FEATURE SELECTION TECHNIQUES TO ANALYSE STUDENT ACADAMIC PERFORMANCE USING NAÏVE BAYES CLASSIFIER

Conference Paper · February 2016

CITATIONS
31

READS
1,863

2 authors, including:

Velmurugan Thambusamy
Dwaraka Doss Goverdhan Doss Vaishnav College
127 PUBLICATIONS **1,442** CITATIONS

SEE PROFILE

The 3rd International Conference on Small & Medium Business 2016
January 19 - 21, 2016, Nikko Saigon Hotel, Hochiminh, Vietnam

FEATURE SELECTION TECHNIQUES TO ANALYSE STUDENT ACADAMIC PERFORMANCE USING NAÏVE BAYES CLASSIFIER

C.Anuradha[1], T.Velmurugan[2]
[1]Research Scholar, Bharathiar University, Coimbatore, India.
[2]Associate Professor, PG and Research Dept. of Computer Science, D.G.Vaishnav College, Chennai-600106, India.
[1]anumphil14@yahoo.co.in; [2]velmurugan_dgvc@yahoo.co.in

Abstract: **Data mining provides educational institutions that the capability to explore, visualize and analyze large amounts of data in order to reveal valuable patterns in students' learning behaviors. Turning raw data into useful information and knowledge also enables educational institutions to improve teaching and learning practices, and to facilitate the decision-making process in educational settings. Thus, educational data mining is becoming an increasingly important with a specific focus to exploit the abundant data generated by various educational systems for enhancing teaching, learning and decision making. In EDM, Feature Selection is to choose a subset of input variables by eliminating irrelevant features. Feature Selection Algorithm has proven to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. The primary objective of this research work is to investigate the most relevant subset features for achieving high performance accuracy by adopting Correlation based feature Subset Attribute evaluation and Gain-Ratio Attribute evaluation feature selection techniques. For classification, the Naïve Bayes classifier is implemented by using WEKA tool. The outcome shows the effectiveness in the predictive accuracy with minimum number of attributes. Also the results reveals that the selected data features have found to be influenced the classification process of the student performance model.**

Keywords: Educational Data Mining (EDM), Classification algorithm, Naïve Bayes Algorithm, Feature Selection, Prediction.

I. INTRODUCTION

Nowadays the field of data analytics and data mining (DM) is taking a new role. The role that is undertaking is as an enabler of educational institutions to improve key performance indicators. The importance of data analytics is growing and a new sub-field of studies is in its infancy. This young field is called Educational Data mining and its main purpose is to analyze data by using a different number of techniques. EDM integrates different approaches as database systems, data warehousing, statistics, machine learning and others. Moreover an experiment will be conducted with this educational data, the experiment will start with the description of the state of the art of EDM and it will continue with the development of a method for exploring data and predicting trends that will contribute to improve educational data or to analyze current problems to increase organizational performance. Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.

Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. As we know large amount of data is stored in educational database, so in order to get required data & to find the hidden relationship, different data mining techniques are developed & used. There are varieties of popular data mining task within the educational data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. We can use the data mining in educational system as: predicting drop-out student, relationship between the student university entrance examination results & their success, predicting student's academic performance, discovery of strongly related subjects in the undergraduate syllabi, knowledge discovery on academic achievement, classification of students performance in computer programming course according to learning style, investing the similarity & difference between colleges and schools. EDM develops methods and applies techniques from statistics, machine learning, and data mining to analyze data collected during teaching and learning. EDM tests learning theories and informs educational practice. As a result, researchers try to determine the variables that are related to academic achievement of students and may affect the registration process. Therefore, one of the most important challenges that higher education faces is recognizing the pattern of loyal students.

The effective feature selection techniques are required to analyze the efficient classification algorithms. This research work attempts to foretell the students academic failure by reviewing the field of various feature selection algorithms based on the Naïve Bayes classifier. This research work is structured as follows. Section 2 illustrates the research work that has been conducted in EDM. In section 3 consist of methods and materials of the domain of study will be defined. The description of the process of building a model includes data collection and used tools are given in Section 4. Then Section 5 presents the experimentation and results obtained. Finally, conclusion is given in Section 6.

