International Journals of Advanced Research in Computer Science and Software Engineering
ISSN: 2277-128X (Volume-8, Issue-5)

Research Article | **May 2018**

An Enhancement of Feature Selection Algorithm for EDM: A Review

Manpreet Kaur
Research Scholar, Guru Kashi University, Talwandi Sabo, Punjab, India

Er. Chamkaur Singh
Assistant Professor, Guru Kashi University, Talwandi Sabo, Punjab, India

Abstract: Educational Data Mining (EDM) is an emerging research area help the educational institutions to improve the performance of their students. Feature Selection (FS) algorithms remove irrelevant data from the educational dataset and hence increases the performance of classifiers used in EDM techniques. This paper present an analysis of the performance of feature selection algorithms on student data set. .In this papers the different problems that are defined in problem formulation. All these problems are resolved in future. Furthermore the paper is an attempt of playing a positive role in the improvement of education quality, as well as guides new researchers in making academic intervention.

Keyword: EDM, Data, mining, feature, accuracy etc.

I. INTRODUCTION

The improvement in the quality of education is one of the most significant aspects of forming a successful member of society. The data stored in educational institutions repository plays an important role in order to extract hidden and interesting patterns to assist every stakeholder of an educational process [1]. There are many techniques being anticipated to assess the student academic performance in way of making fruit full future of a student. Predicting performance of student has been continued to a hot topic in the Educational data mining domain. Data mining is considered to be one of the best choices for the researchers to analyse student's performance. The techniques of data mining are extensively used on educational data now- a day's [2, 3]. It is called educational data mining. Educational Data Mining (EDM) explores the educational data to better understand the issues of student's performance using the fundamental nature of data mining techniques [4]. EDM manipulates educational data to help educational institutions to plan educational strategies, in order to improve the educational quality. Prediction is one of the main areas in EDM. Prediction and analysis of student academic performance are essential for student academic growth. Identifying the factors affecting the student academic performance is complicated research task [5].The original of academic data contains many irrelevant and redundant data .This redundant data effects the results of prediction. Feature selection methods minimize the redundancy and maximize relevancy of features without any loss of crucial data [6]. Feature Selection is very dynamic and productive field and research area of machine learning and data mining. The main goal of feature selection is to choose a subset by eliminating non-predictive data. Furthermore, it increases the predictive accuracy and reduces the complexity of learned results .The effectiveness of student performance prediction models can be increased in connection with feature selection techniques. Feature Selection techniques can be classified in to three groups: filter, wrapper, and embedded models. Filter method depends upon general characteristics of training data, this method is done on pre-processing stage and not dependent on a learning algorithm. Wrapper method uses learning algorithms to evaluate the features. Embedded methods are specific to some given learning algorithms, and these methods are performed on training process of classifiers. Previously, alot of work is done to predict the performance of student using different feature selection techniques [5]. In recent studies, researchers use different feature selection techniques and the combination of classifiers to produce efficient prediction models. A research is required to identify the performance analysis in terms of prediction accuracy in combination of different feature selection algorithms with differently classifiers. This paper is a step towards identifying the prediction accuracy of different available feature selection algorithm in the context of classifiers being used on educational data.

II. FEATURE SELECTION

Feature selection has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities [2, 3]. The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results [4, 5]. Feature selection in supervised learning has a main goal of finding a feature subset that produces higher classification accuracy. As the dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is intractable and problems related feature selections have been proved to be NP-hard [6]. At this juncture, it is essential to describe traditional feature selection process, which consists of four basic steps, namely, subset generation, subset evaluation, stopping criterion, and validation [7]. Subset generation is a search process that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation. If the new subset turns to be better,

it replaces best one. This process is repeated until a given stopping condition is satisfied. Ranking of features determines the importance of any individual feature, neglecting their possible interactions. Ranking methods are based on statistics, information theory, or on some functions of classifier's outputs [8]. Algorithms for feature selection fall into two broad categories namely wrappers that use the learning algorithm itself to evaluate the usefulness of features and filters that evaluate features according to heuristics based on general characteristics of the data [7]. Several justifications for the use of filters for subset selection have been discussed [6] and it has been reported that filters are comparatively faster than wrappers. Many student performance prediction models have been proposed and comparative analyses of different classifier models using Decision Tree, Bayesian Network, and other classification algorithms have also been discussed. But, they reveal only classifier accuracy without performing the feature selection procedures.

