Available at: www.researchpublications.org

Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification

Tina R. Patil, Mrs. S. S. Sherekar
Sant Gadgebaba Amravati University, Amravati
tnpatil2@gmail.com, ss_sherekar@rediffmail.com

ABSTRACT

Classification is an important data mining technique with broad applications to classify the various kinds of data used in nearly every field of our life. Classification is used to classify the item according to the features of the item with respect to the predefined set of classes. This paper put a light on performance evaluation based on the correct and incorrect instances of data classification using Naïve Bayes and J48 classification algorithm. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree. The paper sets out to make comparative evaluation of classifiers NAIVE BAYES AND J48 in the context of bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool.

The experiments results shown in this paper are about classification accuracy, sensitivity and specificity. The results in the paper on this dataset also show that the efficiency and accuracy of j48 is better than that of |Naïve bayes.

Keywords: True positive rate, False positive rate, Naïve bayes, J48 Decision tree

I. INTRODUCTION

Data mining is growing in various applications widely like analysis of organic compounds, medicals diagnosis, product design, targeted marketing, credit card fraud detection, financial forecasting, automatic abstraction, predicting shares of television audiences etc. Data mining refers to the analysis of the large quantities of data that are stored in computers.

Data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files.

Different functions of data mining are mainly classified as classification, clustering, feature selection and association rule mining.

In this paper, we focus on the data classification and the performance measure of the classifier algorithms based on TP rate, FP rate generated by the algorithms when applied on the data set [1][2].

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

II. DATA CLASSIFIER

In this paper, two classifiers, naive bayes algorithm and J48 decision tree algorithm are used for comparison. Comparison is made on accuracy, sensitivity and specificity using true positive and false positive in confusion matrix generated by the respective algorithms. Also we can use the correct and incorrect instances that give us a most efficient method for classification by using the confusion matrix [2].

A. Decision tree algorithm J48:

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple[1][3].

```
Algorithm [1] J48:
INPUT:
        D     //Training data
OUTPUT
        T     //Decision tree
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and label;
For each arc do
        D= Database created by applying splitting predicate to D;
        If stopping point reached for this path, then
                T'= create leaf node and label with appropriate class;
        Else
                T'= DTBUILD(D);
        T= add T' to arc;
}
```

While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them [4][5].

B. Naive Bayes classifier:

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems [6].

Naïve Bayesian classifier is based on Bayes' theorem and the theorem of total probability. The probability that a document d with vector x = < x_1,...,x_n> belongs to hypothesis h is[7][1]

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi|h1)P(h1) + P(xi|h2)P(h2)}$$

Here, P(h1|xi) is posterior probability, while P(h1) is the prior probability associated with hypothesis h1.

For m different hypotheses, we have

$$P(xi) = \sum_{j=1}^{n} P(xi|hj)P(hj)$$

Thus, we have

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi)}$$

III. MEASURING PERFORMANCE

The performance of classification algorithm is usually examined by evaluating the accuracy of the classification. However since classification is often a fuzzy problem, the correct answer may depend on the user. Traditional algorithms evaluation approaches such as determining the space and time overhead can be used but these approaches are usually secondary. Determining which better best is depends on the interpretation of the problem by users.

Classification accuracy is usually calculated by determining the percentage of tuples placed in a correct class. This ignores the fact that there also may be a cost associated with an incorrect assignment to the wrong class. This perhaps should also determine [1][8].

An OC(operating characteristics) curve or ROC(receiver operating characteristic) curve or ROC(relative operating characteristic) curve shows the relationship between false positives and true positives. An OC curve was originally used in communication area examined false alarm rates. It has also been used in information retrieval to examine fall out (percentage of retrieved that are not relevant) VS recall (percentage of retrieve that are relevant) [1][8].

A. Confusion matrix:

A confusion matrix illustrates the accuracy of the solution to a classification problem. Given n classes a confusion matrix is a m x n matrix, where $C_{i,j}$ indicates the number of tuples from D that were assign to class $C_{i,j}$ but where the correct class is C_i. Obviously the best solution will have only zero values outside the diagonal [1].

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study [8]:

1. a is the number of correct predictions that an instance is negative,
2. b is the number of incorrect predictions that an instance is positive,
3. c is the number of incorrect of predictions that an instance negative, and
4. d is the number of correct predictions that an instances positive [9].

Some standards and terms:

1. True positive (TP): If the outcome from a prediction is p and the actual value is also p, then it is called a true positive.
2. False positive (FP): However if the actual value is n then it is said to be a false positive.
3. Precision and recall: Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. Recall is nothing but the true positive rate for the class [10][11].

