Fuzzy Association Rule Mining based Model to Predict Students’ Performance

Sushil Kumar Verma¹, R.S. Thakur², Shailesh Jaloree³
¹Department of Computer Applications, SATI Vidisha, India
²Department of Computer Applications, MANIT Bhopal, India
³Dept. of Applied Maths and Computer Science, SATI Vidisha (MP), India

ABSTRACT
The major intention of higher education institutions is to supply quality education to its students. One approach to get maximum level of quality in higher education system is by discovering knowledge for prediction regarding the internal assessment and end semester examination. The projected work intends to approach this objective by taking the advantage of fuzzy inference technique to classify student scores data according to the level of their performance. In this paper, student’s performance is evaluated using fuzzy association rule mining that describes Prediction of performance of the students at the end of the semester, on the basis of previous database like Attendance, Midsem Marks, Previous semester marks and Previous Academic Records were collected from the student’s previous database, to identify those students which needed individual attention to decrease fail ration and taking suitable action for the next semester examination.

1. INTRODUCTION
Traditional statistical techniques and data base management tools are application oriented and are not suitable for analysis of large amount of data, because in the world a huge quantity of data are collected and stored daily. Study of such data is an important need. This need can be fulfill by the use of data mining [1]. Data mining has become the area of growing significance because it helps in analyzing data and summarizing it in to useful information. The notion of data mining is the technique of extracting previously unknown information with the widest relevance from database, in order to use it in the decision making process [2]. Data mining is the process of extracting previously unknown, valid, potentially useful and hidden patterns from large datasets [3].

Data mining is a new concept in education domain. This concept can be helpful as a work of bridge between this lacks of knowledge. Data mining is used to apply two different processes of knowledge discovery and prediction. Knowledge discovery provides information which has a readable form and the concept of prediction gives the prediction of future events [4]. Data mining provides many techniques that can be used to study the performance of students. There are increasing research interests in education. This is a new field, called “educational data mining”, involves with developing techniques that discover new information from the data originating from educational environments [5]. Data mining techniques are analysis tool that can be used to extract meaningful knowledge from large data sets.

Benefits of Educational Data mining are following [1-6].

a. Improving the student’s performance based on extracting knowledge.
b. Analyzed student academic data and enhance the quality of educational system.
c. Improving student’s learning process.
d. Improve research capability in the educational field.
e. Time saving during knowledge extraction for educational field.
f. More effective sharing of information in educational area.
g. Select student groups with similar characteristics and reactions to learning strategies.
h. Lack of deep and enough knowledge in educational system may prevent system management to achieve quality objectives, Data mining do the work of bridge for this knowledge gaps in education system.
i. Data mining techniques can be utilized effectively in selecting course, managing and improving students’ attendance, providing the knowledge for supplementary classes where necessary.
j. Data mining is useful for analyzing students’ data for predicting their learning behavior and to warn students at risk before their final exams.

The present work intends to approach this objective by taking the advantage of fuzzy inference technique in order to classify student scores data according to the level of their performance. In this paper, student’s performance is evaluated using fuzzy association rule mining. For fuzzy association mining we used modified Apriori like method which is simply based on matrix and vector multiplication approach and ensures that all patterns can be discovered in a fast manner. This paper is organized as follows. Section 2 explains related work in education with the help of data mining. Section 3 explains rules generated by fuzzy association mining. Section 4 explains performance parameters. Section 5 shows results and section 6 is conclusion.