II. RELATED WORKS

This section discusses about some of the research work carried out by various researchers in the same field. A work done by Humera Shaziya et al. has presents an approach to predict the performance of students in a semester exams. This approach is based on a Naive Bayes classifier. The objective is to know what grades students may obtain in their end semesters results. This helps the educational institute, teachers and students i.e., all the stakeholders involved in an education system. Students and teachers can take necessary actions to improve the results of those students whose result prediction is not satisfactory. A training dataset of students is taken to build the Naive Bayes model. The model is then applied on the test data to predict the end semester results of students. In this study, number of attributes is considered to predict the grade of a student [1].

Another work done by Tajunisha and Anjali have discussed about Predicting Student Performance Using MapReduce. Authors introduced the MapReduce concept to improve the accuracy and reduce the time complexity. In this work, the deadline constraint is also introduced. Based on this, an extensional MapReduce Task Scheduling algorithm for Deadline constraints (MTSD) is proposed. It allows user to specify a job's (classification process in data mining) deadline and tries to make the job to be finished before the deadline. Finally, the proposed System has higher classification accuracy even in the big data and it also reduced the time complexity [2]. Another study focused on Predicting Students Final GPA Using Decision Trees by Mashael and Muna [3].

Authors applied the J48 decision tree algorithm to discover classification rules. They extracted useful knowledge and identified the most important courses in the students study plan based on their grades in the mandatory courses. A work carried out by Karthikeyan and Thangaraju had proposed a work in genetic algorithm and particle Swarm optimization search techniques and correlation based feature selection is used for evaluation and naïve Bayes classifier for classification purpose. Accuracy and time is the outcome of the classification model and also various measures like sensitivity, specificity, precision and recall are also calculated [4]. A work carried out by Lumbini and Pravin [5] have proposed an experiment attempts the detection of student's failure to improve their academic performance. They have applied different approaches to resolve the problem of high dimensionality and using classification algorithm on engineering students data set.

Predictive Analytics Using Data Mining Technique [6] by Hina Gulati has presents the work of data mining is predicting the dropout feature of students. Author also applied some feature selection algorithms. Tool used for feature selection and mining is weka. Another work by Jai and David discussed about Analysis of Influencing Factors in Predicting Students Performance Using MLP-A Comparative Study [7]. This paper mainly focused on analyzing the prediction accuracy of the academic performance using influencing factors by Multi Layer Perception algorithm and compares it with the prediction accuracy. Another research work carried out by Anal and Devadatta have discussed about Application of Feature Selection Methods in Educational Data Mining. Different feature selection algorithms are applied on this data set and the results are obtained by Correlation Based Feature Selection algorithm with 8 features. Then classification algorithms may be applied on this feature subset for predicting student grades [8]. Another work by the same authors have discussed about Early Prediction of Students Performance using Machine Learning Techniques. In this paper a set of attributes are first defined. Then feature selection algorithms are applied on the data set to reduce the number of features. Five classed of Machine Learning Algorithm (MLA) are then applied on this data set and it was found that the best results were obtained with the decision tree class of algorithms [9].

III. METERIALS AND METHODS

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods. Wrapper methods use a predictive model to score feature subsets. Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, while still capturing the usefulness of the feature set. Embedded

methods are a catch-all group of techniques which perform feature selection as part of the model construction process [10].

A. Correlation Based Feature Subset Selection

CFS is a correlation-based filter method CFS from [11]. It gives high scores to subsets that include features that are highly correlated to the class attribute but have low correlation to each other Let S be an attribute subset that has k attributes, rcf models the correlation of the attributes to the class attribute, rff the intercorrelation between attributes.

$$\text{meritS} = k\ rcf / \text{sqrt}\ (k+k\ (k-1)\ rff)$$

B. Gain Ratio Attribute Evaluator

Gain Ratio Attribute Evaluator is simple individual attribute ranking mechanism. In this technique, each attribute is assigned a score where the score is delineated by means of the difference of attributes entropy and its class conditional entropy [12].

GainR (Class, Attribute) = (H (Class) - H (Class | Attribute)) / H (Attribute).

Classification is a data mining task that predicts group membership for data instances. In this research work classification techniques are used to predict the class of the graduate student and how the other attributes affects the performance. The classifier used in this study is Naïve Bayesian algorithm.

C. Naïve Bayes

The Naïve Bayes classifier technique is used when dimensionality of the inputs is high. This is a simple algorithm but gives good output than others. This classifier is used to predict the dropout of the students by calculating the probability of each input for a predictable state [13].

