ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(16), DOI: 10.17485/ijst/2015/v8i16/62055, July 2015* ISSN (Online) : 0974-5645

A Naïve Bayesian Classifier for Educational Qualification

S. Karthika* and N. Sairam

School of Computing, SASTRA University, Thanjavur, India; karthikaraghavendrar@gmail.com

Abstract

Manual classification of the individuals into different categories based on their educational qualification is a tedious task and it may vary respective to the considered scenario. This paper proposes a classification methodology utilizing the benchmark Naïve Bayesian classification algorithm for the classification of persons into different classes based on several attributes representing their educational qualification. The experimental results are appreciable indicating that the proposed classification method can be a promising one and can be applied elsewhere. The proposed method has been experimentally verified to be 90% accurate with a high kappa value thus proving its efficiency. This classification methodology can reduce the mundane manual labor and can easily assist in categorization.

Keywords: Classification, Data Mining, Educational Qualification, Kappa, Naïve Bayesian

1. Introduction

There are quite a large number of instances where a person is initially judged or analyzed by his/her educational qualification he/she has gained in his life. Under such cases, the categorization of the persons according to their educational qualification would be of much help and the decision made with the help of technical assistance would be free from any kind of biases and hence can be universally applicable.

This paper proposes a method to categorize the educational qualification utilizing the benchmark Naïve Bayesian Classification Algorithm. This method can be used in a variety of applications such as segregation based on educational relevance, short listing a candidate for recruitment based on his/her degree of education, etc. The organization of this paper is given below: Section 2 contains the literature survey. Section 3 explains the Naïve Bayesian Algorithm and the proposed classification method. Section 4 analyses the experimental results based on the listed tabulations and Section 5 concludes the paper.

Naïve Bayesian algorithm is a classical classification algorithm which has proved its simplicity and efficiency in various applications and a few articles exhibiting the efficiency of the classifier are discussed here.Mauricio A. Valle et al.[10]paper discusses the method of predicting the determining attributes in case of a Naïve Bayesian classification algorithm involving a testing method based on cross-validation. It is verified experimentally that the socio-demographic attributes are not contributing to the prediction of future performance of the sales agent in a call center.

Dunja Mladenic et al.[7] research deals with choosing the features contributing for the classification using certain specifications and the learning ability of the classifier over a text data whose distribution is uneven. It is found that when the domain and the characteristics of the classification algorithm istaken into account, the performance of the classifier increases. Dong Tao et al.[2]paper proposes an improved Naïve Bayesian algorithm by combining the classical method with a feature selection method based on Gini Index. This hybrid method improves the performance of text categorization.

Kabir Md Faisal et al.[4] research deals with combining k-means clustering method with Naïve Bayesian classification algorithm to increase the accuracy. The clustering method groups the training samples into

* *Author for correspondence*

similar categories after which all the groups are trained under Naïve Bayesian classifier. This method is verified to improve the accuracy. Santra A.K. et al.[8] research proves that the time taken for classification and the memory utilized are reduced in case of the web usage mining while utilizing Naïve Bayesian classifier rather than using decision trees. Liangxiao Jiang et al.[5] paper suggests that the conditional independence nature of attributes in the original Naïve Bayesian algorithm seems to be weak in certain cases and proposes a local weightage method which outperforms the classical algorithm in terms of accuracy. Pradeepta K. Sarangiet al.[12] paper describes the feature extraction using LU factorization followed by the usage of Naïve Bayesian classifier for pattern recognition. This proves the universal applicability of the classifier.

Yildrim P. and Birant D.[11] research paper discusses the experimental verification of the effect of various distributions on the attributes. It is found that the application of distributions based on the nature of attributes increases the accuracy rather than using a single distribution across all the attributes. AbeerBadr El Din Ahmed and Ibrahim SayedElarabarticle[1] discusses the application of classification algorithms to predict the final grade of the students. Ron Kohavi[6] article proposes a hybrid classifier combining Naïve Bayesian and Decision Tree which is termed as NBTree to increase the accuracy of the classifier. It is also found that the class conditional independence is passive in case of small data sets but in case of large data sets, this assumption leads to misclassification and reduction in the accuracy.

