Special Issue Published in International Journal of Trend in Research and Development (IJTRD),
ISSN: 2394-9333, www.ijtrd.com

Analysis of the Performance of the Intelligent Training Data in Prediction

Jai Ruby
Research Scholar, Research & Development Centre,
Bharathiar University,
Tamilnadu, India

K. David
Assistant Professor, Department of Computer Science
H. H Rajahs College, Perambalur,
Tamilnadu, India

Abstract— In modern days, Educational Data Mining has grown into a research realm. The need of the Educational Institutions has grown to an extent that it has to study and analyze its stake holder's quality to be competitive. Among the various aspects of educational field, the academic performance of students and the results produce are the vital parameters which determine the quality of the Institution. In higher education institutions a significant amount of information is concealed and need to be extracted using Knowledge Discovery process. Data mining helps to extract the knowledge from available dataset and should be created as Knowledge Intelligence for the benefit of the institution. Various Data mining classification algorithms help to extract the knowledge. The study model is mainly focused on deriving a intelligent training dataset and use it for prediction with well known classification Algorithms like MLP and ID3.

Keywords *- Educational Data Mining, Academic Performance, Higher Education, Prediction, Classification, Multi Layer Perceptron, ID3, Training Dataset.*

I. INTRODUCTION

Education sector in India offers a lot of opportunities for the researchers. The higher education system of India is the largest in the world. The growing number of educational institutions demands quality. The quality is determined by various factors and the most important factor is the performance of the students which directly decide the growth and popularity of the institution. The institutions do have their own hidden potentials which can be brought forth by using convincing computing techniques. In today's scenario data mining over educational data is the accepted method of research. Educational Data Mining refers to techniques, tools, and research designed for automatically extracting the pattern from large repositories of data generated by or related to people's learning activities in educational settings. Key uses of EDM include learning and predicting student performance in order to recommend improvements to current educational practice. EDM can be considered as one of the learning sciences, as well as an area of data mining [1].

The technique behind the extraction of the hidden knowledge is a Knowledge Discovery process that extracts the knowledge from available dataset and creates a knowledge base for the benefit of the institution. Knowledge Discovery or data mining comes up with a number of classification algorithms for prediction of students' academic performance and it depends on several factors like different socio-economic, psychological and environmental factors.

The mined information that describe student performance can be stored as intelligent knowledge and can be used by the institutions principally for predicting the students' academic performance in advance. To improve the prediction accuracy is one of the key issues in the Educational Data mining field. In this paper, the researcher derived a new method for generating intelligent dataset to improve the performance of prediction on educational domain dataset for predicting students' academic performance. The researcher also made an effort to compare the prediction accuracy of two data mining classification algorithms ID3 and MLP. The experiment strengthens the fact that various factors identified in the study model [2] are really influencing in predicting the results.

This paper makes a new attempt to look into the higher educational domain of data mining to improve the prediction accuracy by using a hybrid model. Section 2 gives the methodology and techniques. Section 3 provides the general account of the model and the dataset under study. Section 4 predicts the academic performance of students using the intelligently derived dataset and the comparative analysis. Conclusion and a discussion on future work are in the final section.

A. Related Work

Han and Kamber [3] describe data mining as a tool that help the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. Brijesh Kumar Baradwaj, Saurabh Pal, in [4] conducted a study on a data set of size 50 Post Graduate students for mining educational data to analyze students' performance. Decision tree method was used for classification and to predict the performance of the students. Different measures that are not taken into consideration were economic background, technology exposure etc.

El-Halees. A [5] did a work on describing student behaviour with a dataset of size 151 that includes only personal and academic details of students. Classification based on Decision tree was done followed by clustering and outlier analysis. Z. J. Kovacic, in [6] presented a study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success by using CHAID and CART data mining algorithms. Mohammed M. Abu Tair and Alaa M. El-Halees [7] applied the data mining for discovering knowledge from data that come from educational environment. Students' data had been collected from the college of Science and Technology for a period of 15 years [1993-2007]. The collected data was pre-processed and data mining techniques are applied for analysing graduate students' performance. MuslihahW.et.al.[8] have compared Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting and classifying student's academic performance. Students' data were collected from the data of the National Defence University of Malaysia

(NDUM). H. W. Ian and F. Eibe gave a case study that used educational data mining to identify the behaviour of failing students who are at risk before the final exam [9].

