Mining Educational Data for Students' Placement Prediction using Sum of Difference Method

Ramanathan.L
Assistant Professor(Senior)
SCSE, VIT University
Vellore, Tamindadu-632014, India

Swarnalatha P
Assistant Professor (S.G)
SCSE, VIT University
Vellore, Tamindadu-632014, India

D. Ganesh Gopal
Assistant Professor(Senior)
SCSE, VIT University
Vellore, Tamindadu-632014, India

ABSTRACT
The purpose of higher education organizations have to offer superior education to its students. The proficiency to forecast student's achievement is valuable in affiliated ways associated with organization education system. Students' scores which they got in an exam can be used to invent training set to dominate learning algorithms. With the academia attributes of students such as internal marks, lab marks, age etc., it can be easily predict their performance. After getting predicted result the performance of the student to engage with desirable assistance to the students will be improved. Educational Data Mining (EDM) offers such information to educational organization from educational data. EDM provides various methods for prediction of student`s performance, which improve the future result of students. In this paper, by using the attributes such as academic records, age, and achievement etc., EDM has been used for predicting the performance about placement of final year students. Based on the result, higher education organizations can offer superior education to its students.

Keywords
Data Mining, Educational Data Mining, Sum of Difference, Prediction.

1. INTRODUCTION
The basic idea behind data mining means obtaining the knowledge from the immense set of data, which is useful and favorable. There are various methods used for knowledge discovery from huge data such as classification, clustering, prediction, association rule etc. In data mining, classification is the simple employ data mining technique that utilizes a group of pre-sorted example to create a model which can classify the immense set of data. Using Decision tree and neural network, classification is performed on huge amount of data. Clustering is the other technique in data mining, which are used to find similar type of object. It can easily determine the small and deep patterns. It is not so cheap. So in preprocessing step, clustering can be used for attribute selection. KNN algorithm is used for clustering. For clustering , prediction can be used with regression technique. It defines a relationship between single or more than one dependent variables and independent variables. For prediction and classification, same model can be used like decision tree or neural network. EDM is an application of data mining, which offer necessary information to educational organization, which hides in educational data of students. From the past and operational data inherit in the database of educational organization, the data can be gathered. EDM introduces a various fields such as prediction, clustering, Distillation of human judgment, Discovery with model, Relationship mining. Using techniques such as Decision trees, Multilayer Perceptron, Neural networks, Bayesian Network, Support vector regression and Naive Bayes Simple algorithm are used to describe many types of knowledge like association rules, classification and clustering. The main objective of this paper is to predict the placement of student by using similarity measure with mathematical method which is called sum of difference (SOD). SOD is used to analyze the performance of students using the attributes such as academic records and to find a method is more accurate result for prediction.

2. LITERATURE SURVEY
In data mining, to predict the performance of student there are various data mining tools, where the work of the paper is to set the default parameter of various algorithms to gain the highest accuracy, which is hard for a non-technical person. In recent year, some works have done on automatic parameter tuning. They have taken 14 different educational dataset and focused how to increase precision of the result by using parameter tuning of J48 algorithm, and also made comparison with accuracy when using three basic characteristics (number of instances, number of attributes and number of classes) numerical attributes over categorical attributes [1]. For performance prediction, temporal data of student is important for performance prediction. The performance of student is improved after studying relevant skill, which is not counted in recent work. Pardos have done work on data leakage by using two prediction algorithms: linear regression and Random forest. By including K-means clustering technique he claimed to improve prediction accuracy [2].In other research work, individual information of a student's performance is used to forecast rather than historical data of other student. Obviously it is cleared that learning level of each level are different, so it is not used to fit to judge performance of a student based on others. In proposed model , the paper is used with two skills: how to adopt a student when the student is in execution of the task and what is the difficulty level of the task [3].Additionally B. Sen [4] is used with CRSIP-DM to predict test score of student with four prediction model, and found that C5 algorithm is best among rest of algorithm. By using various attributes of student, the paper is analyzed with the sensitivity of attributes, which are much important for prediction. Baker [5] introduced a model which is used to predict the preparation for future learning of student. Students have ability to learn new thing with the help of his existing knowledge, and they have skill to learn quickly new skill. He used knowledge engineering methods with data mining. This model needs small data of student. Heffeman [6] compared between single algorithm and ensemble method with RMSE and Correlation value with respect to accuracy. Each algorithm generates better result with different aspect. So he mixed more than single algorithms and predicts post score of student to achieve better accuracy. The idea behind the model is called tabling, in which checked the student's response to

