International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 4 Issue 1 January 2013

Performance Evaluation of Classification Techniques Based on Mean Absolute Error

Mikanshu Rani[1], Vikram Singh[2] and Bharat Bhushan[3]
[1]Lecturer, Comuter Science and Applications Deptt. M. M. P. G. College, Fatehabad, India
[2]Proffessor, Comuter Science and Applications Deptt. Chaudhary Devi Lal University, Sirsa, India
[3]Asstt. Proffessor, Government College for Women, Bhodia Kheda, Fatehabad, India
E-mail: mikanshu6406@gmail.com

***Abstract* – Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), is the process that results in the discovery of new patterns in large data sets. Classification is a data mining technique used to map a data item into one of several predefined classes. There are many classification methods to classify instances, but we don't know which classification method is suitable for our dataset i.e. which classification algorithm will give less error. This article evaluates the performance of different classification techniques and compares them, based on the parameter – "Mean Absolute Error". Classification methods covered in this work include Bayesian Networks, Neural Networks, Support Vector Machines, and Nearest Neighbor. To render more credibility to the results, the target algorithms have been tested on five datasets taken from UCI Machine Learning Repository. This comparison will show which algorithm is best, in terms of mean absolute error i.e. which will give less error. The performance of classification techniques are evaluated by using open source software named "WEKA" (Waikato Environment for Knowledge Analysis).**

***Keywords*- Classification, Data Mining, Machine Learning, Neural Networks, Bayesian Networks, Nearest Neighbor, Support Vector Machines, WEKA.**

I. INTRODUCTION

Knowledge Discovery and Data Mining are rapidly evolving areas of research that are at intersection of several disciplines, including statistics, databases, AI, visualization, and parallel computing. KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process (Fayyad et al., 1996). Thus data mining is extraction of new, implicit, valid and previously unknown patterns from the vast amount of data available in the data sets). The goal of data mining is to extract knowledge from a data set in a human understandable structure. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity (Phyu, 2009). Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects) (Gupta et al., 2011).

Classification is a data mining technique with roots in machine learning, is used to map a data item into one of several predefined classes. For example, an email program might attempt to classify an email as legitimate or spam. In this task the goal is to predict the value of a user-specified goal attribute based on the values of other attributes, called the predicting attributes (Bishnoi, 2011). 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.

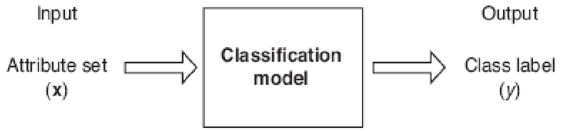

Fig. 1 Classification maps an input attribute x into its class label y

Classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes related to that object. The "Classification Problem" involves data which is divided into two or more groups, or classes. The data mining software is asked to tell us which of the groups a new example falls into. Classification analysis is the organization of data in given classes. Classification is a supervised machine learning procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (referred to as traits, variables, characters, etc) and based on a training set of previously labeled items (Bishnoi, 2011). Aim of this paper is to evaluate the performance of classifiers on the basis of mean absolute error. For this, we test four classification techniques on five datasets using WEKA.

II. LITRATURE REVIEW

Fayyad et al. (1996) introduced the concepts of data mining and knowledge discovery in databases. This article provided an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases.

Phyu (2009) has presented the review of basic classification techniques namely decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques.

Gupta et al. (2011) have summarized various review and technical articles on breast cancer diagnosis and prognosis. In this paper they presented an overview of the current research being carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis.

Thakker et al. (2011) have evaluated the performance of different classification techniques including Multi-Layer Perceptron, Naive Bayes, K-Nearest Neighbor, Decision tree, Support Vector Machine to have most suitable classification technique for the cashew grading system.

Othman & Yau (2007). They have presented the comparison of different classification techniques including Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm using breast cancer data.

Li et al. (2008) have presented a comprehensive comparative study of both five feature selection methods including expert judgment, CFS, LVF, Relief-F, and SVM-RFE, and fourteen algorithms from five distinct kinds of classification methods including decision tree, artificial neural network, support vector machines(SVM), Bayesian network and ensemble learning.

Justin et al. (2010) have presented the comparison of efficiency of classification techniques for the task of classifying a speaker's emotional state into one of two classes: aroused and normal. Their aim was to differentiate the efficiency of classification techniques so that speaker's emotional state can be classified into aroused or normal class.

