ISSN:2348-2079

Volume-5 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Prediction of student performance

[1]M.Mohammed Imran, [2]R.Swaathi, [2]K.Vasuki, [2]A.Manimegalai, [2]A.Tamil arasan
[1]Assistant Professor, Dept of IT, Nandha Engineering College, Erode
[2]UG Scholar, Dept of IT, Nandha Engineering College, Erode
Email: swaathirajmohan@gmail.com

Abstract: ***Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on Student datasets using multiple classifiers and feature selection techniques. Many of them show good classification accuracy. The existing work proposes to apply data mining techniques to predict Students dropout and failure. But this work d oesn't support the huge amount of data. It also takes more time to complete the classification process. So the time complexity is high. To improve the accuracy and reduce the time complexity, the MapReduce concept is introduced. In this work, the deadline constraint is also introduced. Based on this, an extensional MapReduce Task Scheduling algorithm for Deadline constraints (MTSD) is proposed. It allows user to specify a job's (classification process in data mining) deadline and tries to make the job to be finished before the deadline. Finally, the proposed system has higher classification accuracy even in the big data and it also reduced the time complexity.***

Keywords: Data Mining, Big Data, Classification, MapReduce.

1. Introduction

The main objective of higher education institutes is to provide quality education to its student and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the student's performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve student's academic performance and trim down failure rate, to better understand student's behaviour, to assist instructors, to improve teaching and many other benefits.

Educational data mining [25] uses many techniques such as decision tree, rule induction, neural networks, k-nearest neighbour, naïve Bayesian and many others. By using these techniques, many kinds of knowledge can be discovered such as association rules, classifications and clustering. It showed what kind of data could be collected, how to pre-process the data, how to apply data mining methods on the data, and finally how to get benefited from the discovered knowledge. Many kinds of knowledge can be discovered from the extracted data.

In this era of big data, huge amounts of structured and unstructured student data are being produced daily. Big Data is difficult to work with and requires massively parallel software running on a large number of computers. Previous works investigated the most common ones which are attribute selection, classification such as decision tree method.preliminary work proposes to apply data mining techniques to predict student's dropout and failure. But this work doesn't support the huge amount of data. And also it takes more time to complete the classification process. So the time complexity is high. To improve the accuracy and reduce the time complexity, the MapReduce concept is introduced.

MapReduce is a recent programming model that simplifies distributed applications that handle Big Data. For implementing the concept of MapReduce, it has to divide the student data and perform the data mining process. After that the aggregated result is produced. Consequently, the performance of MapReduce strongly depends on how evenly it distributes the student data. In MapReduce [27], workload distribution depends on the algorithm that partitions the data. To improve the accuracy of the system and supports the big data, the novel technique is proposed in this work. In this work, the deadline constraint is also introduced. Based on this, an

extensional MapReduce Task Scheduling algorithm for Deadline constraints (MTSD) is proposed. It allows user to specify data mining process's deadline and tries to make the data mining process to be finished before the deadline.

Increasing digitization of student records means predictive analytics is expected to transform teaching and become a key tool in learning more about students. Predictive analytics is a process in which data collected about the student, typically attendance, subjects taken, assessment is used to understand learning patterns, identify skill gaps, predict performance and identify learning opportunities. The effective feature selection method is required to analyze the efficient classification.

2. Proposed Methodology

This work proposes an extensional MapReduce Task Scheduling algorithm for Deadline constraints (MTSD). It allows user to specify a classification process in data mining deadline and tries to make the data mining process be finished before the deadline. This algorithm classifies the student data into several levels. Under this algorithm, first it illuminates a novel data distribution model which distributes student data according to the student's data capacity level respectively. The experiments show that the student data classification algorithm can improve data locality observably to compare with default scheduler and it can also improve other scheduler's locality. Secondly, it calculates the data mining process's average completion time which is based on the student data level. It improves the precision of classification's remaining time evaluation.

