EVALUATION OF VARIOUS CLASSIFICATION TECHNIQUES OF WEKA USING DIFFERENT DATASETS

Ramesh Prasad Aharwal

Asstt. Prof., Department of Mathematics

Govt.P.G.College, Damoh (M.P.)India

ABSTRACT:

In this paper we have compared various classification methods using UCI machine learning dataset under WEKA. We have used three measuring factors which names are Accuracy, kappa statistics and mean absolute error for execution by each technique is observed during experiment. This work has been carried out to make a performance evolution of J48, Multilayerperceptron, Naïve Bayes and SMO classifier. On Account of this work we have used four type of secondary data.

INTRODUCTION

Knowledge Discovery and Data Mining are rapidly evolving areas of research that are at intersection of several disciplines, including statistics, databases, AI and visualization. KDD refers to the whole process of discovering useful knowledge from data, and data mining refers to a particular step in this process [2]. The aim of this study is to investigate and evolutes the performance of different classification methods using WEKA tools on machine learning dataset. Machine learning covers such a wide range of processes that it is difficult to define precisely. There is a constant stream of new events in the world and continuing redesign of Artificial Intelligent systems to conform to new knowledge is impractical but machine learning methods might be able to track much of it. The evolution and comparisons of classifiers is very useful for selecting the best classifier.

DATA MINING

Data mining is the term used to describe the process of extracting value from a database. According to Moxon "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large volumes of data, using pattern recognition technologies as well as statistical and mathematical techniques. It is a "knowledge discovery process of extracting previously unknown, actionable information from very large databases." It can also be defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [2]. While data mining and knowledge Discovery in Database (KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

DATA MINING TECHNIQUES

There are mainly two types of data mining techniques. Figure 1 shows the types of data mining techniques. In this paper we have used classification techniques.

CLASSIFICATION

Classification is most likely the most frequently used data mining technique. It is the process of finding a set of models that describe and differentiate data classes and concepts, for the purpose of being able to use the model to predict the class whose label is unknown. There are many algorithms that can be used for classification, such as decision trees, neural networks, Naïve Bayes, SMO etc. In this work we are using four classifiers for their evolution.

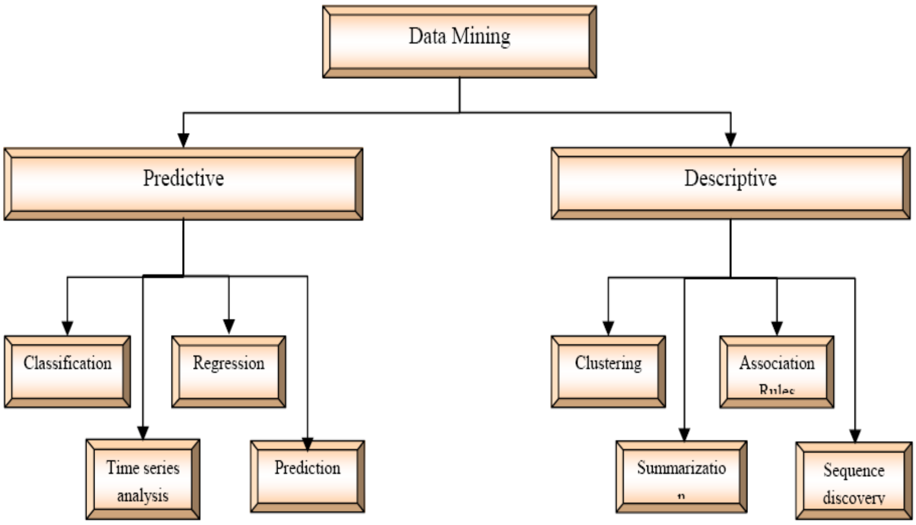


Fig.1 Data mining models and tasks. [Source: Margaret. H.Dunham, 2004 (6)]

WEKA Classifier

In this paper we have evolutes four weka classifiers under secondary dataset. The brief description of each classifier is given in this section.

J48

J48 is an implementation of C4.5 in WEKA. C4.5 uses information entropy concept [11]. The J48 algorithm is WEKA's implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning [11]. It is the classifier according to which we classify our classes it is also known as free classifier who accepts nominal classes only. In this prior knowledge should be there while classifying instances. It is used in the construction of decision tree from a set of labeled training data using the information entropy. Attributes which we use helps in building decision tree by splitting it into subset and normalization information gained can be calculated. Splitting process comes to an end when all instances in a subset belong to the same class. Leaf node is also is also present or being created to choose that class a possibility also can be there that none of the feature provides information gain.J48 creates decision nodes up higher in the tree using expected value of the class.J48 can use both discrete and continuous attributes, attributes with differencing lost and training data with missing attribute values.

