Κανόνες σύνδεσης και ο αλγόριθμος Apriori

Market Basket Analysis [Association Analysis] - YouTube

Το πρόβλημα

Όταν πηγαίνουμε για ψώνια, έχουμε συχνά μια τυπική λίστα με πράγματα που πρέπει να αγοράσουμε. Κάθε αγοραστής έχει μια ξεχωριστή λίστα, ανάλογα με τις ανάγκες και τις προτιμήσεις του. Μια νοικοκυρά μπορεί να αγοράσει υγιεινά υλικά για ένα οικογενειακό δείπνο, ενώ ένας εργένης μπορεί να αγοράσει μπύρα και πατατάκια.

Η κατανόηση αυτών των αγοραστικών μοτίβων μπορεί να βοηθήσει στην αύξηση των πωλήσεων με διάφορους τρόπους. Εάν υπάρχει ένα ζευγάρι αντικειμένων, X και Y, που αγοράζονται συχνά μαζί:

- Τόσο το X όσο και το Y μπορούν να τοποθετηθούν στο ίδιο ράφι, έτσι ώστε οι αγοραστές ενός αντικειμένου να κληθούν να αγοράσουν το άλλο.
- Οι προωθητικές εκπτώσεις θα μπορούσαν να εφαρμοστούν μόνο σε ένα από τα δύο είδη.
- Οι διαφημίσεις στο X θα μπορούσαν να στοχεύουν σε αγοραστές που αγοράζουν το Y.
- Το X και το Y θα μπορούσαν να συνδυαστούν σε ένα νέο προϊόν, όπως το να έχουν το Y σε γεύσεις του X.
-

Ενώ μπορεί να γνωρίζουμε ότι ορισμένα αντικείμενα αγοράζονται συχνά μαζί, το ερώτημα είναι, πώς αποκαλύπτουμε αυτές τις συσχετίσεις;

Εκτός από την αύξηση των κερδών από τις πωλήσεις, οι κανόνες σύνδεσης μπορούν επίσης να χρησιμοποιηθούν σε άλλους τομείς. Στην ιατρική διάγνωση, για παράδειγμα, η κατανόηση των συμπτωμάτων που τείνουν να συννοσηρότητα μπορεί να βοηθήσει στη βελτίωση της φροντίδας των ασθενών και της συνταγογράφησης φαρμάκων.

Ορισμός

Η ανάλυση κανόνων συσχέτισης είναι μια τεχνική για την αποκάλυψη του τρόπου με τον οποίο τα στοιχεία σχετίζονται μεταξύ τους. Υπάρχουν τρεις κοινοί τρόποι μέτρησης της συσχέτισης.

Μέτρο 1: Υποστήριξη. Αυτό λέει πόσο δημοφιλές είναι ένα σύνολο στοιχείων, όπως μετράται από το ποσοστό των συναλλαγών στις οποίες εμφανίζεται ένα σύνολο στοιχείων. Στον Πίνακα 1 παρακάτω, η υποστήριξη του {μήλο} είναι 4 στα 8 ή 50%. Τα σύνολα στοιχείων μπορούν επίσης να περιέχουν πολλά στοιχεία. Για παράδειγμα, η υποστήριξη {μήλο, μπύρα, ρύζι} είναι 2 στα 8 ή 25%.

$$\text{Support } \{🍎\} = \frac{4}{8}$$

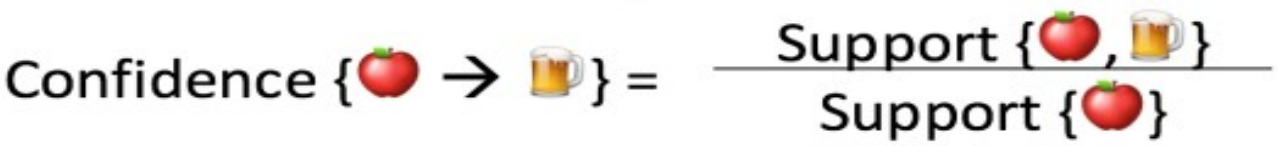
Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Πίνακας 1. Παράδειγμα συναλλαγών

Εάν ανακαλύψετε ότι οι πωλήσεις αντικειμένων πέρα από ένα ορισμένο ποσοστό τείνουν να έχουν σημαντικό αντίκτυπο στα κέρδη σας, μπορείτε να χρησιμοποιήσετε αυτό το ποσοστό ως όριο υποστήριξης. Στη συνέχεια, μπορείτε να προσδιορίσετε σύνολα στοιχείων με τιμές υποστήριξης πάνω από αυτό το όριο ως σημαντικά σύνολα στοιχείων.

Μέτρο 2: Εμπιστοσύνη. Αυτό δείχνει πόσο πιθανό είναι να αγοραστεί το στοιχείο Υ κατά την αγορά του αντικειμένου Χ, εκφρασμένο ως {Χ -> Υ}. Αυτό μετράται από την αναλογία των συναλλαγών με το στοιχείο Χ, στο οποίο εμφανίζεται επίσης το στοιχείο Υ.

Στον Πίνακα πιο κάτω, η εμπιστοσύνη {μήλο -> μπύρας} είναι 3 στα 4, ή 75%.