solve same pattern of question, which he already faced. Vera [7] worked on imbalanced data, in which the number of instances in one class is much greater than the number of instances in another class or other classes, which decreases the accuracy of result. He use various classification algorithm with supervised data filter technique SMOTE with best attributes among all attributes without affecting reliability. Akcapinar [8] use 10-cross fold validation method with two different Random Forest Regression model to find high accuracy. School [9] explored social activities of a group rather than single person activity. This model provides to group a Collaborative learning task and analyzed the group performance and relationship between members. Wang [10] used the knowledge tracing model to find the accuracy of student first response answer and to improve the accuracy of prediction. Student first response may be affected by his skill or by guess about success. On the other hand it may be affected by forget or slip about unsuccessful. Gowda [11] proposed a new model to predict the post score of student with two factors: to detect the time by time learning and to improve the student's slipping and guessing. In this work slipping is the best model because improving the slip model improves the performance of student than other model.

3. METHODOLOGY

To predict the student's placement status, finding is not so easy task, because the placement depends on various factors. Sometimes a student who have good academic record and still not placed, on the other hand a student who has not good in his academic life he got placed. Sometimes it depends on luck and time and the frequency of companies which comes in college campus. It is hard to analyses using above factors to predict the student's placement. Initially, the paper has dataset of students' academic record; to analysis the pattern from given dataset, which affect the placement status. Secondly, to collect the large dataset which is also a difficult task? If the paper have large real dataset, then analysis can be more preciously resulted, which would be more accurate. Third thing is to collect a lot of information about student, which is called attributes, as the paper is not sure which attributes effect the output ,so if the paper have a lot of attributes, then the paper can find precious result ,which will be more accurate our model for prediction. Similarity measure is used to find the pattern in the given object. There are various mathematical methods, which is used to find the pattern from the given data set. Sum of difference is one of them, which is used to find the similarity from the given dataset.

3.1 Attribute List

Gender	- Male, Female
Category	-General, OBC, SC/ST, Other
Academic gap	-0 to 10
Grade in 10th exam (G10E)	-0 to 100
Grade in 12th exam (G12E)	-0 to 100
Number of arrear faced (NOAF)	- 0 to 10.
Grade in BTech exam (GBTECHE)	-0 to 100
English communication skill (EC)	-poor, average, Good
Extra technical course (ECC)	- yes, no
Grade in M.Tech exam (GMTECHE)	-0 to 100.
Placed	- yes, no

3.2 Data Selection

The data have been generated from a Google page by filling students. The initial data contains the performance profile gathered from a number of 50 students with 11 listed attributes which include Gender, Category, Academic gap, grade in 10th exam(G10E), grade in 12th exam, no of arrear (NOAF), grade in Btech exam(GBTECHE) , English communication skill, extra technical course(ECC),grade in M.Tech exam(GMTECHE), Placed. The data contains various types of values either string or numeric value. The data were then processed for generating rules.

3.3 Data Pre-Processing

Now upon initial examination on the data, missing values of the Gender, Category, Academic gap, were found and removed according to the numbers of missing values in one instance as part of the data cleansing process.

3.4 Data Transformation

After the cleaning process, data is converted into a form to perform the data mining process easily. The paper have converted the values in small values because it is hard to deal with large values. After all pre-processing and transformation have been implemented, the data was than ready to be mined.