In this paper, we have used weka (Waikato environment for knowledge analysis) tool for comparison of naive bayes and J48 algorithm and calculating efficiency based on accuracy regarding correct and incorrect instances generated with confusion matrix. We have used here bank-data-train.arff for data classification available on web URL http://www.cs.bme.hu/~kiskat/adatb/bank-data-train.arff. [12]This bank relation consists of attributes age, gender, region, income, married, children, car, mortgage, pep with 300 instances.

IV. Experimental Work And Results

We have performed classification using Naïve Bayes algorithm and J48 decision tree algorithm on bank-data-train.arff dataset in weka tool. Weka tool provide inbuilt algorithms for naïve Bayes and J48.

A. Results for classification using J48 :

Mortgage attribute has been chosen randomly for bank data set. J48 is applied on the data set and the confusion matrix is generated for class gender having two possible values i.e. YES or NO.

Confusion Matrix:

```
 a    b      ← classified as
33   72    | a = YES
25   170   | b = NO
```

For above confusion matrix, true positives for class a='YES' is 33 while false positives is 72 whereas, for class b='NO', true positives is 170 and false positives is 25 i.e. diagonal elements of matrix 33+170 =203 represents the correct instances classified and other elements 25+72 = 97 represents the incorrect instances.

True positive rate = diagonal element/ sum of relevant row

False positive rate = non-diagonal element/ sum of relevant row

Hence,

TP rate for class a = 33/(33+72) = 0.314

FP rate for class a = 25/(25+170) = 0.128

TP rate for class b = 170/(25+170) = 0.871

FP rate for class b = 72/(33+72) = 0.685

Average TP rate = 0.677

Average FP rate = 0.491

Precision = diagonal element/sum of relevant column

Precision for class a = 33/(33+25) = 0.568

Precision for class b = 170/(170+72) = 0.702

F-measures = 2*precision*recall/(precision + recall)

F-measure for class a = 2*0.568*0.314/(0.568+0.314) = 0.404

F-measure for class b= 2*0.702*0.871/(0.702+0.871) = 0.778

Cost/Benefit analysis :

Fig. 1 shows the Cost of J48 for class yes = 105.

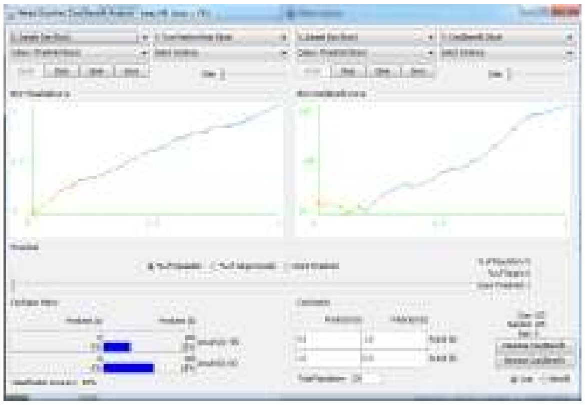

Fig. 1 Cost analysis for class YES

Fig. 2 shows Cost of J48 for class no = 195.

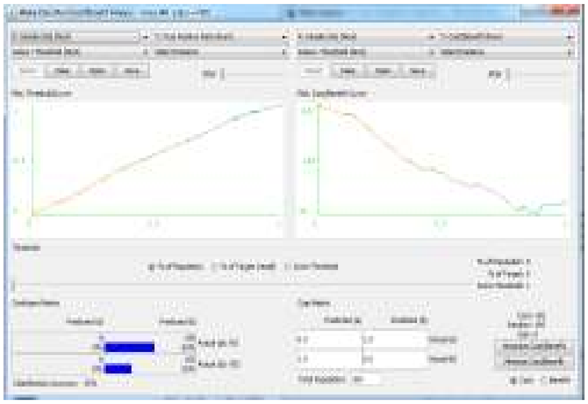

Fig. 2 Cost analysis for class NO

B. Results for classification using Naïve Bayes :

Here same, Mortgage attribute has been chosen for bank data set. Naïve Bayes is applied on the data set and the confusion matrix is generated for class gender having two possible values i.e. YES or NO.

Confusion Matrix:

```
  a    b     ← classified as
 10  95   |  a = YES
 21 174   |  b = NO
```

For above confusion matrix, true positives for class a=’YES’ is 10 while false positives is 95 whereas, for class b=’NO’, true positives is 174 and false positives is 21 i.e. diagonal elements of matrix 10 + 174 =184 represents the correct instances classified and other elements 21+95 = 116 represents the incorrect instances.