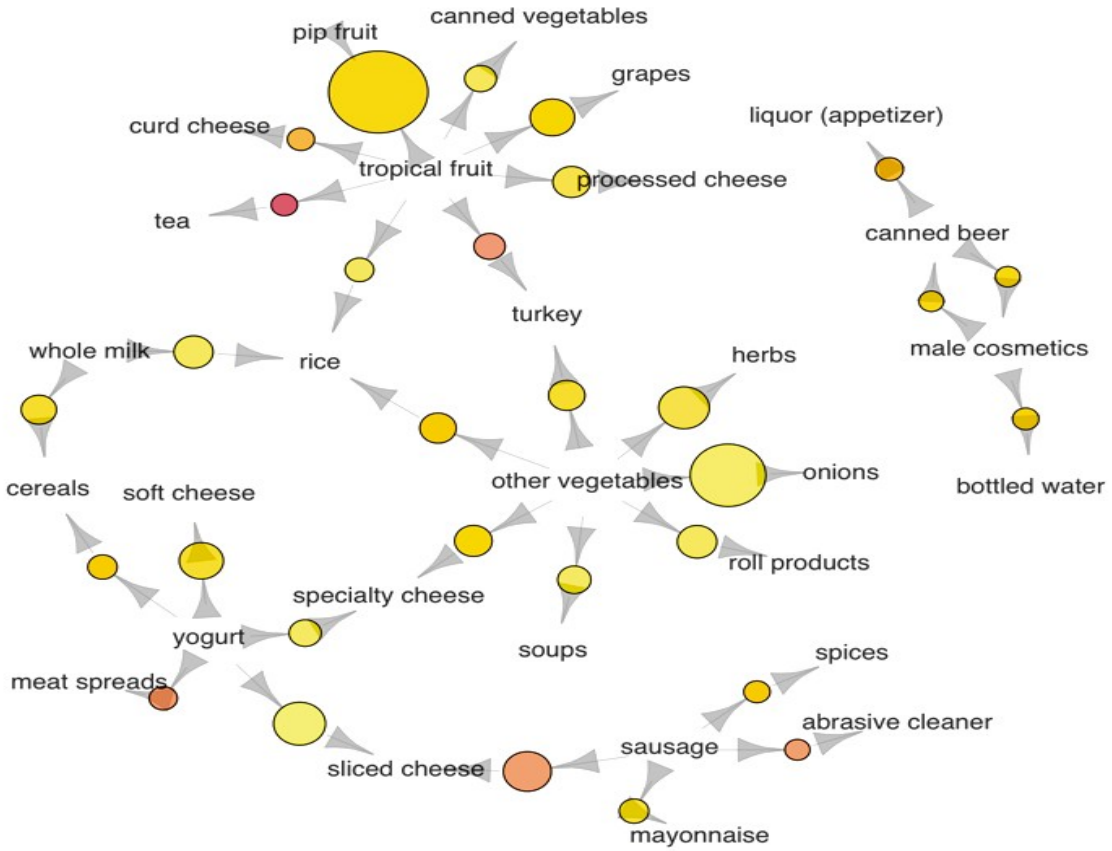
Ένα μειονέκτημα του μέτρου εμπιστοσύνης είναι ότι μπορεί να παραποιήσει τη σημασία μιας ένωσης. Αυτό συμβαίνει επειδή εξηγεί μόνο πόσο δημοφιλή είναι τα μήλα, αλλά όχι οι μπύρες. Εάν οι μπύρες είναι επίσης πολύ δημοφιλείς γενικά, θα υπάρχει μεγαλύτερη πιθανότητα μια συναλλαγή που περιέχει μήλα να περιέχει επίσης μπύρες, διογκώνοντας έτσι το μέτρο εμπιστοσύνης. Για να λάβουμε υπόψη τη βασική δημοτικότητα και των δύο συστατικών στοιχείων, χρησιμοποιούμε ένα τρίτο μέτρο που ονομάζεται **ανελκυστήρας**.

Μέτρο 3: Ανελκυστήρας. Αυτό λέει πόσο πιθανό είναι να αγοραστεί το στοιχείο Υ κατά την αγορά του αντικειμένου Χ, ενώ ελέγχει πόσο δημοφιλές είναι το στοιχείο Υ.

Στον Πίνακα πιο κάτω, η ανύψωση του {μήλο -> μπύρα} είναι 1, πράγμα που δεν σημαίνει συσχέτιση μεταξύ των αντικειμένων. Μια τιμή ανελκυστήρα μεγαλύτερη από 1 σημαίνει ότι το είδος Υ είναι πιθανό να αγοραστεί εάν αγοραστεί το αντικείμενο Χ, ενώ μια τιμή μικρότερη από 1 σημαίνει ότι το είδος Υ είναι απίθανο να αγοραστεί εάν αγοραστεί το στοιχείο Χ.

$$\text{Lift}\{🍎 \rightarrow 🍺\} = \frac{\text{Support}\{🍎, 🍺\}}{\text{Support}\{🍎\} \times \text{Support}\{🍺\}}$$

Χρησιμοποιούμε ένα σύνολο δεδομένων σχετικά με τις συναλλαγές παντοπωλείων.
Περιέχει πραγματικές συναλλαγές σε ένα παντοπωλείο για 30 ημέρες. Το παρακάτω γράφημα δικτύου δείχνει συσχετίσεις μεταξύ επιλεγμένων στοιχείων. Οι μεγαλύτεροι κύκλοι υποδηλώνουν υψηλότερη υποστήριξη, ενώ οι κόκκινοι κύκλοι υποδηλώνουν υψηλότερη ανύψωση:



Συσχετίσεις μεταξύ επιλεγμένων στοιχείων. Οπτικοποιείται χρησιμοποιώντας τη βιβλιοθήκη arulesViz R.

Μπορούν να παρατηρηθούν διάφορα μοτίβα αγοράς. Για παράδειγμα:

- Η πιο δημοφιλής συναλλαγή ήταν του φρούτου pip και των τροπικών φρούτων
- Μια άλλη δημοφιλής συναλλαγή ήταν κρεμμύδια και άλλα λαχανικά
- Αν κάποιος αγοράσει αλείμματα κρέατος, είναι πιθανό να έχει αγοράσει και γιαούρτι
- Σχετικά πολλοί άνθρωποι αγοράζουν λουκάνικο μαζί με τυρί σε φέτες
- Εάν κάποιος αγοράσει τσάι, είναι πιθανό να έχει αγοράσει και φρούτα, πιθανώς εμπνέοντας την παραγωγή τσαγιού με γεύση φρούτων.

