Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparative analysis of Naive Bayes and J48 Classification Algorithms

Sunita Joshi* MTECH Scholars India
Bhuwaneshwari Pandey MTECH Scholars India
Nitin Joshi MTECH Scholars India

Abstract*— Data mining technology is today's need. Due to advancement in technology and population huge amount of data available in different field. This data may be in the form of documents, may be graphical formats, may be the video, and may be records. We need to extract the useful information from this data so that it can be used for human welfare. Data Mining is the hot research area to solve various problem and classification is one of the main problems in the field of data mining. Classification is a data mining technique based on machine learning which is used to categorize the data item in a data set into a set of predefined classes. In this paper we use two classification algorithm J48 and Naïve bayes. Naïve bayes algorithm is based on probability and J48 algorithm is based on decision tree. We make comparative evaluation of Naïve Bayes and J48 in the context of diabetes dataset. The results of comparison shown in this paper are about classification accuracy and cost analysis. The result shows that efficiency and accuracy of Naïve Bayes is better than that of J48.

Keywords*— Data Mining, Classification, J48, Naïve Bayesian, True Positive Rate, False Positive Rate, Recall, Precision.

I. INTRODUCTION

Classification is a technique of classifying the data into certain class based on some criteria or similarity. This requires extraction and selection of feature that is well describes to a given class. Classification is also called supervised learning, as the instances are given with known class labels, contrast to unsupervised learning in which labels are unknown. Each instance in the dataset represented by set of features or attributes which may be categorical or continuous [7] [8]. Classification is the process of building the model from the training set. The resulting model is then used to predict the class label of the testing instances [9].

II. CLASSIFICATION ALGORITHMS

This section explains classification algorithm J48 and Naïve Bayes.

A. J48 decision tree classifier:

J48 is the decision tree based algorithm and it is the extension of C4.5. With this technique a tree is constructed to model the classification process in decision tree the internal nodes of the tree denotes a test on an attribute, branch represent the outcome of the test, leaf node holds a class label and the topmost node is the root node. Model generated by decision tree helps to predict new instances of data [3].

Algorithm [1] J48:
INPUT
D // Training data
OUTPUT
T // Decision tree
DTBUILD (*D)
{
T = Null;
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 tree J48 ignores the missing value. J48 allows classification via either decision tree or rules generated from them [5][6].

B. Naïve Bayesian classifier:

Bayesian classification represents a supervised learning method as well as statistical method for classification. It is simple probabilistic classifier based on Bayesian theorem with strong independence assumption. It is particularly suited when the dimensionality of input is high. They can predict the probability that a given tuple belongs to a particular class. This classification is named after Thomas Bayes (1702-1761) who proposed the bayes theorem. Bayesian formula can be written as [4]:

$$P(H \mid E) = [P(E \mid H) * P(H)] / P(E)$$

The basic idea of Bayes's rule is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed from the Bayes's rule.

WEKA TOOL

Weka (Waikato environment for knowledge analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The Weka suite contains a collection of visualization tool and algorithm for data analysis. In Weka dataset should be formatted to the ARFF format. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file. The pre-process panel has facilities for importing data from database and for pre-processing this data using filtering algorithm. These filters can be used to transform the data [1] [2].

III. PERFORMANCE INVESTIGATION AND RESULTS

Experiment are performed on Weka with 10 fold cross validation. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier [7]. The first step is to find the number of instances of diabetes dataset using both Naïve Bayes and j48 classification algorithm. In the next step experiment calculates the classification accuracy and cost analysis.

Confusion Matrix: - Confusion matrix contain information about actual and predicted classification. *Standard terms defined for this matrix [10].*

- *True positive* –if the outcome of prediction is p and the actual value is also p than it is called true positive(TP).
- *False positive*-if actual value is n than it is false positive(FP)
- *Precision* – precision is measure of exactness and quality
  Precision = tp/(tp + fp)
- *Recall*- measure of completeness and quantity
  Recall = tp / (tp + fn)

Dataset information: - In this study we are taking dataset diabetes and for making comparison of two classifier Naïve bayes and j48. Diabetes dataset have total no. of 768 instances. When algorithms are applied to the dataset the confusion matrix is generated.

