International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X;pISSN:2321-2241 Volume: 2; Issue: 5; May -2014

Application of a Naïve Bayesian Classification Algorithm for Students' Performance Evaluation and Improvement in Education Sector

T.M Lakshmikanth[1], B.K. Sangareddy[2]
[1]P.G. Student, Department of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikmagalur, Karnataka. Email : lkwodeyar@gmail.com
[2]Asst.Professor, Department of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikmagalur, Karnataka Email: sbkurtakoti@gmail.com

Abstract: Nowadays there is an evaluation of educational systems and there is a great importance to the educational field. The amount of data stored in educational database is increasing rapidly. These databases contain hidden information for evaluation and improvement of students' performance. The ability to predict a student's performance is very important in educational environments. The feasible technique to achieve this prediction is Data Mining. The academic performance of a student is based upon diverse factors like personal, social, psychological and other environmental variables. The attributes are collected from individual students to predict the performance. In this paper, we have given an overview of Naïve Bayesian Classification Algorithm, to predict students' performance.

Keywords **Data Mining, Educational database, Bayesian algorithm.**

1. INTRODUCTION

Data mining is called knowledge discovery in databases (KDD). Data mining can be used to find the existing relationships and patterns. Data mining combines machine learning, statistics and visualization techniques to discover and extract knowledge. Data mining techniques are used to operate on large amount of data to discover hidden patterns and relationships helpful in decision making. It can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students Mining in educational environment is called Educational Data Mining.

Student retention has become an indication of academic performance and enrollment management. One of the most useful data mining techniques for e-learning is classification. The ability to predict a student's performance is very important in educational environments. Students' academic performance is based upon diverse factors like personal, social, psychological and other environmental variables. A very promising tool to attain this objective is the use of Data Mining.

Education is an essential element for the betterment and progress of a country. It makes the people of a country civilized and well mannered. Mining in educational environment is called educational data mining. Educational data mining is an upcoming field related to several well-established areas of research including e-learning, adaptive hypermedia, intelligent tutoring systems, web mining and data mining etc. As we know, large amount of data is stored in educational database;

data mining is the process of discovering interesting knowledge from these large amounts of data stored in database, data warehouse or other information repositories. To assist the low academic achievers in higher education the objectives are [1]:

(a) Generation of data source of predictive variables
(b) Identification of different factors, which effects a student's learning behavior and performance during academic career
(c) Construction of a prediction model using classification data mining techniques on the basis of identified predictive variables
(d) Validation of the developed model for higher education students studying in Indian Universities or Institutions. Predictive classification enhances the quality of higher education system to increase number of loyal students to evaluate students' data to study the main attributes that may affect the enrollment factor.

A number of data mining techniques have already been done on educational data mining to improve the performance of students like Regression, Genetic algorithm, Bayes classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Classification is one of the most frequently studied problems by data mining and machine learning researchers. It consists of predicting the value of a categorical attribute based on the value of other attributes.

Classification methods like decision trees, rule mining, Bayesian network etc. can be applied on the educational data for predicting the students behavior, performance in examination etc. Decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles and the leaf nodes are denoted by ovals. It the most commonly used algorithm because of its ease of implementation and easier to understand compared to other classification algorithms. The outcome of the decision tree predicted the number of students who are likely to pass, fail or promoted to next year. Decision tree can be constructed relatively fast compared to other methods of classification. Trees can be easily converted into SQL statements that can be used to access databases efficiently. Decision tree classifiers obtain similar and sometimes better accuracy when compared with other classification methods. Decision tree algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm.

The C4.5, ID3, CART and Naïve Bayesian, decision tree algorithms are applied on the data of students to predict their performance. The prediction of students' performance with high accuracy is more beneficial for identifying low academic achievements students at the beginning. To improve their performance the teacher will monitor the students' performance carefully.

2. Literature Survey

Umesh Kumar Pandey, Brijesh Kumar Bhardwaj and Saurabh pal [2] in their paper Data Mining as a Torch Bearer in Education Sector, discussed the different type of researches used in the education sector using data mining. Students, educators and academic responsible person can use these findings to improve the quality of education.

Brijesh Kumar Bhardwaj and Saurabh Pal [3] in their paper Data Mining: A prediction for performance improvement using classification, Discussed an experimental methodology those are used to generate a database, where the raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute selection. They only proved that the academic performances of the students are not always depending on their own effort.

Boumedyen Shannaq, Yusupov Rafael and V. Alexandro [4] in their paper Student Relationship in Higher Education Using Data Mining Techniques, aimed to improve the current trends in the higher education systems to understand from the system which factors might create loyal students. Their research work concentrated only on the number of students will enroll in the upcoming years.

Decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically, each branch of the tree is a classification question, and the leaves of the tree are partitions of the data set with their classification. Decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles and the leaf nodes are denoted by ovals. It is the most commonly used algorithm because of its ease of implementation and easier to understand compared to ther classification algorithms. The outcome of the decision tree predicted the number of students who are likely to pass, fail or what will be the marks one can score.