III. METHODS

The data mining community inherits the classification techniques developed in the diversity of disciplines. Four distinct methods were examined and described below.

1) Bayesian Networks: BNs are probabilistic graphical models that encode probabilistic dependence relations among variables. A Bayesian network, Bayes network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via DAG. This classifier learns from training data the conditional probability of each attribute Ai given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A1…..An and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes (Othman & Yau, 2007). The Bayesian network structure S is a directed acyclic graph (DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the lack of possible arcs in S encodes conditional independencies (Phyu, 2009).

Naive Bayes is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable (Othman & Yau, 2007). The Naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. This is called conditional independence. It is obvious that the conditional independence assumption is rarely true in most real-world applications. A straightforward approach to overcome the limitation of Naive Bayes is to extend its structure to represent explicitly the dependencies among attributes.

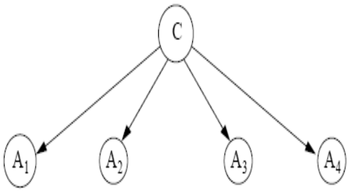

Fig. 2 An example of Naive Bayes

Figure 2 shows an example of naive Bayes. In naive Bayes, each attribute node has no parent except the class node. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) independence assumptions (Li et al., 2008).

2) Neural Networks: Neural networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input. A neural network starts with an *input layer*, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a *hidden layer*. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an *output layer*. The output layer consists of one or more response variables.

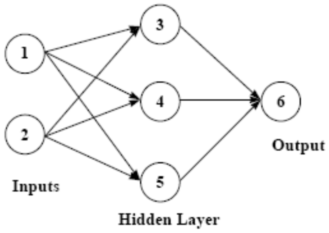

Fig. 3 A neural network with one hidden layer.

Multi-Layer Perceptron (MLPs) also called Feed Forward Neural Networks, is defined as “a network in which the directed graph establishing the interconnections has no closed paths or loops” (Soman et al., 2008). MLPs are trained with the standard back propagation algorithm. They are supervised networks so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input–output map. They have been shown to approximate the performance of optimal statistical classifiers in difficult problems. The most popular static network is the MLP.

3) Support Vector Machines: Support Vector Machines are among the most robust and successful classification algorithms. Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. These are based upon the idea of maximizing the margin i.e. maximizing the minimum distance from the separating hyper plane to the nearest example (Thakkar et al., 2011).

SVMs are well suited to dealing with interactions among features and redundant features. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two data sets (Burges, 1998). To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyper-plane, which is “pushed up against” the two data sets. A good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier (Burges, 1998). This hyper-plane is found by using the support-vectors and margins.

4) Nearest Neighbors classifiers: Among the various methods of supervised learning, the Nearest Neighbor rule achieves consistently high performance, without a priori assumptions about the distributions from which the training examples are drawn. It involves a training set of both positive and negative cases. A new sample is classified by calculating the distance to the nearest training case; the sign of that point then determines the classification of the sample. A very simple classifier can be based on a nearest-neighbor approach. Nearest neighbor algorithm is considered as statistical learning algorithms and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbor does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest neighbor classifier finds the closest training-point to the unknown point and predicts the category of that training point accordingly to some distance metric (Darrell et al., 2006).

K-NN is a type of instance-based learning, or lazy learning. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. The neighbors were taken from a set of objects for which the correct classification is known (Bishnoi, 2011).

IV. WEKA

WEKA is open source data mining software written in java, is a collection of machine learning algorithms for data mining tasks. The Waikato Environment for Knowledge Analysis (WEKA) is a unified workbench that allows researchers easy access to state-of-the-art techniques in machine learning. It was developed by the University of Waikato in New Zealand and implements data mining algorithms using the JAVA language. The WEKA project provides a comprehensive collection of machine learning algorithms and data preprocessing tools. The workbench includes algorithms for regression, classification, clustering, association rule mining and attribute selection. Moreover, it also includes tools for data visualization. The data is usually imported from the ARFF file format, which consists of special tags to indicate different attribute names, attribute types, attribute values and the data itself (Justin et al., 2010).