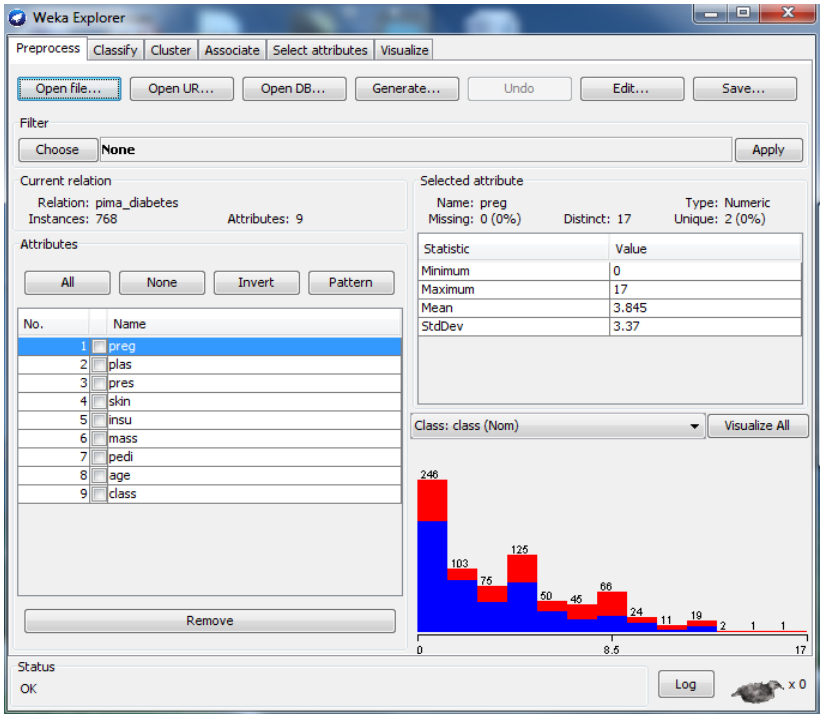

Figure 1: Diabetes dataset open in Weka

RESULT FOR CLASSIFICATION USING J48

J48 is a module for generating a pruned or unpruned C4.5 decision tree. When appplying J48 on diabetes dataset result are as given below:

