Εργασία 5η

Διονύσης Κοτζαΐτσης

Μελετήστε την εργασία

> *A. Merceron, K. Yacef, "Interestingness measures for association rules in educational data", Proc. Int. Conf. Educ. Data Mining, pp. 57-66, 2008*

και απαντήστε τις παρακάτω ερωτήσεις. (Αποφεύγετε να αντιγράφετε σε μεγάλο βαθμό αυτούσια τμήματα της εργασίας ως απαντήσεις στις ερωτήσεις. Οι απαντήσεις να είναι ακριβείς, περιεκτικές. Αποφεύγετε αυτόματες μεταφράσεις όπως από Google translate).

1. ***Διατυπώστε το σκοπό της εργασίας***

   Επειδή κατά την διάρκεια της διαδικασίας εξόρυξης εκπαιδευτικών δεδομένων, είναι δύσκολη η αποτίμηση και σύγκριση των μετρικών και των μεθόδων, εκ των υστέρων και για όλα τα σετ δεδομένων. Έτσι, αναζητήθηκαν τεχνικές και μετρήσεις ώστε να μπορέσουμε να εξάγουμε association rules με «ασφάλεια». Σκοπός της εργασίας λοιπόν, είναι η μελέτη του συνημίτονου και του αθροίσματος (lift) ως μετρικές για τα εκπαιδευτικά δεδομένα και την εξαγωγή κανόνων συσχέτισης, και κατά πόσο οι δάσκαλοι μπορούν να χρησιμοποιήσουν αυτές τις τεχνικές για να εξάγουν ασφαλή συμπεράσματα.

<u>Association Rules, cosine, Added Value and Lift</u>

2. ***Μελετήστε την ενότητα 2.4. Ο Table 1 τι αποτυπώνει; Εξηγείστε τι εκφράζει στη 2η και 7η γραμμή το (a, b, c)***

   Στην ενότητα 2.4 οι συγγραφείς μελετούν association rules και πιο συγκεκριμένα, τις τυπικές τιμές για το συνημίτονο και το lift. Έτσι μελετάνε διάφορα ποσοστά ύπαρξης X και Y υποσυνόλων σε m transactions. Τα m transactions είναι αυτά τα οποία περιέχουν το X, το Y ή και τα δύο. Πιο συγκεκριμένα, στον πίνακα 1 αποτυπώνονται ως:

   - α: το ποσοστό του υποσυνόλου X and Y που περιέχονται στα m transactions.
   - β: το ποσοστό του υποσυνόλου X που περιέχεται στα m transactions.
   - c: το ποσοστό του υποσυνόλου Y που υπάρχει στα m transactions.

   Στις γραμμές 2 και 7 του πίνακα 1, λοιπόν, παρατηρούμε:

   - Το (90,100,90): Από τα m το 90% περιέχει και το X και το Y. Όλα τα m περιέχουν το υποσύνολο X, ενώ ένα 10% περιέχει το X και δεν περιέχει το Y.
   - Το (60,80,80): Από τα m το 60% περιέχει και το X και το Y. Τα X και τα Y αυτόνομα υπάρχουν στο 80% των παρατηρήσεων, συμμετρικά. Αλλά, συνδυάζονται με τέτοιο τρόπο, ώστε τέμνονται μόνο στο 60% των m transactions.

3. ***Με βάση τον Table 1 γίνονται διαπιστώσεις αναφορικά με τις τιμές του cosine και lift όπως και προτάσεις για την εκτίμηση των κανόνων συσχέτισης. Να τις διατυπώσετε***

   Στο τμήμα Discussion του 2.4, οι συγγραφείς αναφέρουν κάποια συμπεράσματα στα οποία κατέληξαν με την μελέτη των δεδομένων για association rules. Αρχικά, αναφέρουν ότι στην περίπτωση που έχουμε ισχυρές συμμετρικές σχέσεις συσχέτισης (association rules που το ο αριθμός του X, του Y και του X and Y είναι όλα κοντά στον συνολικό αριθμό των transactions) τα cosine και lift διαφέρουν, με το cosine να είναι καλύτερο από το lift. Επιπλέον, μπορούμε να δούμε το lift σε αντίθεση με το συνημίτονο, εξαρτάται από τον αριθμό των transactions που δεν περιέχουν τα X ούτε τα Y, κάτι που δεν ξέρουμε κατά πόσο παίζουν ρόλο στα μαθησιακά δεδομένα. Γενικό συμπέρασμα των συγγραφέων είναι ότι πρέπει αρχικά να ελέγχουμε το cosine και μετά

κάνουμε δεύτερο έλεγχο στο lift. Τέλος, το 0.66 είναι το κατώτατο κατώφλι συνημίτονου που ένα association rule γίνεται δεκτό.