Nguyen N et.al [10], compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of students of Under Graduate and Post Graduate students. The decision tree classifier provided better accuracy than the Bayesian network classifier. Different measures that were not taken into consideration were economic background, technology exposure etc. Bengio Y. et.al [11], discussed that neural networks are suitable in data-rich environments and are typically used for extracting embedded knowledge in the form of rules, quantitative evaluation of these rules, clustering, self-organization, classification and regression. Neural networks have an advantage over other types of machine learning algorithms. Vasile P. B [12] presented that the use of classification models such as decision tree and Naïve Bayes in the education field to predict students' behaviour. M. Wook, et.al [13] compared two data mining techniques which are: Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting students' academic performance. Jai Ruby & David [14] compared various data mining algorithms using the student dataset considering only the influencing factors and proved that MLP and ID3 best classification algorithms.

Romero and Ventura [15], had a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it had a specific requirement not presented in other domains. Kuyoro' et.al [16] worked on identifying the optimal algorithm suitable for predicting first-year tertiary students' academic performance based on their family background factors and previous academic achievement. Five decision tree algorithms, five rule induction algorithms and an artificial neural network function were taken for the study. It was discovered that random tree performance was better than that of other algorithms used in this study.

II. DATA MINING AND TECHNIQUES

Knowledge Discovery in Databases (KDD) refers to extracting or "mining" knowledge in the form of rules, patterns or models from large amount of data. It involves various process like Data pre-processing, data mining, pattern evaluation in extracting knowledge from data. Various tasks such as association, clustering, classification, prediction, etc are involved in Data mining. Two steps are involved in classification. In the first step, a model that describes a predetermined set of classes or concepts is made by examining a set of training dataset. The learning is known as supervised learning as the class labels of all the records of the dataset are known. The models are usually in the form of classification rules or decision tree. In the second step, the model is put to test using a different data set that is used to estimate the predictive accuracy of the model. Various methods like holdout, random sub sampling, k-fold cross validation, stratified cross validation, bootstrapping are used to estimate the accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify the dataset for which the class label is not known in advance [3]. Basic techniques for classification are decision tree induction, Bayesian classification and neural networks. A number of well - known data mining classification algorithms such as ID3, REPTree, Simplecart, J48, NB Tree, BFTree, Decision Table, MLP, Bayesnet, etc., exist. A neural network is a network that has the ability to learn from its background and improve its performance through learning. Multilayer Perceptron algorithm is one of the most widely used and common neural network.

Multilayer Perceptron is a feed forward artificial neural network model trained with the standard back propagation algorithm that maps sets of input data onto a collection of acceptable output. They learn how to transform input data into a desired response, so they are widely used for pattern classification and prediction.

ID3 - Iterative Dichotomiser 3 is a decision tree Induction algorithm and induces decision trees from data. It is a supervised learning algorithm that is trained by examples for different classes. After being trained, the algorithm should be able to predict the class of a new item. It uses the statistical property of entropy. Entropy measures the amount of information in an attribute.

III. DATASET, TOOL AND METHODOLOGY

The dataset used for this study was taken from PG Computer Application course offered by an Arts and Science College between 2007 and 2012. The data of 165 students were collected. Student personal and academic details along with their attendance were collected from the student information system. The collected information was integrated into a distinct table. Among the different attributes initially present using feature selection techniques like chi square, info gain, gain ratio, correlation and regression it was found that the high impact attributes that contribute for the performance of the students are Theory, Medium of Study, Previous Course studied, UG Percentage, Stay, Extra Curricular Activities and Family Income[2]. The influencing attributes are selected and are used to classify and predict the student performance using weka data mining tool.

