Research Journal of English Language and Literature (RJELAL)
A Peer Reviewed (Refereed) International Journal
http://www.rjelal.com

Vol.3.2.2015 (Apr-June)

RESEARCH ARTICLE

INTERNATIONAL STANDARD SERIAL NUMBER INDIA
2395-2636 (Print):2321-3108 (online)

PREDICTING STUDENTS' SCORE IN THE ENGLISH SECTION OF THE BA ENTRANCE EXAM BY APRIORI ALGORITHM

MAYSAM HOSSEINI[1*], OMID AKBARI[2], ALIREZA RAHMANI[3], ABOLQASEM SHAKERI[4]

[*1,3&4]MA Student, Department of English Language, Imam Reza International University, Mashhad, Iran

[2]Assistant Professor, Department of English Language, Imam Reza International University, Mashhad, Iran

Article Info:
Article Received:05/06/2015
Revised on:20/06/2015
Accepted on:30/06/2015

ABSTRACT

Assessment is essential not only to guide the development of individual students but also to monitor and continuously improve the quality of programs and students' performance, and inform their parents. The purpose of this study is predicting students' score in the English section of the BA entrance exam. For this purpose, 485 male high school students studying English at several different high schools were chosen from a total number of 520 through their performance on a pilot test including 25 multiple choice questions of Kanoon Farhangi Amoozesh. A questionnaire adopted from Gardner's Attitude/Motivation Test Battery was used and the students were asked to complete it. Some attributes with their values were extracted from the responses of the questionnaire and the students' score in the pilot test. A database was prepared based on the attributes and data preprocessing was applied. After preprocessing, Apriori algorithm was used for extracting association rules. The result of this study showed 15 interesting extracted rules which predict students' score in the English section of the BA entrance exam.

Keywords: assessment; English testing; Apriori algorithm

1. INTRODUCTION

One of the biggest challenges that education faces today is how to have an accurate system for assessing students' performance. Despite the importance of assessments in education today, few teachers receive much formal training in assessment design or analysis. Lacking specific training, teachers rely heavily on the assessments offered by the publisher of their textbooks or instructional materials. When no suitable assessments are available, teachers construct their own in a haphazard fashion, with questions and essay prompts similar to the ones that their teachers used. They treat assessments as evaluation devices to administer when instructional activities are completed and to use primarily for assigning students' grades. Therefore, there should be a tool to fill the gaps between students' performance and its evaluation.

Data mining is the method of analyzing data from different perspectives to discover interesting and helpful information. There are some differences between statistics and data mining. In statistics, the knowledge is not hidden rather you directly are able to observe the knowledge on your own. Statistics only lets you prove your observation (hypothesis) scientifically so that the community accepts your hypothesis. On the contrary, data

mining is an exploratory tool. You have no idea about the hidden knowledge inside the data. Data mining lets you to "discover" that invisible knowledge.

The information gained through data mining has been effectively used in various sectors ranging from finance, agriculture to health and education. There are many data mining tools available that allow users to analyze data from many different aspects, categorize it, and discover the identified relationships. Technically, data mining is a technique of finding correlations or patterns among many fields in large databases. Educational data mining is fast becoming an interesting research area which allows researcher to extract useful, previously unknown patterns from the educational databases for better understanding, improved educational performance and assessment of the student learning process (Chan, Chow, & Cheung, 2008).

The selection of data mining tools and techniques mostly depends on the scope of the problem and the expected results from the analysis. For example, a classification approach is used (Minaei Bidgoli et al., 2003) to classify students to predict their final year performance based on different parameters derived from the data in an educational web-based system. A clustering algorithm is used (Tsai, Tseng, & Lin, 2009) to categorize students with similar behavioral characteristics. Association rule mining techniques have frequently been used to solve educational problems and carry out critical analysis in an academic environment for improving the learning process of student. These efforts are carried out in order to raise the standards and administration of educational processes by investigating the learning systems, learning resources arrangements, and students' results, curriculum restructuring, and institutional websites (Damasevicius, 2009; Talavera, & Gaudioso, 2004; Erdogan, & Timor, 2005).