Improving Teacher Support: Case Study

3 ***Τι πληροφορίες μας δίνουν για τη μελέτη περίπτωσης που διενήργησαν. Η απάντηση σας να είναι επιγραμματική (απαντήστε στα παρακάτω σημεία)***

- ***a. Σύστημα διαχείρισης μάθησης***
  Χρησιμοποιήθηκε το σύστημα διαχείρισης μάθησης Moodle
- ***b. Μάθημα & πληροφορίες για το μάθημα***
  Στο μάθημα Εισαγωγή στην πληροφορική του Computer Science and Media.
- ***c. Ιδρυμα/βαθμίδα***
  Στο πανεπιστήμιο University of Applied Sciences TFH Berlin.
- ***d. Χρονική διάρκεια***
  Η μελέτη έγινε το χειμερινό εξάμηνο του 2007/08. Η συγκεκριμένη μελέτη χρησιμοποιεί δεδομένα μέχρι την πρώτη εξέταση, δηλ. 8 εβδομάδες του εξαμήνου.
- ***e. Πληθυσμός***
  84 μαθητές.
- ***f. Αξιολόγηση (εξετάσεις)***
  Κανονικά στο μάθημα υπήρχαν 2 εξετάσεις για να περαστεί. Απλά λόγω του χρονικού περιορισμού της μελέτης, καταγραψαμε μόνο την 1.
- ***g. Συμμετέχοντες (πλήθος (#) εγγεγραμμένων στο μάθημα, # συμμετείχαν στις εξετάσεις, # πέρασαν το μάθημα, # δεν το πέρασαν, # δε προσήλθαν στις εξετάσεις)***
  Από τους 84 φοιτητές, οι 81 έκαναν εγγραφή στο Moodle. Απ' αυτούς 52 πέρασαν το μάθημα, 8 κόπηκαν και 21 δεν έλαβαν μέρος.

4 ***Ποιοι ήταν οι εκπαιδευτικοί πόροι που χρησιμοποιήθηκαν; Η χρήση του Moodle και των επιπλέον εκπαιδευτικών πόρων ήταν υποχρεωτική;***

Οι εκπαιδευτικοί πόροι που χρησιμοποιήθηκαν, αν και δεν ήταν υποχρεωτικό αλλά συστήνεται, ήταν:

- Βιβλίο.
- DP: Επιπλέον υλικό για διάβασμα (Design Patterns for finite automata)
- JFlap: Λογισμικό για την παραγωγή automata.
- Εργασίες 1..7: 7 σετ εβδομαδιαίων ασκήσεων για αυτό βαθμολόγηση.
- 2 Παραδείγματα εξετάσεων και λύσεις.

5 ***Η απάντηση στην ερώτηση «Did students do the exercises?» που αποτυπώνεται στην εργασία; Τι ξέρουμε για την 1η και 7η εργασία; Ποιο είναι το συμπέρασμα αναφορικά με την εκπόνηση των εργασιών;***

Στον πίνακα 2 εμφανίζονται τα αποτελέσματα από τις αποστολές των εργασιών των μαθητών. Για την πρώτη και την έβδομη εργασία έχουμε:

- 1-> 46 δεν την έκαναν, 21 την έκαναν και πέρασαν και 14 την έκαναν και κόπηκαν.
- 7->71 δεν την έκαναν, 8 την έκαναν και πέρασαν και 2 την έκαναν και κόπηκαν.

Το συμπέρασμα που βγαίνει είναι ότι όσο κυλούσε το εξάμηνο, τόσο λιγότεροι μαθητές έκαναν τις εργασίες.

6 ***Η απάντηση στην ερώτηση «Did they access other resources?» που αποτυπώνεται στην εργασία; Ποιο είναι το συμπέρασμα;***

Στον πίνακα 3 έχουμε την σύνδεση και προβολή του λοιπού υλικού (κάτι που καταγράφεται από log files του συστήματος Moodle). Το συμπέρασμα που καταλήγουμε είναι ότι οι περισσότεροι μαθητές έκαναν τα sample exams με λιγότερους απ' τους μισούς να χρησιμοποιούν το υπόλοιπο υλικό και μόλις 23 να διαβάζουν από το βιβλίο. Επιπλέον 38 είδαν τουλάχιστον 1 άσκηση. Γενικότερα, ότι η στάνταρ προετοιμασία για τις εξετάσεις ήταν τα sample exams και οι λύσεις τους.

