Optimization Feature Selection for classifying student in Educational Data Mining

R. Sasi Regha
Assistant professor, *Department of computer science*
SSM College of Arts & Science,
Kumarapalayam, Tamil nadu, India

Dr R. Uma Rani
Associate professor, *Department of computer science*
Sri Sarada College for Women,
Salem, Tamil nadu, India

Abstract- Technology has revolutionized the field of education. As a result, the education related data is been increasing rapidly. This made data mining approaches to spot over educational data ended in Educational data mining (EDM). The regulation focuses on investigating educational data to build models for enhancing learning experiences and improving institutional effectiveness. In this paper, the data mining techniques is used for predicting the student performance in different educational levels. Irrelevant features, along with redundant features, rigorously influence the accuracy of the classification of student performance. Therefore, feature selection should be able to detect and eliminate both irrelevant and redundant features as hard as possible. After feature selecting process, two effective classification techniques i.e., Prism and J48 is used for predicting the student performance. Experimentation result is shown that the feature selection method is well effective.

Keywords – Educational Data Mining (EDM), Feature Selection, Symmetric Uncertainty (SU) and Classification Artificial fish swarm, Cuckoo search optimization.

I. INTRODUCTION

Educational Data Mining (EDM) is defined as an field of scientific inquiry centered on the growth of approches not only for building discoveries within the exclusive kinds of data that come from academical settings but also for using those methods subsequently to understand successfully the students and the settings which they learn in, has emerged as an independent research area in recent years [1]. One of the key areas of applications of EDM is development of student models that would estimate student's characteristics or academic performances in educational institutions. Education Data mining is used to predicting students' performance with the aim of recommend improvements in academics. The past several decades have witnessed a speedy growth in the use of data and knowledge mining as a tool by which academic institutions take out useful unknown information in the student result repositories with the purpose of improves students' learning processes. As the end products of the models would be presented regularly to students in a comprehensive form, these end products would facilitate reflection and self-regulation during their study.

In prediction, the main aim is to improve a model, which can deduce a single aspect of the data (dependent variable) from some mixture of other aspects of the data (independent variables). Actually, prediction needs having labels for the output variable for a limited data set, where a label corresponds some trusted "ground truth" information about the output variable's value in specific cases. At the same time, prediction has two key uses within educational data mining. As far as first type of usage is concerned, prediction methods can be used to study the features of a model that are important for prediction and giving information about the underlying construction. This is a familiar technique in research that attempt to detect student educational performance, without predicting intermediate factors first. In a second type of usage, predictions approaches are used with the intention of detect the output value that would be in context and attain a label for that construction.

The prediction model performance is extremely depends on the option of selection of most relevant features from the list of features which is used in student data set. This can be attained by means of utilizing various feature selection methods on data set. In fact, percentage of accuracy is usually not chosen for classification, as values of accuracy are highly according to the base rates of different classes. Additionally, many factors affect the success of data mining algorithms on a given task. The worth of the data is defined as the parameter of analyzing the information is irrelevant or redundant, or the data is noisy and unreliable, then knowledge discovery during training is more difficult. In general, Attribute subset selection is the process of detecting and eliminates both irrelevant and redundant features as hard as possible.

Learning approaches differ for emphasis they place on attribute selection. At one intense are approaches such as the simple nearest neighbour learner. This is used for classifying the novel examples by attaining the nearest stored training example, using all the presented features in its distance evaluations. At the other extreme are algorithms that clearly try to focus on relevant features with pay no attention to irrelevant features. Despite whether a learner attempts to select attributes itself or ignores the issue, attribute selection prior to learning can be useful for enhancing the accuracy. Reducing the elements of the data reduces the size of the hypothesis space and permits techniques to operate more rapidly and more in effect.