Shasha Wang et al.[9] paper proposes an upgraded version of the NBtree hybrid classifier and named it as multinomialNBTree (MNBTree), where a multinomial naïve Bayesian classifier is applied to the leaf nodes of a decision tree. Further, to increase the performance, another improvisation is made by the inclusion of multiclass classification and the system is called as multiclass version of MNBTree (MMNBTree). With reference to the above stated research articles, the pros of Naïve Bayesian classification algorithm are studied thoroughly and found that this algorithm will best suit the nature of data used for the experiment comprising of both numerical and text data which are independently contributing to the classification.

2. Proposed System

2.1 Naïve Bayesian –an Overview

Naive Bayesian classification algorithm is very simple which assumes that the classification attributes are independent and they do not have any correlation between them. Many researchers have found that this assumption of independence do not work in all cases for which other alternative methods are proposed to increase the performance. One such alternate method is proposed by Liangxiao Jiang[5].

The original Naïve Bayesian technique is based on the conditional probability and the maximum likelihood occurrence. The Naive Bayesian algorithm based on the description provided in [3] is given as follows:

- Let G be the training set with N tuples where each tuple is represented as a 'k' dimensional attribute vector M, where $M=\{M_1,M_2,\ldots M_k\}$
- Let there be 'p' classes $C_1, C_2,\ldots C_p$. According to this Naive Bayesian classifier, a tuple T belongs to class C_x only when it has a higher conditional probability than any other class C_y, where $x \neq y$.
  $P(C_x \mid T) > P(C_y \mid T)$ and
  $P(C_x \mid T) = (P(T \mid C_x * P(C_x)) / P(T)$
- Since class conditional independence is assumed,
  $P(M \mid C_x) = \prod_{i=1}^{k} P(M_k \mid C_x) = P(M_1) * P(M_2) * P(M_3)$
  $\ldots . * P(M_k)$
- Class C_x is predicted as the output class when $P(M \mid C_x) * P(C_x) > P(M \mid C_y) * P(C_y)$, where $1 \leq x,y \geq p$ and $x \neq y$

2.2 Data Set Description

The scenario considered here deals with the details of the educational qualification of the persons gained in his/her educational career. Based on these details, the persons are classified into different classes. Here, three classes are taken into consideration and they are:

- LOW.
- MEDIUM.
- HIGH.

The attributes taken under consideration are given below:

- Number of years spent totally for education.
- Educational degrees obtained.
- Major Subjects / Area of Specialization relevant to the scenario under consideration.

Number of years spent totally for education is considered as a determining attribute because of the depth of the knowledge gained is directly proportional to the time invested in learning.

Educational degrees obtained by a person play a vital role in deciding the proficiency of the person.

The areas of specialization of the persons are considered as the determining attribute because of the flexibility of the system to be used in a variety of scenarios and this attribute is application dependent.

The data set comprises of about 25000 entries where 20000 entries are used for training and the remaining 5000 entries are used for testing the efficiency of the classifier and in determining how well it has learnt to classify the data. The representation of the data set is as follows:

- Let the entire data be represented as $D = \{D_1, D_2, \ldots D_{25000}\}$.
- Let the training data be represented as Train = $\{D_1, D_2, \ldots D_{20000}\}$.
- Let the test data be represented as Test = $\{D_{20001}, D_{20002}, \ldots D_{25000}\}$.

A person with same educational qualification may be differently categorized under different scenarios because the classification is application dependant.

2.3 Training and Test Data Set Samples

The training set sample corresponding to the selection of a candidate for teaching under the Chemistry department is given below. The table clearly shows the different attributes used for classification and the application dependant nature. The test data is also similar to the training data without the class column which will be predicted with the help of the algorithm implementation given below.

Table 2.

Academic Degree	Numbers of years	Area of Specialization	Class
B.Sc	15	Chemistry	Medium
M.Com	17	Commerce	Low
M.Sc	17	Chemistry	High
Ph.D	25	Chemistry	High
M.B.A	18	Administration	Low
B.Sc	15	Physics	Low
Ph.D	22	Biology	Low

Note: Though a person holding a PG degree is considered to be a person with high educational qualification, the person is labeled as belonging to class LOW because of the attribute Area of Specialization which is application dependant.

2.4 Classification Methodology

The training data is used to categorize the data under consideration into three classes as mentioned in the above section based on the determining attributes. The architecture for such a classification methodology is given in Figure 1.

