Predicting Instructor Performance using Naïve Bayes Classification Algorithm in Data Mining Technique

Priya Subhash Patil
P.G.Student, Department of Computer Engineering,
GF's GCOE, Jalgaon, Maharastra, India

Nilesh Choudhary
Assistant Professor, Department of Computer Engineering,
GF's GCOE, Jalgaon, Maharastra, India

ABSTRACT
Data mining applications are becoming a more common tool in understanding and solving educational and administrative problems in higher education. Generally, research in educational mining focuses on modeling student's performance instead of instructors' performance. One of the common tools to evaluate instructors' performance is the course evaluation questionnaire to evaluate based on students' perception. In this study, classification algorithm of Naïve Bayes, K-Means clustering and C5.0 are used to build classifier models. Their performances are compared over a dataset composed of responses of students to a real course evaluation questionnaire and students final examination results using accuracy, precision, recall, and specificity performance metrics. Although all the classifier models show comparably high classification performances, Naïve Bayes classifier is the best with respect to accuracy, precision, and specificity.

Keywords
Performance evaluation, students final examination results, C5.0, Naïve Bayes classifier, K-Means Clustering.

1. INTRODUCTION
Nowadays Data Mining (DM) has attracted a lot attention in data analysis area, and it became recognizable new tool for data analysis that can be used to extract valuable and meaningful knowledge from data. DM offers promising ways to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be used to predict future behavior. Accordingly, DM has been adopted by many researchers to solve real-world problems in various domains such as marketing, stock market, telecommunication, industrials, health care, medical and customer relationship. Recently a reasonable number of researches have been conducted to apply DM techniques in the education area in ordered to classify and predict student performance in numerous education institutes. Employing DM techniques in education is promising because of the tremendous opportunities in this area[2]. Recent national policies on higher education mandating high stakes evaluation of instructors and the learning system coupled with the quest for an optimal algorithm for evaluation of instructors' performance in higher institutions of learning especially in the developing countries are primary motivation for this work.

Higher education institutions are interested in predicting the paths of students and alumni, thus identifying which students will join particular course programs and which students will require a large number of debates. Nowadays, one of the biggest challenges that educational institutions face is the sudden growth of educational data and to use this data to improve the quality of managerial decisions. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from these large data sets[4]. Moreover, education systems claim new approaches which improve quality, efficiency, and achievement. Mostly DM is utilized in education to investigate the impact of pedagogical strategies on students, and how students understand the course. The academic performance of students based on several factors. The most important factors are the attributes such as the previous academic records, economic status, family background, and demographic data, and the prediction methods. Thus most of the research in this area relayed on the attributes specified student data[2]. The students feedback is an indirect assessment measuring tool which is extensively being used as an evaluation of teaching in the field of higher education[3]. This kind of feedback is not only beneficial for addressing students concerns but also facilitates appropriate enhancement activities undertaken by the institution. A variety of formal and informal procedures based on qualitative and quantitative methods are commonly used with the aim of identifying a variety of issues concerning faculty, curriculum, teaching methodology and essential support services for resolving the identified issues and for enhancing the overall quality of academic programs and services provided by the institution.

This paper attempts to investigate the data associated with the student result and feedback for the instructors to improve the quality of education and indicate the factors that affect the student performance. The prediction of student performance is mainly related to the quality of teaching process. In this paper, some data classification algorithms are applied to Student Evaluation dataset to predict student achievement, investigate instructor's performance, and find the best classification algorithm in accordance with high accuracy.

2. IMPLEMENTED SYSTEM
In the implemented system, Naïve Bayesian classification approach, K-Means Clustering and C5.0 are used to predict the instructor performance. The implemented Naïve Bayesian classifier is best with respect to accuracy, precision, recall and specificity. The system is designed by collecting datasets of the students result and students evaluation of the instructor's performance from the senior students of the institution.

3. CLASIFICATION MODEL
Classification is the separation or ordering of objects into classes. There are two phases in classification algorithm: first, the algorithm tries to find a model for the class attribute as a function of other variables of the datasets. Next, it applies previously designed model on the new and unseen datasets for determining the related class of each record[5]. It is done by using a classifier model, which is built by applying a learning algorithm on a training set composed of past instances having the same variable set as the unseen instance. However, the class label of each instance in the training set is clearly known before training. After learning phase, the classification

performance of the classifier model built is evaluated on an independent test set before used[1].