II. Proposed Algorithm

A. Feature Selection Techniques –

Feature selection to try to identify which feature has the greatest effect on our output variable (academic status). There are an extensive range of attribute selection methods that can be grouped in various behaviors. One famous categorization is one in which the approaches differ in the way they estimate attributes and are classified as: filters, which select and analyse features autonomously of the learning process and wrappers, which utilize the classifier performance to estimate the desirability of a subset.

i) Correlation-based Feature Selection (CFS)

CFS estimates and ranks the subset of features quite than individual features. It chooses the set of attributes which are highly associated with the class in addition those attributes are low intercorrelation [15]. With CFS several heuristic searching approaches such as hill climbing and bestfirst are often functional to search the feature subsets space in reasonable time. CFS first computes the feature-class matrix and feature to feature correlations from the training data after that searches the feature subset space using a bestfirst.

$$M_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \qquad (1)$$

In above equ 1, Ms is the correlation between the summed feature subset S, k is the number of subset feature, $\overline{r_{cf}}$ is the average of the correlation between the subsets feature and the class variable, and $\overline{r_{ff}}$ is the average inter-correlation between subset features [16].

ii) Information Gain (IG)

The IG estimates attributes by calculating their information gain according to the class. It discretizes numeric attributes first using MDL based discretization technique [15]. Let C be set consisting of c data samples with m distinct classes. The training dataset ci contains sample of class I. Expected information needed to classify a given sample is calculated by equ 2:

$$I(c1, c2, \ldots, cm) = -\sum_{i=1}^{m} \frac{c_i}{c} \log_2 \left(\frac{c_i}{c}\right) \qquad (2)$$

Where $\frac{c_i}{c}$ is the probability that an arbitrary sample belongs to class Ci. Let feature F has v distinct values { f1, f2, …, fv } which can split the training set into v subsets {C1,C2, …, Cv } where Ci is the subset which has the value fi for feature F. Let Cj contain Cij samples of class i. The entropy of the feature F is given by equ 3

$$E(F) = \sum_{j=1}^{m} \frac{c_{ij} + \cdots + c_{mj}}{c} \times I(c_{ij} + \cdots + c_{mj}) \qquad (3)$$

Information gain for F can be calculated as equ 4:

$$\text{Gain } (F) = I(c1,...,cm) - E(F) \qquad (4)$$

iii) Gain Ratio (GR)

The information gain is helpful to select attributes which having a large number of values. The gain ratio is an extension of info gain, attempts to overcome this bias. Gain ratio applies normalization to info gain using a value defined as below equ 5,

$$\text{SplitInfo}_F(C) = \sum_{i=1}^{v} \left(\frac{|C_i|}{|C|}\right) \log_2 \left(\frac{|C_i|}{|C|}\right) \qquad (5)$$

The above value represents the information generated splitting the training data set C into v partitions corresponding to v outcomes of a test on the feature F [17].

The gain ratio is defined as below equ 6,

$$\text{Gain Ratio}(F) = \text{Gain } (F) / \text{SplitInfo}_F(C) \qquad (6)$$

iv) ReliefF

Relief is an instance based attribute selection method introduced by Kira and Rendell [18] and later improved by Kononenko [19]. Relief was formerly supported to two-class issues and after it was extended (ReliefF) which is used for handle noise and multi-class data sets. ReliefF smoothes the inuence of noise in the data by averaging the contribution of k nearest neighbours from the same and opposite class of each sampled instance instead of the single nearest neighbour. By determining nearest neighbours from each class, Multi-class data sets are supported.

v) Consistency-based Subset Evaluation

This feature subset selection technique look for mixtures of attributes whose values segregate the data into subsets containing a strong single class majority. Generally the search is subjective in the intension of obtaining the small feature subsets with high class consistency. Consistency-based subset evaluator utilizes by Liu and Setiono's [20] which is defied as equ 7:

$$\text{Consistency}_s = 1 - \frac{\sum_{i=0}^{J} |D_i| - |M_i|}{N} \qquad (7)$$

Where, s is an subset of attribute, J is the number of distinct combinations of attribute values for s, $|D_i|$ is the number of occurrences of the ith attribute value combination, $|M_i|$ is the cardinality of the majority class for the ith attribute value combination and N is the total number of instances in the data set.

B. Artificial Fish Swarm-Cuckoo Search Optimization Based Feature Selection–

Our system introduces a novel technique for the purpose of feature or attributes selection which is called as hybrid of Artificial fish swarm-Cuckoo search optimization. The effectiveness of feature selection is achieved by our proposed technique which incorporated the two coupled components of irrelevant and redundant feature elimination. In our existing method NMFC uses the Symmetric Uncertainty (SU) to remove the irrelevant features. The SU should maintain the mutual information and entropy. To avoid this limitation, in this paper, Artificial Fish Swarm-Cuckoo Search Optimization to remove the irrelevant features or to obtain the relevant features rapidly. After selecting the relevant features, we have to find the redundant features which are presented in the relevant features. For removal of redundant features, we are employing a novel Non-negative Matrix Factorization based

Clustering technique. After feature selection procedure, two classification methods such as Prism and J48 are used to predict the student's performance.

