

A Survey on Clustering and Feature Selection Algorithm for Quickly Predicting Engineering Students' Academic Performance

Susheelamma K H¹, Dr. Brahmananda S H²

¹Assistant Professor Department of ISE SJGIT, Chickballapur, Karnataka, India

²Professor and Head Department of CSE GITAM School of Technology Bengaluru, Karnataka, India

ABSTRACT

Data mining(DM) is the process of extracting hidden and useful information in large data repositories. Knowledge Discovery in large data repositories is time consuming hence the need for fast algorithms for extracting useful knowledge from large database. This knowledge can be used to increase the quality of education. Educational Data Mining(EDM) is concerned with developing new methods to discover knowledge from educational/academic database and can be used for decision making in educational/academic systems. In our survey paper we would like to focus and analyze the various Data clustering and feature selection algorithms to bring the clarity in students' results and faculties contribution to make this one as success.

Keywords: DM, EDM, CFOA, Clustering, Feature Selection.

I. INTRODUCTION

The main aim of engineering institutes is to give quality education to its students and to improve the student's performance. One way to meet the top level of excellence in higher education system is by discovering knowledge that is by analyzing the main attributes that may affect the students' performance. The discovered knowledge can be used to stretch forth helpful and beneficial recommendations to the academic planners in education institutes to predict student academic performance, helps in identifying weak students. Thus, the management of engineering institute takes essential steps to improve weak student's performance and many other benefits.

Clustering and feature selection are data mining techniques which are used to extract knowledge from educational data set.

The objective of data mining is to define the kind of knowledge. There are two categories of data mining tasks: descriptive and predictive. Clustering is the descriptive data mining techniques used to extract hidden patterns from large data sets. Feature selection is the process of selecting a subset of relevant features to avoid the curse of dimensionality and to get enhanced generalization by reducing over fitting.

The main objective of this survey paper is to identifying and improving weaker students' performance of a particular subject in undergraduate engineering programs and collecting and analyzing behaviours of students in any one particular subject. This analysis help a lot in predicting and measuring the interest of a student towards particular subject and deciding the steps to create interest in that particular subject and hence reducing the failure rates in that subject.

The paper is organized as: section II describes the EDM, section III gives a brief description of literature survey of different clustering and feature selection techniques used in EDM, section IV concludes the paper, section V is the future scope, finally section VI gives references.

II. EDM

Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn. Educational Data Mining in data mining retrieves hidden knowledge by applying various techniques of data mining like clustering, classification, feature selection and many others which gives us a final result or if required, the output data can be altered or filtered according to need [1]. Data collected from education institute can be aggregated over large numbers of students and can contain many variables that can be explored through intelligent data mining algorithms for model building.

A. Aim of EDM

EDM is used to predict students' behavior, improving students performance, helping students to get placed in corporate companies and many more areas of study which help in to intensify the quality of skill and attitudes [2].

New model can be discovered by application of EDM techniques or improving the existing techniques by combining the concepts of two existing techniques or slightly altering the previous technique [2].

In EDM process, students, teachers, administrators, and researchers are involved because at each level of EDM from at all these four stakeholders play major role in making decisions, changes in the learning process, development and allocating the resource for institute so that the teaching and learning process in the institute may give good result in improving the performance of students thereby improving grade of the institution.

B. Stages in EDM

As Figure 2.1 shows EDM generally consists of four steps [2]:

During the first step of the EDM process relationships in between data is found by searching through training set of data. This can be done through several algorithms of data mining like clustering, classification, regression, association rule mining.

During the second step of EDM the relationship found in the first step is checked for correctness.

In the third step several validated relationships are used for prediction in the learning environment.

Finally, in fourth step decisions are taken on the basis of prediction results to make some policy for institute. During phases 3 and 4, various human judgments are taken to filter data[2].

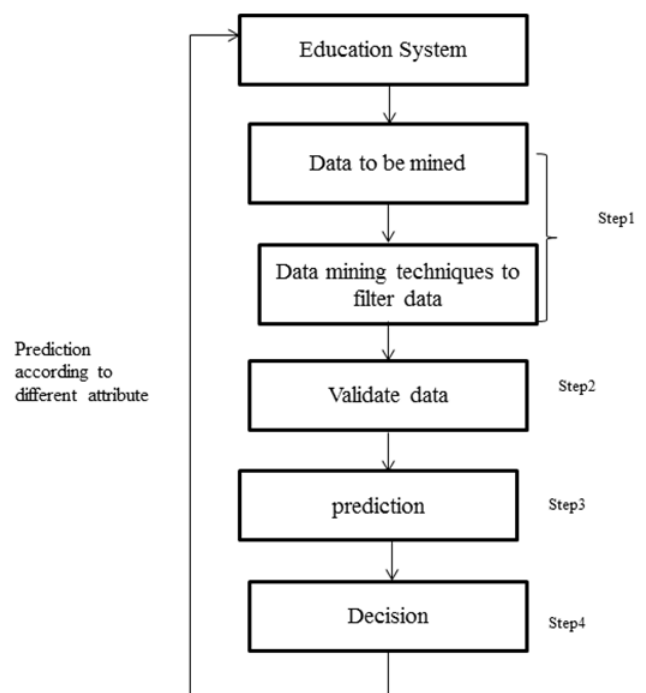


Figure 1

C. Related Work

Analyzing different literature, it is found that predicting and analyzing students' interest towards the subject, prediction of improvement in performance plays a big role to measure the accurate

result which includes student hobby, learning patterns, language through which he can easily understand the subject and many more criteria related to student's habits are co-related to analyze the prediction of various fields [3][4].

Some literature survey areas of different authors are mentioned below:

Vandamme.et.al in 2007 worked on decision tree and neural network for Academic performance prediction of students [5].