A very comprehensive review of data mining in education from 1995 to 2005 is published in 2007 by Romero and Ventura. One of the educational problems that are solved with data mining is the prediction of students' academic performances, whose goal is to predict an unknown variable (outcome, grades or scores) that describes students. The estimation of students' performances includes monitoring and guiding students through the teaching process and assessment.

Since data mining represents the computational data process from different perspectives, with the goal of extracting implicit and interesting samples, it can greatly help every participant in the educational process in order to improve the understanding of the teaching process, and it centers on discovering, detecting and explaining educational phenomenon's (El-Halees, 2008).

The present study adopts Apriori algorithm to mine rules for predicting students' score in the English section of the BA entrance exam.

1. Literature review

Data mining has attracted a great deal of attention in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining can be viewed as a result of the natural evolution of information technology.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information.

There have been done some studies in the area of data mining in education. Each of them is trying to enhance the educational system by discovering patterns among the great deal of data.

Kumar and Vijayalakshmi (2011) applied decision tree algorithm on student's internal assessment data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to fail or pass.

Zekić-Sušac, Frajman-Jakšić and Drvenkar (2009) created a model for predicting students' performance using neural networks and classification trees decision-making, and with the analysis of factors which influence students' success.

Very few studies have focused on predicting students' score in the exam. The present study tries to fill this research gap by focusing on predicting the score through Apriori algorithm.

3. Material and methods

3.1 *Participants*

In this study, 485 out of 520 male high school students studying English in the field of experimental sciences at several different high schools in Razavi Khorasan were randomly chosen through a pilot test including 25 multiple choice questions of Kanoon Farhangi Amoozesh (Ghalamchi Educational Foundation). All the students were in the 12th grade of high school.

3.2 Data Collection

To collect data, a questionnaire adopted from Gardner's Attitude/Motivation Test Battery (AMTB) was used. Gardner (1985) designed a test battery known as the Attitude and Motivation Test Battery (AMTB). It includes some items measuring all factors that affect attitude and motivation. In AMTB, the concept of attitude is incorporated in motivation meaning that positive attitudes increase motivation. Integrative and Instrumental Orientation scales of the original 6-point Likert Scale format of Gardner's AMTB (Gardner, 1985) were used, ranging from 'Strongly Agree' to 'Strongly Disagree'. The questionnaire has 104 items. The AMTB is reported to have good reliability and validity (Gardner, 1985). The students were asked to complete

the questionnaire in the class during a session. They were also asked to check the questions carefully, read them thoroughly and if there were some questions regarding the comprehension of the questions, they were allowed to ask them. Respondents had enough time to complete the task and all the questionnaires were collected at the end of the session. There was no missed or distorted questionnaire. Respondents were informed that the information they gave would be kept confidential and used only for research purposes.

3.3 Data Extraction

Having completed the questionnaire, data is extracted from them. At this stage, attributes equivalent and proportional to the questions are extracted based on the questions on the questionnaire and the students' score in the pilot test. To further organize and create a database from a set of questionnaires, one record for each completed questionnaire is defined in an Excel file and this file is completed according to the attributes. After completing this section, a database is prepared in the form of an Excel file. The attribute set before data preprocessing is shown in Table 1. Since we need scales ranging from 'very high' to 'none', 6-point Likert scales were equalized like this:

Strongly agree = very high
Moderately agree = high
Slightly agree = medium
Slightly disagree = low
Moderately disagree = very low
Strongly disagree = none

Table 1: The attribute set before data preprocessing

Attributes	Values					
Anxiety in situations of without accountability	Very High	High	Medium	Low	Very Low	None
Communication with foreigners						
Interested in the future of employment						
Interested in reading English magazines						
The importance of updating knowledge						
Academic success						
Interested in the English language						

Parent's encouragement for learning English						
The importance of English homework						
Aptitude in learning English						
Teacher's encouragement for learning English						
Having an English teacher fluent in English						
Not using an English teacher for learning English						
The importance of the aspects of language learning						
Lack of interest in teachers of English						
Interested in gaining the highest score in the International English Language Test						
Interested in the English-speaking people						
Decreasing amount of learning with the passage of time						
Interested in going abroad						
Not interested in learning English						
Interested in speaking English						
Score of the pilot test of the Evaluation Organization	0-20	21-40		41-60	61-80	81-100

1.1 Data Preprocessing

After completing the database and understanding concepts related to data, preprocessing is commenced. In this section, a sequence of operations is performed which removes various problems related to the collected data. Preprocessing includes several steps stated in the following.