7 ***Τι παρουσιάζουν στο Table 4. Οι φοιτητές με τις μεγαλύτερες και μικρότερες επιδόσεις ποιους εκπαιδευτικούς πόρους χρησιμοποίησαν στη μελέτη τους; Από τα αποτελέσματα των Table 3 & 4 τι συμπεράσματα διαπιστώνουν.***

Στο πίνακα 4 εμφανίζονται ο min και max βαθμός ανά κατηγορία μαθητών που χρησιμοποίησαν κάθε υλικό (συμπεριλαμβανομένου τουλάχιστον την γενική εικόνα, 1 τουλάχιστον άσκηση και καμία άσκηση), καθώς και ο μέσος βαθμός και η τυπική διακύμανση. Από τα αποτελέσματα του 3 και 4 πίνακα προκύπτει ότι:

- Η στάνταρ προετοιμασία για τις εξετάσεις ήταν τα sample exams και οι λύσεις τους.
- Μαθητές που χρησιμοποίησαν παραπάνω υλικό τα πήγαν καλύτερα.
- Το καλύτερο θετικό πρόσημο ως προς τον βαθμό το είχε το DP.

8 ***Στο Table 5 τι παρουσιάζουν; Διατυπώστε με λόγια 2 κανόνες συσχέτισης. Ποιος είναι ο κανόνας συσχέτισης; Τα αποτελέσματα του Table 5 αφορούν όλους τους φοιτητές που έκαναν χρήση των εκπαιδευτικών πόρων;***

Το table 5 παρουσιάζει τα association rules για τα sample exams αλλά και τις μετρικές αποτίμησης τους το support, confidence, cosine και lift.

- **TrEx01S -> TrEx01:** Με support 0.59, confidence 0.92, cosine 0.87 και lift 1.27 όσοι μαθητές κοιτούσαν τις λύσεις του πρώτου sample exam1 κοιτούσαν και το sample exam1. Αφορά το σύνολο των μαθητών.
- **TrEx02S -> TrEx02:** Με support 0.56, confidence 0.96, cosine 0.9 και lift 1.46 όσοι μαθητές κοιτούσαν τις λύσεις του πρώτου sample exam2 κοιτούσαν και το sample exam2. Αφορά το σύνολο των μαθητών.

Τα αποτελέσματα του table5 διαχωρίζονται σε όλους τους μαθητές (με κανονική γραφή) και στους μαθητές που πήγαν στις εξετάσεις, και αναφέρονται στα sample exams.

9 ***Στο Table 6 τι παρουσιάζουν; Διατυπώστε με λόγια 2 κανόνες συσχέτισης. Προκύπτουν κάποια συμπεράσματα; Υπάρχει ομοφωνία μεταξύ της cosine και lift για όλους τους κανόνες συσχέτισης;***

Το table 6 παρουσιάζει τα association rules για τα attempted exercises αλλά και τις μετρικές αποτίμησης τους το support, confidence, cosine και lift.

- **Ex3->Ex2:** Με support 0.47, confidence 1, cosine 0.8 και lift 1.36, όσοι μαθητές έκαναν την Τρίτη άσκηση είχαν κάνει και την δεύτερη.
- **!Ex5,!Ex6-> Ex1:** Με support 0.63, confidence 1, cosine 0.8 και lift 1, όσοι μαθητές δεν έκαναν τις ασκήσεις 5 και 6, είχαν κάνει την 1.

Υπάρχει ομοφωνία για τα cosine και lift, για όλους τους κανόνες εκτός του !Ex5,!Ex6-> Ex1, δηλαδή το lift=1 που σημαίνει ότι τα γεγονότα πριν και μετά την -> είναι ανεξάρτητα μεταξύ τους, ενώ το cosine=0.8 που δείχνει μια ισχυρή σχετικά συσχέτιση. Όλα τα υπόλοιπα, δείχνουν ότι εφόσων τα lift >1 υπάρχει συσχέτιση, ενώ το cosine είναι πολύ υψηλό, που σημαίνει και αυτό συσχέτιση του X με το Y.