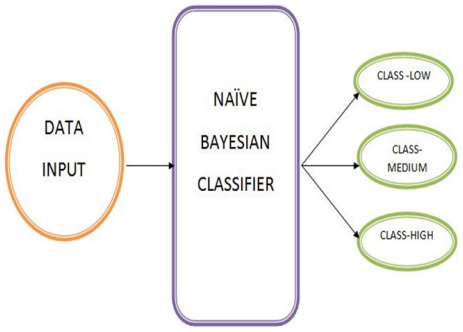

Figure 1. Architecture of the proposed classifier.

2.5 Algorithm

Procedure : Naïve Bayesian Algorithm for Educational Qualification

- Begin
- Initialization

nc → Number of classes

na → Number of attributes

N → Number of samples

- for each class C_i do

Calculate prior probability $P(C_i) = \Sigma C_i / \Sigma N$, $i \in \{1,nc\}$

- for each class C_i do
  For each attribute A_j do

Calculate the conditional probability $P(A_j|C_i) = \Sigma C_i$ *with* $A_j / \Sigma C_i$, $i \in \{1,..,nc\}$ *and* $j \in \{1,..,na\}$

- for each class Ci do

Calculate the conditional probability of the tuple K i.e $P(K|C_i) = P(A_1|C_i) * P(A_2|C_i) * \ldots * P(A_{na}|Ci)$

- for each class Ci do

Calculate the posterior probability of the tuple K i.e $P(C_i) * P(K|C_i)$

- Prediction

If$((P(C_p) * P(K|C_p)) > (P(C_q) * P(K|C_q)))$

Prediction $\rightarrow C_p$

Else

Prediction $\rightarrow C_q$

Where $p,q \in \{1,..,nc\}$ *and* $p \neq q$

- *End*

3. Experimental Results and Analysis

The classification output showing the number of test samples getting classified into the three different classes are tabulated in Table 1. Based on this table, the performance measures are calculated and are tabulated in Table 2 and Table 3. Table 2 gives the Accuracy and Kappa value of the classifier. Table 3 lists the Sensitivity, Specificity and Prevalence of the three different classes considered.

Table 1. Number of samples classified into different classes

PREDICTED\ EXPECTED	CLASSES LOW	MEDIUM	HIGH
LOW	1233	126	0
MEDIUM	100	1716	126
HIGH	0	133	1566

Table 2. Accuracy & Kappa measures of the classifier

PARAMETERS	VALUE
Accuracy	0.903
Kappa	0.8528

Table 3. Sensitivity & Selectivity measures for the three classes

PARAMETERS	LOW class	MEDIUM class	HIGH class
Sensitivity	0.9250	0.8689	0.9255
Specificity	0.9656	0.9253	0.9598
Prevalence	0.2666	0.3950	0.3384

3.1 Performance Measures

Traditionally, accuracy was the main measure to determine the efficiency of the classifier. But, accuracy is unreliable when the data set is completely skewed because of its inclination towards the dominant class. Hence, various other measures are introduced which can clearly portrays the true ability of a classifier.

The performance measures used to determine the efficiency of the classifier are described below:

3.1.1 Sensitivity

Sensitivity denotes the ability of the classifier to identify the positive class as positive and negative class as negative correctly. Sensitivity depicts the proportion of True Positive Values.

Sensitivity = truepos / (truepos + falseneg)

3.1.2 Specificity

Specificity denotes the ability of the classifier to exclude the values of the positive class from negative class and negative class from the positive one appropriately. Specificity depicts the proportion of True Negative Values.

Specificity = trueneg / (trueneg + falsepos)

3.1.3 Accuracy

Accuracy is the measure to determine how much of the positive class values correctly classified as positive and how much negative class is labeled exactly as negative. Accuracy depicts the proportion of the correctly classified values

Accuracy = (truepos + trueneg) / (totpos + totneg)

3.1.4 Kappa

Kappa value denotes the inter-rater agreement and a perfect agreement is denoted by a kappa value of 1. As the value decreases, the agreement level decreases. Kappa depicts the normalized statistical value which compares the accuracy of the user defined system to that of a random hypothetical system. It has been proved to the most effective measure in determining the efficiency of the classifier because of its comparison of the system to that of an ideal one. It also determines whether the classifier has learnt how to classify or has memorized the training values i.e the classifier has understood the relation between the attributes and the exact classification methodology and not just replicating the class of the duplicated values.