1.1.1 Data Cleaning

One of the most important factors that decrease the quality of the data is missing values. Some values of attributes may be null for some reasons. These values are called missing values. For handling missing values in this step, records with null attributes are excluded.

1.1.2 Data Reduction

The extra attributes such as full name, address and telephone number not influencing variables are removed.

1.1.3 Discretization

All continuous attributes in the dataset are transformed to discrete attributes. For example, the attribute of score is a continuous attribute which transformed to three intervals including {0-14}, {15-21}, and {22-30}.

1.1.4 Transforming Attribute Values

In this step, all attributes are transformed to binary attributes. In this case, each attribute is broken to some binary attributes. In fact, after discretization, continuous attributes are broken to several class ones. Then, each attribute has several class values and each class value has a binary attribute. Finally, a new database is created in which

all attributes are transformed to binary ones. Each attribute is binary and has a certain concept.

2. Data Attributes after Preprocessing

2.1 discovery of association rules

Association rules are one of the main techniques of data mining and almost the most important technique of discovery and extraction of learning patterns (Han and Kamber, 2006). Association rules discover interactive communication between a large collection of data and this communication helps decision-makers. In fact, association rules demonstrate situations in which frequent datasets occur with each other. Indeed, the extracted rules describe the existence of some attributes based on other attributes.

In this study, Apriori algorithm is used for extracting association rules.

2.2 Apriori Algorithm

Apriori is one of the most important findings in the history of mining association rules since its introduction (Geetha and Mohiddin, 2013). A series of basic concepts should be defined before the introduction of algorithms for mining association rules.

1. Item sets included in a database is shown by Itemset={X1, X2, …}.
2. For each rule which is X→Y, two values of support and confidence is determined.

2.1 Support is the continuously simultaneous possibility of X and Y in the transaction. $Sup(x \rightarrow y) = p(x \cup y)$

2.2 Confidences are the conditional probability in that the transaction with X has Y, too.

$Conf(x \rightarrow y)=p(y|x)=sup(x \cup y)/supp(x)=p(x \text{ and } y)/p(x)$

Therefore, the rule X→Y with (S=50%, C= 66.7%) signifies that X and Y continuously exist in 50% of all transactions and in 66.7% of all transactions, wherever there is X in the transaction, there is Y, too.

The working process of algorithm includes: First, all frequent item sets with one member are found. Then, according to them, all frequent item sets with two members are found. Next, frequent item sets with three members are found based on frequent item sets with two members and this process continues like this until no larger frequent item sets are found. The flowchart of this algorithm is shown in Figure 1.

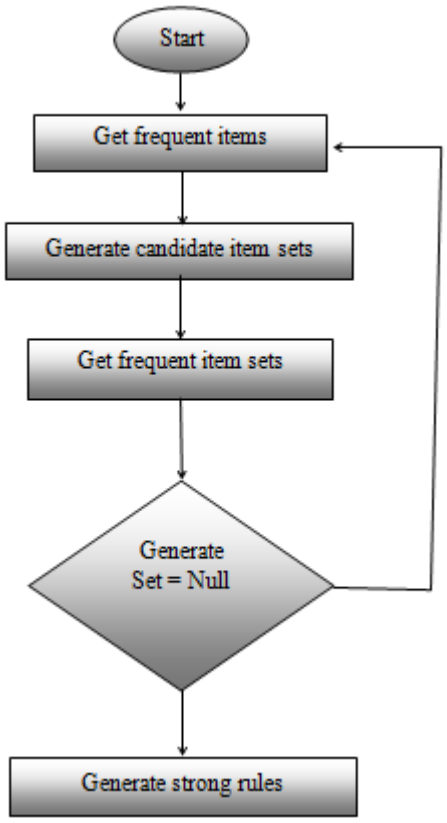

Figure 1: Flowchart of the algorithm

This algorithm operates in two steps including join and prune. These two steps will be discussed in details below.