WEKA is a comprehensive tool bench for machine learning and data mining. It is open source data mining software written in java and widely tested in all operating systems. Weka is a collection of machine learning algorithms for data mining tasks. The Waikato Environment for Knowledge Analysis (WEKA) is a unified workbench that allows researchers easy access to state-of-the-art techniques in machine learning. It was developed by the University of Waikato in New Zealand and implements data mining algorithms using the JAVA language (Pushpa, 2010).

WEKA is a landmark system in the history of the data mining and machine learning research communities because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time (the first version of WEKA was released 11 years ago). Other data mining and machine learning systems that have achieved this are individual systems, such as

C4.5, not toolkits. The WEKA GUI Chooser window is used to launch WEKA's graphical environments. At the bottom of the window are four buttons: Simple CLI, Explorer, Experimenter, and Knowledge Flow (Kirkby & Frank, 2004). The main interface in WEKA is the Explorer. It has a set of panels, each of which can be used to perform a certain task. Once dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis.

V. EXPERIMENT AND RESULTS

A. Datasets from UCI Repository

WEKA expects the data to be in ARFF format because it is necessary to have type information about each attribute, which cannot be automatically deduced from the attribute values. An ARFF file is developed for WEKA machine learning software. Therefore before applying any algorithm to your data, it must first be converted to ARFF form or datasets in ARFF format were taken from UCI Repository (Soman et al., 2008).

We take five datasets from UCI Machine Learning Repository named heart-statlog, diabetes, hepatitis, labor and vote to evaluate the performance of classification techniques using WEKA. Weka expects the datasets to be in ARFF format, so firstly we change the format of these datasets into ARFF format and then test classification techniques on every data set. After the conversion of datasets into ARFF format, their detailed information is given in the table below:

TABLE I
DATASETS IN ARFF FORMAT

Sr. No.	Dataset name	Instances	Attributes
1	heart-statlog.arff	270	14
2	diabetes.arff	768	9
3	labor.arff	57	17
4	hepatitis.arff	155	20
5	vote.arff	435	17

B. Experimental Results

To investigate the performance of the selected classification methods on many datasets, we use the same experiment procedure as suggested by WEKA. In WEKA, all data is considered as instances and features in the data are known as attributes. Performance of each classifier tested on datasets is represented by table and chart. "Mean Absolute Error" is the only parameter for performance evaluation of classification techniques.

Mean Absolute Error: The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. Mean absolute error, MAE is the average of the difference between predicted and actual value in all test cases; it is the average prediction error. The lower value of mean absolute error is considered as good.

Performance based on Mean Absolute Error: Any classification technique is considered as good if it has lower value of mean absolute error. The performance of classification techniques on the basis of mean absolute error is shown in the following table and graph:

The performance of each classifier is different for different datasets. These classification algorithms discussed above have been applied on these five datasets. Table II show the comparative result of the four classification algorithm that applied on five data sets in terms of mean absolute error.

TABLE III
VALUE OF MEAN ABSOLUTE ERROR TAKEN BY CLASSIFIERS ON DATASETS

Dataset Name	Classification Techniques			
	Bayesian Networks (Naive Bayes)	Neural Network (MLP)	Support Vector Machine (SMO)	Nearest Neighbor (KNN-IBK)
heart-statlog	0.1835	0.2328	0.1593	0.2502
diabetes	0.2841	0.294	0.2253	0.3027
labor	0.094	0.1752	0.1404	0.1404
hepatitis	0.1754	0.1928	0.1484	0.1979
vote	0.0995	0.0553	0.0414	0.073

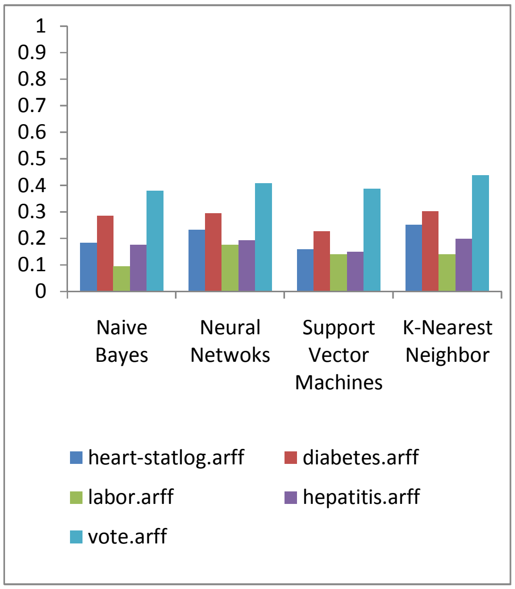

Fig. 4 Graphical representation of Mean Absolute error of classifiers on five data sets

Figure 4 is the graphical representations of performance of classifiers evaluated on five datasets. This chart compares all the classification techniques tested on five datasets on the basis of mean absolute error.