III. LITERATURE SURVEY

Maryam Zaffar et.al.[2017] have studied Student's academic performance is the main focus of all educational institutions. Educational Data Mining (EDM) is an emerging research area help the educational institutions to improve the performance of their students. Feature Selection (FS) algorithms remove irrelevant data from the educational dataset and hence increases the performance of classifiers used in EDM techniques. This paper present an analysis of the performance of feature selection algorithms on student data set. The obtained results of the different FS algorithms and classifiers will also help the new researchers in finding the best combinations of FS algorithms and classifiers. Selecting relevant features for student prediction model is very sensitive issue for educational stakeholders, as they have to take decisions on the basis of results of prediction models. Furthermore our paper is an attempt of playing a positive role in the improvement of education quality, as well as guides new researchers in making academic intervention.[1]

R. Sasi Regha et.al.[2016] have studied 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.[2]

Dr.M.Chidambaram et.al.[2016] have studied Feature Selection is a fundamental problem in machine learning and data mining . Feature Selection is an effective way for reducing dimensionality, removing irrelevant data increasing learning accuracy. Feature Selection is the process of identifying a subset of the most useful features that produce compatible results as the original entire set of features .A Feature Selection techniques may be evaluated from both efficiency and effectiveness point of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of the subset of features. Feature Selection is different from dimensionality reduction. Both methods search for to reduce the number of attributes in the dataset. But dimensionality reduction method creating new combination of attributes. Feature Selection methods include and exclude attributes present in the data without change. The central assumption when using a Feature Selection technique is that the data contains many redundant or irrelevant features. This paper actually a survey on various technique of feature selection and its advantages disadvantages.[3]

Mital Doshi et.al.[2014] have studied Feature selection is a preprocessing step to machine learning which is effective in reducing dimensionality, helps in removing irrelevant data, increasing learning accuracy, and improving result. In this paper we have shown different feature selection approaches, applications and the relation between them and the various machine learning algorithms. [4]

Anal Acharya et.al.[2014] have studied web based learning has emerged as a new field of research due to growth of network and communication technology. These learning systems generate a large volume of student data. Data mining algorithms may be applied on this data set to study interesting patterns. As an example, student enrollment data and his past examination records could be used to predict his grades in the term end examination. However this prediction could mean examining a lot of features of the student data resulting in creation of a model with high computational complexity. In this context this work first defines a student data set with 309 records and 14 features collected by a survey from various graduation level students majoring in Computer Science under University of Calcutta. Different feature selection algorithms are applied on this data set. The best results are obtained by Correlation Based Feature Selection algorithm with 8 features. Subsequently classification algorithms may be applied on this feature subset for predicting student grades.[5]

M. Ramaswami et.al.[2009] have studied Educational data mining (EDM) is a new growing research area and the essence of data mining concepts are used in the educational field for the purpose of extracting useful information on the behaviors of students in the learning process. In this EDM, feature selection is to be made for the generation of subset of candidate variables. As the feature selection influences the predictive accuracy of any performance model, it is essential to study elaborately the effectiveness of student performance model in connection with feature selection techniques. In this connection, the present study is devoted not only to investigate the most relevant subset features with minimum cardinality for achieving high predictive performance by adopting various filtered feature selection techniques in data mining but also to evaluate the goodness of subsets with different cardinalities and the quality of six filtered feature selection algorithms in terms of F-measure value and Receiver Operating Characteristics (ROC) value, generated

by the NaïveBayes algorithm as base-line classifier method. The comparative study carried out by us on six filter feature section algorithms reveals the best method, as well as optimal dimensionality of the feature subset. Benchmarking of filter feature selection method is subsequently carried out by deploying different classifier models. The result of the present study effectively supports the well known fact of increase in the predictive accuracy with the existence of minimum number of features. The expected outcomes show a reduction in computational time and constructional cost in both training and classification phases of the student performance model.[6]

IV. PROBLEM FORMULATION

In the research work Feature Selection Algorithm for Educational Data Mining different problems are faced that are given below:

- As the dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is intractable and problems related feature selections have been proved to be NP-hard.
- It is to be noted that Receiver Operating Characteristic (ROC) parameter is not used here since it applies to a 2-class problem.
- Many kinds of redundancy, irrelevant and noisy feature as well as Relief algorithm limitations in high datasets.
- Another problem is high rate feature selection and accuracy problem.