Let L be a set and L_k is each member of L with size K. Therefore, L_{K-1} is used to obtain L_k. These item sets are C_K and the step in which it is gained is called join step. The next step is to add these items to the previous sets; of course, if the items are frequent and all its subsets are frequent. This step is called prune step.

2.2.1 Join Step

It should be ensured that items are arranged in alphabetical order. L_{K-1} its members are shown by l_i and $l_i[j]$. Here, i represents the number of the set and j is the number of items in the set. If K-2 of the first two sets is equivalent, two sets of L_{K-1} can be joined together. It means that:

$(l1[1] = l2[1]) \& (l1[2] = l2[2]) \& \ldots (l1[k-2] = l2[k-2]) \& (l1[k-1] < l2[k-1])$

Note that two items are arranged in alphabetical order and it prevents existing frequent items.

Therefore, the combination of the obtained set is the last K member of the second set in terms of their order.

P in L_{K-1} = (1 2 3)

Q in L_{K-1} = (1 2 4)

Join: Result in C_K = (1 2 3 4)

2.2.2 Prune Step

C_K (a candidate item set) is the collection of L_K with members either frequent or non-frequent; but, there are all frequent members in it. All items of C_K must be investigated whether they are frequent or not. However, there may be a majority of them. Thus, the origin of Apriori is used in order to reduce the amount of computation. It means that if one of the subsets is not frequent, that item set will not be frequent, too. Now, non-frequent item sets must be separated in order to find frequent item sets, so that if a member of C_K has K-1 subsets in which there is not L_{K-1} , that member of C_K will not be frequent.

This algorithm separately produces C_K by joining large frequent items obtained the previous step and excluding other existed items in the previous steps. Thus, the number of C_K decreases significantly. C_K is calculated according to pseudo code provided in Figure 2.

Apriori-gen (L_{k-1})
1. *Join step*
2. *insert into C_k*
3. *select p.item1, p.item2 ,… , p.item$_{k-1}$, q.item$_{k-1}$*
4. *from L_{k-1} p, L_{k-1} q*
5. *where p.item1=q.item1, …,p.item$_{k-2}$ = q.item$_{k-2}$, p.item$_{k-1}$ < q.item$_{k-1}$*
6. *Prune step*
7. *forall itemsets c C_k do*
8. *forall (k-1)-subsets s of c do*
9. *if (s L_{k-1}) then*
10. *delete c from C_k*

Fiqure 2: Pseudo code of CK

Example: assume that L_3 is like this:

L_3 = {{1 3 5} {2 3 4} {1 3 4} {1 2 4} {1 2 3}}

After join step, we have:

C_3 = {{1 3 4 5} {1 2 3 4}}

And after prune step, we have:

L_4 = {1 2 3 4}

After calculating $C_{k,}$ values of supports for its each member are calculated and those with minimum supports are placed in $L_{k.}$ Figure 3 shows pseudo code of the algorithm in general.

The Apriori Algorithm
1. *L1 = {large 1–itemsets}*
2. *for (k = 2; $L_{k-1} \neq \varnothing$; k++) do begin*
3. *Ck = apriori – gen (L_{k-1});*
4. *forall transactions t ∈D do begin*
5. *Ct = subset (C_k,t)*
6. *for all candidates c ∈Ct do*
7. *c.count++;*
8. *end*
9. *end*
10. *Lk = {c ∈C_k | c.count ≥ minsup}*
11. *End*
12. *Answer=$U_K L_K$*

Figure 3: Pseudo code of Apriori algorithm

After extracting large item sets with acceptable frequency, association rules must be extracted.