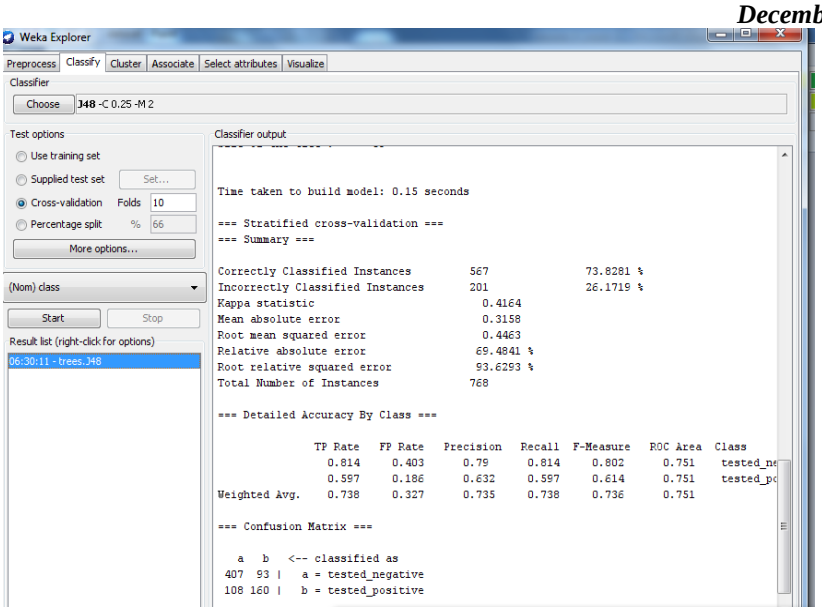

Figure 2: J48 Tree

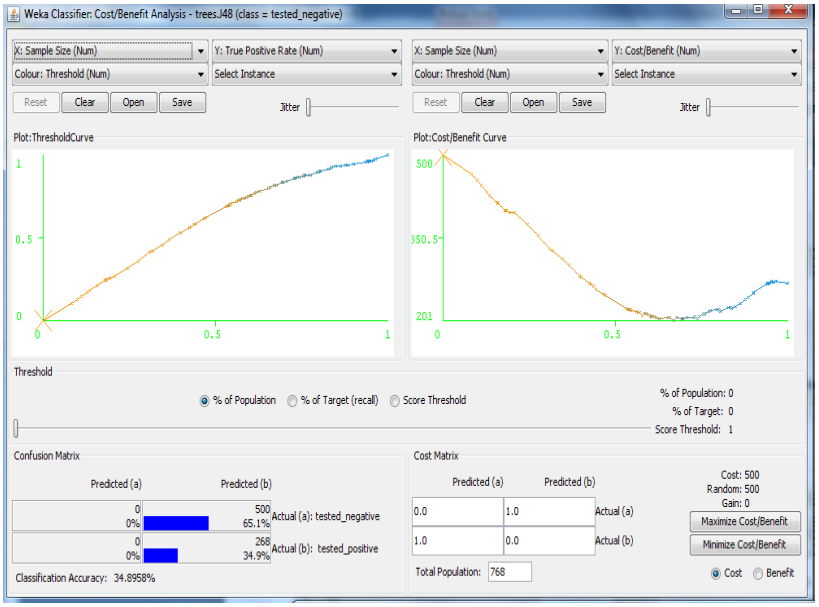

Figure3 : Cost analysis of J48 for class negative

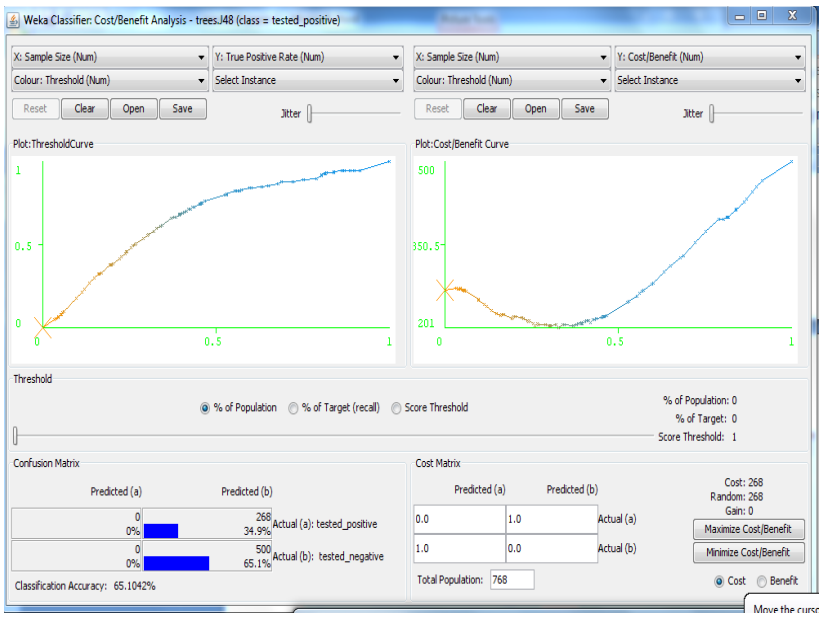

Figure 4: Cost analysisof J48 for class positive

RESULT FOR CLASSIFICATION USING NAÏVE BAYES

When Naïve Bayes algorithm is applied on diabetes dataset, we got the result shown as below on figure .