TP rate for class a = 10/(10+95) = 0.095

FP rate for class a = 21/(21+174) = 0.108

TP rate for class b = 174/(21+174) = 0.892

FP rate for class b = 95/(10+95) = 0.905

Average TP rate = 0.613

Average FP rate = 0.626

Precision for class a = 10/(10+21) = 0.323

Precision for class b = 174/(174+95) = 0.647

F-measure for class a = 2*0.323*0.095/(0.323+0.095) = 0.147

F-measure for class b = 2*0.892*0.647/(0.892*0.647) = 0.75

Cost/Benefit analysis :

Fig. 3 shows the Cost of Naïve Bayes for class YES = 105

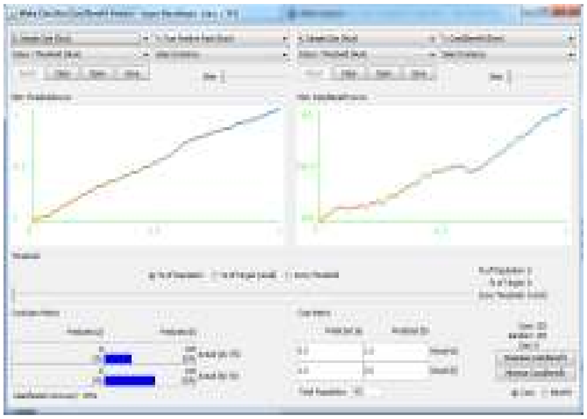

Fig. 3 Cost analysis for class YES

Fig. 4 shows the Cost of Naïve Bayes for class NO = 195

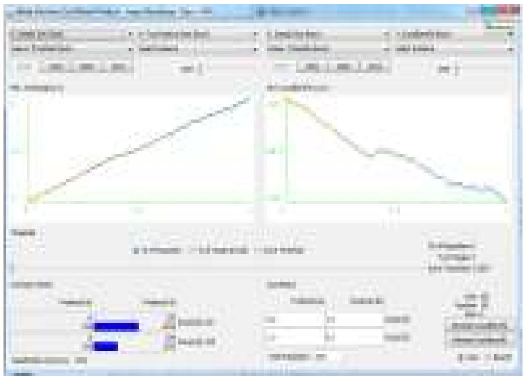

Fig. 4 Cost analysis for class NO

Though results of cost and benefit analysis for mortgage is same for J48 and Naïve Bayes, but in case of gender cost/benefit analysis for J48 is lesser than that of Naïve Bayes as shown below.

Cost of J48 for class Male = 142

Cost of J48 for class Female = 144

Cost of Naïve Bayes for class Female = 155

Cost of Naïve Bayes for class Female = 147

V. Conclusion

From above experimental work we can conclude that correct instances generated by J48 are 203 and Naïve Bayes are 184, as well as performance evolution on the basis of mortgage is:

Classification Accuracy			Cost analysis	
Mortgage	Naïve Bayes	J48	Naïve Bayes	J48
YES	9 %	31%	105	105
NO	89 %	87%	195	195

This proves that the, J48 is a simple classifier technique to make a decision tree. Efficient result has been taken from bank dataset using weka tool in the experiment. Naive Bayes classifier also showing good results. The experiments results shown in the study are about classification accuracy and cost analysis. J48 gives more classification accuracy for class mortgage in bank dataset having two values Yes and No. Though here in this example, cost analysis valued same for both the classifier, with gender attribute, we can prove that J48 is cost efficient than the Naïve Bayes classifier.

References

[1] Margaret H. Danham,S. Sridhar, " Data mining, Introductory and Advanced Topics", Person education , 1st ed., 2006

[2] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok, "A Data Mining Framework for Building Intrusion Detection Models"

[3] Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895

[4]http://www.jstor.org/discover/10.2307/40398417?uid=3738256&uid=2134&uid=368470121&uid=2&uid=70&uid=3&uid=368470111&uid=60&sid=21101751936641

[5]http://stackoverflow.com/questions/10317885/decision-tree-vs-naive-bayes-classifier

[6] George Dimitoglou, James A. Adams, and Carol M. Jim," Comparison of the C4.5 and a Naïve Bayes Classifier for the Prediction of Lung Cancer Survivability"

[7] Seongwook Youn, Dennis McLeod, "A Comparative Study for Email Classification"

[8] Anshul Goyal, Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", IJAER, Vol. 7, No. 11, 2012, pp.

[9] Xiang yang Li, Nong Ye, "A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, Vol. 36, No. 2, 2006, pp. 396-406

[10] Hong Hu, Jiuyong Li, Ashley Plank, "A Comparative Study of Classification Methods for Microarray Data Analysis", published in CRPIT, Vol. 61, 2006.

[11] Milan Kumari, Sunila Godara, "Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction", IJCST, Vol. 2, Issue 2, 2011, pp. 304-308

[12]http://www.cs.bme.hu/~kiskat/adatb/bank-data-train.arff