Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220586

Discovering Rules from a National Exam Repository: A Use Case for Data Analysis from Iranian Medical Schools Entry Exam

Hanieh ZEHTAB HASHEMI[a,b] , Somayeh ABEDIAN[c,d], Parvane PARVASIDEH[a,1], Zahra BAHREVAR[a] and Sina MADANI[e]

[a] *National Center of Medical Education Assessment, Ministry of Health and Medical Education, Tehran, Iran*
[b]*Department of Health Informatics, Virtual University of Medical Sciences, Tehran, Iran*
[c]*Department of Information Technology Management, Islamic Azad University, Qazvin, Iran*
[d]*Ministry of Health and Medical Education, Tehran, Iran*
[e] *Department of Health IT, Vanderbilt University Medical Center, Nashville, TN, USA*

Abstract Many methods have been studied to analyze and interpret patterns and relationships that are embedded in the database to discover new knowledge in educational systems. Association rule mining is a type of data mining that identifies relationships among elements of the dataset. However, because these methods often generate various rules including non-significant ones, it is important to identify the most useful rules. Therefore, evaluating and ranking rules has become a topic of interest in the decision-making process in order to represent the level of usefulness of rules. We incorporated Apriori and Eclat algorithms on an educational dataset of a national medical exam in Iran. The aim of this study is to identify the usefulness of the extracted rules. This method can reliably discover new knowledge by interpreting the prioritized rules. The results show that those who have Scored in the highest category, i.e. [407,493], are accepted and who have scored in the lowest category, i.e. [150,236), are not accepted in the exam regardless of others features. Although, the rules that implication Accept=0 occurs, find out with high confidence, due to a large number of samples in this case. The ranking rules show this method is effective in the identification of insignificant rules that have no effect on decision making.

Keywords. Educational data mining, Association rules, Residency Education, Data Envelopment Analysis

1. Introduction

Data mining techniques have been used in educational systems to analyze student behavior. The amount of data that is stored in these environments is huge and merits investing in the analysis and interpretation of the information(1). Due to this, we were able to extract important knowledge about socioeconomic diversity that may have been hidden otherwise in the dataset (2). Discovering relationships among variables in data may reveal hidden concepts that can be used next for making better decisions(3).

[1] Corresponding Author, Parvane Parvazideh, National Center of Medical Education Assessment, Ministry of Health and Medical Education, Tehran, Iran.Email : parvane.parvasideh@gmail.com

Association Rule Mining(ARM) is a technique for discovering relations in the dataset(4). The items that are associated with a given rule constitute an item set and if this item set occurs frequently in the database, then we can refer to it as a safe rule.
Finding all frequently occurring item sets involves searching all possible item sets that are a power set on the set of all items i.e. 2^(n-1) possible item set. Thus, with increasing the number of items in the dataset, the power set grows exponentially. Therefore, there is a need to determine a threshold value such as the support to calculate the frequency of item sets(5).

ARM discovers the total number of rules that have certain criteria such as confidence and support. It is important to rank rules since the decision-maker must select appropriate rules based on the relevant business application(6). The mined rules are ranked based on support, confidence, and domain-related measures(7).

In this paper, we investigated an educational dataset used in a national exam repository (medical science exams) with two Association Rule algorithms; Apriori and Eclat. We also used the Data Envelopment Analysis to rank and evaluate extracted rules. Our method designates the most important rules to obtain knowledge in order to make better decisions for politicians in the Ministry of Health and Medical Education. Such an approach enables the analyst to have a better perception in order to have a better understanding of the educational domain. In the following, ARM including Apriori and Eclat algorithms are discussed.

2. Methods and Materials

2.1. Association rules

Association Rule Mining discover the relationships between items from the set of transactions. The interestingness of associaton rules are measured by support and confidence. Association rules are regarded as interesting if their support and confidence are greater than the user-specified minimum support and minimum confidence(8).
Several algorithms are available for mining frequently item sets. The Apriori algorithm is an algorithm for mining frequent item sets and uses the minimum support criterion to eliminate infrequent itemsets(9). Since the Apriori algorithm is the first algorithm that was proposed in the domain, it has been improved upon in terms of computational efficiency.

The Eclat algorithm is a more efficient and scalable version of the Apriori algorithm. While the Apriori algorithm works in a horizontal sense imitating the Breadth-First Search of a graph, the Eclat algorithm works in a vertical manner just like the Depth-First Search of a graph. This vertical approach of the Eclat algorithm makes it a faster algorithm than the Apriori algorithm(10).