Algorithm:

```
Step 1 : Derive Dataset
Step 2 : Check preprocess needed
Step 3 : If needed, do pre-process
Step 4 : Find out intelligent subset of data
Step 5 : Prepare Training dataset
Step 6 : Select test dataset
Step 7 : Classify using ID3
Step 8 : Classify using MLP
Step 9 : Compare results
```

The programming tool used for preprocessing and for generating intelligent dataset is MATLAB. It stands for MATrix LABoratory. It is a high-performance language that integrates computation, visualization, and programming environment. It also has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. It has powerful built-in routines that enable a very wide variety of computations. Specific applications are collected in packages referred to as toolbox. There are toolboxes for signal processing, symbolic computation, control theory, simulation, optimization, and several other fields of applied science and engineering.

It is noted that the performance of the algorithms is influenced by the quality of the dataset we select for training and testing. The given dataset is checked for the need of pre-processing. Various flaws like missing data, inconsistencies are removed by a user defined algorithm. To produce a good quality training dataset, it is proposed to derive the unique records present in the dataset and is compelled to be the part of the training dataset. The algorithm for generating the intelligent sub dataset is developed using MATLAB.

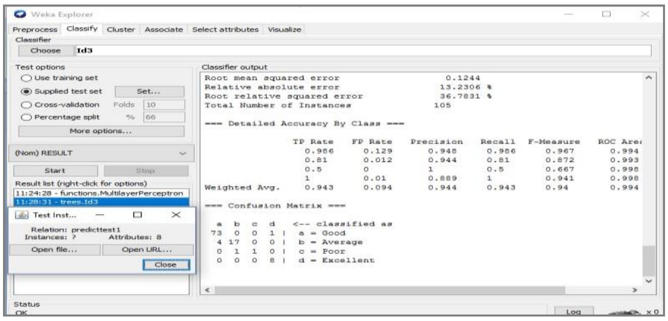
The initial dataset was split up into two sets. Two thirds of the data are allocated to be the training set where it is confirmed that the derived intelligent subset be a part of it. The remaining one third is allocated to be test set. The training set helps in building the model and it is used for classification. The classify panel in the weka tool facilitates applying classification algorithms and to estimate the accuracy of the predictive model. Two different classifiers ID3 and MLP (Multi Layer Perceptron) were used in the model.

The given dataset is scrutinized for the need of pre-processing. If needed, it is pre-processed and it is subjected to feature extraction. This minimises the number of attributes that are not so relevant in prediction of academic results in a student dataset. The model extracts the subset which is the intelligent subset of the dataset. From the original dataset derive for about two third of the dataset which strictly includes the intelligent subset. This is termed as training set and other the remaining one third of the dataset is consider to be the test. Then the model is tested with ID3 and MLP classification algorithms and is used for predicting the result of the unknown dataset.

IV. EXPERIMENTAL RESULT ANALYSIS

ID3 and MLP were the two classification algorithms used for analysis. The student records were pre-processed. The preprocess step involves filling up of missing values, data is normalized and generalized. Using the algorithm developed in MATLAB, unique records are retrieved and is generated as a intelligent subset data. This intelligent set is clubbed with a part of the initial dataset and derived as a training dataset. The remaining is considered as test dataset.

The data which is in the form of excel file is converted to Attribute Relation File Format and a final format of training and testing datasets were prepared. The Fig. 1 shows the classification of the training data set using ID3 algorithm and Fig. 2 shows the classification of test data using ID3 algorithm via weka tool. If the accuracy of the model is acceptable then it is used for the prediction of data for which the class label is unknown. The same procedure is followed for testing with MLP algorithm.