V. METHODOLOGY

The main aim of the research is to evaluate the performance of different FS algorithms on different classification algorithms using student dataset. The comparison between different FS algorithms give a deep insight to new educational data miners about the performance of different feature selection algorithms on student data .To achieve the objective of the research , a student dataset is taken from a valid sources, and then different FS algorithms are applied on it , which was not used earlier on this dataset. Different classification algorithms are applied by using selected FS algorithms, and furthermore evaluated to check the best performance among all the combinations applied on student data set.

Data set Description:

The dataset used in this study is taken from the source www.kaggle.com, and is comprised of 500 students 16 features. This dataset has been used in the study [11], to check the learner's interactivity with e- learning management system, bagging and boosting methods are applied on the given dataset, however, only information gain based feature selection algorithm is used previously. In this paper, the main aim of using the dataset is to identify the best combinations of FS algorithms and classifiers, in order to identify the key performance factors on the academic achievements of students.

WEKA (Waikato Environment for Knowledge Analysis) is used as a data mining tool. It has a rich source of Machine learning algorithms. WEKA is developed by the University of Waikato in New Zealand. It is an open source software developed in JAVA language, that provides facility for developing machine learning techniques for data mining tasks.

Feature Selection Algorithm and Classifiers

In this research work six FS algorithm CfsSubsetEval, ChiSquaredAttributeEval, FilteredAttribute Eval, GainRatioAttributeEval, Principal Components, and ReliefAttributeEval are evaluated. The classification algorithm BayesNet(BN), Naïve Bayes(NB), NaiveBayesUpdateable(NBU), MLP, Simple Logistic(SL), SMO, Decision Table(DT), Jrip, OneR, OneR, DecsionStump(DS), J48, Random Forest(RF), RandomTree(RT), REPtree(RepT) are evaluated through the educational data set.

VI. CONCLUSION

Educational Data Mining (EDM) explores the educational data to better understand the issues of student's performance using the fundamental nature of data mining techniques. EDM manipulates educational data to help educational institutions to plan educational strategies, in order to improve the educational quality. Prediction is one of the main areas in EDM. Prediction and analysis of student academic performance are essential for student academic growth. In this paper different problems like dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is intractable and problems related feature selections have been proved to be NP-hard. Future all the problems are resolved with Feature Selection Algorithm on student data with Naïve Bayes (NB), Decision tree and decision table.

REFERENCES

[1] Maryam Zaffar et.al. "Performance Analysis of Feature Selection Algorithm for Educational Data Mining" IEEE Conference on Big Data and Analytics (ICBDA)-2017.

[2] R. Sasi Regha et.al. "Optimization Feature Selection for classifying student in Educational Data Mining" International Journal of Innovations in Engineering and Technology (IJIET) , Volume 7 Issue 4 December 2016.

[3] Dr.M.Chidambaram et.al. "A Survey on Feature Selection in Data Mining " International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-4, Issue-1, January 2016.

[4] Mital Doshi et.al. "Survey of Feature Selection Algorithms in Higher Education" International Journal of Computer Applications in Engineering Sciences [VOL IV, ISSUE I, MARCH 2014].

[5] Anal Acharya et.al. "Application of Feature Selection Methods in Educational Data Mining" International Journal of Computer Applications © 2014 by IJCA Journal Volume 103 - Number 2 Year of Publication: 2014.

[6] M. Ramaswami et.al. "A Study on Feature Selection Techniques in Educational Data Mining" Journal Of Computing, Volume 1, Issue 1, December 2009.

[7] E. Osmanbegović, M. Suljić, and H. Agić, "DETERMINING DOMINANT FACTOR FOR STUDENTS PERFORMANCE PREDICTION BY USING DATA MINING CLASSIFICATION ALGORITHMS," Tranzicija, vol. 16, pp. 147-158, 2015.

[8] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414-422, 2015.

[9] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, pp. 601-618, 2010.

[10] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," arXiv preprint arXiv:0912.3924, 2009.

[11] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," International Journal of Modern Education and Computer Science, vol. 8, p. 36, 2016.

[12] W. Punlumjeak and N. Rachburee, "A comparative study of feature selection techniques for classify student performance," in Information Technology and Electrical Engineering (ICITEE), 2015 7th International Conference on, 2015, pp. 425- 429.

