Comparison of Different Classification Techniques Using WEKA for Breast Cancer

Mohd Fauzi bin Othman,Thomas Moh Shan Yau

Control and Instrumentation Department, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, Malaysia

Abstract— **The development of data-mining applications such as classification and clustering has shown the need for machine learning algorithms to be applied to large scale data. In this paper we present the comparison of different classification techniques using Waikato Environment for Knowledge Analysis or in short, WEKA. WEKA is an open source software which consists of a collection of machine learning algorithms for data mining tasks. The aim of this paper is to investigate the performance of different classification or clustering methods for a set of large data. The algorithm or methods tested are Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm. A fundamental review on the selected technique is presented for introduction purposes. The data breast cancer data with a total data of 6291 and a dimension of 699 rows and 9 columns will be used to test and justify the differences between the classification methods or algorithms. Subsequently, the classification technique that has the potential to significantly improve the common or conventional methods will be suggested for use in large scale data, bioinformatics or other general applications.**

Keywords— **Machine Learning, Data Mining, WEKA, Classification, Bioinformatics.**

I. Introduction

The aim of our work is to investigate the performance of different classification methods using WEKA for breast cancer. A major problem in bioinformatics analysis or medical science is in attaining the correct diagnosis of certain important information. For the ultimate diagnosis, normally, many tests generally involve the clustering or classification of large scale data. All of these test procedures are said to be necessary in order to reach the ultimate diagnosis. However, on the other hand, too many tests could complicate the main diagnosis process and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers.

Machine learning covers such a broad range of processes that it is difficult to define precisely. A dictionary definition includes phrases such as to gain knowledge or understanding of or skill by studying the instruction or experience and modification of a behavioral tendency by experienced zoologists and psychologists study learning in animals and humans [1]. The extraction of important information from a large pile of data and its correlations is often the advantage of using machine learning.New knowledge about tasks is constantly being discovered by humans and vocabulary changes. 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 [1].

There is a substantial amount of research with machine learning algorithm such as Bayes Network, Radial Basis Function, Decision tree and pruning, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm.

II. Methods

A. Bayes Network Classifier

Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute A_i given the class label C [2,3]. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A_1…..A_n 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 [4]. In particular, the naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent [3,4].

B. Radial Basis Function

Radial basis function (RBF) networks have a static Gaussian function as the nonlinearity for the hidden layer processing elements. The Gaussian function responds only to a small region of the input space where the Gaussian is centered [5]. The key to a successful implementation of these networks is to find suitable centers for the Gaussian functions [6,7]. The simulation starts with the training of an

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 520-523, 2007

unsupervised layer. Its function is to derive the Gaussian centers and the widths from the input data. These centers are encoded within the weights of the unsupervised layer using competitive learning [7]. During the unsupervised learning, the widths of the Gaussians are computed based on the centers of their neighbors. The output of this layer is derived from the input data weighted by a Gaussian mixture. The advantage of the radial basis function network is that it finds the input to output map using local approximators. Usually the supervised segment is simply a linear combination of the approximators. Since linear combiners have few weights, these networks train extremely fast and require fewer training samples.

C. Decision Tree and Pruning

A decision tree partitions the input space of a data set into mutually exclusive regions, each of which is assigned a label, a value or an action to characterize its data points. The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made [8]. A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is associated with a label or value that characterizes the given data that leads to its being visited. However, many decision tree construction algorithms involve a two - step process. First, a very large decision tree is grown. Then, to reduce large size and overfiting the data, in the second step, the given tree is pruned [9]. The pruned decision tree that is used for classification purposes is called the classification tree.

D. Single Conjunctive Rule Learner

Single conjunctive rule learner is one of the machine learning algorithms and is normally known as inductive learning. The goal of rule induction is generally to induce a set of rules from data that captures all generalizable knowledge within that data, and at the same time being as small as possible [10]. Classification in rule-induction classifiers is typically based on the firing of a rule on a test instance, triggered by matching feature values at the left-hand side of the rule [11]. Rules can be of various normal forms, and are typically ordered; with ordered rules, the first rule that fires determines the classification outcome and halts the classification process.

E. Nearest Neighbors Algorithm

Nearest neighbors 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 [12]. The distance metric used in nearest neighbor methods for numerical attributes can be simple Euclidean distance.

