ISSN No. 0976-5697

Volume 1, No. 4, Nov-Dec 2010

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Analysis of Classification and Clustering Algorithms using Weka For Banking Data

K.Vanitha
Department of Computer Studies,
Saranathan College of Engineering,
Trichy,Tamilnadu, India
rkvanithamca@gmail.com

G.Roch Libia Rani*
Department of Computer Studies,
Saranathan College of Engineering,
Trichy,Tamilnadu, India
roch_libia@yahoo.com

Abstract: In this paper, we investigate the performance of different classification and clustering algorithms using weka software. The J48,Naive Bayes and Simple CART Classification algorthims are evaluated based on accuracy, time efficiency and error rates. The K-means, DBScan and EM clustering algorithms are evaluated based on accuracy of clustering. We run these algorithms on large and small data sets to evaluate how well they work.

Keywords: Classification, Clustering, Naïve Bayes, Simple CART, K-Means, DBScan, EM

I. INTRODUCTION

This paper shows the comparative study of various clustering and classification algorithms for banking data. Classification and clustering are important data mining techniques that partition objects into meaningful disjoint subgroups.

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Classification is a two step process. In the first step, training data are analyzed by a classification algorithm. In the second step, test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.The learning of classifier is supervised in that it is told to which class each training tuple belongs. It contrasts with unsupervised learning or clustering, in which the class label of each training tuple is not known in advance.[1,9].

This paper evaluates the performance of classification algorithms based on accuracy, time efficiency and error rates. We examine various clustering algorithms based on accuracy.

A. Classification algorithms

[a] J48:

J48 [QUI93] implements Quinlan"s C4.5 algorithm [QUI92] for generating a pruned or unpruned C4.5 decision Tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J48 can be used for classification. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation.[2,4]

[b] Naïve Bayes:

Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. It is made to simplify the computation and in this sense considered to be Naive. This assumption is a fairly strong assumption and is often not applicable. However, bias in estimating probabilities often may not make a difference in practice -- it is the order of the probabilities, not their exact values that determine the classifications. Studies comparing classification algorithms have found the Naive Bayesian classifier to be comparable in performance with classification trees and with neural network classifiers. They have also exhibited high accuracy and speed when applied to large databases.

[c] CART (Classification and Regression Trees):

A Classification and regression tree (CART) is a set of techniques for classification and prediction. The technique is aimed at finding a rule(s) which could predict the value of a dependent variable **Y** from known values of **n** explanatory variables X_i,i=1,….,n (predictors). The predictor variables X_i may be a mixture of categorical and continuous variables. The initial data represent a set of objects with known values of the dependent variable Y and predictors X_i. CART builds trees - i.e. formulates simple if/then rules for recursive partitioning (splitting) of all the objects into smaller subgroups. Each such step may give rise to new "branches". The goal of this process is to maximize homogeneity of the

values of the dependent variable Y in the various subgroups. All the CART techniques are essentially non-parametric - they do not rely on any particular assumptions about the type of dependence of the dependent variable Y on predictors X_i (in contrast to various regression techniques) and about statistical properties of the data. This is an essential practical advantage for the cases when apriori information about the data is limited.

There are two main approaches in CART - classification trees (used to predict the class or category of records) and regression trees, (used to predict a continuous value) [5].

B. Clustering algorithm:

[a] Simple K-Means Method

K-means is an algorithm to classify or to group objects based on attributes/features into K number of groups. The grouping is achieved by minimizing the sum of squares of distances between data and the corresponding cluster centroids.

The k-Means method may be described as follows:

Select the number of clusters (k).

Pick k means of the k clusters repeat

Allocate each object to the cluster which has the closest mean

Calculate new mean for each cluster until cluster membership is unchanged

The time complexity of K-Means is O (tkN), where t is the number of iterations, k is the number of clusters and N is the size of the dataset. [7, 8]

[b] DBScan Method

DBScan (Density Based Spatial clustering of application with noise) is to create clusters with minimum size and density. Density is defined as the minimum number of points within a certain distance of each other. DBScan requires two parameters: epsilon (eps) and minimum points (minPts)

Following is the pseudo code of DBScan method to explain how it works.

```
C = 0
for each unvisited point P in dataset D
N = getNeighbors (P, epsilon)
if (sizeof(N) < minPts)
mark P as NOISE
else
++C
mark P as visited
add P to cluster C
recurse (N)
```

DBScan does not require you to know the number of clusters in the data a priori. DBScan does not have a bias towards a particular cluster shape or size. DBScan is resistant to noise and provides a means of filtering for noise if desired.

DBScan does not respond well to high dimensional data. As dimensionality increases, so does the relative distance between points making it harder to perform density analysis. DBScan does not respond well to data sets with varying densities [1].