Here are some of the interesting things about the decision tree.

- It divides up the data on each branch point without losing any of the data (the number of total records in a given parent node is equal to the sum of the records contained in its two children).
- It is pretty easy to understand how the model is being built (in contrast to the models from neural networks or from standard statistics).
- It can be constructed relatively fast compared to other methods of classification.
- Trees can be easily converted into SQL statements that can be used to access database efficiently.

2.1. ID3 Algorithm

Iterative Dichotomiser3, it is a decision tree algorithm introduced in 1986 by Quinlan Ross. It is based on Hunts algorithm. The tree is constructed in two phases. The two phases are tree building and pruning.

Pruning means to change the model by deleting the child nodes of a branch node. Pruned node is regarded as a leaf node, leaf node cannot be pruned. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used.

To build decision tree [9], information gain is calculated for each and every attribute and select the attribute with the highest information gain to designate as a root node. Label the attribute as a root node and the possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances. Continuous attributes can be handled using the ID3 algorithm by discretizing or directly, by considering the values to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

2.2. C4.5 Algorithm

This algorithm is a successor to ID3 developed by Quinlan Ross. It is also based on Hunt's algorithm. C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes,

C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child [8]. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

2.2.1. Limitations of C4.5 Algorithm:

Some of the limitations of C4.5 Algorithm are listed below:

- **Empty branches**: Constructing tree with meaningful value is one of the crucial steps for rule generation by C4.5 algorithm. Raj Kumar and DR. Rajesh Verma [5] in their experiment, they have found many nodes with zero values or close to zero values. These values neither contribute to generate rules nor help to construct any class for classification task. Rather it makes the tree bigger and more complex.
- **Insignificant branches**: Numbers of selected discrete attributes create equal number of potential branches to build a decision tree. But all of them are not significant for classification task. These insignificant branches not only reduce the usability of decision
- **Over fitting**: Over fitting happens when algorithm model picks up data with uncommon characteristics. This cause many fragmentations is the process distribution. Statistically insignificant nodes with very few samples are known as fragmentations [6]. Generally C4.5 algorithm constructs trees and grows it branches 'just deep enough to perfectly classify the training examples'. This strategy performs well with noise free data. But most of the time this approach over fits the training examples with noisy data.

2.3. CART

CART stands for Classification and Regression Trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values.

CART uses Gini Index as an attribute selection measure to build a decision tree .Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

The pruning process is completed in one pass through the tree C4.5's tree-construction algorithm differs in several respects from CART [7]

- Tests in CART are always binary, but C4.5 allows two or more outcomes.
- CART uses the Gini diversity index to rank tests, whereas C4.5 uses information-based criteria.
- CART prunes trees using a cost-complexity model whose parameters are estimated by cross-validation. C4.5 uses a single-pass algorithm derived from binomial confidence limits.

3. METHODOLOGY

3.1. System architecture for student performance evaluation.

System architecture for the proposed system is shown in figure 1. The data is collected from the database and analyzed. The necessary preprocessing steps are applied on the data. Then the Bayesian Classification Algorithm is applied for the preprocessed data. The performance of the result obtained is evaluated and the pattern is extracted to predict the student performance in the further examination.

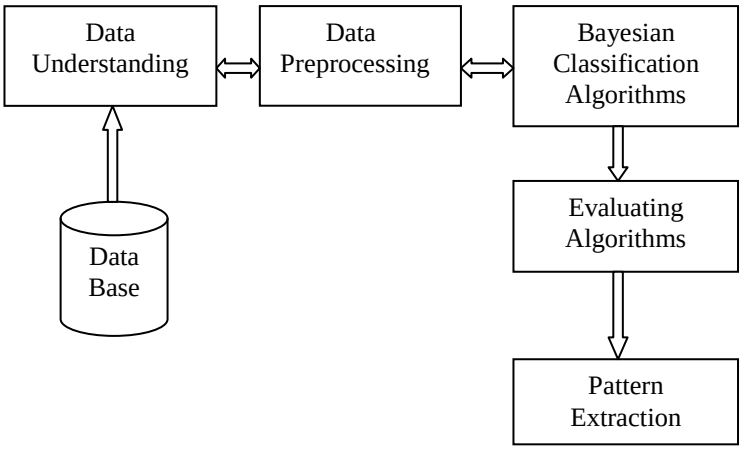

Figure.1.System architecture for Performance Evaluation.

A Naive Bayes algorithm is one of the most effective methods in the field of text classification, but only in the large training sample set can it get a more accurate result. The requirement of a large number of samples not only brings heavy work for previous manual classification, but also puts forward a higher request for storage and computing resources during the computer post-processing [10]. Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular

feature of a class is unrelated to the presence (or absence) of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficiency of naive Bayes classifiers. Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. Classification is enduring to be one of the most researched problems due to continuously-increasing amount of electronic documents and digital data. Naive Bayes is an effective and a simple classifier for data mining tasks [11]. The naive Bayes classifier's beauty is in its simplicity, computational efficiency, and good classification performance. In fact, it often outperforms more sophisticated classifiers even when the underlying assumption of (conditionally) independent predictors is far from true. This advantage is especially pronounced when the number of predictors is very large.