Θυμηθείτε ότι ένα μειονέκτημα του μέτρου εμπιστοσύνης είναι ότι τείνει να παραποιεί τη σημασία μιας ένωσης. Για να το αποδείξουμε αυτό, επιστρέφουμε στο κύριο σύνολο δεδομένων για να επιλέξουμε 3 κανόνες συσχέτισης που περιέχουν μπύρα:

Transaction	Support	Confidence	Lift
Canned Beer → Soda	1%	20%	1.0
Canned Beer → Berries	0.1%	1%	0.3
Canned Beer → Male Cosmetics	0.1%	1%	2.6

Πίνακας 2. Μέτρα σύνδεσης για τους κανόνες που σχετίζονται με την μπύρα

Ο κανόνας {μπύρα -> σόδα} έχει την υψηλότερη εμπιστοσύνη στο 20%. Ωστόσο, τόσο η μπύρα όσο και η σόδα εμφανίζονται συχνά σε όλες τις συναλλαγές (βλέπε πίνακα 3), οπότε η σχέση τους θα μπορούσε απλώς να είναι μια αναταραχή. Αυτό επιβεβαιώνεται από την τιμή ανύψωσης της {μπύρας -> σόδα}, η οποία είναι 1, υπονοώντας ότι δεν υπάρχει σχέση μεταξύ μπύρας και σόδας.

Transaction	Support
Canned Beer	10%
Soda	20%
Berries	3%
Male Cosmetics	0.5%

Πίνακας 3. Υποστήριξη μεμονωμένων αντικειμένων

Από την άλλη, ο κανόνας {μπύρα -> ανδρικά καλλυντικά} έχει χαμηλή εμπιστοσύνη, λόγω των λίγων αγορών ανδρικών καλλυντικών γενικά. Ωστόσο, κάθε φορά που κάποιος αγοράζει ανδρικά καλλυντικά, είναι πολύ πιθανό να αγοράσει και μπύρα, όπως συνάγεται από την υψηλή τιμή ανύψωσης 2,6. Το αντίστροφο ισχύει για {μπύρα -> μούρα}. Με τιμή ανύψωσης κάτω από 1, μπορούμε να συμπεράνουμε ότι αν κάποιος αγοράσει μούρα, πιθανότατα θα ήταν αντίθετος στην μπύρα.

Είναι εύκολο να υπολογίσετε τη δημοτικότητα ενός μόνο συνόλου αντικειμένων, όπως {μπύρα, σόδα}. Ωστόσο, ένας ιδιοκτήτης επιχείρησης δεν θα ρωτούσε συνήθως για μεμονωμένα σύνολα αντικειμένων. Αντίθετα, ο ιδιοκτήτης θα ενδιαφερόταν περισσότερο να έχει μια πλήρη λίστα δημοφιλών ειδών. Για να λάβετε αυτήν τη λίστα, πρέπει να υπολογίσετε τις τιμές υποστήριξης για κάθε πιθανή διαμόρφωση στοιχείων και, στη συνέχεια, να επιλέξετε τα σύνολα στοιχείων που πληρούν το ελάχιστο όριο υποστήριξης.
Σε ένα κατάστημα με μόλις 10 αντικείμενα, ο συνολικός αριθμός πιθανών διαμορφώσεων προς εξέταση θα ήταν επιβλητικός 1023. Αυτός ο αριθμός αυξάνεται εκθετικά σε ένα κατάστημα με εκατοντάδες αντικείμενα.

Μια εισαγωγή στις διασταυρούμενες πωλήσεις χρησιμοποιώντας την ανάλυση καλαθιού αγοράς στο Excel

Market Basket Analysis στο Excel

Ξέρουν πώς γεμίζεις το καλάθι σου...
Πριν σας δείξω πώς να κάνετε μια ανάλυση καλαθιού αγοράς στο Excel, πρέπει να βεβαιωθούμε ότι κατανοείτε τις υποκείμενες έννοιες. Σε αυτό το σημείωμα, φανταστείτε ότι μια μέρα πηγαίνετε στο κοντινό σούπερ μάρκετ για να αγοράσετε κάποιο δηλητήριο αρουραίων για τον αρουραίο που έχει εισβάλει στο σπίτι σας.

Ενώ ψάχνατε για το προϊόν που θέλατε, παρατηρήσατε συσκευασίες mortein και μπουκάλια εντομοκτόνου κουνουπιών, ακριβώς δίπλα στο ποντικοφάρμακο. Καταλήξατε να αγοράσετε ένα μπουκάλι σπρέι κουνουπιών, παρόλο που δεν είχατε καμία πρόθεση να το αγοράσετε όταν μπήκατε στην αγορά. Δεν είχατε την επείγουσα ανάγκη να το αγοράσετε, οπότε αν δεν είχατε δει τα εντομο απωθητικά αντικείμενα, σίγουρα δεν θα τα είχατε αγοράσει. Οι λιανοπωλητές πιθανότατα γνώριζαν ότι όσοι αγοράζουν αντι-επιβλαβή αντικείμενα αγοράζουν επίσης έντομο απωθητικά τις περισσότερες φορές, οπότε αποφάσισαν να τα τοποθετήσουν το ένα δίπλα στο άλλο, για να ενθαρρύνουν τους αγοραστές.

Για τους εμπόρους λιανικής, η κατανόηση αυτού του είδους συμπεριφοράς πελατών μπορεί να οδηγήσει σε αύξηση των πωλήσεων. Ωστόσο, δεν μπορεί κανείς να διαισθάνεται κάτι τέτοιο.
Ενώ μερικοί άνθρωποι αγαπούν το ψωμί και το βούτυρο για πρωινό, μερικοί αγαπούν επίσης τα αυγά και το ψωμί, αλλά ποιο θα σας δώσει περισσότερα κέρδη; Αυτά τα ερωτήματα αποτελούν τη βάση πίσω από την έννοια της ανάλυσης καλαθιού αγοράς.

Η ανάλυση καλαθιού αγοράς είναι το σκεπτικό πίσω από την τέχνη της τακτοποίησης αντικειμένων σε ένα κατάστημα. Οι τοποθετήσεις προϊόντων θα πρέπει να γίνονται με τέτοιο τρόπο ώστε τα είδη που αγοράζονται συχνά μαζί να διατηρούνται το ένα δίπλα στο άλλο, έτσι ώστε οι πελάτες να ενθαρρύνονται να τα αγοράσουν και έτσι ώστε αυτό να έχει ως αποτέλεσμα την αύξηση των πωλήσεων.