10 ***Στο Table 7 τι παρουσιάζουν; Προκύπτουν κάποια συμπεράσματα; Αναφέρουν «Surprisingly, their average mark is smaller than for all students who have attempted at least 1 exercise» από ποια αποτελέσματα προκύπτει αυτό το συμπέρασμα; Ποιοι είναι οι ΜΟ των βαθμολογιών αυτών των φοιτητών;***

Στο table7 παρουσιάζονται τα στοιχεία για τους βαθμούς των φοιτητών που δοκίμασαν τις εργασίες 4-7. Φαίνεται η min και max βαθμολογίες, το standard deviation και ο μέσος όρος = 38.76/50. Η πτώση των αποτελεσμάτων ανάμεσα σ' αυτούς που έκαναν τις 3 πρώτες ασκήσεις και τις υπόλοιπες, έδωσε την ώθηση στους συγγραφείς να εξετάσουν τον πληθυσμό του πίνακα 7. Αυτό που παρατηρήθηκε είναι ότι, παραδόξως, οι μαθητές που έκαναν τουλάχιστον μία άσκηση, είχαν υψηλότερο μέσο όρο βαθμού από τους μαθητές που έκαναν τις ασκήσεις 4 έως 7.

11 ***Στο Table 8 τι παρουσιάζουν; Προκύπτουν κάποια συμπεράσματα; Υπάρχει συμφωνία ή διαφωνία μεταξύ cosine και lift; Ναι ή όχι και που την αποδίδουν οι συγγραφείς;***

Το table 8 παρουσιάζει τα association rules για τα άλλα εκπαιδευτικά υλικά, πέρα από ασκήσεις και sample exams, αλλά και τις μετρικές αποτίμησης τους το support, confidence, cosine και lift. Σ 'αυτό τον πίνακα οι δύο μετρικές διαφέρουν αισθητά, αφού το cosine, είναι κάτω από το κατώφλι που θεωρείται σημαντική η συσχέτιση του X με το Y (0.66, όπως αναφέρθηκε), ενώ το lift είναι μεγαλύτερο του 1 υποδεικνύοντας συσχέτιση μεταβλητών. Οι συγγραφείς αναφέρουν και δεν γνωρίζουμε, αν το υλικό χρησιμοποιήθηκε από τους ίδιους μαθητές.

12 ***Τι συμπεράσματα προκύπτουν για τον τελευταίο κανόνα συσχέτισης του Table 6. Το 60% από ποια γραμμή του Table 1 προκύπτει; και 43% πως προκύπτει από το Table 2; Εξηγήστε. Ο δάσκαλος με βάση πιο μέτρο ενδιαφέροντος θα αποδεχθεί ως χρήσιμο αυτόν κανόνα συσχέτισης;***

Όπως αναφέραμε και πριν, ο κανόνας αυτός είναι «προβληματικός», καθώς εμφανίζει διαφορετικά αποτελέσματα ανάμεσα στο cosine και το lift. Το cosine (0.8) λέει ότι το 60% των μαθητών που δεν έκαναν τα ex5, ex6 έκαναν το ex1, κάτι που προκύπτει από την γραμμή (60,100,60) του πίνακα 1 -κοιτώντας το αντίστοιχο κατώφλι cosine. Από την άλλη το lift (1) λέει ότι το ποσοστό των μαθητών που δεν έκανε τα ex5,ex6 δεν είναι μεγαλύτερο από αυτούς που έκαναν το ex1, το οποίο αντιστοιχεί από το 43% των μαθητών σύμφωνα με τον πίνακα 2 και πιο συγκεκριμένα από το (21+14) που έκαναν την άσκηση / 81 συνολικοί εγγεγραμμένοι μαθητές (από την στήλη Ex1). Έτσι, ο δάσκαλος καλείται να επιλέξει ανάμεσα σε ποια μετρική του δίνει τα καλύτερα αποτελέσματα για την εκπαιδευτική διαδικασία, και έτσι αν πρέπει να επιλέξει να κρατήσει το association rule. Εν τέλει οι συγγραφείς αναφέρουν ότι το cosine (αρά κρατάμε το rule) είναι σημαντικότερο εν προκειμένω καθώς για την ανάλυση δεν μας ενδιαφέρουν αυτοί που δεν έκαναν την πρώτη εργασία.