3.1 Architecture

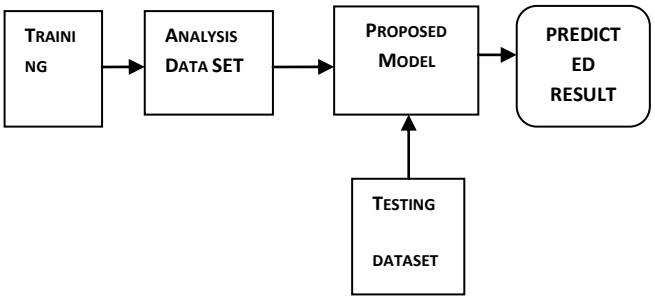

Fig. 1. Architecture of proposed model

In the architecture the paper will prepare input dataset for statistical analysis. Based on similarity measure, the paper can find a pattern in given attribute, which is used for prediction for placement. There are various mathematical methods to find pattern for prediction. In this work, the paper have chosen sum of difference method to find pattern for prediction of student's placement. Based on mathematical method the paper is implemented in C sharp language, which will predict the student placement. And the paper is used with most important attribute which can be useful for prediction and set the priority of each attribute. For each attribute the paper has chosen a reference point in the range of 0.0 to 1.0 based on its priority. Now the paper have chosen other reference point for each attribute in the range of value of given attribute and subtract it with each value respectively. Then the paper work is dealt with the addition of each subtracted value. Based on sum of difference value, the paper can find a cut-off value where placement value can be altered.

3.5Algorithm

Algorithm [P, I, j, Z, T, SOD]

1. Analysis the most important parameter P from the input dataset. P = {P1, P2…….Pi} where i = number of most important parameter.
2. Choose appropriate reference value Z for each attribute Pi based on priority for normalization. Where Z = {Z1, Z2……..Zi} where i=number of most important parameter.
3. Multiply by reference number Zi to each parameter value Pij.
   Where J = number of record.
4. Choose another reference value T in the range of parameter value for each attribute Pi based on priority.

Where T= {T1, T2…….. Ti} where i=number of most important parameter.

5. Subtract by reference number Ti to each parameter value Pij.
   Where J = number of record.
6. Calculate SODi , Sum of differences of Pi.
7. Based on SODi value analysis the result and set the value for prediction.
8. End.

3.6 Implementation

Implementation is done using C# a programming language as front End. The C# is used to create a user interactive page and MS Visual Studio 2012 as the run time environment. First of all, run the application. A user interface screen will be generated as given figure 2. In left top corner, a browse button is used for selecting the input file. If there is no file selected, a warning message will be displayed that select an input file. In right top corner, a browse button is used for set the path for output file. Below of it, there is a field for set the name for output field. After selecting input file, chose the path for the output file where user want to save the output file. User can give the name for the output file. If the user does not provide name and set the path for the output file, output file will be saved in C drive by default. After selecting the input file and set the path for the output file, a notepad file is generated which contains the predicted result.