Πώληση και στη συνέχεια διασταυρούμενη πώληση

Ανάλυση καλαθιού αγοράς στο Excel. Ένα πολύ διάσημο παράδειγμα ανάλυσης καλαθιού αγοράς έχει να κάνει με αγορές μπύρας και πάνες. Μια ενδιαφέρουσα μελέτη σχετικά με τους παντρεμένους άνδρες που έρχονται στα καταστήματα για την αγορά πάνες έδειξε ότι συνήθως αγόραζαν ένα σετ μπύρες για να πάνε με τις πάνες.

Το σούπερ μάρκετ το πήρε ως ευκαιρία διασταυρούμενων πωλήσεων και αποφάσισε να κρατήσει τις πάνες μωρών στο διάδρομο δίπλα στο διάδρομο μπύρας. Είδαν αμέσως μια απότομη αύξηση των ανδρών που ήρθαν να αγοράσουν πάνες μωρών αγοράζοντας επίσης μπύρα. Αυτή η μέθοδος διασταυρούμενων πωλήσεων προέρχεται απευθείας από την ανάλυση καλαθιού αγοράς.

Όταν εντοπίζεται μια συσχέτιση δύο ή περισσότερων προϊόντων, η διατήρησή τους το ένα δίπλα στο άλλο θα έχει ως αποτέλεσμα την αύξηση των πωλήσεων λόγω της διασταυρούμενης πώλησης των προϊόντων. Ωστόσο, η διατήρηση της μπύρας και των πάνες μαζί μπορεί να μην οδηγήσει σε αύξηση των πωλήσεων πάνας. Αυτό οφείλεται στο γεγονός ότι η ανάλυση έδειξε ότι οι άνδρες που

αγοράζουν πάνες είναι πιθανό να αγοράσουν μπύρα και όχι το αντίστροφο, υπονοώντας μια μονόδρομη σχέση διασταυρούμενων πωλήσεων.

Για να επιλυθεί αυτή η σύγχυση, η ανάλυση καλαθιού αγοράς περιλαμβάνει αυτό που είναι γνωστό ως «κανόνες σύνδεσης», για να προσδιορίσει ποια προϊόντα θα υποστούν διασταυρούμενες πωλήσεις όταν δύο ή περισσότερα προϊόντα διατηρούνται μαζί.
Ας κατανοήσουμε επίσημα τις έννοιες της ανάλυσης καλαθιού αγοράς (MBA) και τους σχετικούς όρους. Μία από τις πιο δημοφιλείς τεχνικές διασταυρούμενων πωλήσεων, το MBA περιλαμβάνει τον εντοπισμό του συνδυασμού προϊόντων που βοηθούν στην ενίσχυση των πωλήσεων. Αυτός ο συνδυασμός αντικατοπτρίζει τα προϊόντα που αγοράζονται συχνά από τους πελάτες μαζί.

Αυτός είναι ο λόγος για τον οποίο προέκυψε ο όρος ανάλυση καλαθιού αγοράς.
Σημαίνει ανάλυση του περιεχομένου που περιλαμβάνεται στις λίστες αγοράς διαφορετικών ατόμων. Οι συνδυασμοί σημειώνονται με τη χρήση κανόνων συσχέτισης με τη μορφή A → B όπου το A και το B αντιπροσωπεύουν ένα ή περισσότερα προϊόντα.

Αυτός ο κανόνας σημαίνει ότι οι πελάτες που αγοράζουν τον A είναι πιθανό να αγοράσουν τον B και έτσι η διατήρηση και των δύο μαζί θα έχει ως αποτέλεσμα ο B να πουλήσει περισσότερα. Το ερώτημα που έρχεται στο μυαλό είναι: «Πώς ξέρουμε αν υπάρχει ένας τέτοιος συσχετισμός;». Ο αλγόριθμος apriori είναι μια δημοφιλής μέθοδος για τον εντοπισμό τέτοιων συσχετίσεων. Πριν εξηγήσω τον αλγόριθμο, θα περάσω από «υποστήριξη» και «εμπιστοσύνη» δείχνοντας ένα παράδειγμα.

Ένα παράδειγμα ανάλυσης καλαθιού αγοράς

Παίρνοντας ένα παράδειγμα παιχνιδιού. Ας υποθέσουμε ότι υπάρχουν δύο προϊόντα A και B που είναι τα προϊόντα με τις μεγαλύτερες πωλήσεις σε ένα κατάστημα. Σχεδόν το 70% των πελατών αγοράζουν το A ως ένα από τα προϊόντα τους και το 80% των πελατών αγοράζουν το B ως ένα από τα προϊόντα τους. Οι άνθρωποι που αγοράζουν τόσο το A όσο και το B μαζί είναι 60%. Έχουμε τη συσχέτιση A → B ή B → A; Γνωρίζουμε ότι όλοι οι πελάτες δεν αγοράζουν A. Επίσης, όλοι οι πελάτες που αγοράζουν το A δεν αγοράζουν B. Ωστόσο, δεν μπορούμε να είμαστε σίγουροι αν μια συσχέτιση από το A στο B ή από το B στο A υπάρχει μόνο γνωρίζοντας τους αριθμούς.

Εδώ ορίζουμε την υποστήριξη και την εμπιστοσύνη.
Η υποστήριξη καθορίζει τον αριθμό των συναλλαγών που αφορούν ένα συγκεκριμένο προϊόν ή σύνολο προϊόντων. Αυτό σημαίνει ότι η υποστήριξη για τον A σε αυτό το παράδειγμα είναι 70%, για τον B είναι 80% και για τον A και για τον B είναι 60%. Η εμπιστοσύνη υποδηλώνει την ακρίβεια με την οποία μπορούμε να πούμε ότι μια συναλλαγή είναι αληθινή. Συνήθως εξετάζουμε την εμπιστοσύνη για να υπολογίσουμε την εγκυρότητα ενός κανόνα συσχέτισης και στη συνέχεια εξετάζουμε την υποστήριξη για να δούμε αν είναι βιώσιμο να χρησιμοποιήσουμε τον κανόνα σύνδεσης με κάποιο τρόπο. Ας ελέγξουμε τους κανόνες έναν προς έναν