Alaa el-Halees in 2009 described data mining as Data Mining can be used in the educational field to enhance understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students [6].

In 2011 Bhardwaj and Pal[3] describes the student performance analysis by using the classification techniques.

In February 2012 Mohd m Abul Tair; Alaa M.El-Hales [7] shows for the improvement of performance in graduate student by techniques like association rule, classification, clustering.

III. REVIEW OF VARIOUS SIGNIFICANT CLUSTERING ALGORITHMS, FEATURE SELECTION TECHNIQUES APPLIED TO EDUCATIONAL DATASET

[8] For evaluating undergraduate student academic performance. S Chen and X Liu., "An integrated approach for modeling learning patterns of students in web-based instruction: A cognitive style perspective", used a combination of DM methods like ANN (Artificial Neural Network), Farthest First method and Decision Tree as a classification approach. They have used Student data of the Computer Science department at Faculty of Science and Defence technology, National Defence university of Malaysia (NUDM).

[9] In 2013 A Bovo, S Sanchez, O Heguy and Y Duthen., "Clustering model data as a tool for profiling students", used Expectation Maximization, Hierarchical Clustering, Simple K-Means and X-Means as provided in the WEKA software has been used to predict the potentiality of students' performance who can fail during an online curriculum in a Learning Management System (LMS) on the Real life dataset provided by Juris campus accessible at <http://www.juriscampus.fr/>.

[10] H Grob, F Bensberg, and F Kaderali., "Controlling open source intermediaries - a web log mining approach", during 2004 used Apriori Algorithm is applied to academic records of students to obtain the best association rules, which help in student profiling. K-means clustering is used to group students categorically. To show the applications of various DM techniques on student academic data on student academic record file.

[11] T Peckham and G McCalla., "Mining student behavior patterns in reading comprehension tasks", during 2012 used Markov Clustering (MCL) algorithm for clustering the students' activity and a Simple K-Means algorithm for clustering the courses to analyze the web log data files of a Learning Management System (LMS) on dataset was collected from the Technological Education Institute (TEI) of Kavala (Greece) that uses the Open e-Class e-learning platform (GUNet 2009). The data are from the spring semester of 2009 from the Department of Information management and involve 1199 students and 39 different courses. The data are in ASCII form and are obtained from the Apache server log file.

A meta-heuristic bio-inspired optimization algorithm which is called Cuttlefish Optimization Algorithm (CFOA) [12] [13] mimics the mechanism of colour changing behavior of the cuttlefish to solve numerical global optimization problems. The colours and patterns of the cuttlefish are produced by reflected light from three different layers of cells.

IV. CONCLUSION

In this survey paper partially it is analyzed that different techniques are used in literature survey to find the performance of students who require extra attention to improve their academic performance. This predictive model predicts the future scope of each student. Good performance of student helps placement the officer to guide students to choose their right career in which their talent or commands are good. Most of the companies hire the students according to the skill like technical, communication skill, CGPA marks, his behavior and many more features. Performance improvement in academic helps the students to update and getting placed in good companies. This will improve the overall productivity.

Through the literature survey, we found that in DM, the combination of two algorithms/techniques and nature inspired genetic algorithms always gives good result. Hence, whenever quick decision is to be taken, it is always better to use a combination/mixture of two algorithms to get improved result.

V. FUTURE SCOPE

As the future aspect the student academic performance can be improved through continuous observation by giving module wise tests, finding out and analyzing students' difficulties towards the subject and preparing the students to come out of their difficulties by providing extra inputs, resources required to better understand the subject.

Hence helping the student to keep good academic score, getting placed in good companies and building their career.

VI. REFERENCES

- [1]. Kumar J., "A Comprehensive Study of educational Data mining", IJEECS, 2015.
- [2]. http://en.wikipedia.org/wiki/Educational_data_mining.
- [3]. Pal B. K., "Mining Educational Data to Analyze Students Performance", IJACSA, 2015.
- [4]. S. Pal., U. A., "Data Mining: A prediction of performer or underperformer using classification", IJCSIT, 2011.
- [5]. Dutta R. J., "A survey on educational data mining and research trends department". International Journal of Database Management system, June 2013.
- [6]. el-Halees A., "Mining student's data to analyze e learning behaviour: A case Study", 2009.
- [7]. Jay Ruby K., "Predicting the performance of students in higher education using Data Mining Classification:A Case Study". IJRASET, 2014.
- [8]. S Chen and X Liu, "An integrated approach for modeling learning patterns of students in web-based instruction: A cognitive style perspective"., ACM Trans. Comput. Interact., vol. 15, no. 1, 2008.
- [9]. A Bovo, S Sanchez., O Heguy and Y Duthen., "Clustering moodle data as a tool for profiling students"., in Proc. Second Int. Conf. E-Learning E-Technologies Educ., pp. 121-126, 2013
- [10]. H Grob, F Bensberg, and F Kaderali., "Controlling open source intermediaries - a web log mining approach"., IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews, vol. 1, pp. 233-242, 2004.
- [11]. T Peckham and G McCalla., "Mining student behavior patterns in reading comprehension tasks"., Int. Educ. Data Min. Soc., pp. 87-94, 2012.
- [12]. Adel Sabry Eesa, Zeynep Orman and Adnan Mohsin Abdulazeez Brifceni., "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems"., Expert Systems with Applications, Vol.42, pp.2670-2679, 2015.
- [13]. Adel Sabry Eesa, Adnan Mohsin Abdulazeez Brifceni and Zeynep Orman., "Cuttlefish Algorithm – A Novel Bio-Inspired Optimization Algorithm"., International Journal of Scientific & Engineering Research, Vol. 4, Issue 9, September 2013.