For each frequent set L, all subsets, all non-empty subsets are written. For each set obtained by S, the below rules are written, too.

"s $\rightarrow$ (L - s)"

Then, the value of confidence is calculated and if it is more than the acceptable value, it will be accepted.

Confidence (A $\rightarrow$B) = P (B |A) = support_count(A$\cup$ B) / support_count(A)

3. Findings

In this study, 15 interesting rules were extracted including:

Rule 1 (support = 0.2514, confidence= 1) indicates that learners who have low interest in learning English, medium aptitude in learning English, their score will be 21-40.

Rule 2 (support = 0.2153, confidence= 1) indicates that learners who have very low interest in not using an English teacher for learning English, medium interest in aspects of learning English, their score will be 41-60.

Rule 3 (support = 0.2617, confidence= 1) indicates that learners who have very high interest in using a teacher for learning English, receive their parents' very high encouragement, their score will be 81-100.

Rule 4 (support = 0.2641, confidence= 1) indicates that learners who give very high importance to their

academic success, have medium interest in learning English, their score will be 61-80.

Rule 5 (support = 0.2126, confidence= 0.98) indicates that learners who give medium importance to aspects of learning English, very high importance to have a teacher fluent in English, their score will be 41-60.

Rule 6 (support = 0.1628, confidence= 0.98) indicates that learners whom their high amount of learning decreases with the passage of time, have low interest in using a teacher for learning English, their score will be 0-20.

Rule 7 (support = 0.1557, confidence= 0.98) indicates that learners who have medium anxiety in situations of without accountability, receive teacher's high encouragement for learning English, their score will be 61-80.

Rule 8 (support = 0.1891, confidence= 0.98) indicates that learners who have medium aptitude in learning English, receive teacher's very high encouragement for learning English, their score will be 61-80.

Rule 9 (support = 0.1685, confidence= 0.96) indicates that learners who have low interest in communicating with foreigners, very low interest in reading English magazines, their score will be 0-20.

Rule 10 (support = 0.1914, confidence= 0.95) indicates that learners who have very high Interest in the future of employment, low interest in their English teacher, their score will be 41-60.

Rule 11 (support = 0.1952, confidence= 0.95) indicates that learners who have low aptitude in learning English, very high interest in reading English magazines, their score will be 21-40.

Rule 12 (support = 0.1924, confidence= 0.94) indicates that learners who have high interest in gaining the highest score in the International English Language Test, give low importance to aspects of learning English, their score will be 81-100.

Rule 13 (support = 0.1857, confidence= 0.93) indicates that learners who receive their parents' high encouragement, have medium aptitude in learning English, their score will be 61-80.

Rule 14 (support = 0.1691, confidence= 0.92) indicates that learners who give medium importance of updating their knowledge, have low interest in reading English magazines, their score will be 21-40.

Rule 15 (support = 0.1935, confidence= 0.90) indicates that learners who have medium interest in going abroad, give high importance to doing their English homework, their score will be 81-100.

4. Discussion

The present study aims to predict the students' score in the English section of the BA entrance exam by using Apriori algorithm. Apriori is the classical and most famous algorithm. Objective of using Apriori algorithm is to find frequent item sets and association between different item sets i.e. association rule. It means that this algorithm finds a relationship among the responses of the questionnaire and the students' score in the pilot test. Therefore, 15 rules were extracted based on this relationship.

According to the Finding section, rules 1, 2, 3, and 4 have confidence 1. It means that if A occurs, there will certainly be B. Thus, if a student has a low interest in learning English and a medium aptitude in learning English, his score will certainly be 21-40 (rule 1). According to the rule 2, if a student has a very low interest in not using an English teacher for learning English and a medium interest in aspects of learning English, his score will certainly be 41-60. Rule 3 shows that if a student has a very high interest in using a teacher for learning English and receives his parents' very high encouragement, his score will certainly be 81-100. If a student gives very high importance to his academic success and has a medium interest in learning English, his score will certainly be 61-80.