If customers purchase A, then they also purchase B (A → B)
Εάν οι πελάτες αγοράσουν A, τότε αγοράζουν επίσης B (A → B)

Γνωρίζουμε ήδη όλους τους αριθμούς υποστήριξης. Η εμπιστοσύνη για το A → B μπορεί να υπολογιστεί χρησιμοποιώντας τον τύπο:

Confidence(A → B) = support(A and B)/support(A)
Εμπιστοσύνη(A → B) = υποστήριξη(A και B)/υποστήριξη(A)

Χρησιμοποιώντας υποστήριξη (A και B) = 60% και υποστήριξη (A) ως 70%, έχουμε την εμπιστοσύνη ως 6/7 που είναι 85.7%. Όπως προκύπτει από τον τύπο, η εμπιστοσύνη απεικονίζει το

κλάσμα των περιπτώσεων στις οποίες ο κανόνας του συσχετισμού ισχύει από εκείνες στις οποίες οι πελάτες αγόρασαν τα προϊόντα στην αριστερή πλευρά

If customers purchase B, then they also purchase A (B → A)
Εάν οι πελάτες αγοράσουν B, τότε αγοράζουν επίσης A (B → A)

Σε αυτή την περίπτωση, η εμπιστοσύνη για το B → A μπορεί να υπολογιστεί χρησιμοποιώντας τον τύπο:

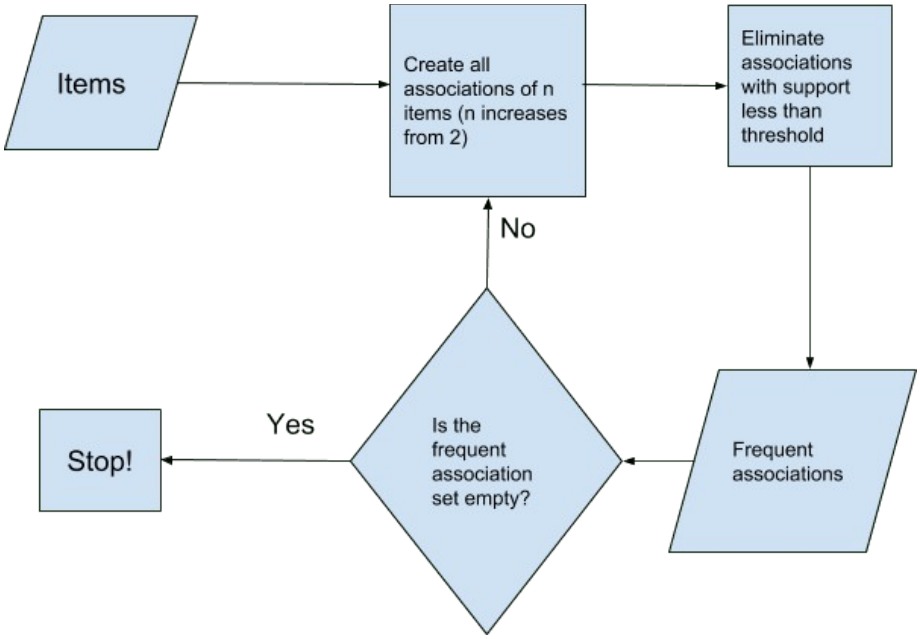
Confidence(B → A) = support(A and B)/support(B)
Εμπιστοσύνη(B → A) = υποστήριξη(A και B)/υποστήριξη(B)

Χρησιμοποιώντας υποστήριξη (A και B) = 60% και υποστήριξη (A) ως 80%, έχουμε την εμπιστοσύνη ως 6/8 που είναι 75%. Συνολικά βλέπουμε ότι η εμπιστοσύνη για το B → το A είναι χαμηλότερο από το A → το B, οπότε αν πρέπει να βάλουμε τα στοιχήματά μας σε ένα από τα δύο, θα πάμε για A → B. Ωστόσο, δεδομένου ότι η συσχέτιση είναι αρκετά υψηλή και με τους δύο τρόπους, η διατήρηση των δύο προϊόντων μαζί μπορεί να οδηγήσει σε αύξηση των πωλήσεων και των δύο προϊόντων.

Γιατί δύο επαληθεύσεις;

Στην πραγματική ζωή, δεδομένου ότι δεν μπορούμε να συμπεριλάβουμε όλους τους κανόνες σύνδεσης, συνήθως αποφασίζουμε κατώτατα όρια τόσο για την υποστήριξη όσο και για την εμπιστοσύνη κάτω από τα οποία η ένωση δεν θα ληφθεί υπόψη. Γιατί όμως και υποστήριξη και εμπιστοσύνη; Η εμπιστοσύνη από μόνη της δείχνει την εγκυρότητα ενός κανόνα συσχέτισης, αλλά δεν είναι τόσο χρήσιμο εάν ισχύει μόνο για λίγους πελάτες ή με άλλα λόγια, η υποστήριξη είναι χαμηλή.

Εδώ έρχεται στην εικόνα ο αλγόριθμος apriori. Όταν έχετε πολλά αντικείμενα σε ένα κατάστημα, υπάρχουν πολλές πιθανές συσχετίσεις. Επιπλέον, οι ενώσεις πρέπει να ελέγχονται για συνδυασμούς δύο προϊόντων, στη συνέχεια τριών προϊόντων και ούτω καθεξής καθώς το επίπεδο ανεβαίνει.
Ο αλγόριθμος apriori εξαλείφει τις δυνατότητες πολύ πριν χρειαστεί να εκτελέσουμε όλους αυτούς τους υπολογισμούς. Λειτουργεί επαναληπτικά από την επιλογή συνδυασμών δύο προϊόντων έως το ανοδικό επίπεδο συνδυασμών. Στο πρώτο επίπεδο, καταργούνται όλες οι ενώσεις όπου η υποστήριξη και η εμπιστοσύνη είναι χαμηλότερες από τα καθορισμένα όρια. Από εκεί, τα ανώτερα επίπεδα υπολογίζονται συνδυάζοντας μόνο τις έγκυρες συσχετίσεις. Σύμφωνα με τον κανόνα apriori, ένας κανόνας σύνδεσης δεν μπορεί να έχει υψηλότερη υποστήριξη και εμπιστοσύνη από οποιοδήποτε από τα υποσύνολά του. Έτσι, το επόμενο επίπεδο έχει πολύ λιγότερες συσχετίσεις προς δοκιμή. Το ίδιο μπορεί να εξηγηθεί χρησιμοποιώντας το διάγραμμα ροής:

