Incremental Learning Helps to Improve Student Skills in Educational Field

Vrushali A. Sungar

ME Scholar, Department of Computer Engineering, Dr. D.Y. Patil School of Engineering and Technology, Lohegaon, Pune, Maharashtra, India

***Abstract*— Data extraction plays a vital role in data mining. While extracting data it will consider some factors: statistics data, applications, algorithm and so on. Extraction of data means extraction of features, knowledge from particular domain and learning data from systems perceptive. Considering these entire scenario we build system according unlearn data to learn data while data entering into system as input. Based on these we get fruitful knowledge by using some standard algorithms of data mining. Here we apply student's dataset to our system to learn specific pattern namely strong, week and excellent category of students. Our system help to identify these patterns and helps to those students are weak in particular field. For this purpose we take experimental result on Decision tree algorithm like J48 and Random Tree.**

***Keywords*— Extraction, Incremental Learning, Classification**

I. INTRODUCTION

Predicting the student's performance is an important aspect now a day. Predicting student's performance becomes trickier due to the huge volume of data in educational databases. Students performance is based on their inter personality skills, which type of environment aspects having round them, psychological is most important factor of students life. One of the crucial problems in data mining is how to classify the real world dataset online or offline if it is stream data. Stream data are an ordered sequence of instances that can be read only once and to store result in well manner. Machine learning, is a part of artificial intelligence, it can learn the object and build the object after the system has been studied on that object. For example, machine learning system could be trained on robot to identify different shape of object of square, rectangle and quadrilateral. Machine Learning, is a knowledge representation of instances. Knowledge representation is to describe or represent information in such manner to solve the complex tasks such as diagnosing a medical condition. For example if particular patient having cancer diseases or not. But to diagnoses diseases, we need previous information of that patient to make some future decision so how it can be possible? To solve this problem, machine learning introduces a new term called Incremental Learning [3]. Considering this incremental learning is divided into 3 parts. First parts is it starts to learn the instance but without keeping that instance in system [4]. Meanwhile whenever new data is captured, old data are discarded. Because of this new data are incorporated into the classification model. Due to this, whatever data is previously learned will forget by classification model. So same training data will produce by different classification rules because the order of obtaining data is different. Arrival of new data is combined with old data as training data to modify the classification model. That means it is started learning data with partial instance memory. AQ-PM learning method proposed by Maloof and Michalski [5] which stores data located near the rule boundary. In addition to this, streaming ensemble algorithm for classification developed by Street and Kim [7]. Firstly the whole algorithm is divided into small fix sized continuous chunks. Then it build the classification model for each individual chunk. By combining to this individual classification model, an ensemble classification model is constructed.

The basic aim behind this research work is to identify if any structure or entity is required to predicate students classification that could be depend on university or colleges characteristics' as well as their personal growth. Using predication colleges or university will easily identify that which different factors are required to build the students carrier. For example if particular student is strong in basket boll but along with this he/she might be strong in other game too. So our database will identity if is any hidden knowledge will they have and this thinks will improve student's skill, confidence and personal skills too. It will indirectly beneficial for colleges. That means system collect the data continuously as offline for making reliable predications, after collecting data if any changes required to transform the data for particular pattern identification and how to improve it, which type of data will collect to enlarge the usability of the analysis results [2].

The performance is measured under the some criteria. For analysis and classify the sample following rules applied : above 10 % is high , 2 to 3 % is normal , 8-1 % is low, below 1 is middle, 2-11 % is excellent, 7-9 is ordinary, 6-18 is extraordinary. Student performance is basically calculated by according to 1-7 points (1-Strongly Agree and 5-Strongly Disagree) and evaluation score. There is several classification techniques apply to classify the data whether it will be based on 2 classification problem or multiclassification problem.

Once data is apply on classification techniques it shows us confusion matrix, It has the training set, percentage split and cross-validation methods. It shows FP rate, TP rate and ROC values too. It is also used to check the correct and incorrect classified instances of the given input.

This study is more useful for identifying weak students and the identified students can be individually assisted by the educators so that their performance is better in future. This study investigates the accuracy of some classification techniques for predicting performance of a student.

II. BACKGROUND

Classified students by using genetic algorithms to predict their final grade Minaei-Bidgolim, et al. (2003). Predicted a student's marks (pass and fail classes) using the regression methods, Kotsiantis and Pintelas (2005). Predicted a student's academic success (classified into low, medium, and high risk classes) using different data mining methods (decision trees and neural network) Superby, Vandamme and Meskens (2006). Al-Radaideh, Al-Shawakfa and Al-Najjar (2006) applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan. Romero et al. (2008). Predicting students' performance using neural networks and classification trees decision-making, and with the analysis of factors which influence students' success by Zekić-Sušac, Frajman-Jakšić and Drvenkar (2009). Kumar and Vijayalakshmi (2011) using the decision tree predicted the result of the final exam to help professors identify students who needed help, in order to improve their performance and pass the exam[8].

Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme [9] have applied machine learning techniques to improve the prediction results of academic performances for two the real case studies. Three methods have been used to deal with the class imbalance problem and all of them show satisfactory results. They first re balanced the datasets and then used both cost-insensitive and sensitive learning with SVM for the small datasets and with Decision Tree for the larger datasets. The models are initially deployed on the local web.

Cortez and Silva [10] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset

Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [11]. Mining in educational environment is called Educational Data Mining.

Han and Kamber [12] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

III. METHODOLOGY

3.1. Decision Tree: J48

Decision tree algorithm comes under supervised learning algorithm. Decision tree used to solve classification and regression problem. By using decision tree algorithm we can create training model which can be used to predicate class or value of target variables. It has two factors – internal node and leaf node. Internal node corresponds to an attribute and leaf node corresponds to a class label. For predicting the class label for record we start from the root of the tree. Then we can compare the values of the root attribute with record's attribute. We have to continue comparing our record's attribute values with other internal nodes of the tree until we reach a leaf node with predicted class value. The attribute selection measures are: information gain and Gini Index. If dataset having 'n' attributes and we do not know which attribute should be selected as root node. We can do this as selecting any random node as root node but this cannot be worked. By following this it will give us low accuracy. For solving this attribute selection problem, researchers worked and devised some solution.

3.2. Random Tree

This algorithm was created by Leo Brieman and Adele Cutler in 2001. Random tree looks like decision tree which involve several trees, then combining their output to improve generalization ability of the model. The method of combining trees is known as an ensemble method. To produced strong learner combine weak learner is know as ensembling. This algorithm used for classification and regression problem. This algorithm will handle the missing values. It is split in two stages. First random tree creation pseudo code. Secondly, it creates the predication based on first part.

IV. DATASET DESCRIPTION

This psychometric test is divided into 3 parts. The first part of test is depends on how student will react in difficult situation, Is he/she control their emotions in bad or good conditions? Is they really concentrate on the task at hand in spite of disturbances? First part is analysis their emotional stability, Self Motivation, empathy and Self Awareness. The second part focuses on students' inter personality skills, managerial skills, Self Development and value orientation. Third part is

focus on commitment and altruistic behavior. The final part of the questionnaire depicts resource utilization. The Emotional Intelligence Scale (EIS) is measured by Evaluation score by 1-Strongly Agree and 5-Strongly Disagree.

The research is conducted from the Dr. D.Y.Patil School of Engineering and Technology, Pune. The student's dataset of 10 attributes along with 1325 instances. This dataset contains numeric values. This dataset having 7 classes which represents performance levels of student. This class label assigned to samples based upon some rules. The result dataset contains raw fact about the types of personality, informal learning factors and overall behavior. Selected dataset is preprocessed and converted into normalize form to the input of data mining tools.

In final preprocessed stage we used cross validation with 10 fold parameter. In this technique divide and conquer method is used. At the training part k-data is used to train and remaining subsequent k-1 part used for testing purpose. The main advantage of this technique is all the samples are used for validation and training. Each sample is used for validation exactly once.

V. RESULTS AND COMPARISON

This section shows that performance of of each algorithm on student dataset. Weka's Knowledge Flow tool is used to build the model and test the system of algorithms on student evaluation application. This is done for simulation purpose. Number of classifier is plotted on x-axis and Accuracy is plotted on y-axis. Combined graph show the performance in terms of accuracies. Name of the algorithm is shown above the graph.

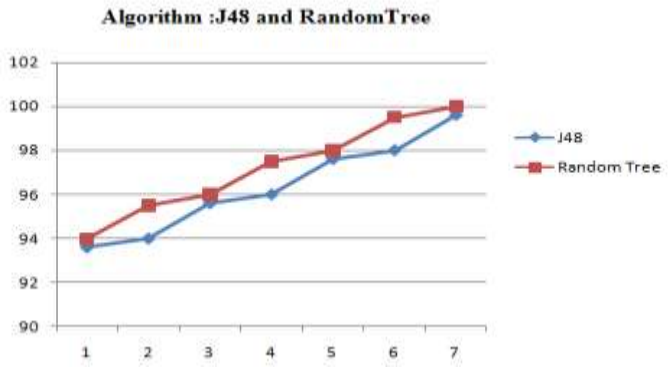

Figure 1: Summary graph of J48 and Random Tree

Table 1 show that the Random Tree performs well as compare to J48 algorithm. J48 gives less accuracy as data increased. It works on 1325 samples. Random tree work well on larger data too. The performance of algorithm shown in terms of different kind of errors accrued like incorrectly classified instances, mean absolute error, root mean squared error. This overall summary support that Random Tree algorithm gives much better results compare to J48 and it proves that Random Tree algorithm is best in such environment in student data prediction application. These algorithms are tested in this experiment on student dataset, and prediction is done. In the same way it is desirable to test them in another stream data application. Incremental algorithm used here which gives good results