Classifier: Id3

Test options
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

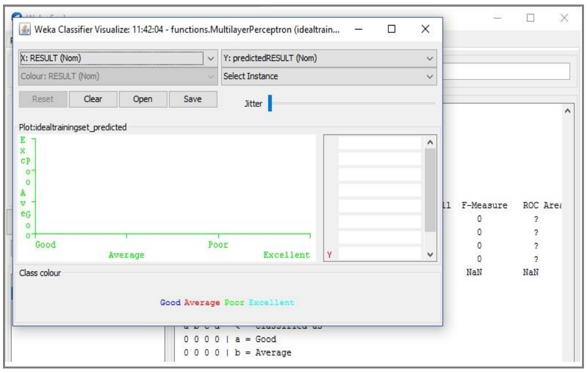
(Nom) RESULT

Result list (right-click for options)
11:24:28 - functions.MultilayerPerceptron
11:28:31 - trees.Id3

Test Inst...
Relation: predicttest1
Instances: ? Attributes: 8

Classifier output
```
Root mean squared error                  0.1244
Relative absolute error                 13.2306 %
Root relative squared error             36.7831 %
Total Number of Instances              105

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area
                 0.986    0.129     0.948     0.986    0.967     0.994
                 0.81     0.012     0.944     0.81     0.872     0.993
                 0.5      0         1         0.5      0.667     0.998
                 1        0.01      0.889     1        0.941     0.998
Weighted Avg.    0.943    0.094     0.944     0.943    0.94      0.994

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 73  0  0  1 |  a = Good
  4 17  0  0 |  b = Average
  0  1  1  0 |  c = Poor
  0  0  0  8 |  d = Excellent
```

Status
OK

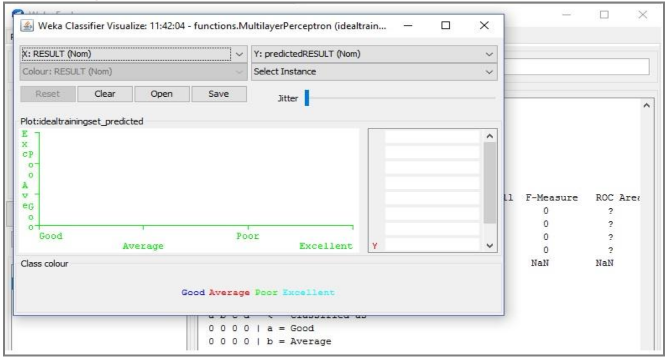


Weka Classifier Visualize: 11:42:04 - functions.MultilayerPerceptron (idealtrain...

X: RESULT (Nom)
Y: predictedRESULT (Nom)
Colour: RESULT (Nom)
Select Instance

```
ll  F-Measure  ROC Area
       0          ?
       0          ?
       0          ?
       0          ?
      NaN        NaN

0 0 0 0 | a = Good
0 0 0 0 | b = Average
```

Fig. 2 Classification of Training Dataset using



Weka Classifier Visualize: 11:42:04 - functions.MultilayerPerceptron (idealtrain...

X: RESULT (Nom)
Y: predictedRESULT (Nom)
Colour: RESULT (Nom)
Select Instance

```
ll  F-Measure  ROC Area
       0          ?
       0          ?
       0          ?
       0          ?
      NaN        NaN

0 0 0 0 | a = Good
0 0 0 0 | b = Average
```

Fig. 3 Prediction of unknown dataset using MLP

The sample set was divided into 4 sets of distinct one-third records. They are the data set used in Run1 through Run4 respectively. In each run, four new sets of data whose class labels were unknown was given for prediction. Fig. 3 shows the prediction of new test data whose label is unknown using MLP algorithm via weka. The same experimental setup is repeated using ID3 classification algorithm and prediction accuracy obtained during different runs are recorded. Fig.4 shows the comparison of prediction accuracy of different datasets using ID3, MLP where the training set is a intelligently derived set. The prediction accuracy percentage using intelligent dataset shows a better result using ID3 and MLP.

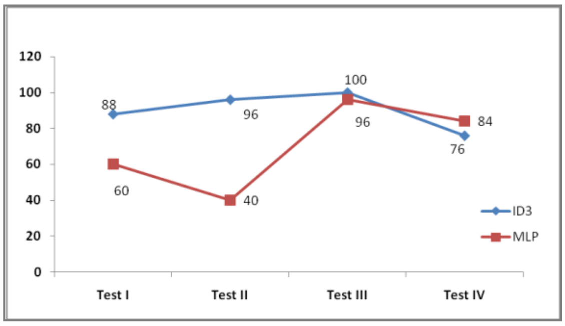


Fig. 4 Comparison of Prediction Accuracy using ID3, MLP using intelligent Dataset

Table I shows prediction accuracy of the Classification Algorithms in percentage for a different combination of training data set with new unknown dataset.