Although the confidence of the other rules is not 1and their occurrence is not assured, they received a confidence between 0.98 and 0.90 which is remarkable.

The rules discovered in this perspective do confirm some findings from previous studies (Patrick, 2007; Walberg, 2011).

The advantage of these rules is that the students can assess themselves based on the attributes in this study. So, we can conclude that our paper is an appropriate database to be used by all participants in the BA entrance exam.

5. Conclusion

As was mentioned previously the present study was conducted in order to predict the students' score in the English section of the BA entrance exam. In so doing, Apriori algorithm was used because it is most widely used algorithm in the history of association rule mining that uses efficient candidate generation process and represents the computational data process from different perspectives, with the goal of extracting implicit and interesting samples. The result of this study is 15 extracted rules which predict the students' score based on the given attributes and their values. This study is a special database and all participants in the university entrance exam can use it to be aware of their English score before taking the test and improve their performance.

6. Acknowledgements

The authors would like to express their special gratitude to a number of the colleagues without whom this research could not have been accomplished.

7. References

[1] Chan.A.Y.K, Chow. K.O, and Cheung. K.S. (2008). "Online Course Refinement through Association Rule Mining", *Journal of Educational Technology Systems*, Volume 36, Number 4/2007 – 2008, pp 433 – 444.

[2] Damasevicius. R. (2009). "Analysis of Academic Results for Informatics Course Improvement using Association Rule Mining". Information Systems Development towards a Service Provision Society. ISBN 978-0-387-84810-5 (print), pp 357 – 363, published by Springer US.

[3] El-Halees, A. (2008), Mining students data to analyze learning behavior: a case study, Available-on:http://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/30/papers/f158.pdf.

[4] Erdogan. S. Z, m. Timor. (2005) "A Data Mining Application in a Student Database". *Journal of Aeronautics and Space Technologies*, Vol. 2, Number 2., pp 53 – 57.

[5] Gardner, R.C. (1985). Social psychology and language learning: The role of attitudes and motivation. London: Edward Arnold.

[6] Geetha. K, Mohiddin. Sk. (2013). "An Efficient Data Mining Technique for Generating Frequent Item Sets", In: Proceeding of IJARCSSE, ISSN 2277-128X, Vol. 3, Issue 4.

[7] Han J, Kamber M . (2006). Data Mining: Concepts and Techniques. 3rd ed. UK: Morgan Kaufmann.

[8] Kumar, S. A., & Vijayalakshmi, M. N. (2011, October). Efficiency of decision trees in predicting student's academic performance. In First International Conference on Computer Science, Engineering and Applications, CS and IT (Vol. 2, pp. 335-343).

[9] Minaei Bidgoli. B, Kashy. B.A, Kortemeyer. G, and Punch. W. F. (2003). "Predicting Students Performance: an application of data mining methods with the educational web-based system LON-CAPA", in proceedings of ASEE/IEEE Frontier in Education Conference, Boulder, CO: IEEE.

[10] Patrick, A. O. (2007). Motivation Effects on Test Scores of Senior Secondary School. International Journal of Educational Sciences, pp. 57-64.

[11] Romero, C. & Ventura, S. (2007), Educational Data Mining: a Survey from 1995 to 2005, Expert Systems with Applications, Elsevier, pp. 135-146.

[12] Talavera, L., Gaudioso, E. (2004). "Mining Students Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces". In proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI , Valencia, Spain.

[13] Tsai. G.J, Tseng. S.S, and Lin. C.Y. "A Two Phase Fuzzy Mining and Learning Algorithm for Adaptive Learning Environment". In proceedings of the Alexandrov, V.N. et al. (eds.)

[14] Walberg, S. M. (2011). Motivational Effects on Test Scores of Elementary Students. The Journal of Educational Research, 133-136.

[15] Zekić-Sušac M., Frajman-Jakšić A. & Drvenkar N. (2009), Neuron Networks and Trees of Decision-making for Prediction of Eficiency in Studies, Ekonomski vjesnik, Vol.No.2, pp. 314-327.