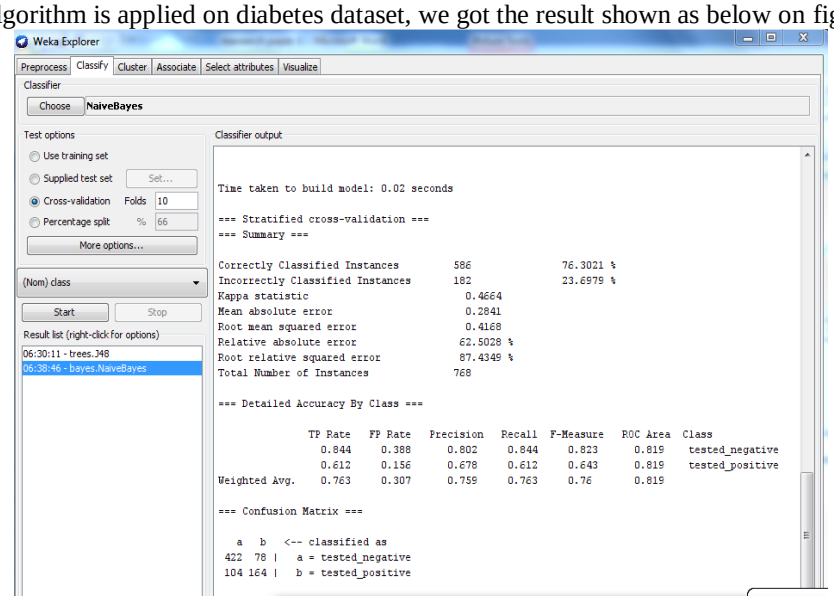

Figure 5: Bayes Net

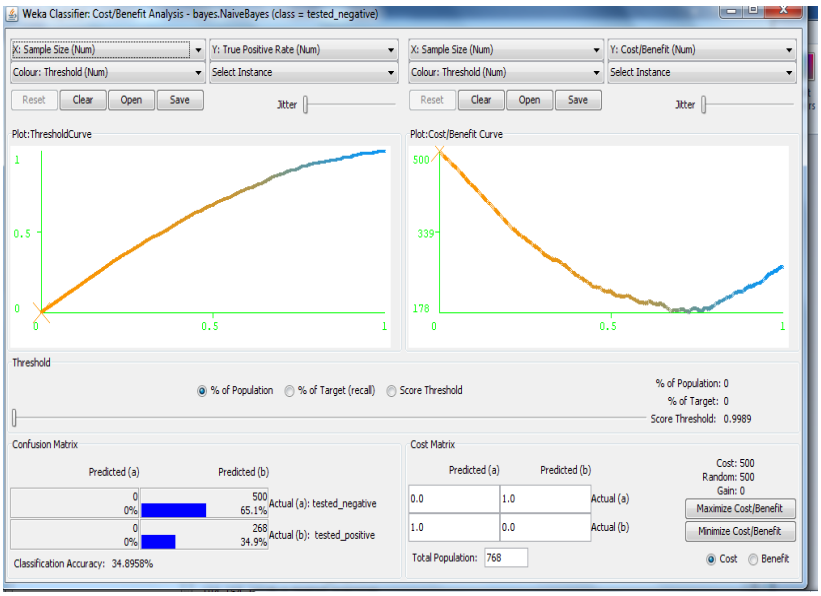

Figure 6: Cost analysis of Naïve Bayes for class negative

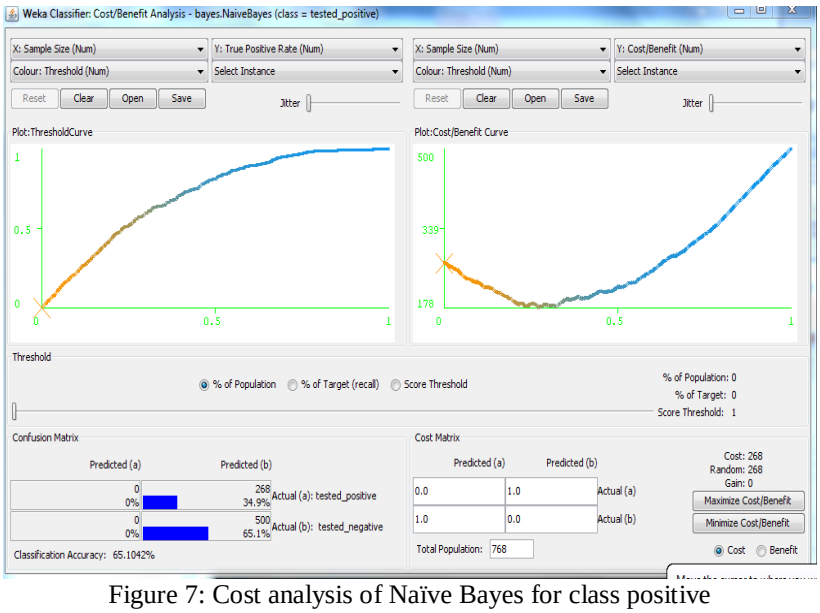

Figure 7: Cost analysis of Naïve Bayes for class positive

IV. CONCLUSIONS & FUTURE WORK

Both the algorithms are applied on the diabetes dataset and the results are given.

Evaluation Criteria		**J48**	**Naïve Bayesian**
Time to Build Model(in secs)		0.06	0.00
Correctly Classified Instances		567	586
Incorrectly Classified Instances		201	182
Prediction Accuracy		73.82	76.30
Cost	**Negative**	500	500
	Positive	268	268

From the result we see time to build the model is less when using Naive Bayes and correctly classified instances are more when using Naive Bayes and prediction accuracy is also greater in Naive Bayes than of J48. Hence it is concluded that Naïve Bayes perform better than of J48 on diabetes dataset.

FUTURE WORK

Classification is important data mining technique used to make sense of data. In this study we focused on comparison of two classification techniques and few issues like accuracy and cost. There are still many issues that can be taken into consideration for further research which are as follows:

- Different algorithms which are not included in Weka can be tested.
- The real dataset from the industry can be taken.
- These algorithms can be compared using Tanagra and Matlab tool.

