Hybrid Intelligent System of Heterogeneous Classifiers for Breast Cancer Diagnosis

A. Sorna Gowri

Assistant Professor, The M.D.T Hindu College, Tirunelveli

gowrimurugan74@yahoo.co.in

Dr. K. Ramar

Principal, Einstein College of Engineering, Tirunelveli

Kramar.einstein@gmail.com

Abstract

Effective machine learning tools can assist in early detection of diseases such as breast cancer. Investigating novel approaches to diagnose breast cancer based on machine learning tools and involves development of new techniques to construct and process missing features values, investigate different feature selection methods and how to employ them into diagnosis process.Thus,the fusion of heterogeneous classifiers are used to get the classification accuracy of the medical image dataset.

Keywords: K-NN,Naïve bayes,Random Tree

Introduction

.The new technologies produced a huge variety of data which are over human capability of analysis. Therefore, the intelligent systems found to help human arrange and utilize huge data effectively. For the sake of obtaining full benefits of the use of intelligent systems, it is became a common in machine learning to look for intelligent systems advantages and mix and match them to produce a new approach that maximize the advantages and minimize the overhead. *Fusion Intelligent systems* or Hybrid intelligent systems, in which two or more machine leaning algorithms are combined in one new approach are often effective and can overcome the limitations of individual approaches. Classification is one branch of machine learning and the process of integrating two or more classifiers is usually referred to as multi-classifications or *fusion classification*. There are two main paradigms in combining different classification algorithms: Classification Selection and classification Fusion. Classification Selection paradigms uses a single model to predicate the new case. However, fusion classification merges two or more outputs of all models to produce a single output. The paper will introduce different types of fusion classification on three well-known classifiers on breast cancer datasets of integrating two

or more classifiers enhanced the classification accuracy in some cases. However, there is no single combination that suits all datasets.

Multi-Classification Approach

The process of combining two or more classifiers is called multi-classification approach. The purpose of multi-classification is based on the argument that no single classifier that suites all learning problems [1]. Multi-classification can be divided into two types, classifier selection and fusion classifier.

Classifier Selection

Classifier selection is one of the simplest methods for combining learning algorithms or classifiers. The idea is to evaluate two or more classifiers on the training dataset and then make use of the best performed classifiers on the testing dataset. This method is simple, straight forward, no output combination, and performs well in compare to more complex classifiers [2].

Fusion Classifier

Fusion classifier is a group of classifiers whose individual predictions are combined in some way to classify new cases (highest average ranking, average probability, or voting). It is became one of the active areas of research in supervised learning that study new ways of constructing classifiers for more accurate outcome. Voting is the simplest method for multi-classification in heterogeneous and homogeneous models. The voting method is divided into two types, weighted voting and un-weighted voting. In un-weighted voting, all classifiers are treated equally with no priority over other classifiers. Therefore, each classifier outputs a class value and the class with the most votes is the final outcome of the multi-classifier. Note that this type of voting is in fact called Plurality Voting, in contrast to the frequently used term Majority Voting, as the majority voting implies that at least 50%+1 (the majority) of the votes should belong to the winning class. In weighted voting, classifiers may have different weight according to the user believe in the classifier performance and classification accuracy. For example, the user can put more weight in *k*-NN (60%) over the Naïve Bayes (40%) in a certain multi-classifiers problem. The weighted voting is usually discriminating between classifiers based on classifier reputation and ranking among others classifiers [2].

Classifiers Combination Strategies

Machine learning and data mining are rich of classification tools and algorithms. In the context of combining classification there is uncertainty which classifier works best with other classifiers. Kuncheva and Whiteaker [3] stated that the best combination of a set of classifiers depends on the application and on the classifiers characteristics. However, there is no best combination of classifiers. However, one approach is to generate a large number of classifiers and then to select the best combinations to use based on the classification accuracy or other criteria set by the researcher. On the other hand, this approach is costly and time consuming since n classifiers may be combined in 2n combinations, and it is difficult to obtain this amount of experiments except in simple or restricted circumstances. To solve this, several search techniques have been used to find the best combination of classifiers, including forward and backward search, Tabu search, and genetic algorithms.In general, more powerful technique to find the best possible combination of classifiers is needed.