2. LITERATURE REVIEW

This section presents various existing works in the area of educational data mining. Most of existing works is based on machine learning and mining techniques. In 2000 Han and Kamber [6] describes data mining software that allows the users to analyze data from different dimensions, categorize it and summarize the relationship which are identified during the mining process. They explained classification technique for predict the related subject in a course curriculum. This information can be used to improve the syllabus of any course in their educational institute. In 2011 Pandey et.al. [7] explained their study based on the performance of 60 students. They used Bayes classification on belongs to category, language and background details such as qualification. They were found that whether new students will perform or not. In 2012 Azhar Rauf et.al. [8] suggested a method known as k- means clustering algorithm for students’ data, it calculates initial centroids instead of random selection, due to which the number of iterations is reduced and elapsed time is improved for find the students’ cluster. In 2005 Khan [9] described a performance study on 400 students selected from the senior secondary school of Aligarh Muslim University, Aligarh, India. Their objective was to establish the predictive value of different measures of cognition, individuality and demographic variables for success at higher secondary level in science stream. This was based on cluster sampling technique in which the all population of interest was divided into homogenous groups and a random sample of these groups was selected for further analysis. They found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general stream. In 2007 Galit et.al. [10] gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams.

3. RESEARCH METHOD

The fuzzy association rule mining is divided into three steps. Fuzzy sets are generated first, followed by discovering fuzzy frequent Itemsets form the newly constructed database [11-13]. Finally, fuzzy association rules are generated and evaluated. Figure 1 shows the schematic view of frequent closed itemset discovery.

![Diagram of Fuzzy Association Rule Mining Process]

IJECE Vol. 7, No. 4, August 2017 : 2223 – 2231
3.1. Constructing Fuzzy Sets

During the mining attributes were considered as linguistic variables which include Previous Academic Record (PAR), Previous Semester Marks (PSM), Midsem Marks (MSM), Attendance (ATT) and End Semester Marks (ESM). ESM is the output linguistic variable and others serve as input linguistic variables. For each linguistic variable fuzzy set was defined. This comprises of linguistic values. In order to normalize the data value, fuzzy membership expressions are defined for each linguistic value.

For PAR score value (let J) fuzzy membership expressions using triangular membership function (trimf) will be as:

\[
\mu_{first}(J) = \begin{cases} 
1 & J \geq 60 \\
\frac{60 - J}{60 - 55} & 55 \leq J \leq 60 \\
0 & J < 50 
\end{cases}
\]

\[
\mu_{second}(J) = \begin{cases} 
1 & J \geq 60 \text{ & } J < 65 \\
\frac{65 - J}{65 - 60} & 60 \leq J \leq 65 \\
0 & J < 45 \text{ or } J > 65 
\end{cases}
\]

For PSM score value (let K) fuzzy membership expressions using triangular membership function (trimf) will be as:

\[
\mu_{first}(K) = \begin{cases} 
1 & K \geq 65 \\
\frac{65 - K}{65 - 60} & 60 \leq K \leq 65 \\
0 & \text{otherwise} 
\end{cases}
\]

\[
\mu_{second}(K) = \begin{cases} 
1 & K \geq 65 \text{ & } K < 70 \\
\frac{70 - K}{70 - 65} & 65 \leq K \leq 70 \\
0 & K < 55 \text{ or } K > 70 
\end{cases}
\]

For ATT score value (let L) fuzzy membership expressions using triangular membership function (trimf) will be as:

\[
\mu_{good}(L) = \begin{cases} 
1 & L \geq 80 \\
\frac{80 - L}{80 - 75} & 75 \leq L \leq 80 \\
0 & \text{otherwise} 
\end{cases}
\]

\[
\mu_{average}(L) = \begin{cases} 
1 & L \geq 80 \text{ & } L < 85 \\
\frac{85 - L}{85 - 80} & 80 \leq L \leq 85 \\
0 & L < 60 \text{ or } L > 85 
\end{cases}
\]

For MSM score value (let M) fuzzy membership expressions using triangular membership function (trimf) will be as:
The decision on the right fuzzy sets is crucial for the success of a data mining project, therefore this easy method is not accurate enough and sets should be researched more carefully. However, it will give users a quick start for experimenting with the idea of fuzzy association rules. For use in a real project, fuzzy sets will have to be defined a priori or a more sophisticated algorithm has to be used for finding them.