Table 1 - Prediction Accuracy Of The Various Classification Models In Percentage For 4 Different Dataset

	Test I	Test II	Test III	Test IV	Average
ID3	88	96	100	76	90
MLP	60	40	96	84	70

The average accuracy percentage of 4 different datasets is found to be 70 % for MLP and 90% for ID3. Also the prediction by ID3 shows consistency inspite of different test set. There is a drastic improvement in the prediction percentage if intelligent dataset is used. The experimental results show that the prediction accuracy highly depends on the classification algorithm and the generated training dataset.

CONCLUSION

This experimental model is mainly focused on analyzing the prediction accuracy of the academic performance of the students using different classification algorithms like ID3 and MLP. The proposed model using an intelligent dataset proved to be a better performing model as the prediction rate is notably high. This analysis helps the institution to know the academic status of the students in advance and can concentrate on weak students to improve their academic results. To represent the knowledge created in a standard form would be a future work.

References

[1] Monika Goyal & Rajan Vohra, “Applications of Data Mining in Higher Education” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, pp.130-120.

[2] Jai Ruby & K. David, “A study model on the impact of various indicators in the performance of students in higher education“, IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, May-2014, pp.750-755.

[3] Han. J & Kamber. M, “Data mining concepts and techniques”, San Francisco, USA, Morgan Kaufmann, 2001.

[4] Brijesh Kumar Baradwaj, Saurabh Pal,” Mining Educational Data to Analyze Students‘ Performance” IJACSA, Vol.2, No.6,2011

[5] El-Hales-A.(2008),‟Mining Students Data to Analyze Learning Behavior: A Case Study‟, The 2008 International Arab Conference of Information Technology(ACIT2008)- Conference Proceedings, University of Sfax, Tunisia,Dec 15-18.

[6] Kovacic Z. J., “Early prediction of student success: Mining student enrollment data”, Proceedings of Informing Science & IT Education Conference 2010.

[7] Mohammed M. Abu Tair & Alaa M. El-Halees, ‘Mining Educational Data to Improve Students’ Performance: A Case Study’, 2012

[8] Muslihah W., Yuhanim Y., Norshahriah W., Mohd Rizal M., Nor Fatimah A., & Hoo Y. S., ‘Predicting NDUM Student’s Academic Performance Using Data Mining Techniques’, In Proceedings of the Second International Conference on Computer and Electrical Engineering, IEEE computer society, 2009.

[9] Ian H. W. & Eibe F., “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,” California: Morgan Kaufmann, 2005

[10] Nguyen N., Paul J., & Peter H., ‘A Comparative Analysis of Techniques for Predicting Academic Performance’. In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. pp. 7-12, 2007.

[11] Bengio Y., Buhmann J. M., Embrechts M., & Zurada J. M., "Introduction to the special issue on neural networks for data mining and knowledge discovery," IEEE Trans. Neural Networks, vol. 11, pp. 545-549, 2000

[12] Vasile P. B., “Analysis and Predictions on Students’ Behavior Using Decision Trees in Weka Environment”. Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, IEEE, (2007).

[13] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. Seong, “Prediction NDUM student's academic performance using data mining techniques,” presented at the International Conference on Computer and Electrical Engineering, 2009.

[14] Jai Ruby & K. David, “Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study “, IJRASET International Journal for Research in Applied Science & Engineering Technology, Volume 2 Issue XI, November 2014

[15] Romero, C. and Ventura, S. (2007) ‘Educational data Mining: A Survey from 1995 to 2005’, Expert Systems with Applications (33), pp. 135-146.

[16] Kuyoro 'shade o, Nnicolae Goga, Oludele Awodele and Samuel Okolie “Optimal algorithm for predicting students’ academic performance” International journal of computers & technology volume 4 no. 1, JAN-FEB, 2013 ISSN 2277-3061