III. Experiment and Result

The performance of the classification with and without feature selection method in terms of classification accuracy, True Positive rate (TP rate) and True Negative rate (TN rate) is compared

A. Accuracy rate

Accuracy is defined as the overall accuracy rate or classification accuracy and is calculated as equ 16

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}} \quad (16)$$

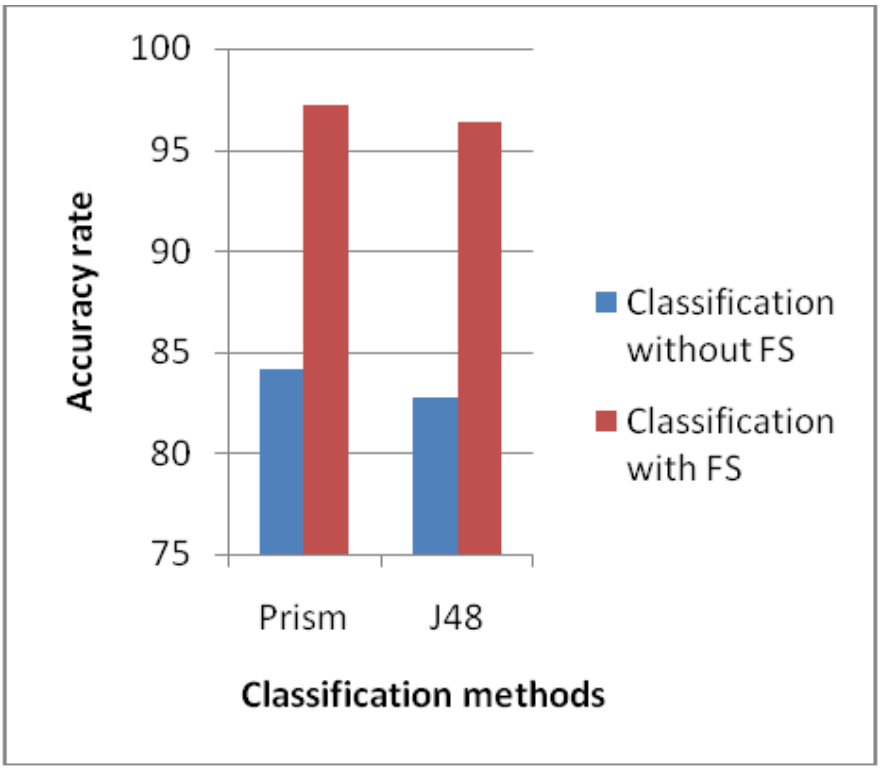

Fig.1. Accuracy comparison

Fig.1. showed that comparison of the accuracy parameter between classification without feature selection and classification without feature selection. Accuracy rate is mathematically calculated by using formula. As usual in the graph X-axis will be classification methods such as Prism, J48 and Y-axis will be accuracy rate.

B. True Positive rate

True Positive rate (TP rate), also called sensitivity or recall, is the proportion of actual positives which are predicted to be positive and is calculated as equ 17