Multilayer Perceptron

Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a back propagation neural network with one or more layers between input and output layer. The following diagram illustrates a perceptron network with three layers. Artificial Neural Networks (ANN) are one of the common classification methods in data mining. To employ Neural Network based classifiers, Multi Layer Perceptron (MLP) were used in this work.

SMO (Sequential Minimal Optimization)

SMO is a new algorithm for training Support Vector Machines (SVMs). Training a support vector machine requires the solution of a very large quadratic programming optimization(QP) problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets. Because matrix computation is avoided, SMO scales somewhere between linear and quadratic in the training set size for various test problems, while the standard chunking SVM algorithm scales somewhere between linear and cubic in the training set size. SMO's computation time is dominated by SVM evaluation; hence SMO is fastest for linear SVMs and sparse data sets [8]. Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all missing the values and transforms nominal attributes into binary ones. A single hidden layer neural network uses exactly the same form of model as an SVM.

NAÏVE BAYS

Naïve Bayes classifier has relatively simple interface in WEKA. It allows one to select the kernel estimator for numeric attributes rather than a normal distribution and used Supervised Discretization while converting numeric attributes to normal ones. The output of Naïve Bayes classifier has text form. The Naïve Bayes is a simple probabilistic classifier. It is based on an supposition about mutual independency of attributes. Typically this supposition is far away from being true and this is the reason for the naivety of the method. The probabilities applied in the Naïve Bayes algorithm are calculated. According to the Bayes' Rule: The probability of hypothesis H can be calculated on the basis of the hypothesis H and evidence about the hypothesis E according to the following formula:

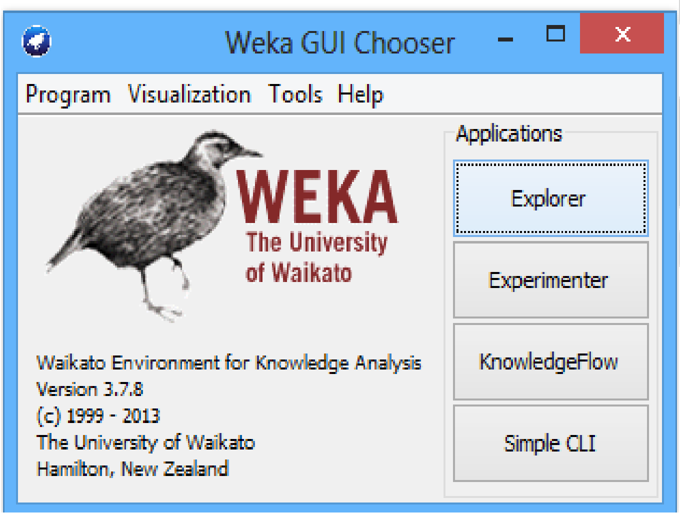
$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

METHODOLOGY

We have used the popular, open-source data mining tool Weka (version 3.6.6) for this analysis. Four different data sets have been used and the performance of a comprehensive set of classification algorithms (classifiers) has been analyzed.

WEKA

WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, NewZealand. Weka is open source software which consists of a collection of machine learning algorithms for data mining tasks [14]. Weka is a milestone in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption. Weka is a bird name of Newzealand. WEKA is a modern feature for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data pre processing tools to researchers [9]



Waikato Environment for Knowledge Analysis
Version 3.7.8
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Fig.1. WEKA GUI

Table 1 Description of dataset

Name of dataset	Number of instances
Breast-W	699
diabetes	768
Heart-statlog	155
hepatitis	270

Experimental work and result

In this experiment we have used four types of medical dataset. Dataset have extract from UCI machine learning data repository. The details of each datasets are shown in Table 1. After experimental work we have observed that the classification accuracy of SMO is higher than other three classifiers. Accuracy of each classifier is shown in table 2 and figure 2. Figure 3 shows the mean absolute error of each classifier. We found that the mean absolute error of SMO classifier is less comparatively to other three classifiers.

Table 2 Accuracy of correctly classification in %

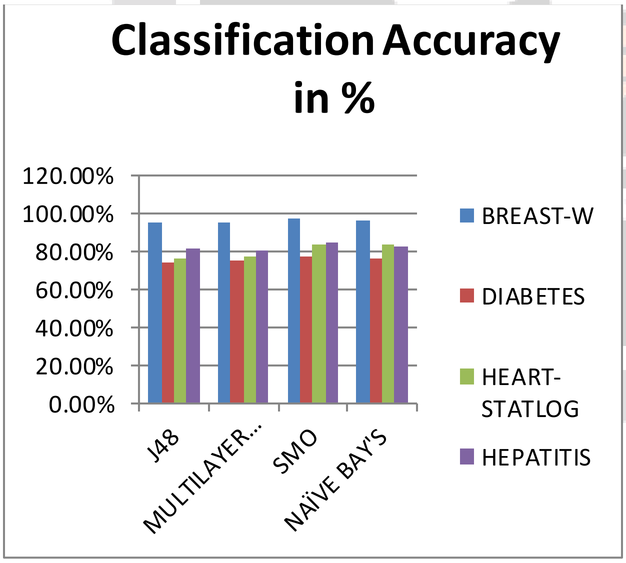
CLASSIFIER/DATASETS	**BREAST-W**	**DIABETES**	**HEART-STATLOG**	**HEPATITIS**
J48	95.1359 %	74.2188 %	76.6667 %	81.2903 %
MULTILAYER PERCEPTRON	95.279 %	75.1302 %	77.4074 %	80 %
SMO	**96.9957 %**	**77.474 %**	**84.0741 %**	**85.1613 %**
NAÏVE BAYES	95.9943 %	76.3021 %	83.7037 %	82.5806 %