In classification, there are many different methods and algorithms possible to use for building a classifier model. Some of the most popular ones can be counted as Decision tree algorithms, Support vector machines (SVM), Artificial neural networks (ANN), Discriminant analysis (DA), Logistic regression,Naïve Bayes and Genetic Algorithm. In this paper we will use Decision Tree, Naïve Bayes, and K-Means Clustering.

3.1 K-Means clustering algorithms:

K-Means clustering algorithms build a hierarchy of quality clusters. One of the main problems with the K-Means clustering is that the documents put together in the early stage of the algorithm will never be changed. In other words, K-Means clustering tries to preserve the local optimization criterion but not the global optimization criterion [TSK05]. If we somehow correct these misplaced documents in the generated clusters, we can try to preserve the global optimization criterion.

Our algorithm uses both the top-down (Bisect K-means) and bottom-up (UPGMA) agglomerative K-Means clustering algorithms to address this problem. We pass the K' cluster information (centroids) computed from the bisect K-means algorithm to the UPGMA algorithm to correct the inconsistencies occurred due to the wrong decision made while merging or splitting
a cluster.

First, we ran the bisect K-means algorithm on the document collection for a particular value of the K0 (in this case K0 =pN, Appendix B) until K0 number of document clusters were generated. One cluster with more number of documents or highest intra-cluster similarity value is chosen at each step to split. The generated document clusters should not be empty. Then, we calculated the centroids for each of the resulting clusters. Each of these centroids represents a document cluster and all of its documents.

1. Pick a cluster to split. (Initially the whole document collection is used as a single cluster)
2. Find 2 sub clusters using k-means algorithm .
3. Repeat Steps 1 (Initialization step) and 2 (bisecting step) until the K' > K number of clusters are generated .
4. Compute the centroids (cluster prototypes) for each of the K' clusters such that each document in a collection be longs to one of these centroids .
5. Construct a K' X K' similarity matrix between these centroid clusters .
6. Merge two similar centroid clusters (i.e. , place these centroids in the same cluster) .
7. Update the centroid clusters similarity matrix .
8. Repeat Steps 6 (Merging s t ep) and 7 (Updating step) until the K clusters of centroids are generated .
9. If two centroids belong to same centroid clusters , then the document clusters of these centroids will go together as a final cluster (Merging step) .

In Steps 5 & 8, we ran the UPGMA agglomerative K-Means clustering algorithm on the centroids of these document clusters for a given value of K (given in the algorithm) to generate a set of K centroid clusters. We used the term centroid clusters to avoid possible confusion with the document clusters. Like document cluster is a cluster of documents, centroid cluster is a cluster of centroids. The resulting centroid clusters are used as a reference in merging the document clusters to obtain the final K clusters as shown in the Step 9

3.2 C5.0 Classifier:

C5.0 is based on the information gain ratio that is evaluated by entropy. The information gain ratio measure is used to select the test features at each node in the tree. Such a measure is referred to as a feature (attribute) selection measure. The attribute with the highest information gain ratio is chosen as the test feature for the current node. Let D be a set consisting of (D1… Dj) data instances. Suppose the class label attribute has m distinct values defining m distinct classes, Ci(for I = 1,…,m). Let Djbe the number of samples of D in class Ci. The expected information needed to classify a given sample is given by

$\text{Splitinfo}_A(D) = -\sum (|Dj|/|(D|)*\log((|Dj|/|D|))$
$\text{Gail ratio}(A) = \text{Gain}(A)/\text{Splitinfo}_A(D)$

Where
$\text{Gain} = \text{Info}(D) - \text{InfoA}(D)$
$\text{Info}(D) = -\sum Pi \log2(Pi)$

And
$\text{Info}_A(D) = -\sum(|Dj|/(|D|)* \text{Info}(Dj)$

Where pi = probability of distinct class Ci, D =data Set, A=Sub attribute from attribute, (|Dj|/|D|)=act as weight of jth partition. In other words, Gain (A) is the expected reduction in entropy caused by knowing the value of feature A.