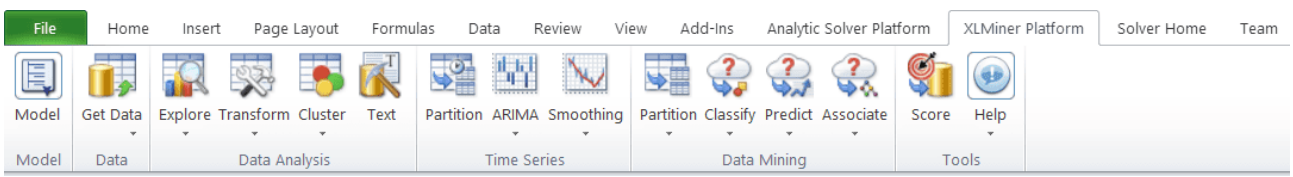


Μεταξύ των τελικών έγκυρων ενώσεων, οι ιδιοκτήτες καταστημάτων αποφασίζουν στρατηγικές ανάλογα με το επίπεδο υποστήριξης και εμπιστοσύνης. Για παράδειγμα, μπορεί κανείς να κρατήσει τα αντικείμενα μαζί για να ενθαρρύνει τους αγοραστές ή εναλλακτικά να πουλήσει προϊόντα ως πακέτο ή να δώσει εκπτώσεις στην αγορά του συνδυασμού μαζί.

Όλα αυτά έχουν νόημα, σωστά; Εντάξει, τέλεια - τώρα ας μπούμε στην επίδειξη του πώς να κάνουμε μια ανάλυση καλαθιού αγοράς στο Excel.

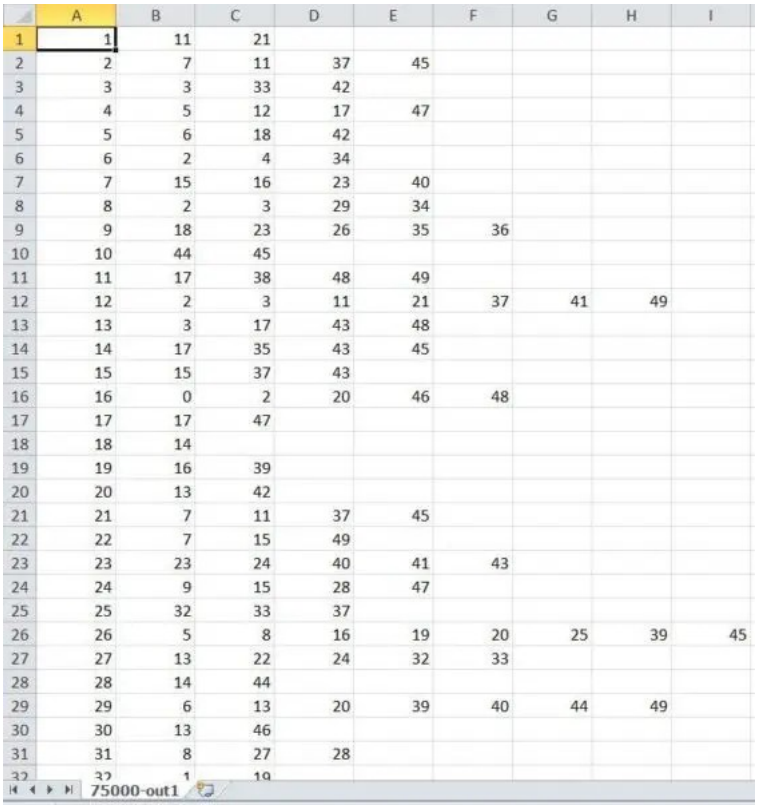
Μια πρακτική επίδειξη της ανάλυσης καλαθιού αγοράς στο Excel

Έχουμε περιγράψει αρκετά την έννοια της ανάλυσης καλαθιού αγοράς, αλλά τα διδάγματα γίνονται καλύτερα με ένα πρακτικό παράδειγμα. Ίσως αναρωτιέστε, "Πώς στο καλό μπορείτε να κάνετε μια ανάλυση καλαθιού αγοράς στο Excel;". Λοιπόν, αυτό είναι απλό! Θα χρησιμοποιήσουμε ένα πρόσθετο του Excel που ονομάζεται XLminer. Μπορείτε να το πάρετε από τον ιστότοπο Solver.com.
(Εναλλακτικά: Analytic Solver Data Mining Add-in For Excel (Formerly XLMiner) | solver, XLMiner® Platform | solver).
Ας εγκαταστήσουμε το XLminer (πρόσθετο) στο Microsoft Excel, το οποίο μπορεί να χρησιμοποιηθεί για την εκτέλεση ανάλυσης καλαθιού αγοράς και την εργασία σε ένα δείγμα συνόλου δεδομένων. Το XLMiner μπορεί να μεταφορτωθεί από solver.com ιστότοπο ως δωρεάν δοκιμή 15 ημερών.
Υπάρχουν ξεχωριστοί σύνδεσμοι για παράθυρα 32 bit και 64 bit και η ρύθμιση συνοδεύεται από μια δέσμη εργαλείων ανάλυσης. Μπορείτε να εγκαταστήσετε το πρόσθετο για το Excel και να δοκιμάσετε την επιλογή κανόνα συσχέτισης. Δείτε πώς φαίνεται το Microsoft Office 2010 μετά την εγκατάσταση του πρόσθετου.



