Modified Associative Algorithm to Determine Frequent Pattern from Student Dataset

Kamalpreet Kaur, Kiranbir Kaur

Abstract*: *The phenomenal advances in Students produces huge amount of data like MOOC data and high throughput information that makes Electronic Student records (ESRs) expensive and complex. For the analysis of such a huge amount of data, AI and data mining techniques have been utilized along with Student services. Today, Data mining is utilized to detect performances using various informational datasets along with machine learning algorithms. There are many techniques available which are utilized for diagnosis of student performance like FP growth, Apriori and Associative algorithm etc. These techniques discover unknown patterns or relationships from large amount of data and these are utilized for making decisions for preventive and suggestive medicine. The main disadvantage of these techniques is it discovers fewer patterns. In this paper we proposed modified associative algorithm that discovers patterns to detect performance accurately. The results will help in predicting the performance quicker and more accurately, so that it leads to timely aware the students.

Keywords* : *data mining, associative, accuracy

I. INTRODUCTION

The critical approach of filtering of data set used to discover normal and abnormal patterns from the database is step by step analysis process. Filtering of data set is the process of extraction of useful information from large database. The fetched information must be converted into user understandable form for future use. Mining approaches used at different places vary according to the size and complexity of the problem in hand. Mining approaches which are useful for detecting patterns[1] from the database includes web, text, sequential and temporal mining. Step by step analysis process is employed to discover patterns that are frequent within the database. The interest in pattern mining has been grown due to its ability to discover the hidden patterns within the database[2], that are useful for the users and cannot be extracted manually. Patterns category discovery is vital for successful interpretation of the performance. The step by step analysis process finds out frequent pattern from the sequence database. The well-known pattern mining methods[3] are utilized for web-log analysis, student record analysis and performance prediction. It identifies strong student-performance correlations which can be valuable information for the diagnosis and preventive medicine.

There are various types of classification of step by step analysis process algorithm that are based on following criteria:

- It considers the sequence that is generated and stored. It minimizes the number of sequences for decreasing the overall cost.
- It also supports the sequence of frequency that are counted and tested. The frequency check is maintained in order to remove any noisy data from the dataset.

The Apriori based algorithms are classified as given below:

- GSP (Generalized Sequential Pattern): It identifies the patterns that are common within the large dataset are discovered using the algorithm and then anomalies are highlighted. Hence noisy data can efficiently be handled by this algorithm. The data is scanned from top to bottom to discover patterns for checking abnormalities. .
- ASSOCIATIVE (Sequential Pattern Discovery using Equivalence Classes): In associative method[3], [4] vertical id list of dataset is achieved and then intersection of IDs has been obtained. This intersection is used for reducing scan of database and also decreasing overall execution time. It counts the sequence of each id and vertical representations are then converted into horizontal. The algorithm stops when there is no sequence found. It utilizes breath first search and depth first search to uncover the sequencing.
- Pre-fix span: Prefix span projection method[5], [6] is used to discover patterns and it uses sub sequences generation. It mines complete dataset into pattern using candidate sequence and then gives efficient processing of dataset.

II. RELATED WORK

Alzahrani(2016), proposed data mining method for performance prediction[7]. Sequential data mining is used in order to accomplish the data preprocessing mechanism. After applying the preprocessing mechanism, attributes are analyzed using passes on student data. The first pass determines whether support for each performance is present or not. At the end of this phase, the frequent performance within the database is identified and a counter is maintained to count the occurrence of each performance within the dataset. Next phase determines the second sequence of performances present within the dataset. The overall process yields the performances which can cause the occurrence of other performances. The performance resulting in another performance is termed as candidate generation and for declaring that it is generated from the previous level, Pruning is used.

Revised Manuscript Received on April 27, 2020.
* Correspondence Author

Kamalpreet Kaur*, Computer and Science Department, Guru Nanak Dev University, Amritsar, India. Email: kamaldosanjh2305@gmail.com

Kiranbir Kaur, Computer and Science Department, Guru Nanak Dev University, Amritsar, India. Email: kiran.dcse@gndu.ac.in

CHENG et al. (2017), proposed a sequential mining approach for early assessment of chronic performance. The student database is considered [8]. A dataset of students is derived from Taiwan that has richest of risk patterns. Data preprocessing is performed to rectify the problem if found but missing values are not considered. Sequential pattern mining is used to observe the risk pattern and generate the results. The problem with this approach is that no precautions have been suggested. The classification accuracy is 80% and further improvement in classification is needed.

Kunjir,et al.(2017), proposed multiclass Naive Bayes algorithm that is used for prediction of a particular performance. The dataset used for operation is taken from UCI machine learning website The discussed approach [9] deals with prediction accuracy corresponding to particular performance. The result in terms of confusion matrix is also presented.