Experimental Methodology

Different sets of experiments were performed to evaluate the multi-classification approach on well-known publicly available breast cancer datasets from UCI machine learning repository. Recently, this research found that there are three versions of breast cancer diagnosis and will be introduced in this chapter. The datasets are Wisconsin Breast Cancer (Original), Wisconsin Diagnosis Breast Cancer (WDBC), and Wisconsin Prognosis Breast Cancer (WPBC). Table 1shows the statistical details of the used datasets. Table 1 shows that the three datasets are different of sizes. The smallest dataset contains 11 attributes and the largest dataset contains 34 attributers. Number of instances also ranges from 198 to 699 instances while all the datasets contain 2 classes. The study has considered a heterogeneous classifier from three machine learning categories. The first algorithm is *k*-nearest neighbours (*k*-NN) from lazy learning category. *k*-NN is an instance-based classifier where the class of a test instance is based upon the class of those training instances alike to it. Distance functions are common to find the similarity between instances. Examples of distance functions are Euclidean and Manhattan distance functions. The second algorithm is Naïve Bayes classifier (NB) form Bayes category. NB is a simple probabilistic classifier based on applying Bayes' theorem. NB is one of the most efficient and effective learning algorithms for machine learning and data mining because the condition of independency (no attributes depend on each other) [5]. The last machine learning algorithm is Random Tree (RT) from the tree classification category. RT is used to classify an instance to a predefined set of classes based on their attributes values. RT is frequently used in many fields such as engineering, marketing, and medicine. The study used k-fold cross validation technique to separate the training set from test set with k=10. The environment of experiment is the well-known machine learning software, WEKA.

Table 1: Statistics of Breast Cancer Datasets

Dataset	**Number of Attributes**	**Number of Instances**	**Number of Classes**
Wisconsin Breast Cancer (Original)	11	699	2
Wisconsin Diagnosis Breast Cancer(WDBC)	32	569	2
Wisconsin Prognosis Breast Cancer(WPBC)	34	198	2

In this experiment, this work used the confusion matrix to measure classifiers performance. The classification accuracy is the main criteria to estimate the effectiveness of the classification model based on the number of correct and incorrect classification cases.

Experimental Results

Three experiments were performed on three different datasets of breast cancer data. The first experiments were performed on single classifier model to set a base line of classification accuracy and how it can be enhanced. The second experiment was performed using a combination of two classifiers while the last experiment was performed on fusion of three classifiers.