13 ***Τι συμπεράσματα προκύπτουν για τον 2ο κανόνα συσχέτισης του Table 8. Το 40% από ποια γραμμή του Table 1 προκύπτει; Εξηγήστε. Αναφέρουν "…10 students who satisfied …" από πού προκύπτει ο αριθμός 10; Εξηγήστε.***

Αντίστοιχα με την ερώτηση 12, έχουμε και εδώ ένα θέμα με τον συγκεκριμένο κανόνα. Σύμφωνα με την απάντηση 11, εμφανίζεται μεγάλη διαφορά στις δυο μετρικές, και σύμφωνα με το cosine του rule έχουμε ότι από τους μαθητές που χρησιμοποίησαν το βιβλίο, το επιπλέον υλικό από το δεύτερο βιβλίο και έκαναν τουλάχιστον μία άσκηση, λιγότεροι από 40% χρησιμοποίησαν το JFlap, κάτι που φαίνεται από την γραμμή (30,100,30) όπου το cosine ποσό μας σε σχέση με αυτό είναι χαμηλότερο με συνέπεια το P(Y)<=30% των παρατηρήσεων. Από την άλλη το lift, μας λέει ότι το ποσοστό μαθητών που χρησιμοποίησαν το JFlap είναι υψηλότερο στους μαθητές που χρησιμοποίησαν τα X (βιβλίο, DP, AtLeast1Ex), σε σχέση με τον συνολικό πληθυσμό κάτι που δείχνει ισχυρή συσχέτιση. Το ποσό 10 προκύπτει το support, που σημαίνει το σύνολο των transactions που περιέχουν X και Y σε σχέση με όλο το σύνολο. Το 10 λοιπόν προκύπτει από το :

$sup(X,Y)=|X,Y|/n => 0.12=|X,Y|/81 => |X,Y|=9.72\approx 10$ με το 81 να είναι το σύνολο των εγγεγραμμένων.

Από αυτό το νούμερο, που είναι πολύ μικρό και έτσι δεν δίνει αρκετή εμπιστοσύνη, προκύπτει πρέπει να ακολουθήσουμε το cosine και να απορρίψουμε το rule. Σε περίπτωση που το |X,Y| ήταν μεγάλο, τότε θα ακολουθούσαμε το lift.

<u>Conclusion</u>

14 ***Πότε ένας κανόνας συσχέτισης θεωρείται ενδιαφέρων με βάση το μέτρο cosine και πότε με βάση το lift;***

Ένα association rule θεωρείται ενδιαφέρων αν:

- Η τιμή του cosine>=0.65, με το κατάλληλο support και confidence.
- Η τιμή του lift>=1, με το κατάλληλο support και confidence.

15 ***Η εκτίμηση κατά πόσο ένας κανόνας συσχέτισης είναι ενδιαφέρων θα πρέπει να γίνεται με ποιο από τα 2 μέτρα;***

Η παραπάνω εκτίμηση θα πρέπει να συνυπολογίζεται και από τα δύο μέτρα. Αρχικά να κοιτάμε αν συμφωνούν ή όχι ως προς το ενδιαφέρον του κανόνα. Ελέγχουμε πρώτα το cosine και μετέπειτα το lift. Αν υπάρχει διαφωνία σ' αυτά τα δύο μέτρα κοιτάμε τα support και confidence. Μικρός πληθυσμός που ακολουθεί τα X,Y δίνει βάρος στο cosine, ενώ μεγάλος πληθυσμός δίνει βάρος στην επιλογή του lift ως μετρική αποτίμησης.

16 ***Τι διαπιστώνουν σε σχέση με τα δεδομένα που μπορούμε να αντλήσουμε από ένα LMS όπως του Moodle; Τι εργαλείο θα βοηθούσε να ενσωματωθεί στα LMS σε σχέση με τη διερεύνηση των δεδομένων με τη βοήθεια κανόνων συσχέτισης;***

Αυτό που προκύπτει από την μελέτη είναι ότι τα συμβατικά και εμπορικά LMS είναι μακριά από το να είναι εύκολα προς χρήση για εξόρυξη δεδομένων, παρότι εδώ βοήθησαν για την παραγωγή συμπερασμάτων για τα association rules. Παραδείγματα αυτής της δύσκολης χρήσης είναι ότι τα log δεδομένα δεν αποθηκεύονται με παρόμοιο τρόπο και χρειάζονται ειδική μεταχείριση, τα στατιστικά που παρουσιάζονται είναι λίγα και δύσκολα ως προς την χρήση. Προτάσεις που γίνονται από τους συγγραφείς, είναι:

- Η ύπαρξη ειδικής πλατφόρμας στα LMS για την εξόρυξη δεδομένων.
- Τα LMS να είναι data mining friendly.
- Να παρέχουν πληροφορίες για association rules με μετρικές όπως το lift και cosine.