Alamanda, et al.(2017), proposed sequence pattern mining in order to detect the time duration used for promotion. The sequence or pattern is checked within the database. The weight of each sequence in each database is achieved from the interval of the successive element in the sequence and the mining is performed on the basis of weight considering time interval[10]. Time interval based pattern is used in this case. In preprocessing missing values are not considered.

Ghosh,et al.(2015), proposed a technique that extracts sequential patterns from hypotensive student groups. These patterns are further utilized to inform student decisions and randomized student trials. It further extended by including various student features and also includes some sequential patterns[11]. It also does not consider missing value during the preprocessing phase.

Zhang,et al.(2016), proposed a technique named ConSgen that is used to identify the contiguous sequential generator and also minimizes the redundant patterns, It utilizes the divide and conquer technique to find the sequential generator with contiguous constraints[12]. But it does not consider the gapped alignments and also not discover the binding sites.

M. Zihayat et al.(2016), identified a problem of top- k utility based regulation pattern which is used to find out meaning in biology. Firstly proposed a utility model called TU-SEQ which is used to find top-K high utility gene regulation sequential patterns[13]. It considers the relation between the various patterns and interactions in student data studies.

Abbasghorbani et al.(2015), analyzed various pattern mining techniques and the features of all the algorithms. It introduced various minimizing support counting which is used for minimizing search space[14]. They have generated small search space which includes earlier candidate sequence pruning.

Then database is analyzed with compression technique.

Image set name	Parameters	Existing (%)	Proposed (%)
Level 1 Student(Mild)	Accuracy	85	95
	Specificity	84	94
	Sensitivity	84	92
Level 2 Student(Moderate)	Accuracy	85	95
	Specificity	86	96
	Sensitivity	87	97
Level 3 Student(Severe)	Accuracy	86	91
	Specificity	87	94
	Sensitivity	87	96

III. PROPOSED SYSTEM

The proposed algorithm uses the prefix span algorithm for determining patterns which can be grouped together to form clusters. Pre-processing mechanism includes most probable value replacement with the missing value.

A. Algorithm

- Input: Dataset
- Output:Classification Accuracy, Performance Prediction

```
• Input Dataset
    Data=$Dataset_i$
    Where I are the number of rows within the dataset
• Apply Pre-processing mechanism to resolve the missing values
    MPV=mean
    (Values ($Person_{id_i} = dataset(person_{id_i})$))
• Repeat while all the missing values are tackled
    If (Missing$_i$)
    Missing$_i$=MPV
    End of if
    End of loop
• Apply Pre-fix span algorithm for pattern growth determination
• Form clusters
    Repeat until values in dataset are examined
    If(Datset$_{iValue}$==Dataset$_{i+1value}$)
    Cluster$_i$=Datset$_{iValue}$
    End of if
    i=i+1
     End of loop
• Predict performance looking at the pattern clusters
    Result: Accuracy, Performance.
```

IV. RESULTS

The performance of the system is analyzed by the use of parameters such as accuracy, specificity and sensitivity.

Accuracy is obtained by subtracting the actual result from the approximate result. In terms of predictions accuracy is obtained as

$$Accuracy = \frac{Correct_{Pre}}{Total_{Pred}}$$

Equation 1: Accuracy in terms of prediction

Sensitivity is obtained by dividing number of positive predictions to the total true positive rate.

Sensitivity=$\frac{Correct_{Positive_{predictions}}}{Total_{Positives}}$

Retrieval Number: D8321049420/2020©BEIESP
DOI: 10.35940/ijeat.D8321.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Equation 2: Sensitivity evaluation formula

Specificity is another parameter used to evaluate correctness of the proposed system. It is given as under

$$\text{Specificity}=\frac{True_{Negitives}}{TP+FN}$$

Equation 3: Specificity obtaining formula

The performance detection and prediction is given through accurate classification, result in terms of plots is given as under:

Classification accuracy of proposed system appears to be more as compared to existing techniques. Multiple class prediction mechanism showing higher accuracy proving the worth of study.

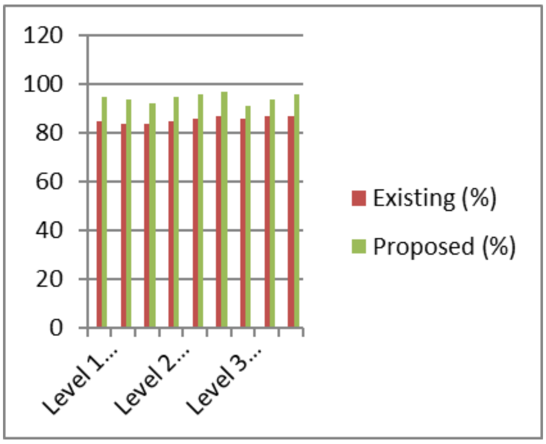

Figure 1 Confusion matrix

Results and performance analysis as indicated through the plot shows that prefix span algorithm along with MPV algorithm yield better result.