REFERENCES

[1] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J.Han. Generalization and decision tree induction: Efficient classification in data mining. In Proc. 1997Int. Workshop Research Issues on Data Engineering (RIDE'97), pages 111-120, Birmingham, England, April 1997.

[2] Remco R. Bouckaert, Eibe Frank, Mark Hall Richard Kirkby, Peter Reutemann, Seewald David Scuse, WEKA Manual for Version 3-7-5, October 28, 2011.

[3] Jiawei Han, Micheline Kamber, "Data Mining : Concepts and Techniques", 2nd edition, Morgan Kaufmann, 2006.

[4] Margaret Dunham, "Data Mining: Concepts and Techniques", Morgan Kaufmann Pub.

[5] 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.

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

[7] lan H. Witten, Eibe Frank, "Data Mining –Practical Machine Learning Tools and Techniques,"2nd Edition, Elsevier, 2005.

[8] Efraim Turban, Linda Volonino, Information Technology for Management: Wiley Publication, 8th Edition 2009.

[9] Mr. Shirdhar kamble, Mr. Aditya Desai, Ms. Priya Vartak, "Evaluation and Performance Analysis of Machine Learning Algorithm", Published in International Journal of Engineering Sciences & Research Technology, ISSN: 2277-9655.

[10] Anshul Goyal , Rajni Mehta, "Performaance Comparison of Naïve Bayes and J48 Classification Algorithms", Published in International Journal of Applied Engineering Research, ISSN: 0973-4562 Vol.7 no.11 (2012).