3.2. Constructing a Dataset for Mining

After having defined the fuzzy sets, a new data set enabling the mining of fuzzy association rules has to be constructed out of the original data. This process is rather simple and intuitive, since the values only need to be fitted into the sets. For every fuzzy set that we have previously defined, there is one column in the new database containing the grade of membership of the single items to the specific set. Figure 2 visualizes the process of getting the membership values of a data point to different fuzzy sets.

As an example, we will look at a sample educational database representing one of the columns of the original database i.e. MSM: t={17,18,14,17,19,15,12,8,14} three fuzzy sets have been defined previously are: good={16,20} , average={10,15} and poor={0,9}. The row will be subdivided into four sub columns, one for each fuzzy set. The new table will only contain the membership values to these fuzzy sets (see Table 1).

![Figure 2. Membership of an Item](image.png)
Table 1. New Database without fuzzy normalization

| MSM:good | MSM:average | MSM:poor |
|----------|-------------|----------|
| 1        | 0.4         | 0        |
| 1        | 0.2         | 0        |
| 0.28     | 1           | 0        |
| 1        | 0.4         | 0        |
| 1        | 0           | 0        |
| 0.14     | 1           | 0        |
| 0.14     | 1           | 0        |
| 0.57     | 1           | 0        |
| 0        | 0           | 1        |
| 0.28     | 1           | 0        |

3.3. Fuzzy Normalization

When we are dealing with quantitative attributes mapped to fuzzy sets we might, depending on the membership function, find that the membership values to the sets of one single entity does not add up to one. The entry in a database in Table 1 serves as an example.

The quantitative attribute MSM represented by fuzzy sets, contributes with 1.4 in case of first row and similar with other rows. It is unreasonable for one transaction to contribute more than others. Here, the fuzzy normalization process takes place. It will further transform the transaction to values of MSM that sum up to 1. The new values can be calculated easily by dividing the value of a single element by the sum of all the fuzzy values corresponding to that attribute. Table 2 shows the modified database after processing of fuzzy normalization.

Table 2. New Database With Fuzzy Normalization

| MSM:good | MSM:average | MSM:poor |
|----------|-------------|----------|
| 0.71     | 0.29        | 0        |
| 0.83     | 0.17        | 0        |
| 0.23     | 0.77        | 0        |
| 0.71     | 0.29        | 0        |
| 1        | 0           | 0        |
| 0.12     | 0.88        | 0        |
| 0.13     | 0.87        | 0        |
| 0.36     | 0.66        | 0        |
| 0        | 0           | 1        |
| 0.23     | 0.78        | 0        |

3.4. Frequent Itemsets Generation: The Apriori-Like Algorithm

The proposed algorithm has similar philosophy as the Apriori TID, which is does not revisit the original table of data, for computing the supports larger Itemsets, but transforms the table as it goes along with the generation of the k-Itemsets.

Our procedure is based on a simple and easily implementable matrix representation of the frequent Itemsets. The idea is to store the data and Itemsets in vectors [12-14]. Then, simple matrix and vector multiplication operations can be applied to calculate the supports of Itemsets efficiently.

**Step1:** First step is to generate the 1-frequent Itemsets, by erasing the columns related to the non-frequent items, to reduce the size of a Bittable and improve the performance of the generation process

**Step2:** In the next step the rows which do not contain frequent Itemsets (the sum of the row is zero) are also deleted from the table.
Step 3: In this step, join the Previous frequent itemset \((k-1)\) to generate candidate itemset \(k\) based on the element wise products of the vectors corresponding to the previously generated \((k-1)\)-frequent itemsets.

Step 4: Next, Support of \((k)\) itemsets obtained by a simple vector product of the two related vectors because when both items appear in a given transaction the product of the two related items can be represented by the AND connection of the two items.

Step 5: Rows that are not containing any frequent itemsets (the sum of the row is zero) are also deleted from Candidate itemset \(k\).