[c] EM (Expectation Maximization) method

EM (Expectation maximization) method consists of a two step iterative algorithm. The first step is called the estimation step, involves estimating the probability distribution of the clusters. The second step called the maximization step, involves finding the model parameters that maximize the likelihood of the solution.

Given a likelihood function $L(\theta; x, z)$, where θ is the parameter vector, x is the observed data and z represents the unobserved latent data or missing values, the maximum likelihood estimate (MLE) is determined by the marginal likelihood of the observed data $L(\theta; x)$, however this quantity is often intractable [9].

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

Expectation step: Calculate the expected value of the log likelihood function, with respect to the conditional distribution of z given x under the current estimate of the parameters $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{Z|x,\theta^{(t)}}\left[\log L(\theta; x, Z)\right]$$

Maximization step: Find the parameter that maximizes this quantity:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

II. METHODOLOGY

The classification and clustering algorithms are evaluated by applying banking dataset in weka software. Weka is a datamining system that implements data mining algorithms using a java language, the algorithms are applied directly to a data set. The new machine language schemes can also be developed with this package. Weka is open software under General public license. The data file normally used by Weka is in ARFF format which consists of special tags to indicate different things in the data file(attribute name, attribue type, attribute value and data) .Once the data has been loaded, one of other panels in the explorer can be used to perform other analysis . The data used in this study is the banking data. It has a total of 600 instances and 13 attributes. Only 67% of the overall data is used for training and the remaining is used for testing the accuracy of the Classification. The small dataset is extracted as a subset of the huge dataset.

III. RESULT AND DISCUSSION

We examine three classification methods namely j48, Naïve Bayes and Simple CART according to the factors such as accuracy, time and error rates.

According to the size of data each of the three algorithms (J48, Naïve Bayes and simple CART) is executed: first by trying a large dataset and then by trying a small dataset with training set and supplied test set.

Table 1 mainly summarizes the result based on time taken for j48, Naïve Bayes and simple CART classification algorithms for datasets with different size. As a result, as the size of dataset becomes greater, the time for j48 and CART becomes higher and the time for Naïve Bayes remains the same.

Table I. Time taken to build a model

	Large dataset(secs)	Small dataset (secs)
J48	0.06	0.02
Naïve Bayes	0.02	0.02
Simple CART	0.38	0.05

As a result, we can say that a Naïve Bayes takes shortest time (0.02 secs) to build a model for huge dataset when compared to others. J48 and Naïve Bayes requires the shortest time (0.02 secs) when using a small data set. Simple cart requires longest time for both huge and small dataset. Fig 1 shows the graphical representation of the simulation result based on time collection.

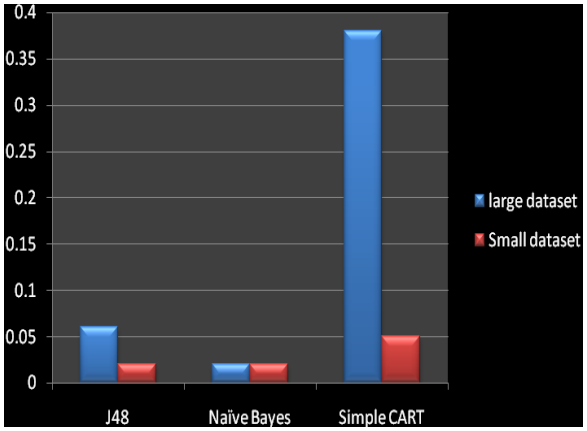

Figure 1. Behavior of both datasets with Time collection

Table 2 and Table 3 summarize the result based on error rates (Mean Absolute error, Root Mean squared error, Relative absolute error, Root relative squared error) for large dataset and small dataset respectively.

The lowest error rate belongs to J48 algorithm for large dataset (23.74%) and small dataset (35.16%). Simple CART has highest average error rate for huge data set (61.635%) and small dataset (45.3635%). Fig 2 shows the error rates for different classification algorithm.

Table II. Error rates for small data set

	Mean Absolute Error	Root Mean squared Error	Relative Absolute Error	Root relative squared error
J48	0.1118	0.2364	31.44	56.13
Naïve Bayes	0.243	0.358	68.35	85.00
Simple Cart	0.3102	0.3938	87.25	93.50

Table III. Error rates for huge data set

	Mean Absolute Error	Root Mean squared Error	Relative Absolute Error	Root relative squared error
J48	0.1457	0.2699	43.34	65.86
Naïve Bayes	0.2834	0.3738	84.30	91.20
Simple Cart	0.2905	0.3811	86.39	92.99