The MTSD algorithm takes the student data locality and cluster heterogeneity into account. The data locality is the key factor that affects the efficiency of MapReduce classification process. The data locality means that the classification's operation code and the classification's input data are on the same computing node or on the same rack. Of course, the efficiency when the code and data are on the same node is higher than on the same rack. If the code and data are on the same node, it would avoid the data transmission in the network and greatly reduce the delay. Therefore, in the large scale data processing applications, shifting

2.1 Data gathering from school

All the information used in this study has been gathered from different sources. In this research work synthetic dataset has been used and a specific survey was designed to administer all students in the middle of the course. Its purpose was to obtain personal and family information to identify some important factors that could affect school performance. From a general survey [14] which is completed when the students register in the National Evaluation Center (CENEVAL) for admission to many institutions of secondary and higher education.

2.2 Data preprocessing

Before applying DM algorithm it is necessary to carry out some pre-processing tasks such as cleaning, integration, discretization and variable transformation. It must be pointed out that very important task in this work was data pre-processing, due to the quality and reliability of available information, which directly affects the results obtained. In fact, some specific pre-processing tasks were applied to prepare all the previously described data so that the classification task could be carried out correctly. First, all available data were integrated into a single dataset. During this process inefficient student's information are eliminated. Some modifications are also made to the values of some attributes. Furthermore, the continuous variables are transformed into discrete variables, which provide a much more comprehensible view of the data.

2.3 MapReduce

In the proposed system the performance of the system is improved by using MapReduce. This is simple yet powerful framework which lets the programmer write simple units of work as **map** and **reduce** functions. In summary, they are:

- "In-mapper combining", where the functionality of the combiner is moved into the mapper. Instead of emitting intermediate output for every input key-value pair, the mapper aggregates partial results across multiple input records and only emits intermediate
    - key-value pairs after some amount of local aggregation is performed.

- The related patterns "pairs" and "stripes" for keeping track of joint events from a large number of observations. The pairs approach keeps

track of each joint event separately, whereas the stripes approach keeps track of all events that co-occur with the same event. Although the stripes approach is significantly more efficient, it requires memory on the order of the size of the event space, which presents a scalability bottleneck.

- "Order inversion", where the main idea is to convert the sequencing of computations into a sorting problem. Through careful orchestration, one can send reducer the result of a computation (e.g., an aggregate statistic) before it encounters the data necessary to produce that computation.

- "Value-to-key conversion", which provides a scalable solution for secondary sorting. By moving part of the value into the key, one can exploit the MapReduce execution framework itself for sorting.

Based on MapReduce concept, an extensional MapReduce Task Scheduling algorithm for Deadline constraints (MTSD) is proposed. It allows user to specify data mining process's deadline and tries to make the data mining process to be finished before the deadline.

2.4 Attribute selection

The attributes are selected using the Feature Selection Techniques [18] called, Correlation-based Feature Selection (CFS).

The Correlation-based Feature Selection (CFS) estimates and ranks the subset of features than individual features. It chooses the set of attributes that are highly associated with the class, in addition to those attributes that are in low inter-correlation.