Η τρέχουσα ανοιχτή καρτέλα είναι η καρτέλα Πλατφόρμα XLMiner και έχει την πλατφόρμα επίλυσης Analytics και την αρχική σελίδα επίλυσης και στις δύο πλευρές της, οι οποίες αποτελούν επίσης μέρος του πακέτου. Θα χρησιμοποιήσουμε την επιλογή συνεργάτη που είναι 3η από τα δεξιά στην καρτέλα για την ανάλυσή μας. Για επίδειξη, θα χρησιμοποιήσω το εκτεταμένο σύνολο δεδομένων αρτοποιίας για 75000 πελάτες. Το σύνολο δεδομένων είναι διαθέσιμο στο διαδίκτυο και μπορείτε να το κατεβάσετε από τον σύνδεσμο: https://wiki.csc.calpoly.edu/datasets/wiki/apriori

Το αρχείο apriori zip περιέχει σύνολα δεδομένων για 1000, 5000, 20000 και 75000 μέλη. Θα χρησιμοποιήσω το αρχείο 75000-out1. Το αρχείο περιέχει 75000 γραμμές όπου κάθε γραμμή υποδεικνύει τη λίστα των στοιχείων που έφερε ένας μοναδικός πελάτης. Έχουμε σχεδόν 50 αντικείμενα για αυτούς τους πελάτες. Ας δούμε ένα στιγμιότυπο του αρχείου.



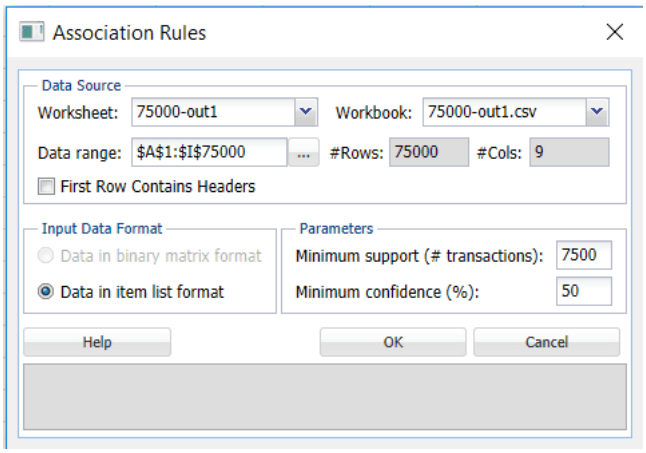
	A	B	C	D	E	F	G	H	I
1	1	11	21						
2	2	7	11	37	45				
3	3	3	33	42					
4	4	5	12	17	47				
5	5	6	18	42					
6	6	2	4	34					
7	7	15	16	23	40				
8	8	2	3	29	34				
9	9	18	23	26	35	36			
10	10	44	45						
11	11	17	38	48	49				
12	12	2	3	11	21	37	41	49	
13	13	3	17	43	48				
14	14	17	35	43	45				
15	15	15	37	43					
16	16	0	2	20	46	48			
17	17	17	47						
18	18	14							
19	19	16	39						
20	20	13	42						
21	21	7	11	37	45				
22	22	7	15	49					
23	23	23	24	40	41	43			
24	24	9	15	28	47				
25	25	32	33	37					
26	26	5	8	16	19	20	25	39	45
27	27	13	22	24	32	33			
28	28	14	44						
29	29	6	13	20	39	40	44	49	
30	30	13	46						
31	31	8	27	28					
32	32	1	19						

75000-out1

Η πρώτη σειρά μας δείχνει τον αριθμό πελάτη και όλοι οι αριθμοί από τη δεύτερη σειρά και μετά δείχνουν το μέγεθος του καλαθιού. Για παράδειγμα, ο πελάτης αριθμός 26 αγόρασε αντικείμενα με ετικέτα 5,8,16,19,20,25,39 και 45 στο καλάθι του. Τα είδη φέρουν παρόμοια ετικέτα από το 0 έως το 49.

Χρησιμοποιώντας το XLMiner – Τα βήματα

Τώρα που έχετε εγκαταστήσει το XLMiner, μπορείτε να κάνετε γρήγορα και εύκολα μια ανάλυση καλαθιού αγοράς στο Excel. Ας κάνουμε κλικ στο κουμπί συσχέτισης και στη συνέχεια στην επιλογή κανόνων συσχέτισης από την καρτέλα XLMiner και να κατανοήσουμε τις επιλογές.



Το μενού μας δείχνει κάποιες λεπτομέρειες σχετικά με την πηγή δεδομένων που λαμβάνονται αυτόματα για να εμφανιστεί το φύλλο εργασίας και το εύρος δεδομένων μας. Δείχνει επίσης ότι ο αριθμός των στηλών είναι 9, πράγμα που σημαίνει ότι το μέγιστο μέγεθος καλαθιού που θα εξετάσουμε θα είναι 9 στοιχεία.

Το πλαίσιο διαλόγου επιλέγει επίσης τα δεδομένα σε μορφή λίστας στοιχείων. Η ενότητα παραμέτρων είναι η πιο σημαντική ενότητα σε αυτό το πλαίσιο διαλόγου για εμάς. Έχουμε τις παραμέτρους της ελάχιστης υποστήριξης και της ελάχιστης εμπιστοσύνης σε αυτό το τμήμα. Η προεπιλεγμένη τιμή για την ελάχιστη υποστήριξη είναι 10% των συναλλαγών και είναι 7.500 εδώ. Έχουμε επίσης μια προεπιλεγμένη τιμή 50% για την εμπιστοσύνη.
Ας τροποποιήσουμε αυτές τις τιμές και ας ορίσουμε την εμπιστοσύνη στο 90%. Δεδομένου ότι υπάρχουν 50 αντικείμενα για 75,000 μέλη, θα χρησιμοποιήσω 1,500 συναλλαγές ως ελάχιστη υποστήριξη (2%). Ο ορισμός αυτών των παραμέτρων και κάνοντας κλικ στο OK θα κάνει τους υπολογισμούς για τους κανόνες συσχέτισης και θα μας δώσει τα αποτελέσματα στο επόμενο φύλλο.

XLMiner : Association Rules

Output Navigator	
Inputs	List of Rules

Elapsed Times in Milliseconds		
AssocRules Time	Report Time	Total
2640	0	2640

Inputs

Data	
# Transactions in Input Data	75000
# Columns in Input Data	9
# Items in Input Data	75001
# Association Rules	95
Minimum Support	1500
Minimum Confidence	0.9

List of Rules

Rule: If all Antecedent items are purchased, then with Confidence percentage Consequent items will also be purchased.