Table 3 Mean Absolute Error of tested classifier under four medical dataset

CLASSIFIER/DATASETS	**BREAST-W**	**DIABETES**	**HEART-STATLOG**	**HEPATITIS**
J48	0.0637	0.3134	0.274	0.2073
Multilayer perceptron	0.0497	0.294	0.2328	0.1928
SMO	**0.03**	**0.2253**	**0.1593**	**0.1484**
Naivebayes	0.0407	0.2841	0.1835	0.1735

Table 4 kappa statistics of tested classifier under four medical dataset

CLASSIFIER/DATASETS	**BREAST-W**	**DIABETES**	**HEART-STATLOG**	**HEPATITIS**
J48	0.893	0.4246	0.5271	0.394
Multilayer perceptron	0.8956	0.4445	0.4438	0.3825
SMO	**0.9337**	**0.4708**	**0.6762**	**0.5309**
Naivebayes	0.9127	0.4664	0.6683	0.5082



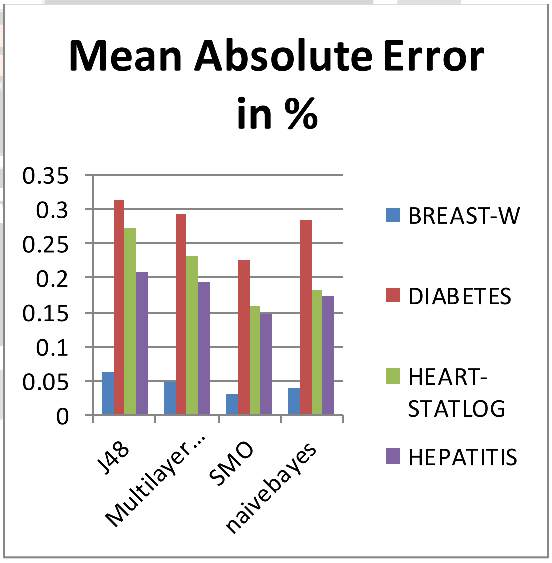


Fig.2 Graphical representation of Classification accuracy of tested classifier

Fig.3 Graphical representation of Mean absolute error of tested classifier

Conclusion

In this study four different classification algorithms under WEKA compared with using UCI data set. Then preprocessed datasets, used to test the four classifiers using 10-folds cross validation. Three different performance measures considered for classifiers. Results of comparison showed that SMO classifier achieved the highest value in accuracy and lowest value of mean absolute error measures which we can see that on table 2 and 3.

Reference

[1] D. A. Avellaneda, et al., "Natural Texture Classification: A Neural Network Models Benchmark," 2009 ,pp. 325-329.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smith. 1996. From Data Mining to Knowledge Discovery: An Overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smith and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pages 1-34. MIT Press, Cambridge, MA

[3]. Gopala Krishna Murthy Nookala, Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013,pp

[4] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann, 2nd , 2006

[5] Meenakshi, Geetika Survey on Classification Methods using WEKA, International Journal of Computer Applications (0975 – 8887), Volume 86 – No 18, January 2014,

[6] Margaret.H.Dunham (2004). Basic Data Mining Tasks. Singapore, Pearson Education.

[7] Olalekan S. Akinola et al, evaluating classification effectiveness of sequential minimal optimization(SMO) algorithm on chemical parametization of granitoids,ijrras 13 (2) November 2012 pp 636-644

[8] Platt, J. (1998), Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,

[9] Purva Sewaiwar, Kamal Kant Verma, Comparative Study of Various Decision Tree Classification Algorithm Using WEKA, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10), 2015 pp 87-91

[10] Rosenblatt and X. Frank. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961.

[11] J. R. Quinlan “C4.5: programs for machine learning” Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[12] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.html

[13] V. Vaithiyanathan1, K. Rajeswari2, Kapil Tajane3, Rahul Pitale3,Comparison Of Different Classification Techniques Using Different Datasets International Journal of Advances in Engineering & Technology, May 2013.
©IJAET ISSN: 2231-1963

[14] Weka: http://www.cs.waikato.ac.nz/~ml/weka/