IV. EXPERIMENTAL DATA

The dataset is a collection of first year students information contains 5 undergraduate degree courses collected from SSBSTAS College, Thiruvalluvar University, Tamilnadu for a period of 2013-2014. The student data set of 257 records with 21 attributes that includes the gender, category of admission, living location, family size, and family type, annual income of the family, father's qualification and mother's qualification. The attributes referring to the students' pre-college characteristics included Students Grade in High School and Students Grade in Senior Secondary School. The attributes describing other college features include the branch of study of the students, place of stay, previous semester mark, class test performance, seminar performance, assignment, general proficiency, class attendance and performance in the laboratory work. Following Table 1 shows the description of attributes.

Table 1: Student Data Set Description

Variables	Description	Possible Values
Gender	Students Sex	{Male, Female}
Branch	Students Branch	{BCA, B.SC, B.COM, B.A}
Cat	Students category	{BC, MBC, MSC, OC, SBC, SC}
HSG	Students grade in High School	{O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 35% - 49%, FAIL - <35%}
SSG	Students grade in Senior Secondary	{O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 35% - 49%, FAIL - <35% }
Medium	Medium of instruction	Tamil, English, others
LLoc	Living Location of Student	{Village, Taluk, Rural, Town, District}
HOS	Student stay in hostel or not	{Yes, No}
FSize	student's family size	{1, 2, 3, >3}
FType	Students family type	{Joint, Individual}
FINC	Family annual income	{poor, medium, high}
FQual	Fathers qualification	{no-education, elementary, secondary, UG, PG, Ph.D}
MQual	Mother's Qualification	{no-education, elementary, secondary, UG, PG, Ph.D. NA}
PSM	Previous Semester Mark	{First > 60%, Second >45 &<60%, Third >36 &<45% Fail < 36%}
CTG	Class Test Grade	{Poor, Average, Good}

SEM_P	Seminar Performance	{Poor , Average, Good}
ASS	Assignment	{Yes, No}
GP	General Proficiency	{Yes, No}
ATT	Attendance	{Poor , Average, Good}
LW	Lab Work	{Yes, No}
ESM	End Semester Marks	{First > 60% , Second >45 &<60% , Third >36 &<45%, Fail < 36%}

For the purpose of designing and evaluating our experiments, we have used WEKA. It is open source software which is freely available for mining data and implements a large collection of mining algorithm. It can accept data in various formats and also has converter supported with it. So we have converted the student dataset into CSV file. Under the "Test options", the 10-fold cross-validation is selected as our evaluation process. The various performance Metrics are discussed as follows.

The Accuracy of the predictive model is calculated based on the True positive rate, false positive rate, and precision and recall values [14]. TP rate(True Positive): A positive test results accurately reflects the test for activity. If the outcome from a prediction is p, and the actual value is also p, then it is called true positive (TP).

TP = TP/P where P= (TP+FN)

TN (True negative): It has occurred when both the prediction outcome and the actual value are n in the number of input data.

TN = TN/N, where N = (TN+FN)

FP rate(False positive): If the outcome from a prediction is p and the actual value is n, then it is said to be false positive (FP).

FP = FP / (FP+TN)

Precision: It is the fraction of retrieved instances that are relevant.

Precision = TP/ (TP+FP)

Recall: It is a fraction of relevant instances that are retrieved. TP/ (TP+FN)

V. RESULTS AND DISCUSSION

The present investigation focuses on two feature selection techniques namely cfsSubsetEval and GainRatioAttributeEval, which is one of the important and frequently used in data preprocessing in data mining. Using these attribute selection algorithms we can select the best attributes out of huge number of attributes of students that affect the student's performance. And the results are obtained with Naïve Bayes classifier. Table 2 shows the results of applying two feature selection algorithms.

Table 2: Best Selected Attributes

Algorithm	Attributes Selected
cfsSubsetEval	Branch,SSG,FINC,PSM,GP,ATT
GainRatio AttributeEval	Age,branch,cat,SSG, medium,ATT,GP,FINC, FQUAL,MQUAL, HSG,SEM_P,LOC

A. Results of cfsSubset Evaluator

In this experiment Correlation Based Feature selection algorithm is used with 6 attributes along with Naïve Bayes classifier was implemented on the data set and the results are presented in Table 3. It shows that classification results for Naïve Bayes correctly classifies about 84.2% for 10 fold cross validation. Also True Positive rate is high for the class Second and first, Whereas TP rate is very low for the class Third. Fig.1 shows the graphical representation of the classifier.