3. Results

The dataset that we used for our study comes from a national medical exam database that contains students' information. The participants of the exam are students who have passed the primary exam and are moving to the next level. This exam is held annually in Iran with 14,000 students competing for top seats. The students must first sit in this admission exam in order to be eligible for entering university and higher-level education.

Students who were absent from the exam are eliminated from our study. Exam score and several other factors, that determine the students acceptance in the test, were the subject of our study. This exam has two steps: first, students have to answer 200 multiple-choice questions. Each correct and incorrect answer earns 3 positive and 1 negative point, respectively. The volunteers who gain at least 150 out of the total raw score, can select multiple fields of study in their favorite order. Participants who were absent are removed from the next step of the exam. Also, people who have not received an acceptable score (at least 150) cannot move to the next step(11). Both of these two groups were eliminated from our study as well.

Table 1. Selected features

No.	Feature name
2	Quota
3	Sex
6	Military
7	Internship
10	Booklet code
11	Score
12	Pre-internship
13	Average
14	Military staff
15	Military staff scholarship
21	Elite quota
22	Army scholarship

Since the national data repository was composed from several local databases it had to be pre-process first. Missing and duplicate values were removed as well. Data types include both integer and string. The final dataset for our study contained information from 7,723 participants who were accepted in the first phase of the national exam. We have conducted our experiments in three steps. First, As(12) has selected a list of features that are important for this study and has shown in table 1. Then, on such smaller dataset, we used Apriori and Eclat algorithms to generate association rules based on a predefined minimum support and confidence values. Then we used Eclat algorithm since it was more suitable for our sequential data compared to the Apriori algorithm. Some of the relationships, that can be discovered by Eclat approach, may not be found by Apriori approach due to the methodological differences in their algorithms. The results Eclat algorithm is shown in figure 1,2 respectively.

The Apriori algorithm was compiled in Weka environment, whereas for Elcat we used a rules package in R programming language. During data pre-processing, all numeric variables were converted into nominal values, and for some like Score we did more classification (4 categories for Score). The last step was to interpret association rules extracted by Eclat algorithm. The minimum support and confidence values set to 0.1 and 0.9, respectively. We extracted rules with the value of 1 in the class label (Accept=1) and the top 15 values are depicted in figure 1. The first rule shows that when Score belongs to [322,407) range and Army scholarship is 1, then Accept=1 with a confidence score of 0.96% , i.e. students who Score belong to this range and don't have Army scholarship, but have been accepted in this exam. The second rule implies that when Score belongs to [322,407) interval, it is also Accept=1. According to the principle of "rules for proof by cases" logic, we can infer that the first rule can be ignored since the second one is suffice. The person who has Score belonging to [322,407) has been accepted with confidence 0.96%. The next rule implies when Score belongs to this range and Army scholarship=1 and Elite quota=1, i.e. students do not have Army scholarship and Elite quota, then it concludes Accept=1. Subsequent rule shows that by removing unimportant feature like Army scholarship=1, we can infer who has Score in this range and has not Elite quota, and been accepted. Also, we found who has a Score in this range with Quota =1, and has been accepted.

By reviewing these rules with Accept=0 in Eclat, the following results are found: When the Score belongs to [150,236) and Quota=1 then Accept=0. This means those who are Score at a lower level and do not use any Quota, are not accepted in exam. There is another rule that says Score belongs to [150,236) and Quota=1 and Army scholarship=11 then Accept=0. This means even if the Army scholarship is used, the result does not change.

rules	support	confidence
{Score=[322,407),Army.scholarship=1} => {Accept=1}	0.104752	0.964243
{Score=[322,407)} => {Accept=1}	0.108119	0.964203
{Score=[322,407),Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.103587	0.963855
{Score=[322,407),Elite.quota=1} => {Accept=1}	0.106953	0.963827
{Quota=1,Score=[322,407),Army.scholarship=1} => {Accept=1}	0.103457	0.963812
{Score=[322,407),Military.staff=0,Military.staff.scholarship=1,Army.scholarship=1} => {Accept=1}	0.103328	0.963768
{Score=[322,407),Military.staff=0,Army.scholarship=1} => {Accept=1}	0.103328	0.963768
{Score=[322,407),Military.staff.scholarship=1,Army.scholarship=1} => {Accept=1}	0.103328	0.963768
{Score=[322,407),Military.staff=0,Military.staff.scholarship=1} => {Accept=1}	0.106694	0.963743
{Score=[322,407),Military.staff.scholarship=1} => {Accept=1}	0.106694	0.963743
{Score=[322,407),Military.staff=0} => {Accept=1}	0.106694	0.963743
{Quota=1,Score=[322,407)} => {Accept=1}	0.106694	0.963743
{Quota=1,Score=[322,407),Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.102292	0.963415
{Score=[322,407),Military.staff=0,Military.staff.scholarship=1,Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.102162	0.96337
{Score=[322,407),Military.staff=0,Elite.quota=1,Army.scholarship=1} => {Accept=1}	0.102162	0.96337