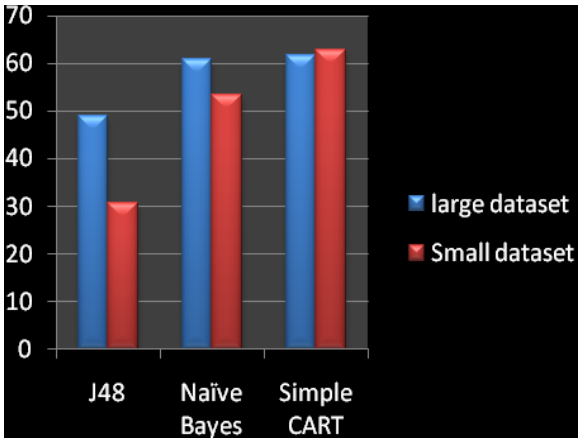

Figure 2. Average Error rates

According to the accuracy j48 has more accuracy than other algorithms. Simple CART has lowest accuracy than others for both data sets. As the dataset size increases the accuracy becomes lower for j48 and Naïve bayes. As the dataset size increases the accuracy becomes higher for Simple CART.

Table IV. Accuracy

	Small dataset	Large dataset
J48	84.8	79.8
Naïve Bayes	61.6	59.9
Simple CART	55.6	59.5

In Fig 3, we can see the accuracy of three classification algorithms.

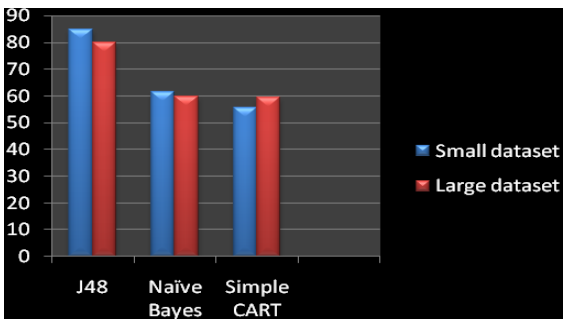

Figure 3. Accuracy

The three clustering algorithm are compared according to the size of data

Table V. Percentage of Incorrectly Classified Instances

	Large dataset	Small dataset
EM	64.14 %	67.6768 %
DBSCAN	55.68%	52.53%
K-MEANS	66.59%	61.62%

According to the accuracy (Table.5), DBScan shows more accuracy in clustering the objects than other algorithms while using huge dataset and small dataset. The EM has less accuracy than others for small dataset. The K-Means shows smallest accuracy than others for large dataset**.** Fig 4 shows the percentage of accuracy.

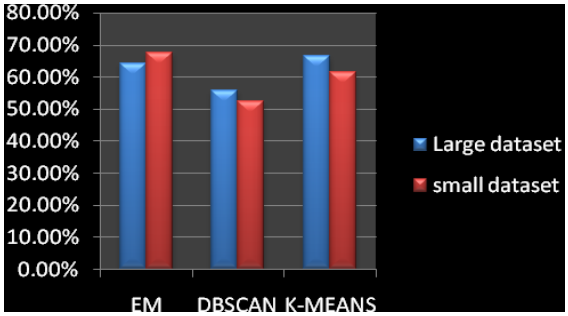

Fig ure 4. Percentage of Incorrectly Classified Instances

IV. CONCLUSION

This study is aimed to find the best classification algorithm for banking data process. Our results show that, the j48 has the highest classification accuracy performance with the lowest error rate. But the j48 requires more time to build the model. On the other hand, we also found that the simple cart has the lowest accuracy and highest error rates. It also has one of the highest time requirements to build the classification model. Therefore, the j48 should be preferred over Naïve Bayes and simple CART for bank information, based on accuracy and error rate.

V. REFERENCES

[1] Osama Abu Abbas, "Comparsion between data clustering algorithms," in Proceedings of International Arab Journal of IT, vol. 5, No 3, July 2008.

[2] Mohd Fauzi bin Othman, Thomas moh shan yaw, "Comparsion of Different Classification Techniques using WEKA for Breast Cancer," in IFMBE Proceeding 15, pp. 520 – 523, 2007.

[3] Gayatri.N,Nicholas.s,Reddy.A.v,Chitra.k, "Performance Analysis of Datamining Algorithms for software Quality Prediction," in Proceedings of International Conference on Advances in Recent Technologies in communication and computing," 2009.

[4] Christos Tjortjis and John Keane, H-Yin et al," T3: A Clasification algorithm for Data Mining," Ideal 2002, LNCS 2412, Springer_verlag Berlin Heidelberg pp.50-55, 2002.

[5] Lin Zhang,Yan chen, Yan liang,Nanli, " Application of Data Mining Classification algorithm in Customer membership Card Classification model",pp.211– 215, 06 Jan 2009.

[6] Sivaram N K. Ramar. Article: "Applicability of Clustering and Classification Algorithms for Recruitment Data Mining". International Journal of Computer Applications 4(5):23–28, July 2010.

[7] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm ", Proceedings of the World Congress on Engineering 2009 Vol I, WCE 2009, July 1 - 3, 2009, London, U.K.

[8] XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg "Top 10 algorithms in data mining", Knowledge Information System 14:1–37, 2008.

[9] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, 2006.