Table 1: Summary of performance Incremental Learning on Students Dataset

Parameters	J48	Random Tree
Accuracy %	93.69	94.07
Incorrectly classified instances %	6.30	5.92
Time taken to evaluate model (in Sec)	0.03	0.1
Kappa Statistic	0.925	0.929
Mean Absolute Error	0.033	0.029
Root mean squared error	0.129	0.120

Fig 1 shows that instances is coming one after one continuously. In another side we can use instance in batch wise too. Fig 2 shows overall representation of student dataset.

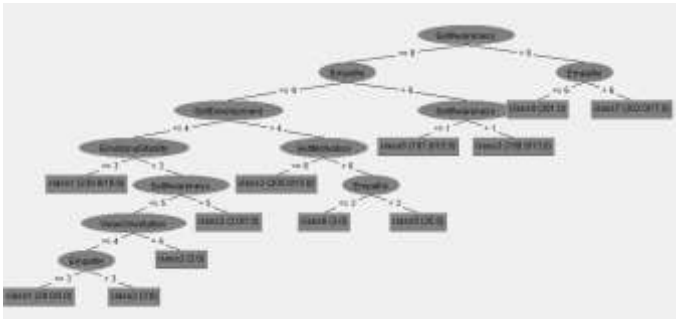

Figure 2: Tree View of Student Dataset

VI. CONCLUSION AND FUTURE SCOPE

Extraction of knowledge from raw data we can do this using data mining techniques that providing us interesting possibilities for the education domain. Classification model is based upon some selected inputs. In testing part, some of most influencing factors were identified and taken to predict the grades. Data mining techniques are applied to predict the performance of the students and found that Random Tree algorithm is best suited to predict the grades. As a result, having the information generated through our experiment, institution would be able to identify students at risk early, and provide better additional training for the weak students. Therefore, it seems to us that data mining has a lot of potential for education. Furthermore, we intent to enlarge the experiments to collect additional features like psychological factors which disturb the students, motivational efforts taken by the teachers and e-learning materials available to the students.

REFERENCES

[1]. A Comparison of Adaboost and Learn++ Distribution Update Rule, G.T. PRASANNA KUMARI, K.SEKAR, R.POORNIMA, International Journal of Engineering Sciences Research-IJESR, ISSN: 2230-8504, Vol 04, Special Issue 01, 2013

[2]. Predicting Student Performance by Using Data Mining Methods for Classification, Dorina Kabakchiev, BULGARIAN ACADEMY OF SCIENCES, ISSN: 1314-4081.

[3]. Vrushali A. Sungar, Roshani Ade," Classification of Student by Using V-Incremental Learning Based on Regression Learning Model", Third Post Graduate Symposium on Computer Engineering cPGCON2014

[4]. Schlimmer, J. C. and Fisher, D. H., "A Case Study of Incremental Concept Induction," Proceedings of the 5th International Conference on Artificial Intelligence, pp. 496 501 (1986).

[5]. Maloof, M. A. and Michalski, R. S., "Incremental Learning with Partial Instance Memory," Foundations of Intelligent Systems, Vol. 2366, pp. 16 27 (2002).

[6]. Jin, R. and Agrawa, G., "Efficient Decision Tree Construction on Streaming Data," Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 571 576 (2003).

[7]. Street,W. and Kim, Y., "AStreaming Ensemble Algorithm (SEA) for Large-Scale Classification," Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining, pp. 377 382 (2001).

[8]. Edin Osmanbegović *, Mirza Suljić **," DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE", Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.

[9]. Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme, "Improving Academic Performance Prediction by Dealing with Class Imbalance", 2009 Ninth International Conference on Intelligent Systems Design and Applications

[10]. P. Cortez, and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[11]. Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009.

[12]. J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[13]. Vrushali A. Sungar, Pooja D. Shinde, Monali V. Rupnar, " Predicating Student's Performance using Machine Learning ", Communications on Applied Electronics (CAE) – ISSN : 2394-4714, Volume 7 – No. 11, December 2017.