Figure 1. Apriori rules extracted

```
Best rules found:

 1. Quota=1 Score='(-inf-235.75]' Military staff=0 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 2. Quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 3. Quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 4. Quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Elite quota=11 Army scholarship=11 3144 ==> Accept=0 3053    conf:(0.97)
 5. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 6. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 7. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 8. Quota=1 Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Elite quota=11 Army scholarship=11 3102 ==> Accept=0 3012    conf:(0.97)
 9. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
10. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
11. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
12. Fighter quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 Elite quota=11 Army scholarship=11 3123 ==> Accept=0 3025    conf:(0.97)
13. Quota=1 Score='(-inf-235.75]' Military staff=0 3260 ==> Accept=0 3155    conf:(0.97)
14. Quota=1 Score='(-inf-235.75]' Military staff=0 Military staff scholarship=1 3260 ==> Accept=0 3155    conf:(0.97)
15. Quota=1 Score='(-inf-235.75]' Military staff=0 Elite quota=11 3259 ==> Accept=0 3154    conf:(0.97)
```

Figure 2. Eclat rules extracted

4. Conclusion

In this study, we focused on an educational dataset containing the results of a national medical school entry exam. The repository included information about students themselves as well as exam results. We used existing algorithms to discover patterns and relationships that may have been hidden from the subject matter experts and exam developers. Therefore, leveraging relationship mining techniques is a vital task for uncovering knowledge from existing data and establishing a learning knowledge management cycle. The results show that those who have Scored in the highest category, i.e. [407,493], are accepted and who have a Score in the lowest category, i.e. [150,236), are not accepted in the exam regardless of others features. Although, the rules that implication Accept=0 occurs, find out with high confidence, due to a large number of samples in this case. The ranking rules show this method is effective in the identification of insignificant rules that have no effect on decision making.

References

[1] Singh A, Rocke S, Pooransingh A, Ramlal CJJIToE. Improving student engagement in teaching electric machines through blended learning. 2019;62(4):297-304.
[2] Gupta A, Garg D, Kumar PJIToE. Analysis of students' ratings of teaching quality to understand the role of gender and socio-economic diversity in higher education. 2018;61(4):319-27.
[3] Swart AJJIToE. Evaluation of final examination papers in engineering: A case study using Bloom's Taxonomy. 2009;53(2):257-64.
[4] Agrawal R, Imielinski T, Swami AJItok, engineering d. Database mining: A performance perspective. 1993;5(6):914-25.
[5] Zhao Y, Zhang C, Cao L. Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction: Techniques for Effective Knowledge Extraction: IGI Global; 2009.
[6] Buchanan RA, Forster DJ, Douglas S, Nakar S, Boon HJ, Heath T, et al. Philosophy of Education in a New Key: exploring new ways of teaching and doing ethics in education in the 21st century. 2021:1-20.
[7] Obata T, Ishii HJEJoOR. A method for discriminating efficient candidates with ranked voting data. 2003;151(1):233-7.
[8] Agrawal R, Imieliński T, Swami A, editors. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data; 1993.
[9] Agrawal R, Srikant R, editors. Fast algorithms for mining association rules. Proc 20th int conf very large data bases, VLDB; 1994: Citeseer.
[10] Zaki MJJItok, engineering d. Scalable algorithms for association mining. 2000;12(3):372-90.
[11] Heidarzadeh A, Zehtab Hashemi H, Parvasideh P, Bahrevar ZJJoM, Cultivation. A Historical Review on the Achievements and Challenges of the Residency exam in a Forty Years Trend (1357-1397) in Iran. 2019;28(3):11-20.
[12] Hashemi HZ, Parvasideh P, Larijani ZH, Moradi F, editors. Analyze Students Performance of a National Exam Using Feature Selection Methods. 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE); 2018: IEEE.