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
1	99.42455243	24 & 40 & 43	23 & 41	1564	2079	1555	35.86744316
2	99.80744544	23 & 24 & 43	40 & 41	1558	2088	1555	35.85037552
3	99.67948718	40 & 41 & 43	23 & 24	1560	2087	1555	35.82156942
4	99.42455243	23 & 41 & 43	24 & 40	1564	2086	1555	35.74708261
5	99.36102236	23 & 40 & 43	24 & 41	1565	2087	1555	35.70712351
6	99.04458599	24 & 41 & 43	23 & 40	1570	2101	1555	35.35623012
7	90.45956952	41 & 43	23 & 24 & 40	1719	1932	1555	35.11629251
8	90.04053272	40 & 43	23 & 24 & 41	1727	1926	1555	35.06251274

Τώρα που ολοκληρώσαμε την ανάλυση του καλαθιού της αγοράς μας στο Excel, φαίνεται ότι η XLMiner βρήκε 95 συσχετίσεις με βάση τις ρυθμίσεις που επιλέξαμε. Όλοι οι κανόνες ορίζονται καθαρά στην ενότητα της λίστας κανόνων. Βλέπουμε ότι η εμπιστοσύνη για ορισμένους κανόνες είναι πολύ υψηλή και είναι πάνω από 99%.

Ο Προηγούμενος είναι το Α μέρος στο Α → Β και το Συνακόλουθο είναι το Β μέρος. Έχουμε επίσης τους αριθμούς υποστήριξης για το Α, C και τα δύο μαζί. Υπάρχουν μερικές ενδιαφέρουσες συσχετίσεις που περιλαμβάνουν 4 στοιχεία ως Α και 3 στοιχεία ως Γ. Μπορούμε επίσης να δούμε ότι η υποστήριξη είναι μεγαλύτερη από 2000 σε ορισμένες περιπτώσεις. Σε αυτήν την περίπτωση, μπορεί να θέλουμε να τους δώσουμε προτεραιότητα. Η μέγιστη υποστήριξη φτάνει τις 3000 συναλλαγές.
Έχουμε επίσης τρεις κανόνες συσχέτισης όπου η εμπιστοσύνη είναι 100%. Οι συσχετισμοί ταξινομούνται με βάση τη φθίνουσα αναλογία ανύψωσης. Μετά τον καθορισμό των ορίων υποστήριξης και εμπιστοσύνης, χρησιμοποιείται συνήθως ο λόγος ανύψωσης καθώς περιγράφει μια εικόνα για το όφελος που προκύπτει από τη χρήση του συσχετισμού σε σύγκριση με όταν τα γεγονότα ήταν ανεξάρτητα.

Αυτό σημαίνει ότι ο λόγος ανύψωσης είναι ο λόγος στήριξης όλων των αντικειμένων που συμβαίνουν μαζί προς τη στήριξη καθενός από αυτά ανεξάρτητα. Με άλλα λόγια, η ανύψωση για έναν συσχετισμό Α→ Β θα είναι υποστήριξη (Α U Β)/(υποστήριξη Α * υποστήριξη Β). Διαισθητικά, ένας λόγος ανύψωσης μεγαλύτερος από 1 σημαίνει ότι εάν οι πελάτες αγοράσουν το Α, είναι πιο πιθανό να αγοράσουν το Β. Ομοίως, ένας λόγος ανύψωσης μικρότερος από 1 σημαίνει ότι εάν οι πελάτες αγοράσουν το Α, είναι λιγότερο πιθανό να αγοράσουν το Β. Ένας λόγος ανύψωσης 1 σημαίνει εντελώς ανεξάρτητη αγορά των αντικειμένων. Η ανάλυση καλαθιού αγοράς στο Excel δεν θα μπορούσε να γίνει πολύ πιο απλή, ε;

Συμπέρασμα

Πώς πρέπει να χρησιμοποιήσει κανείς τους κανόνες συσχέτισης που προκύπτουν από αυτό (ή οποιοδήποτε άλλο, για αυτό το θέμα) αποτέλεσμα; Οι κορυφαίες ενώσεις είναι κάπως περίπλοκες και περιγράφουν τα ίδια 5 προϊόντα. Αλλά αν είναι ικανοποιημένοι, είναι σχεδόν σαφείς ενώσεις καθώς η εμπιστοσύνη τους είναι πάνω από 99%. Επειδή οι ενώσεις ταξινομούνται με βάση τη μείωση της αναλογίας ανελκυστήρων, αυτές οι ενώσεις είναι αξιόπιστες όσον αφορά την εμπιστοσύνη ή την αναλογία ανύψωσης.

Έχουν επίσης παρόμοιο εύρος υποστήριξης και επομένως είναι προϊόντα που πρέπει να διατηρούνται μαζί.
Προς το τέλος της λίστας, έχουμε κάποιες συσχετίσεις όπου το Επακόλουθο είναι ένα ενιαίο στοιχείο. Αυτά τα στοιχεία μπορεί να είναι καλές περιπτώσεις για λόγους ομαδοποίησης. Το Επακόλουθο μπορεί να συνδυαστεί μαζί με το Προηγούμενο. Ή μπορεί να δοθεί κάποια έκπτωση στην αγορά του συνόλου των αντικειμένων μαζί. Με αυτόν τον τρόπο, οι πελάτες ενθαρρύνονται να αγοράσουν το προϊόν. Μερικές φορές, η ενίσχυση της πώλησης του προηγούμενου έχει επίσης ως αποτέλεσμα την αύξηση της πώλησης των επακόλουθων (ειδικά όταν η εμπιστοσύνη είναι κοντά στο 100%).

Σε περιπτώσεις που το προηγούμενο είναι ένα ενιαίο αντικείμενο λίγων αντικειμένων τότε μπορούν να προωθηθούν και να επιτευχθεί περισσότερη πώληση.
Με αυτόν τον τρόπο, υπάρχουν πολλές ευκαιρίες διασταυρούμενων πωλήσεων. Καμία από αυτές τις στρατηγικές δεν είναι σίγουροι κανόνες και δεν είναι εξαντλητικές.