F. The Data

The data used in this investigation is the breast cancer data. It has a total of 6291 data and a dimension of 699 rows and 9 columns. For the purposes of training and testing, only 75% of the overall data is used for training and the rest is used for testing the accuracy of the classification of the selected classification methods.

III. WEKA

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art facility 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 algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is an open source software issued under General Public License [13].

The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file (foremost: attribute names, attribute types, attribute values and the data). 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 a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis.

IV. RESULT

To gauge and investigate the performance on the selected classification methods or algorithms namely Bayes Network

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 520-523, 2007
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007

Classifier, Radial Basis Function, Decision Tree with pruning, Single Conjunctive Rule Learner and Nearest Neighbor, we use the same experiment procedure as suggested by WEKA. The 75% data is used for training and the remaining is for testing purposes.

In WEKA, all data is considered as instances and features in the data are known as attributes. The simulation results are partitioned into several sub items for easier analysis and evaluation. On the first part, correctly and incorrectly classified instances will be partitioned in numeric and percentage value and subsequently Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results of the simulation are shown in Tables 1 and 2 below. Table 1 mainly summarizes the result based on accuracy and time taken for each simulation. Meanwhile, Table 2 shows the result based on error during the simulation. Figures 1 and 2 are the graphical representations of the simulation result.

Table 1 Simulation result of each algorithm.

Algorithm (Total Instances, 175)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (Value)	Time Taken (seconds)	Kappa Statistic
Bayes Net.	89.7143 (157)	10.2857 (18)	0.19	0.7858
Radial Basis Function	87.4286 (153)	12.5710 (22)	0.53	0.7404
Decision Tree and Pruning	85.7143 (150)	14.2857 (25)	0.23	0.7019
Single Conj. Rule Learner	85.1429 (149)	14.8571 (26)	0.15	0.6893
Nearest Neighbors	84.5714 (148)	15.4286 (27)	0.81	0.6860

Table 2 Training and simulation errors

Algorithm (Total Instances, 175)	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Bayes Network	0.1062	0.3217	22.2878	65.1135
Radial Basis Function	0.1999	0.3162	41.9593	63.9903
Decision Tree and Pruning	0.1871	0.3635	39.2681	73.5759
Single Conj. Rule Learner	0.2449	0.3559	51.4069	72.0207
Nearest Neighbors	0.1543	0.3928	32.3840	79.4963

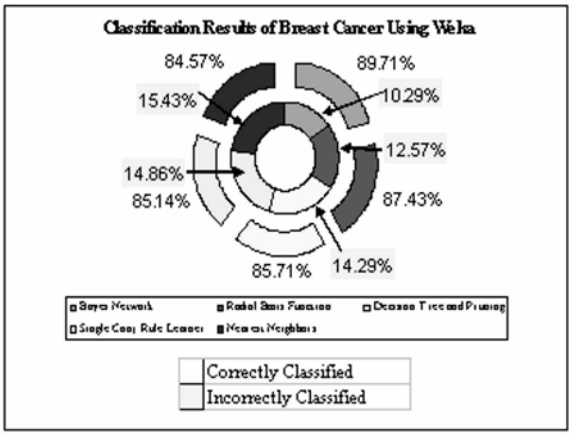

Fig. 1 Results

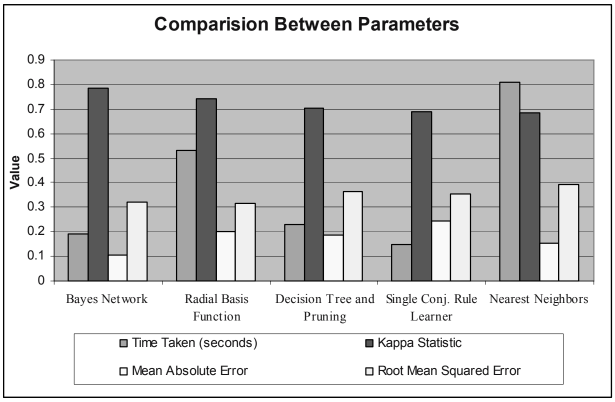

Fig. 2 Comparison between parameters

V. Disscussions

Based on the above Figures 1, 2 and Table 1, we can clearly see that the highest accuracy is 89.71% and the lowest is 84.57%. The other algorithm yields an average accuracy of around 85%. In fact, the highest accuracy belongs to the Bayes network classifier, followed by Radial basis function with a percentage of 87.43% and subsequently decision tree with pruning and single conjunctive rule learner. Nearest neighbor bottom the chart with percentage around 84%. An average of 151 instances out of total 175 instances is found to be correctly classified with highest score of 157 instances compared to 148 instances, which is the lowest score. The total time required to build the model is also a crucial parameter in comparing the classification algorithm. In this simple experiment, from Figure 2, we can say that a single conjunctive rule learner requires the shortest time which is around 0.15 seconds compared to the others. Nearest neighbor algorithm requires the longest model building time which is around 0.81 seconds. The second on the list is Bayes network with 0.19 seconds.

Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity [14]. The average Kappa score from the selected algorithm is around 0.6-0.7. Based on the Kappa Statistic criteria, the accuracy of this classification purposes is substantial [14]. From Figure 2, we can observe the differences of errors resultant from the training of the five selected algorithms. This experiment implies a very commonly used indicator which are mean of absolute errors and root mean squared errors. Alternatively, the relative errors are also used. Since, we have two readings on the errors, taking the average value will be wise. It is discovered that the highest error is found in single rule conjunctive rule learner with an average score of around 0.3 where the rest of the algorithm ranging averagely around 0.2-0.28. An algorithm which has a lower error rate will be preferred as it has more powerful classification capability and ability in terms of medical and bioinformatics fields.

VI. Conclusions

As a conclusion, we have met our objective which is to evaluate and investigate five selected classification algorithms based on Weka. The best algorithm based on the breast cancer data is Bayes network classifier with an accuracy of 89.71% and the total time taken to build the model is at 0.19 seconds. Bayes network classifier has the lowest average error at 0.2140 compared to others. These results suggest that among the machine learning algorithm tested, Bayes network classifier has the potential to significantly improve the conventional classification methods for use in medical or in general, bioinformatics field.

Acknowledgment

We would like to thank Associate Prof. Dr Naomi from the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia for the data. This work is using free software from University of Waikato in New Zealand

References

1. Nils J. Nilsson (1999) Introduction to Machine Learning. California. United Stated of Americas.
2. Bouckaert, R.R. (1994). Properties of Bayesian network Learning Algorithms. In R. Lopex De Mantaras & D. Poole (Eds.), I*n Press of Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 102-109). San Francisco, CA.
3. Buntine, W. (1991). Theory refinement on Bayesian networks. In B. D. D'Ambrosio, P. Smets, & P.P. Bonissone (Eds.), *In Press of Proceedings of the Seventh Annual Conference on Uncertainty Artificial Intelligent* (pp. 52-60). San Francisco, CA
4. Daniel Grossman and Pedro Domingos (2004). Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood. *In Press of Proceedings of the 21st International Conference on Machine Learning,* Banff, Canada.
5. M. D. Buhmann (2003), Radial Basis Functions: Theory and Implementations, 12. Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge.
6. S. V. Chakravarthy and J. Ghosh (1994), Scale Based Clustering using Radial Basis Function Networks, *In Press of Proceeding of IEEE International Conference on Neural Networks*, Orlando, Florida. pp. 897-902.
7. Howell, A.J. and Buxton, H. (2002). RBF Network Methods for Face Detection and Attentional Frames, *Neural Processing Letters* (15), pp.197-211
8. J.S R Jang (1993). ANFIS Adaptive Network Based Fuzzy inference System. IEEE Transaction on Systems, Man and Cybernetics. Vol. 23, no3, pp 665-685
9. Mansour Y (1997). Pessimistic decision tree pruning based on tree size. *In Press of Proc. 14th International Conference on Machine Learning*. Pp.195-201.
10. Cohen, W. (1995) Fast effective rule induction. *In Press of Proceedings 12th International Conference on Machine Learning*, Morgan Kaufmann. Pp. 115–123.
11. Clark, P., Niblett, T. (1989). The CN2 rule induction algorithm. Machine Learning 3. pp. 261–284
12. T. Darrell and P. Indyk and G. Shakhnarovich (2006). Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press.
13. WEKA at http://www.cs.waikato.ac.nz/~ml/weka.
14. Kappa at http://www.dmi.columbia.edu/homepages/chuangj/kappa

Address of the corresponding author:
Author: Dr Mohd Fauzi Othman
Institute: Fakulti Kejuruteraan Elektrik
Universiti Teknologi Malaysia
Street: 81300 UTM Skudai
City: Johor Bahru
Country: Malaysia
Email: fauzi@fke.utm.my

F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 520-523, 2007
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007