The Classifier

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values $a1, a2, ..n$ This results in:

$$V_{nb} = \arg\max_{vj \in V} P(v_j) \prod P(a_i \mid v_j) \quad (1)$$

We generally estimate $P(a_i \mid v_j)$ using m-estimates:

$$P(a_i \mid v_j) = \frac{n\_c + mp}{n + m} \quad (2)$$

n is the number of training examples for which $v = v_j$

n_c is number of examples for which $v = v_j$ and $a = a_i$

p is a priori estimate for $P(a_i \mid v_j)$

m is the equivalent sample size

Table 1.
Student related attributes for Performance Evaluation

VARIABLE	DESCRIPTION	POSSIBLE VALUES
Gender	Students' gender	Male, female
Cat	Students' category	Gm, Obc, Sc, St
Med	Medium of teaching	Kannada, English, Hindi
SFH	Students food habit	Veg, Non-veg
SOH	Students other habit	Smoking, Drinking
LLoc	Living location	Village, Talluk, District
Hos	Where do u stay	Hostel, Room, Pg
FSize	Number of members in a family	2, 3, >3
FStatus	Students family status	Joint, Individual
FAIn	Family annual income status	Bpl, Poor, Medium, High
GSSLC	Students grade in 10th	<40,40-59, 60-80,>80
GPUC	Students grade in 12th	<40,40-59, 60-80,>80
TColl	Students college type	Boys, Girls, Co-ed
FQual	Fathers qualification	No-ed, Elementary, Secondary, Graduate, PG,Doctarate
MQual	Mothers qualification	No-ed, Elementry, Secondary,Graduate, PG,Doctarate
FOcc	Fathers occupation	Farmer, Business, Service, Retired
MOcc	Mothers occupation	Farmer, Business, Service, Retired
IHE	Student interested in higher education	Yes, No
UOM	Do u use mobile?	Yes, No
UOI	Do u use internet?	Yes, No
UOSN	Do u use social network?	Yes, No
SQ	How many siblings & their qualification?	Yes, No
RH	Reading habit	Night, Early morning
UOV	Do u use vehicle	Yes, No

The attributes that are required as the input to the system are shown in the table 1. Information pertaining to 25 attributes is collected from individual student, which include the diverse factors that affect the student behavior. These include personal, social and academic details. The Naïve Bayesian Classification Algorithm is applied to these attributes to predict the performance of the student
.

4. CONCLUSION

Through this paper, any education institute will have the ability to predict the students' performance. This helps the teacher to give more attention towards the weaker students to improve their performance. This helps in reducing the failing ratio and to take actions at the right time. Bayesian classification method is applied on student database to predict the student performance. Prediction, results and recommendation are provided by this information, which help the user to take further decision.

REFERENCES

[1] M. Sukanya, S. Biruntha, Dr.S. Karthik and T. Kalaikumaran, "*Data Mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm"* International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012

[2] Umesh Kumar Pandey, Brijesh Kumar Bhardwaj, Saurabh pal "*Data Mining as a Torch Bearer in Education Sector"* International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 1, January 2012

[3] Brijesh Kumar Bhardwaj, Saurabh Pal *"Data Mining: A prediction for performance improvement using classification"* (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011

[4] Boumedyen Shannaq, Yusupov Rafael, V. Alexandro *"Student Relationship in Higher Education Using Data Mining Techniques"* Vol. 10 Issue 11 (Ver. 1.0) October 2010

[5] Raj Kumar. Dr. Rajesh Verma, *"Classification Algorithm for Data Mining: A Survey"*. International Journal of Innovations in Engineering and Technology (IJIET) Vol. 1 Issue 2 August 2012.

[6] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques"*, Morgan Kaufmann Publish, 2001.

[7] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) "*Classification and regression trees,Wadsworth."*

[8] Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal, "*Mining Education Data to Predict Student's Retention: A comparative Study*", IJCSIS) International Journal of Computer Science and Information Security, Vol. 10, No. 2, 2012.

[9] D.Lavanya, Dr.K.Usha Rani, "*Performance Evaluation of Decision Tree Classifiers on Medical Datasets*", International Journal of Computer Applications (0975 – 8887),Volume 26–No.4, July 2011.

[10] Yuguang Huang Beijing Univ. of Posts &, Telecommunication., Beijing, china Lei Li *"Naïve Bayes classification algorithm based on small sample set*", Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference.

[11] Meena, M.J. Chandran, "*Naïve Bayes text classification with positive features selected by statistical method"* Advanced Computing, ICAC. First International Conference, IEEE 2009.