Lower value of mean absolute error is considered as good. We analyze the performance of classifiers on the basis of average value of mean absolute error for classifier on all datasets. Table III below shows the average value of mean absolute error of all classification techniques.

TABLE II
AVERAGE VALUE OF MEAN ABSOLUTE ERROR FOR CLASSIFIERS

Classification Technique	Average value of mean absolute error
Naïve Bayes	0.1673
Neural Networks	0.19002
Support Vector Machine	0.14296
Nearest Neighbor	0.19284

VI. DISCUSSION

Above figure and tables show the performance of classification techniques in terms of mean absolute error. In Figure 3 and Table II and III, we can see that average of the difference between predicted and actual value in all test cases; i.e. average prediction error or mean absolute error for all these classification techniques is less, but for support vector machines, the value of mean absolute error is lower than other techniques for all datasets.

The value of mean absolute error for other techniques is more than Support Vector Machines and nearest neighbors has highest value of mean absolute error.

VII. CONCLUSION

As a conclusion, we have met our objective which is to evaluate and investigate the performance of four selected classification algorithms using WEKA. The best algorithm on the basis of mean absolute error is support vector machine. On the other hand, nearest neighbor classifier has highest value of mean absolute error among these techniques. These results suggest that among the machine learning algorithm tested, Support Vector Machines has the potential to significantly improve the performance of conventional classification methods for data sets of these types if lower value of average of the difference between predicted and actual value in all test cases; i.e. average prediction error is required.

REFERENCES

Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, AI MAGAZINE, 37-54.

Phyu, Thair Nu (2009), "Survey of Classification Techniques in Data Mining," *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009*, Hong Kong.

Viera, A. J., and Garrett, J. M. (2005), "Understanding Interobserver Agreement: The Kappa Statistic", *Research Series Vol. 37, No. 5*, pp. 360-363.

Gupta, S., Kumar, D., and Sharma, A. (2011), "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis", Indian Journal of Computer Science and Engineering(IJCSE), Vol 2, pp, 188-195.

Bishnoi, Savika (2011), "Comparison of classification techniques," *IJRIM Volume 1, Issue 2*, pp. 107-119.

Othman & Yau (2007), "Comparison of Different Classification Techniques Using WEKA for Breast Cancer," *IFMBE Proceedings 15*, pp. 520-523.

Li, Hongqi, Guo, Haifeng, Guo, Haimin, and Meng, Zhaaoxu (2008), *"Data Mining Techniques for Complex Formation Evaluation in Petroleum Exploration and Production: A Comparison of Feature Selection and Classification Methods,"* Pacific-Asia Workshop on Computational Intelligence and Industrial Application, IEEE, pp. 37-43.

Burges, C (1998), *"A Tutorial on Support Vector Machines for Pattern Recognition"*. Data Mining and Knowledge Discovery, Vol. 2, pp. 121-167.

Thakkar, M., Bhatt, M., and Bhensdadia, C. K. (2011), "Performance Evaluation of Classification Techniques for Computer Vision based Cashew Grading System," *International Journal of Computer Applications (0975 – 8887), Volume 18– No.6*, pp. 9-12.

Darrell, Indyk, P., and Shakhnarovich, G. (2006), "*Nearest Neighbour Methods in Learning and Vision: Theory and Practice,"* MIT Press.

Pushpa (2010), "Copmarision of Clustering Techniques using WEKA," M. Eng. thesis, Guru Jambheshwar University of Science and Technology, Hisar, India.

Justin, T., Gajsek, R., Struc, V. and Dobrisek, S. (2010) "Comparison of Different Classification Methods for Emotion Recognition," MIPRO 2010, Opatija, Croatia, pp. 700-703.

Kirkby & Frank (2004), "WEKA Explorer User Guide for Version 3-4-3", University of Waikato.

Soman, K. P., Diwakar, S., and Ajay, V. (2008), *Insight into Data Mining: Theory and Practice*, PHI, Delhi, India.