓΑ
5	ΓΑΛΑ,ΚΑΡΟΤΑ
6	ΨΩΜΙ,ΚΑΡΟΤΑ
7	ΓΑΛΑ,ΚΑΡΟΤΑ
8	ΓΑΛΑ,ΨΩΜΙ,ΚΑΡΟΤΑ,ΤΥΡΙ
9	ΓΑΛΑ,ΨΩΜΙ,ΚΑΡΟΤΑ

Βήμα 1.
Συχνότητα εμφάνισης κάθε αντικειμένου

Αντικείμενο	Συχνότητα
ΓΑΛΑ	6
ΨΩΜΙ	7
ΚΑΡΟΤΑ	6
ΑΥΓΑ	2
ΤΥΡΙ	2

Βήμα 2.
Συχνότητα εμφάνισης διπλών συνδυασμών

Συνδυασμοί	Συχνότητα
ΓΑΛΑ,ΨΩΜΙ	4
ΓΑΛΑ,ΚΑΡΟΤΑ	4
ΓΑΛΑ,ΑΥΓΑ	1
ΓΑΛΑ,ΤΥΡΙ	2
ΨΩΜΙ,ΚΑΡΟΤΑ	4
ΨΩΜΙ,ΑΥΓΑ	2
ΨΩΜΙ,ΤΥΡΙ	2
ΚΑΡΟΤΑ,ΑΥΓΑ	0
ΚΑΡΟΤΑ,ΑΥΓΑ	1
ΑΥΓΑ,ΤΥΡΙ	0

θέτουμε σαν κριτήριο Support>=2

Αλγόριθμος Apriori – Παράδειγμα #2

Βήμα 3.
Συχνότητα εμφάνισης διπλών συνδυασμών με Συχνότητα>=2

Συνδυασμοί	Συχνότητα
ΓΑΛΑ,ΨΩΜΙ	4
ΓΑΛΑ,ΚΑΡΟΤΑ	4
ΓΑΛΑ,ΤΥΡΙ	2
ΨΩΜΙ,ΚΑΡΟΤΑ	4
ΨΩΜΙ,ΑΥΓΑ	2
ΨΩΜΙ,ΤΥΡΙ	2

Βήμα 4.
Συχνότητα εμφάνισης τριπλών συνδυασμών

Συνδυασμοί	Συχνότητα
ΓΑΛΑ,ΨΩΜΙ,ΚΑΡΟΤΑ	2
ΓΑΛΑ,ΨΩΜΙ, ΤΥΡΙ	2

Δείκτης Lift

$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$\boxed{Lift = \frac{Support}{Supp(X) \times Supp(Y)}}$$

Rule: {butter,bread} => {milk}

Trans_ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

- X= {butter,bread}
- Y= {milk}

- Support =1/5
- Confidence = Support(XUY)/Support(X) = (1/5)/(1/5) =0.2/0.2=1
- Lift= Support(XUY)/Support(X) * Support(Y) = (1/5)/(1/5)*(2/5) =0.2/0.2*0.4=2.5

ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