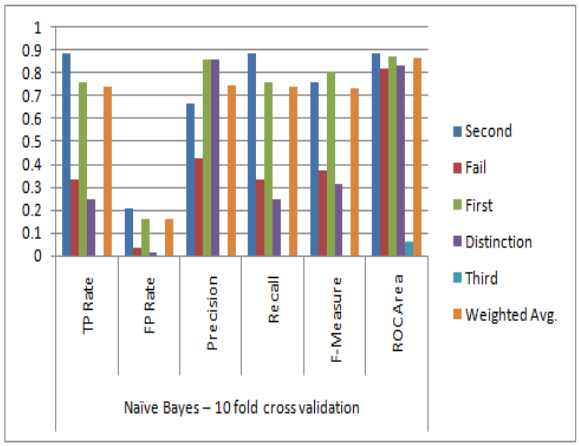

Figure 1: Result of CfsSubset Evaluator

B. Results of GainRatioAttributeEvaluator

The present study implements GainRatio Attribute Evaluator with 13 attributes. The Result of Naïve Bayes classifier is shown in Table 4. It shows that

classifier correctly classifies about 74.4% for 10 fold cross validation. True positive rate is high for the class second and it is very low for the class Third. Fig.2 shows the graphical representation of Naïve Bayes classification algorithm.

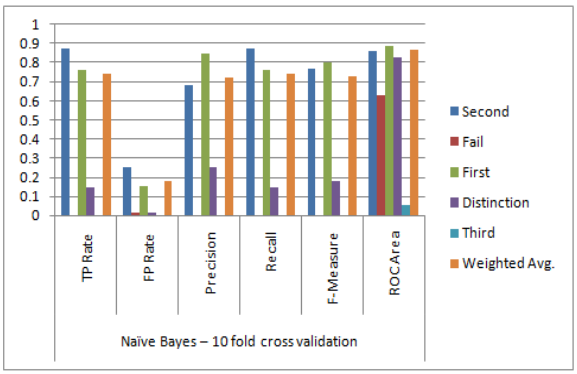

Figure 2: Result of Gain-Ratio Attribute Evaluator

Table 3: Classifier Result for CfsSubsetEvaluator

Class	Naïve Bayes – 10 fold cross validation					
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Second	0.888	0.205	0.664	0.888	0.759	0.885
Fail	0.333	0.034	0.429	0.333	0.375	0.823
First	0.759	0.162	0.859	0.759	0.806	0.875
Distinction	0.25	0.016	0.859	0.25	0.316	0.835
Third	0	0	0	0	0	0.059
Weighted Avg.	0.842	0.359	0.844	0.842	0.835	0.869

Table 4: Classifier Result for GainRatioAttributeEvaluator

Class	Naïve Bayes – 10 fold cross validation					
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Second	0.873	0.25	0.683	0.873	0.767	0.859
Fail	0	0.015	0	0	0	0.631
First	0.764	0.155	0.848	0.764	0.804	0.886
Distinction	0.143	0.015	0.25	0.143	0.182	0.825
Third	0	0	0	0	0	0.053
Weighted Avg.	0.744	0.179	0.72	0.744	0.726	0.867

Table 5: Overall Accuracy of Feature Selection Algorithm

Algorithm	Naïve Bayes					
	Second	Fail	First	Distinction	Third	Weighted Avg.
cfsSEval	0.888	0.333	0.759	0.25	0	0.842
GRAE	0.873	0	0.764	0.143	0	0.744

C. Performance comparison between the Feature Selection Algorithms

The results for the performance of the selected feature selection algorithm on Naïve Bayes classifier is summarized in Table 5. The results of Feature Selection algorithm along with the naïve Bayes classifier reveals that Correlation Based Feature subset Evaluator performs very well with 6 attributes in comparison with Gain Ratio which has 13 attributes. The overall accuracy of CFS algorithm is about 84%. On the other hand Gain Ratio performs less accurate of just 74%. Also the classification accuracy is very good for the class Second and First. In addition, further analysis that the prediction result shows that accuracy is low for the class Distinction and very worst for the class Third.