Step 6: This process is repeated until as only those \((k-1)\) itemsets will be joined whose first \(k-1\) items are identical.

4. PERFORMANCE PARAMETERS

4.1. Processing Time

Processing Time \(T\) defines the time required to complete the execution of proposed method that is required by Decision tree and Fuzzy Association Mining process and measured in Seconds.

4.2. Fuzzy Quality Measures

In order to enable the evaluation of a fuzzy association rule, we use the standard approach for calculating support and confidence, replacing the set-theoretic operations by the corresponding fuzzy set-theoretic operations [12-14]:

\[
supp(A \rightarrow B) = \sum_{(x,y) \in D} T(A(x), B(x)) = \sum_{(x,y) \in D} \text{min}(A(x), B(x))
\]

\[
\text{conf}(A \rightarrow B) = \frac{\sum_{(x,y) \in D} T((A(x), B(x)))}{\sum_{y \in D} T(A(y))}
\]

The usual choice for the t-norm is the minimum, yet the product has also been applied.

5. RESULTS AND ANALYSIS

Proposed technique, are implemented on windows PC having Intel 2.4 GHz processor and 2GB RAM, and run using Matlab 9a. We have considered two different student records dataset obtained from Samrat Ashok Technological Institute, Vidisha, (Madhya Pradesh) of course MCA (Master of Computer Applications) from session 2007 to 2010 and 2011 to 2013. They are dataset50 and dataset154 respectively.

Size of the dataset50 is 50 student records which includes name, scholar no, DOB, 10th, 12th, PAR, MSM, ATT, ESM fields. Dataset154 contains 154 records of other MCA student with same field common. In this experiment datasets are used to identify the performance of student.

Table 3 shows the relationship between the minimum support, minimum confidence, execution time and the generated rule. Figure 3 to 4 gives the graphical representation of minimum support, minimum confidence, execution time and the generated rule. It was observed that the execution time is also inversely proportional to minimum support, since it increases as minimum support decreases, which confirmed increase in system complexity and response time as the minimum support decreases as shown in table III. With all these observations it shows that to have a less complex system and a constructive, interesting and relevant patterns the minimum confidence and support should be large enough to trash out coincidence patterns.

Figure 3 shows the relationship between the minimum support, minimum confidence, execution time and the generated rule. Figure 3 to 4 gives the graphical representation of minimum support, minimum confidence, execution time and the generated rule. It was observed that the execution time is also inversely proportional to minimum support, since it increases as minimum support decreases, which confirmed increase in system complexity and response time as the minimum support decreases as shown in table III. With all these observations it shows that to have a less complex system and a constructive, interesting and relevant patterns the minimum confidence and support should be large enough to trash out coincidence patterns.

From Figure 3, Figure 4, we can observe that increasing in minsup and minconf causes decreasing in the execution time of the algorithm because it causes reduction of dataset size and of the number of candidate fuzzy itemsets. In these figures, we show the number of rules generated and total time when the minsup and minconf is increased. Increasing the number of parameter values leads to decreasing number of rules, most of which are created in the second pass. In our algorithm, the number of the candidate itemsets is growing slowly in comparison with the other algorithm. A correlation can be observed between the number of rules and the execution time for this algorithm. The experimental results show that the running time of our
algorithm is much less than that of the other algorithms. The running time of the algorithms is determined by the number of passes and execution time of each pass. For our algorithm, the number of passes is less than the other algorithm, because it removes many candidate itemsets. With decreasing the number of the candidate itemsets, the size of the largest frequent itemset reduces, which is the number of passes the algorithm has to perform. Furthermore, the execution time of each passes is less than the old algorithm, because of reduction in the number of calculating operations.