$$TP = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (17)$$

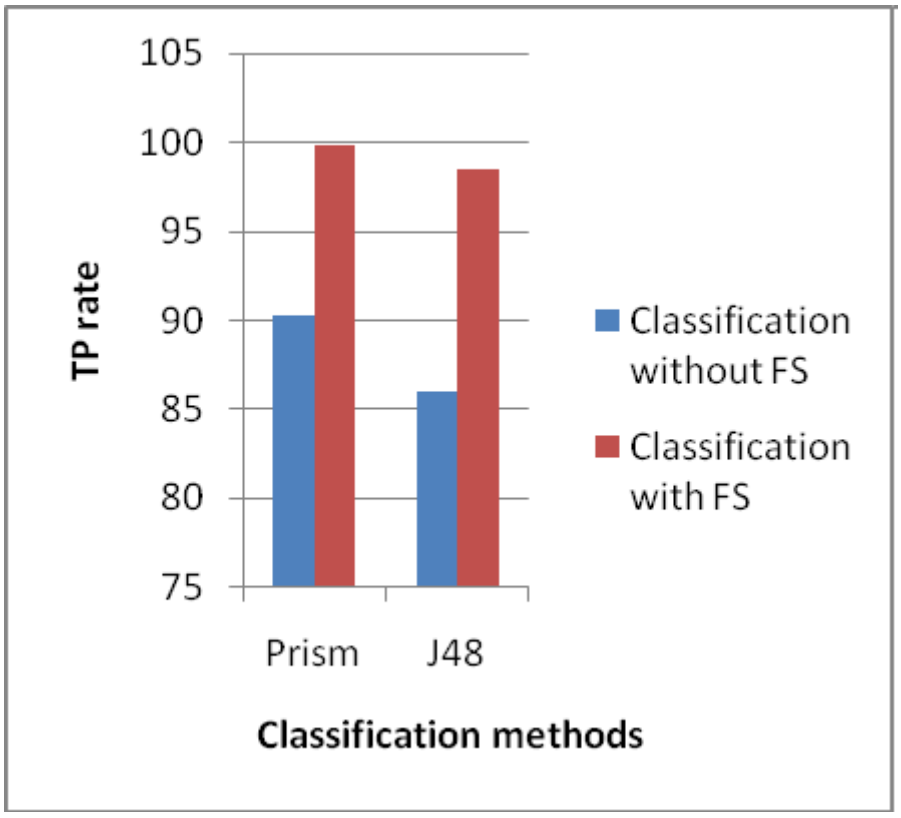

Fig.2. TP rate comparison

Fig.2. showed that comparison of the TP parameter between classification without feature selection and classification without feature selection. TP rate is mathematically calculated by using formula. As usual in the graph X-axis will be classification methods such as Prism, J48 and Y-axis will be TP rate.

C. True Negative rate

True Negative rate (TN rate), or specificity, is the proportion of actual negatives which are predicted to be negative and is calculated as equ 18

$$TN = \frac{\text{True negative}}{\text{True negative} + \text{False positive}} \quad (18)$$

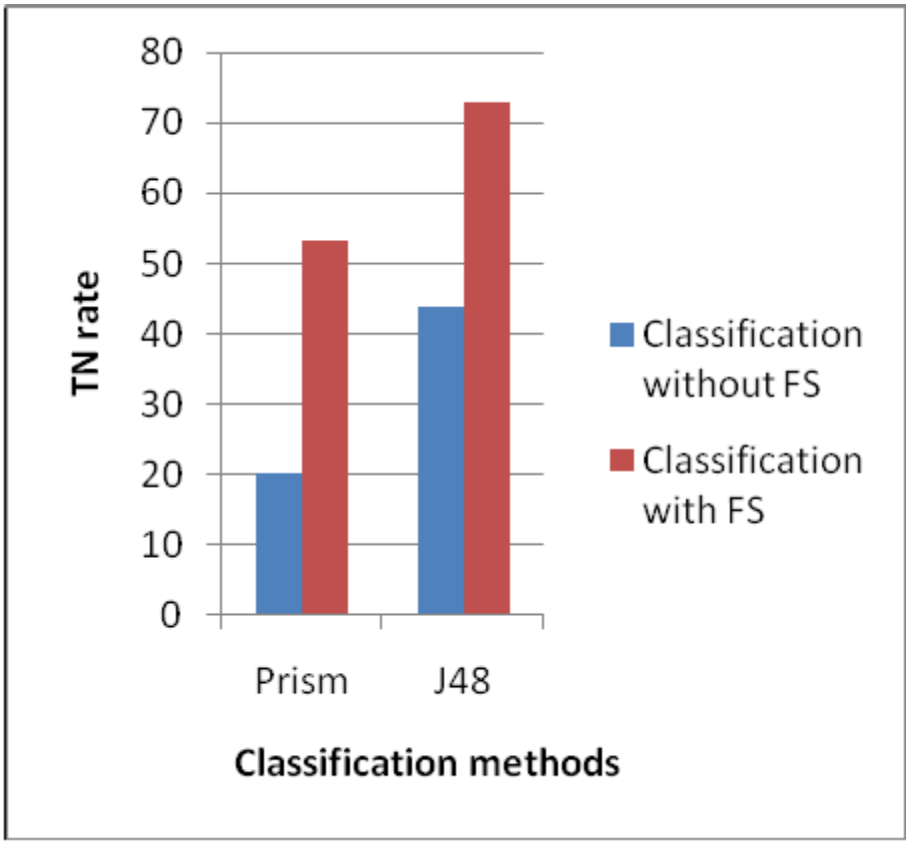

Fig.3. TN rate comparison

Fig.3. showed that comparison of the TN parameter between classification without feature selection and classification without feature selection. TN rate is mathematically calculated by using formula. As usual in the graph X-axis will be classification methods such as Prism, J48 and Y-axis will be TN rate. From view of this TN comparison graph a conclusion has obtain the classification with feature selection has more effective in TN performance comparatively.