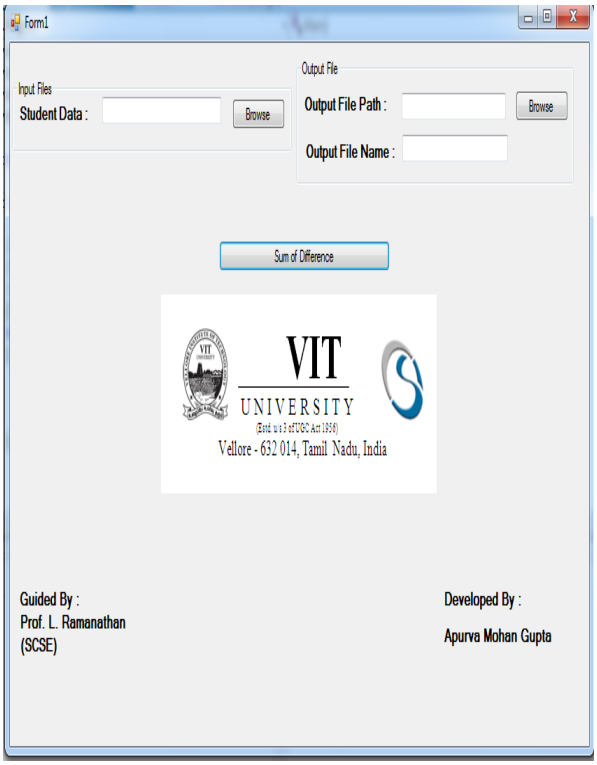

Fig. 2. User interface screen

3.7 Results and Discussion

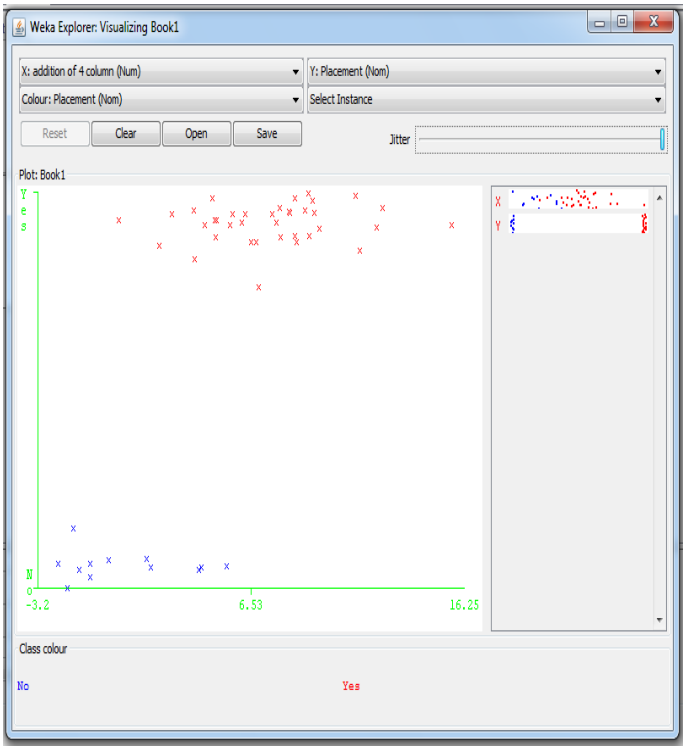

Fig. 3. Sum of difference score vs. Placement

In above figure 3, X-axis is given with the sum of difference value and Y – axis shows the placements value (yes, no).in graph Red Cross indicates yes value and blue cross shows no value of student's placement. It is obvious from the figure 3, when the sum of difference value increase till 4; it indicates no value of placement. When the value goes high than 4, it indicates yes value of placement. So the work of the paper can say that if the SOD (sum of difference) of a student is below 4, he will not be placed, if no then he will get placed in placement.

3.8 Future Enhancements

The future enhancement of the paper is based on applying clustering techniques of [12] [13] [14] [15] [16, 17] on numerical data in an efficient way for improvement of this paper. However, future research may focus more attributes with different dataset.

4. CONCLUSION

In the paper, Sum of difference method has been used to achieve the goal and extract the patterns from the given dataset. The paper is dealt with a proposed model using Sum of difference method to discover most essential attributes between items in a given dataset. Sum of difference is used to find a pattern from the given dataset with statistical analysis. The paper have been simplified with the value of attributes, because it is hard to deal with large value, and then normalize to find a pattern.

The proposed model, which efficiently predicts the placement of a student. It is obvious that placement is not so easy to predict because it depends on many attributes, even the paper is considered with four attributes. The paper has chosen the best combination of attributes. This combination works well for given dataset. From the given dataset, it is enough to predict the placement status of student. Hence, our proposed

model with SOD is found to be more admirable in terms of efficiency.

5. ACKNOWLEDGMENTS

We would like to thank Mr. Apurva who worked with us for the successful completion of the paper and VIT University for their continuous support.