Algorithm for Experimental Model
Input: dataset.
Output: classified output.
1. Take a data set as input.
2. If that set has more features then apply the feature selection technique (PCA) as pre-processing technique
3. Apply parallelism from step 4 to step 6.
4. Evaluate the entropy value and information gain ratio of all three entropies (Shannon, havrda and Charvat's entropy and quadratic entropy).
5. Construct the models individually using C5.0 algorithm based on various entropies.
6. Find the accuracy and execution time of each model and store the value in array.
7. Find a model that has maximum Accuracy.
8. If two have maximum accuracy then
9. Find a minimum execution time of the model that has maximum accuracy.
10. Categorize by that model which has minimum execution time.
11. Else categorization done by the model which has maximum accuracy.
12. End

3.3 Naïve Bayes Algorithm:

Naïve Bayesian classifiers assume that there are no dependencies amongst attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, hence is called "naive". This classifier is also called idiot Bayes, simple Bayes, or independent Bayes.

Algorithm 1 Implemented Naive Bayes Algorithm
1. Let, N is Number of parameters
2. Let, M[N] is Matrix of N

3. Let, P[N]is Probability of N
4. Let, c is classes
5. Let, Pi is Individual Probability
6. Let, Cn is Number of classes
7. Let, Pn is Number of probability
8. Initialize an array M[N] for N no. of parameters
   Where N is real number and $1 < N < 20$
9. Let, P[N] be array of possible values in M[N]
   P[N]= {1; 2; 3; ….};
10. Calculate individual probability Pi for all classes
    Hence, Pi = P (Cn);
    Where 1<i<Cn
11. Calculate group probability for all combinations
    Hence, Pn = P (n | n + c) where, n and c are no .of classes
12. Calculate prediction from individual and group
    Hence, P (Ci | Pn) > P(Cj | Pn)
13. Calculate maximum probability from the prediction
    Hence,
    P(Ci > Pn) = P(N | Ci) P(Cj) | P(N)
    Pmax < P(Ci | Pn)
    Pmax = P(Ci | Pn)

The advantages of Naive Bayes over C5.0 classifier are[5]:

- It uses a very instinctive technique. Bayes classifiers, different from neural networks, do not have several free parameters that must be set. This greatly simplifies the design process.

- Since the classifier returns probabilities, it is easy to apply these results to a wide variety of tasks than if an arbitrary scale was used.
- It does not require large amounts of data before learning can start.
- Naive Bayes classifiers are computationally fast when making decisions.

4. RESULT EVALUATION

Data is collected from one of the randomly selected department of college. A total of 1400 students result and 14000 evaluations are obtained.