Kappa = (accuracy – random) /(1 - random)

*Random = ((trueneg + falsepos) * (trueneg + falseneg) + (falseneg + truepos) * (falsepos + truepos)) / (totpos + totneg) * (totpos + totneg)*

3.1.5 Prevalence

Prevalence which is the proportion of the positive class to that of the total population and is given as

Prevelance = (truepos + falseneg) / (totpos + totneg)

3.1.6 Notations Used

Table 4 contains the notations of the values used for the calculation of the above listed performance measures. These values form the basis for determining the above

listed performance measures. These values are usually obtained from the confusion matrix. A confusion matrix is a two-by-two matrix holding the original and the values of the positive and negative classes as predicted by the classifier after training.

Table 4. Notations used in the calculation of performance measures

- True Positives(TRUEPOS) → Correctly Included Values
- False Positives(FALSEPOS) → Incorrectly Included Values
- True Negatives(TRUENEG) → Correctly Excluded Values
- False Negatives(FALSENEG) → Incorrectly Excluded Values
- Included Values → Class X being identified as Class X and Class Y being identified as Class Y
- Excluded Values → Class Y not being identified as Class X and Class X not being identified as Class Y
- Total Positives(TOTPOS) → Total number of values predicted as Positive Class
- Total Negatives(TOTNEG) → Total number of values predicted as Negative Class

3.2 Analysis

The performance results prove that the classical Naïve Bayesian classification algorithm performs well when the attributes are non-numerical. Because of the classifier's original nature of prediction based on maximum likelihood occurrence, the classification results are satisfactory and such a classifier could be used in various applications where a person can be categorized based on his/her educational degree obtained.

High Accuracy and high kappa value indicates that the classification is made by the knowledge obtained during the training process and hence proves that the system is a promising one. High kappa value also instantiates that the system has learnt to classify rather than memorizing the values. High values of Sensitivity and Specificity denote that the classifier is able to identify a positive class as a positive one and exclude the negative classes from the positive one and vice versa.

Naïve Bayesian algorithm which has proved its performance in a variety of real life situations has again proved its performance in this undertaken scenario.

4. Conclusion

A classification methodology based on the classical Naïve Bayesian classifier is proposed in this paper to categorize the persons into different classes based on different attributes relevant to their educational qualification. Future work may be carried out by increasing the number of attributes related to education to determine the original knowledge of the persons and classify them accordingly. Further, various other classification algorithms may be used to check for the variations in the performance measures and analyze the classification ability of the various classifiers.

5. References

1. El Din Ahmed AB, Elarab IS. Data Mining: A prediction for Student's Performance using Classification Method. World Journal of Computer Application and Technology. 2014; 2(2):43–7.
2. Dong T, Shang W, Zhu H. An improved algorithm of Bayesian text categorization. Journal of Software. 2011; 6(9):1837–43.
3. Han J, Kamber M. Data Mining Concepts and Techniques. 2nd Ed. 2006.
4. Faisal KM, Mofizur RC, Alamgir H, Kesav D. Enhanced classification accuracy on naive bayes data mining models. International Journal of Computer Applications. 2011; 28(3):9–16.
5. Jiang L, Cai Z, Zhang H, Wang D. Naïve Bayes text classifiers: a locally weighted learning approach. Journal of Experimental and Theoretical Artificial Intelligence. 2013; 25(2):273–86.
6. Kohavi R. Scaling up the accuracy of the Naïve Bayes Clasifiers a Decision-Tree Hybrid.Proceedings of the second international conference on knowledge discovery and data mining;1996. p. 202–7.
7. Mladenic D, Grobelnik M. Feature selection for unbalanced class distribution and naive bayes. Proceedings of the 16th International Conference on Machine Learning (ICML); 1999. p. 258–67.
8. Santra AK, Jayasudha S.Classification of web log data to identify interested users using Naïve Bayesian classification. International Journal of Computer Science Issues. 2012; 9(1):381–7.
9. Wang S, Jiang L, Li C. Adapting naive Bayes tree for text classification. Knowledge and Information Systems. 2015; 44(1):77–89.
10. Mauricio AV, Samuel V, Gonzalo R. Job performance prediction in a call center using a naive Bayes classifier. Expert Systems with Applications. 2012; 39(11):9939–45.
11. Yildrim P, Birant D. Naïve Bayes classifier for continuous variables using novel method(NBC4D) and distributions. IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA); 2014;Alberobello. p. 110–5.
12. SarangiPK, AhmedP, RavulakolluKK. Naïve Bayes Classifier with LU Factorization for Recognition of Handwritten Odia. Indian Journal of Science and Technology. 2014; 7(1):35–8.