Table 3. Performance test using fuzzy association rule mining technique

| MinSup | MinConf | No. of Rules | Time(S) |
|--------|---------|--------------|---------|
|        |         | Dataset50    | Dataset154 |
| 05%    | 40%     | 19           | 10       | 0.57555  | 1.6416  |
| 05%    | 60%     | 16           | 08       | 0.43255  | 1.6321  |
| 05%    | 80%     | 8            | 03       | 0.43257  | 1.6215  |
| 05%    | 100%    | 6            | 01       | 0.41324  | 1.6007  |
| 10%    | 40%     | 3            | 06       | 0.23847  | 1.1088  |
| 10%    | 60%     | 1            | 04       | 0.23198  | 1.0515  |
| 10%    | 80%     | 1            | 02       | 0.22876  | 1.0456  |
| 10%    | 100%    | 1            | 00       | 0.21457  | 1.0087  |
| 15%    | 40%     | 1            | 02       | 0.13367  | 0.8073  |
| 15%    | 60%     | 0            | 01       | 0.13088  | 0.8299  |
| 15%    | 80%     | 0            | 01       | 0.11080  | 0.7801  |
| 15%    | 100%    | 0            | 00       | 0.10333  | 0.7614  |
| 20%    | 40%     | 0            | 02       | 0.06449  | 0.6022  |
| 20%    | 60%     | 0            | 01       | 0.06428  | 0.6008  |
| 20%    | 80%     | 0            | 01       | 0.05857  | 0.5794  |
| 20%    | 100%    | 0            | 00       | 0.05392  | 0.5129  |

Table 4 shows the rules generated by fuzzy association rule mining in dataset50 with highest percentage support count and confidence. We can observes from the table that if the student was fail in last semester or he / she got third division and in this session attendance is poor then he / she may fail in upcoming semester. So we have to pay additional attention on those students.

This fuzzy mining result, reveals the characteristic of student who are likely to repeat or likely to have high academic rating. It can also serve as a predictive model to the admission office in an institution to know the performance of their intake right from their year one. With this they can determine a corrective
measure for subsequent admission processes. Also, this will serve as guidance to the level adviser for proper monitoring and effective advice for the students to enhance their academic performance.

| S.No. | Rules Generated by FARM                                                                 | Support Count | Confidence |
|-------|----------------------------------------------------------------------------------------|---------------|------------|
| 1     | if MSM=good then ESM=first                                                             | 8             | 0.57       |
| 2     | if PAR=first and if MSM=good then ESM=first                                            | 5             | 1          |
| 3     | if ATT=poor and if MSM=poor then ESM=fail                                               | 5             | 0.8333     |
| 4     | if PAR=third and if ATT=poor then ESM=fail                                              | 4             | 0.8        |
| 5     | if PSM=fail and if ATT=poor then ESM=fail                                               | 3             | 1          |
| 6     | if PSM=first and if MSM=good then ESM=first                                             | 3             | 1          |
| 7     | if PSM=third and if ATT=average then ESM=third                                         | 3             | 0.75       |
| 8     | if PAR=third and if MSM=poor then ESM=fail                                             | 3             | 0.6        |
| 9     | if PAR=third and if ATT=poor and if MSM=poor then ESM=fail                             | 3             | 0.6        |
| 10    | if PSM=fail and ESM=fail                                                                | 3             | 0.5        |

6. CONCLUSION
There is some more work to do, especially on Fuzzy association mining algorithm. This paper predicts the performance of student data on the basis of Apriori-like algorithm which is applied on fuzzy set. This study shows the potential of the fuzzy association rule mining algorithm for enhancing the effectiveness of academic planners and level advisers in higher institutions of learning. The analysis was done on students of Samrat Ashok Technological Institute, Vidisha (Madhya Pradesh). A total number of 50 tuples for dataset50 and 154 tuples for dataset154 are considered as a case study. The analysis reveals some hidden patterns of students’ poor performance which could serve as bedrock for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students’ performance and reducing failure rate.