Prefix algorithm

Load Dataset

First of all data set is loaded from offline sources and dataset is synthetically prepared. After loading the data set the number of students is displayed in the box.

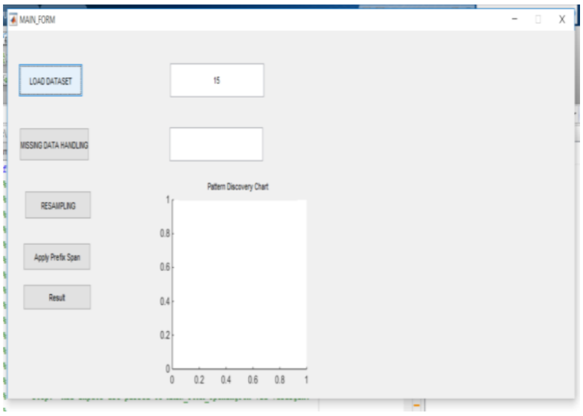

Figure 2 Load dataset

Handle missing data:

For handling missing data ,it has inbuilt mechanism. It will eliminate noisy data from the dataset and display the remaining data.

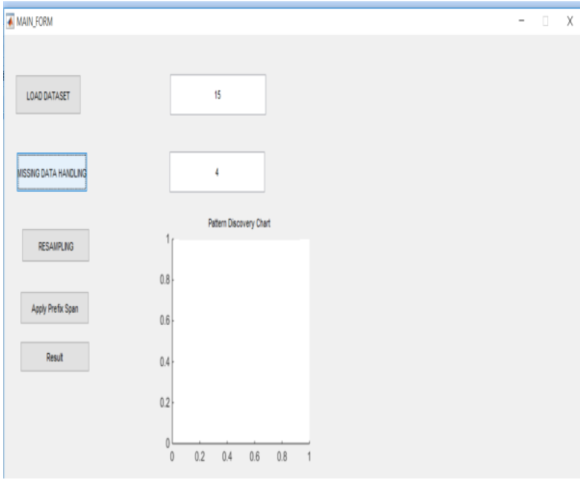

Figure 3 Handle Missing data

Implementation of Prefix Span Algorithm:

It generates various patterns on the basis of comparisons. It will predict pattern along
with another relative patterns.

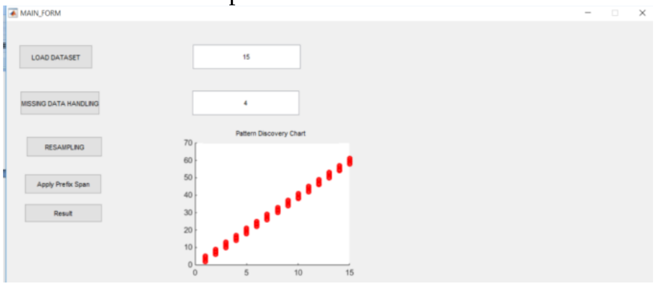

Figure 4 Implementation of Prefix Span Algorithm

Parameter Result

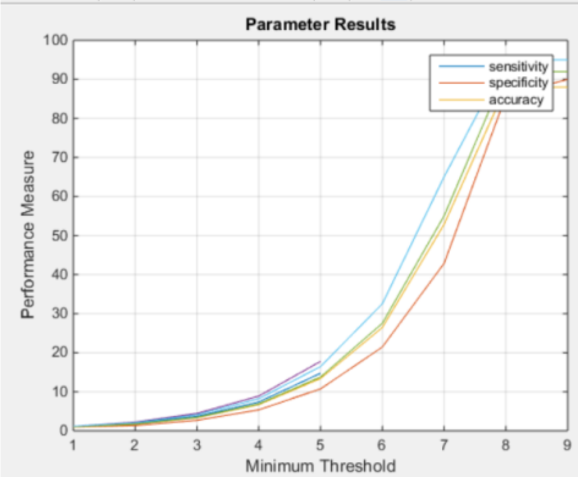

Figure 5 Comparison of Specificity , Sensitivity and Accuracy.

V. CONCLUSION AND FUTURE SCOPE

An automated system that utilizes MPV along with prefix span algorithm for detecting student performance proposed is used. Pre-processing phase is critical and is well defined using noise handling and resizing operation. Obtained data are fed into the trained network for feature extraction using prefix span algorithm and classification is performed using MPV.

Hybrid approach followed gives better results. An effective pattern development technique, Prefix Span, is proposed and contemplated in this paper. The main objective of the work is creating optimized detection of the performance using prefix span for better accuracy. Higher accuracy is achieved with the proposed algorithm. In future, proposed strategy can be examined against the real time datasets for better evaluation of accuracy.