IV.CONCLUSION

Several strategies can be developed and implemented enabling the educational institutions to transform a wealth of information into a wealth of predictability, stability and profits. Feature selection method is applied for classifying the performance of the students in the educational institutions. This technique has several benefits such as removal of irrelevant features and also eliminating the redundant features present in the relevant features as much as possible. This will improve the accuracy of the classification result. The classification technique such that prism and J48 are used for experimentation. The performance of the students' failure and dropout prediction can improve by performing feature selection.

REFERENCE

[1] E. Baker, *International Encyclopedia of Education (3rd edition)*, Oxford, UK: Elsevier, (In Press).
[2] P. Mitra, C. A. Murthy and S. K. Pal. "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002.
[3] Miller, "Subset Selection in Regression," Chapman & Hall/CRC (2nd Ed.), 2002.
[4] H. Almuallim and T. G. Dietterich. "Learning boolean concepts in the presence of many irrelevant features," *Artificial Intelligence*, vol. 69, no. 1-2, pp. 279–305, 1994.

[5] D. Koller and M. Sahami, "Toward optimal feature selection," In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 284–292, 1996.
[6] K. R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos.1-2, pp. 273-324, 1997.
[7] M.Dash and H.Liu, "Feature Selection for Classification," An International Journal of Intelligent Data Analysis, vol. 1, no. 3, pp.131-156, 1997.
[8] W. Duch, T. Winiarski, J. Biesiada, J, and A. Kachel, "Feature Ranking, Selection and Discretization," *Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP)*, pp. 251 – 254, 2003.
[9] P. Langley, "Selection of Relevant Features in Machine Learning," *Proceedings of AAAI Fall Symp. Relevance*, pp. 140-144, 1994.
[10] Isabella Guyon and Andre Elisseeff, "An Introduction to Variable and Feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157 – 1182, 2003.
[11] M. Ramaswami and R. Bhaskaran, "Student Performance Forecasting: A Study of Classification Models," *Intelligent Computting Models*, Narosa Publishing House, New Delhi, pp. 38-45, 2009.
[12] M. K. Cope, H. H. Baker, R. Fisk, J. N. Gorby and R. W. Foster, "Prediction of Student Performance on the Comprehensive Osteopathic Medical Examination Level Based on Admission Data and Course Performance," *Journal of the American Osteopathic Association*, vol. 101, no. 2, pp. 84 – 90, 2001.
[13] N. T. Nghe, P. Janecek and P. Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance," *Paper presented at 37th ASEE/IEEE Frontiers in Education Conference*, Milwaukee, WI, October 10 – 13, 2007.
[14] W. R. Veitch, "Identifying Characterstics of High School Dropouts: Data Mining with a Decision Tree Model," *Paper Presented at Annual Metting of the American Educational Research Association*, San Diego, CA, 2004 (ERIC Document No. ED490086).
[15] I.H.Witten, E.Frank, M.A. Hall "Data Mining Practical Machine Leanrning Tools & Techniques" Third edition, Pub. – Morgan kouffman.
[16] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer Science, University of Waikato.http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.
[17] J.Han ,M Kamber, Data mining : Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers(2001).
[18] K. Kira and L. Rendell, \A practical approach to feature selection," in Proceedings of the Ninth International Conference on Machine Learning. 1992, pp. 249{256, Morgan Kaufmann.
[19] I.Kononenko, \Estimating attributes: Analysis and extensions of relief," in Proceedings of the Seventh European Conference on Machine Learning. 1994, pp. 171{182, Springer-Verlag.
[20] H. Liu and R. Setiono, \A probabilistic approach to feature selection: A _lter solution," in Proceedings of the 13th International Conference on Machine Learning. 1996, pp. 319{327, Morgan Kaufmann.