REFERENCES
[1] Alaa el-Halees, “Mining students data to analyze e-Learning behavior: A Case Study”, 2009..
[2] S. T. Hijazi, and R. S. M. M. Naqvi, “Factors affecting student’s performance: A Case of Private Colleges”, Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.
[3] Q. A. Al-Radaideh, E. W. Al-Shawakfa, and M. I. Al-Najjar, “Mining student data using decision trees”, International Arab Conference on Information Technology (ACTIT'2006), Yarmouk University, Jordan, 2006.
[4] C. Romero, S. Ventura (2007), “Educational data mining: A Survey from 1995 to 2005”, Expert Systems with Applications (33), pp. 135-146, 2007.
[5] J. Han and M. Kamber, ”Data Mining: Concepts and Techniques”, Morgan Kaufmann, 2000.
[6] U. K. Pandey, and S. Pal, ”Data Mining: A prediction of performer or underperformer using classification”, (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2, pp. 686-690, ISSN: 0975-9646, 2011.
[7] Azhar Rauf, Sheeba, “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity”, Middle-East Journal of Scientific Research, Vol. 12, Pp. 959-963, 2012.
[8] Z. N. Khan, ”Scholastic achievement of higher secondary students in science stream”, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.
[9] Galit.et.al, “Examining online learning processes based on log files analysis: a case study”, Research, Reflection and Innovations in Integrating ICT in Education, 2007.
[10] P. Eklund, and J. Zhou, ”Comparison of Learning Strategies for Adaptation of Fuzzy Controller Parameters-Fuzzy Sets and Systems”, pp. 321-333, 1999.
[11] Tizhoosh, "Fuzzy Image Processing" © Copyright Springer, pp. 51-55, 1997.
[12] Zadeh, L. A., ”Fuzzy Sets”, Journal of Information Sciences, Vol. 8, pp 338-353, 1965.
[13] Zadeh, L.A., ”Fuzzy Sets as a Basis for a Theory of Possibility”, Fuzzy Sets and Systems, pp. 3-28, 1965.
[14] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan (2010), ”Data Mining Model for Higher Education System”, European Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.
BIOGRAPHIES OF AUTHORS

Sushil Kumar Verma is an Assistant Professor in the Department of Computer Applications at Samrat ashok technological Institute Vidisha (MP). Obtained Master degree (MCA) from RGPV Bhopal in 2003 and Ph.D. From Barkatullah University Bhopal in 2016. Research area is Data mining and warehousing. Published the research paper in the field of data mining in various international Journals and attended international.
E-mail: sushilverma81@gmail.com

Ramjeevan Singh Thakur is an Associate Professor in the Department of Computer Applications at Maulana Azad National Institute of Technology, Bhopal, India. He had a long carrier in teaching and research, including Three Year Teaching in the Department of Computer Applications at National Institute of Technology, Tiruchirapalli, Tamilnadu, India. At Present he is guiding several Ph.D. Research Scholars and handling Government Research Projects of about Rs. One Crore. He has published more than 75 Research paper in National, International, Journals and Conferences. He has visited several Universities in USA, Hong Kong, Iran, Thiland, Malaysia, and Singapore.
E-mail: ramthakur2000@yahoo.com

Shailesh Jaloree is an Associate Professor in the Department of Applied Math’s and Computer Science at Samrat Ashok Technological Institute (S.A.T.I.), Vidisha, India. He earned his Master Degree from Devi Ahiliya University Indore (M.P.) in 1991 and Ph.D. Degree (Applied Maths) From Barkatullah University, Bhopal (M.P.) in 2002. At Present he is guiding several Ph.D. Research Scholars in Mathematics and Computer Science field. He has published more than 35 Research Paper in National, International, Journals and Conferences. His areas of interest include Special Function, Data Mining, Data Warehousing and Web Mining.
E-mail: shailesh_jaloree@rediffmail.com

Fuzzy Association Rule Mining based Model to Predict Students’ Performance (Sushil Kumar Verma)