REFERENCES

1. C. T. Nadu, and C. T. Nadu, “Heart performance classification and its co-morbid condition detection using WPCA genetic algorithm,”,IEEE, pp. 287–291, 2016.
2. K. Adlakha, “Recapitulation of Ant Colony and Firefly Optimization Techniques,” *AIS*, vol. 1, no. 2, pp. 175–180, 2015.
3. C. Anusha, S. K. Vinay, H. J. Pooja Raj, and S. Ranganatha, “Student data mining and analysis for heart performance dataset using classification techniques,” *Natl. Conf. Challenges Res. Technol. Coming Decad. (CRT 2013)*, pp. 1.09–1.09, 2013.
4. I. Ţăranu, “Data mining in Studentcare: decision making and precision,” *Database Syst. J.*, vol. 5, no. 4, pp. 33–40, 2015.
5. S. Sharma, “Data Preprocessing Algorithm for Web Structure Mining,” pp. 1–5, 2016. S. D. Thepade, S. Vasai
6. kar, N. Bhavsar, R. More, and A. Bhatkhande, “Vehicle Traffic Density Estimation Using Bayes , Rule , Tree Family Data Mining Classifiers Applied On Background Subtracted Traffic Images,” *Ieee Access*, pp. 87–92, 2016.
7. M. Y. Alzahrani, “Discovering Sequential Patterns from Student Datasets,” 2016.
8. Y. CHENG, Y.-F. Lin, K.-H. Chiang, and V. Tseng, “Mining Sequential Risk Patterns from Large-Scale Student Databases for Early Assessment of Chronic Performances: A Case Study on Chronic Obstructive Pulmonary Performance,” *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, 2017.
9. A. Kunjir, H. Sawant, and N. F. Shaikh, “Data mining and visualization for prediction of multiple performances in Studentcare,” *Proc. 2017 Int. Conf. Big Data Anal. Comput. Intell. ICBDACI 2017*, pp. 329–334, 2017.
10. S. Alamanda, S. Pabboju, and N. Gugulothu, “An Approach to Mine Time Interval Based Weighted Sequential Patterns in Sequence Databases,” *2017 13th Int. Conf. Signal-Image Technol. Internet-Based Syst.*, pp. 29–34, 2017.
11. S. Ghosh, M. Feng, H. Nguyen, and J. Li, “Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure,” *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 5, pp. 1416–1426, 2015.
12. J. Zhang, Y. Wang, C. Zhang, and Y. Shi, “Mining contiguous sequential generators in biological sequences,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 855–867, 2016.
13. M. Zihayat, H. Davoudi, and A. An, “Top-k utility-based gene regulation sequential pattern discovery,” *Proc. - 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016*, pp. 266–273, 2017.
14. S. Abbasghorbani and R. Tavoli(2015), “Survey on Sequential Pattern Mining Algorithms,” in 2nd Int. Conf. Knowledge-Based Eng. Innov. 2015, pp. 1153–1164.
15. N. Gandhi and L. J. Armstrong, “A review of the application of data mining techniques for decision making in agriculture,” pp. 1–6, 2016.
16. E. Pinheiro, W. Weber, and L. Barroso, “Failure trends in a large disk drive population,” *Proc. 5th USENIX Conf. File Storage Technol. (FAST 2007)*, no. February, pp. 17–29, 2007.
17. A. Sharma and V. Mansotra, “Emerging applications of data mining for Studentcare management - A critical review,” *2014 Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 377–382, 2014.
18. K. Yan, X. You, X. Ji, G. Yin, and F. Yang, “A Hybrid Outlier Detection Method for Student Care Big Data,” *2016 IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun.*, pp. 157–162, 2016

AUTHORS PROFILE

I am **Kamalpreet Kaur** currently pursuing my post-graduation in the field of computer engineering and technology from Guru Nanak Dev University. With the merit scoring and being in the top 10 toppers in non-medical of Spring Dale Senior School, I persuaded my Bachelors in Computer Engineering and Technology in **Guru Nanak Dev University.**

I am **Mrs Kirabir Kaur** currentl the professor in Guru Nanak Dev University. I am post-graduate in the field Computer Engineering and Technolgy and cleared UGC-NET. With the specialization in the field of cloud computing, I have published my 40 journal accomplishments, 5 papers in conference and writing 4 book chapters.

I have experience of more than 7 years and the following are the titles of paper published in conferences: An Enhanced Hybrid Approach for Reducing Downtime, Cost and Power Consumption of Live VM Migration,Fuzzy based map reduce technique for energy efficiency in multi core cloud computing, A computation offloading scheme for performance enhancement of smart mobile devices for mobile cloud computing, A study of power management techniques for Internet of Things (IoT), Efficiency improvement of Apriori algorithm over dense databases