6. REFERENCES

[1] Molina, M. M., Luna, J. M., Romero, C., & Ventura, S., 2012, "Meta-learning approach for automatic parameter tuning: a case of study with educational datasets", in Proceedings of the 5th international conference on educational data mining ,pp.180-183.

[2] Pardos, Z. A., Wang, Q. Y., & Trivedi, S., 2012,"The real world significance of performance prediction".

[3] Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L., 2011, "Factorization models for forecasting student performance. In Proceedings of the 4th international conference on educational data mining",pp. 11–20.

[4] B.Sen, Ucar E., delen D.,2012,"Predicting and analyzing secondary education placement-test scores.

[5] Baker, R. S. J. D., Gowda, S. M., & Corbett, A. T.,2011, "Automatically detecting a student's preparation for future learning: Help use is key",in Proceedings of the 4th international conference on educational data mining ,pp. 179–188.

[6] Pardos, Z. A., Gowda, S. M., Baker, R. S. J. D., & Heffernan, N. T.,2011, "Ensembling predictions of student post-test scores for an intelligent tutoring system in networks",in Proceedings of the 3rd international conference on educational data mining, pp.299–300.

[7] Marquez-Vera, C., Romero, C., & Ventura, S.,2011, "Predicting school failure using data mining",in Proceedings of the 4th international conference on educational data mining,pp. 271– 275.

[8] J. Akcapinar, G., Cosgun, E., & Altun, A., 2011,"Prediction of perceived disorientation in online learning environment with random forest regression", in Proceedings of the 4th international conference on educational data mining, pp.259–263.

[9] Schoor, C., & Bannert, M.,2012, "Exploring regulatory processes during a computer-supported collaborative learning task using process mining. Computers in Human Behaviour", Vol. 28(4), pp.1321–1331

[10] Wang, Y., & Heffernan, N. T.,2012,"Leveraging first response time into the knowledge tracing model",in Proceedings of the 5th international conference on educational data mining, pp. 176–179.

[11] Gowda, S. M., Rowe, J. P., Baker, R. S.J. D., Chi, M., & Koedinger, K. R.,2011,"Improving models of slipping, guessing, and moment-by-moment learning with estimates of skill difficulty",in Proceedings of the 4th international conference on educational data mining,pp. 199–208.

[12] Swarnalatha, P. and.Tripathy B.K., 2012,"A Centroid Model for the Depth Assessment of Images using Rough Fuzzy Set Techniques" at International Journal of Intelligent Systems and Applications,vol.1.no.3. pp. 20-26.

[13] Tripathy, B.K., Swarnalatha P., et.al.,2013,"Rough Intuitionistic Fuzzy C-MeansAlgorithm and a Comparative Analysis", Proceedings of the 6th ACM India Computing Convention, COMPUTE '13, Aug 22-24, 2013 ACM 978-1-4503-2545- 5/13/08.

[14] Swarnalatha, P. and Tripathy B.K.,2013, "Depth Computation using bit plane with clustering techniques for satellite images", International Journal of Earth Sciences and Engineering, vol.6,no.6(01),pp.1541-1553.

[15] Swarnalatha, P. and Tripathy B.K.,2014, "A Comparative Study of RIFCM with Other Related Algorithms from Their Suitability in Analysis of Satellite Images using Other Supporting Techniques", Kybernetes, Emerald Publications,vol.43. no.1., pp.53-81.

[16] Swarnalatha, P.,. Tripathy B.K, Nithin Prakash Ladda and Debashish Ghosh,2014,"Cluster Analysis Using Hybrid Soft Computing Techniques", Proceedings of International Conference on Advances in Communication, Network, and,Computing, CNC, Elsevier,pp.516- 524.

[17] Swarnalatha, P.,. Tripathy B.K,.Prabu S, Ramakrishanan R. and Manthira Moorthi S.,2014, "Depth Reconstruction using Geometric Correction with Anaglyph Approach for Satellite Imagery", Proceedings of International Conference on Advances in Communication, Network, and Computing, CNC, Elsevier, pp.506-515.