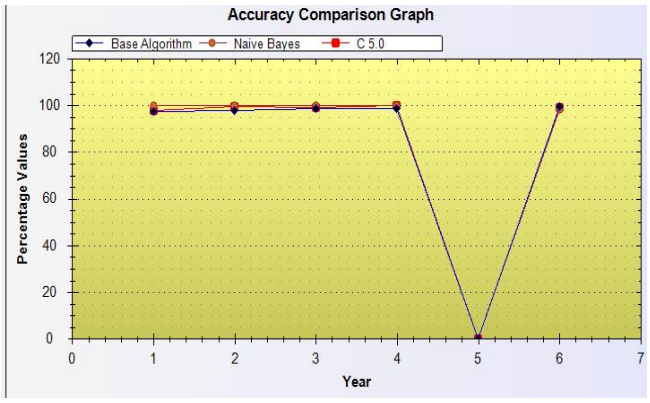

Fig 1: Accuracy Comparison Graph

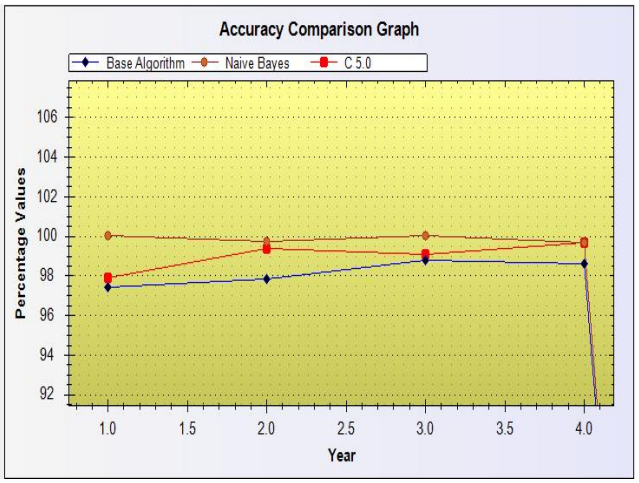

Fig 2: Zoom image of accuracy comparison graph

Figure 1 shows accuracy comparison between naïve bayes classification ,C5.0 decision tree algorithm and K-Means clustering algorithm. Naïve bayes classification shows higher accuracy than C5.0 and K-Means algorithm. Accuracy values, which assess the effectiveness of the models. The equation[1] is used to calculate the accuracy.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \quad \text{----------> [1]}$$

Where

P→ Positive
N→ Negative
TP→ True Positive
TN → True Negative

5. CONCLUSION AND FUTURE SCOPE

The implemented system provides the accurate performance of the staff and performances of classification algorithms used in building a model. This implemented system implement model using naïve bayes classification ,C5.0 decision tree algorithm and K-Means clustering algorithm. Naive bayes gives higher accuracy than C5.0 and K-Means clustering. Naïve Bayes can outperform more sophisticated classification methods. Based on Accuracy the performance of Naïve Bayes is the best. In this system Naïve Bayes outperforms Decision Tree and k-Means clustering.

In existing system only student evaluation is used to predict staff performance but in implemented system student evaluation as well as students result and staff personal details such as experience, education etc. is also used to predict staff performance. So that implemented system gives more accurate staff performance.

In future we implement the system using staff evaluation which include the questionary about his/her (mastery) subject. Questionary is designed by the expert of that subject. So that the system gives more accurate staff performance.

6. REFERENCES

[1]. Mustafa agaoglu" Predicting Instructor Performance Using Data Mining Techniques in Higher Education" volume 4, 2016 IEEE

[2]. Ahmed Mohamed Ahmed, Ahmet Rizaner,Ali Hakan Ulusoy" Using data mining to predict instructor performance" International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016

[3]. Anwar Muhammad Abaidullah, Naseer Ahmed, and Edriss Ali "Identifying Hidden Patterns in Students Feedback through Cluster Analysis" International Journal of Computer Theory and Engineering, Vol. 7, No. 1, February 2015

[4]. Monika Goyal and Rajan Vohra" Applications of Data Mining in Higher Education" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012

[5]. Ahmad Ashari, Iman Paryudi, A Min Tjoa" Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013

[6]. Randa Kh. Hemaid, Alaa M. El-Halees" Improving Teacher Performance using Data Mining" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015

[7]. B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, ``Predicting student performance: An application of data mining methods with an educational Web-based system," in Proc. 33rd Annu. IEEE Frontiers Edu., vol. 1. Nov. 2003, p. T2A-13.

[8]. V. Kumar and A. Chadha, ``An empirical study of the applications of da mining techniques in higher education," Int. J. Adv. Computer. Sci. Appl., vol. 2, no. 3, pp. 80_84, 2011.

[9]. C. Romero and S. Ventura, ``Educational data mining: A survey from 1995 to 2005," Expert Syst. Appl., vol. 33, pp. 135_146, Jul. 2007.

[10]. B. K. Baradwaj and S. Pal, ``Mining educational data to analyze students' performance," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 6, pp. 63_69, 2011.

[11]. S. Calkins and M. Micari, ``Less-than-perfect judges: Evaluating student evaluations," NEA Higher Edu. J., pp. 7_22, Fall 2010.

[12]. J. Sojka, A. K. Gupta, and D. R. Deeter-Schmelz, ``Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences," College Teach., vol. 20, no. 2, pp. 44_49, 2002.

[13]. L. Coburn. (1984). Student Evaluation of Teacher Performance, ERIC/TME Update Series. [Online]. Available: http://ericae.net/edo/ED289887.htm

[14]. S. A. Radmacher and D. J. Martin, ``Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis," J. Psychol., vol. 135, no. 3, pp. 259_269, 2001.