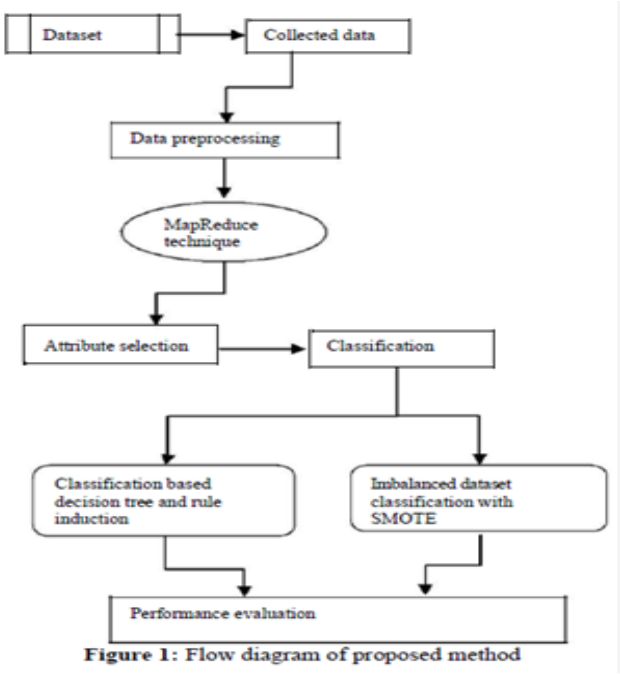

Figure 1: Flow diagram of proposed method

2.5 Decision tree classification

A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. The information gain measure is used to select the test attribute at each node of the decision tree. The information gain measure prefers to select attributes having a large number of values. The basic decision tree induction algorithm ID3 was enhanced by C4.5. C4.5 a successor of ID3 uses an extension of information gain known as gain ratio, which attempts to overcome this bias. The WEKA classifier package has its own version of C4.5 known as J4.8. J4.8 is used to identify the significant attributes.

2.6 Imbalanced dataset classification with Smote

The problem of imbalanced data classification occurs when the number of instances in one class is much smaller than the number of instances in another class or other classes. Traditional classification algorithms have been developed to stage. One way to solve this problem is to act during the pre- processing of data by carrying out a sampling or balancing of class distribution. There are several data balancing or rebalancing algorithms; one that is widely used and that is available in Weka as a supervised data filter is SMOTE (Synthetic Minority Oversampling Technique). In the SMOTE algorithm, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments

joining any or all of the *k* minority class nearest neighbors. Depending on the amount of over-sampling required, neighbors from the *k* nearest neighbors are randomly chosen.

2.7 Performance Evaluation

Finally, the performance of the existing classification and prediction system is measured with proposed grammar based genetic programming approach to derive the pass/failure result. Measure the performance results in terms of the true positive rate (TPR), false positive rate (FPR), False Negative Rate (FNR) and True negative Rate (TNR), accuracy, Time comparison.

3.Conclusion:

The aim of this study is to analyze factors affecting academic achievement that contribute to the prediction of students' academic performance. It is useful in identifying weak students who are likely to perform poor in their studies. An educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics. The various data mining techniques can be effectively implemented on educational data. From the above results it is clear that classification techniques and mapreduce concept can be applied on educational data for predicting the student's outcome and improves their results. The classification accuracy and performance is high in the proposed system. This experiment shows that the proposed system is more efficient than the existing system.

REFERENCES

[1] Angie Parker, "A study of variables that predict dropout from distance education," Int. J. Educ. Technol., vol. 1, no. 2, pp. 1–11, 1999.

[2] Araque F, C. Roldán, and A. Salguero, "Factors influencing university dropout rates," *Comput. Educ.*, vol. 53, no. 3, pp. 563–574, 2009.

[3] Breiman L, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York, USA: Chapman & Hall, 1984.

[4] Cendrowska J, "PRISM: An algorithm for inducing modular rules," *Int. J. Man-Mach. Stud.*, vol. 27, no. 4, pp. 349–370, 1987.

[5] Chawla N. V. and K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[6] Chao Jin and Christian Vecchiolaand Rajkumar Buyya ,"MRPGA: An Extension of MapReduce for Parallelizing Genetic Algorithms", Grid Computing and Distributed Systems (GRIDS) Laboratory, Department of Computer Science and Software Engineering , The University of Melbourne, Australia , 2008.

[7] Chris Miceli and Michael Miceli, Bety Rodriguez-Milla and ShantenuJha, "Understanding performance of distributed data-intensive applications", 2009.

[8] Cristobal Romero and Sebastian Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.

[9] Espíndola E and A. León, "La deserción escolar en américalatina: Un Temaprioritario para la agenda regional," *RevistaIberoamer. Educ.*, vol. 1, no. 30, pp. 39– 62, 2002.

[10] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, "Bigtable: A Distributed Storage System for Structured Data" 2006.