VI. CONCLUSION

In this research work, It is presented a case study in educational data mining. The obtained results show that the feature selection techniques can improve the accuracy and efficiency of the classification algorithm by removing irrelevant and redundant attributes. It was especially used to improve the student performance. The most relevant features are got by using GainRatio and CFS subset evaluator. Naïve Bayes classifiers have been applied on the selected features. From the results, it is concluded that Correlation Based Feature Subset evaluator performs well with the Naïve Bayes classifier as compared with Gain-Ratio Attribute Evaluator. In future, this work extend the experiment with different data mining techniques like clusters can applied with other feature selection algorithms on large data set in the same educational field.

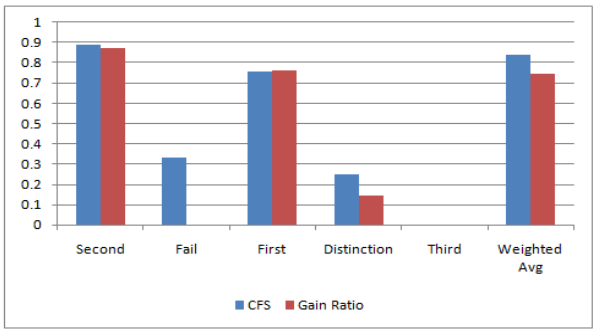

Figure 3: Overall accuracy of Feature Selection Algorithm

REFERENCES

[1] Humera Shaziya, Raniah Zaheer, Kavitha.G, " Prediction of students in Semester Exams using a Naïve Bayes Classifier", Int. Journal of Innovative Research in Science, Engineering and Technology, Vol.4, Issue 10, 2015, pp.9823-9829.

[2] Tajunisha N,Anjali M, " Predicting Student Performance Using MapReduce", Int. Journal of Engineering and Computer Science, Vol.4, Issue 1, 2015,pp.9971-9976.

[3] Mashael A. Al-Barrak, Muna Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study", Int. Journal of Information and Education Technology, Vol.6, No.7, 2016, pp.528-533.

[4] Karthikeyan.T, Thangaraju.P,"Genetic Algorithm based CFS and Naïve Bayes Algorithm to Enhance the Predictive Accuracy", Indian Journal of Science and Technology, Vol.8, No.26, 2015, pp.1-8.

[5] Lumbini P.Khobragade, Pravin Mahadik," Students Academic Failure Prediction Using Data Mining", Int. Journal of Advanced Research in Computer and Communication Engineering, Vol.4, Issue.11,2015,pp.290-298.

[6] Hina Gulati, "Predictive Analytics Using Data Mining Technique", 2nd International Conference on Computing for Sustainable Global Development, 2015, pp.713-716.

[7] Jai Ruby,K. David, "Analysis of Influencing Factors in Predicting Students Performance Using MLP-A Comparative Study", Int. Journal of Innovative Research in Computer and Communication Engineering, Vol.3, Issue.2, 2015, pp.1085-1092.

[8] Anal Acharya, Devadatta Sinha, "Application of Feature Selection Methods in Educational Data Mining", Int. Journal of Computer Applications, Vol.103, No.2,2014,pp.34-38.

[9] Anal Acharya, Devadatta Sinha, "Early Prediction of Students Performance using Machine Learning Techniques", Int. Journal of Computer Applications, Vol.107,No.1,2014,pp.37-43.

[10] Guyon, Isabelle, and Andre Elisseeff, "An introduction to variable and feature selection", The Journal of Machine Learning Research, Vol. 3, 2003, pp. 1157-1182.

[11] Hall, M. A., Smith, L. A, "Practical feature subset selection for machine learning", Australian Computer Science Conference, Springer, 1998, pp.181-191.

[12] Muhammad Naeem,"An Empirical Analysis and Performance Evaluation of Feature Selection Techniques for Belief Network Classification System", Int. Journal of Control and Automation, Vol.8, No.3, 2015, pp.375-388.

[13] Mital Doshi, Setu K Chaturvedi,"Correclation Based Feature Seleciton(CFS) Technique to Predict Student Performance", Int. Journal of Computer Networks & Communications, Vol.6, No.3, 2014,pp.197-206.

[14] P.V.Praveen Sundar," A comparative study for Predicting Students Academic Performance using Bayesian Network Classifiers", IOSR Journal of Engineering, Vol.3, Issue 2,2013,pp.37-42.