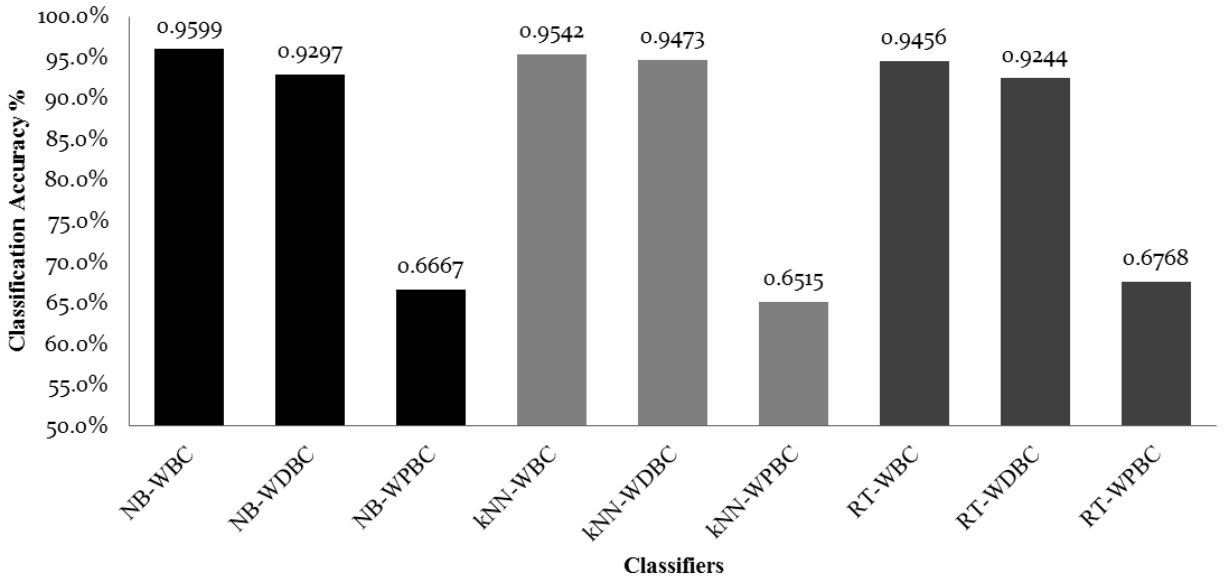

Figure 1 shows the results of performing three single classifiers on three datasets. It shows that Naïve Bayes performed best in regards to classification accuracy on WBC (0.9599) while *k*-NN and Random Tree performed just better on WDBC and WPBC respectively.

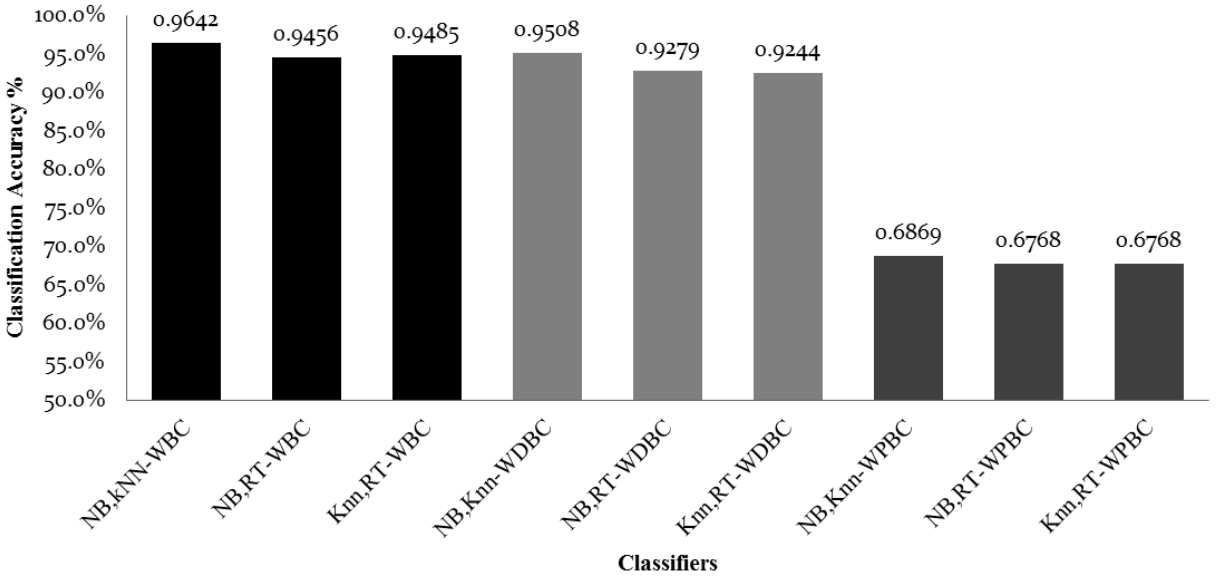

Figure2: Two Classifiers on three datasets WBC, WDBC, and WPBC.

Figure 2 shows the result of combining two classifiers (Naïve Bayes and *k*-NN, Naïve Bayes and Random Tree, and *k*-NN and Random Tree). The results indicate that the fusion between Naïve Bayes and *k*-NN produced the best classification accuracy (0.9642 on WBC, 0.9508 on WDBC, and 0.6869 on WPBC). This may draw an intention that Naïve Bayes and *k*-NN may produce better results when they combined together.

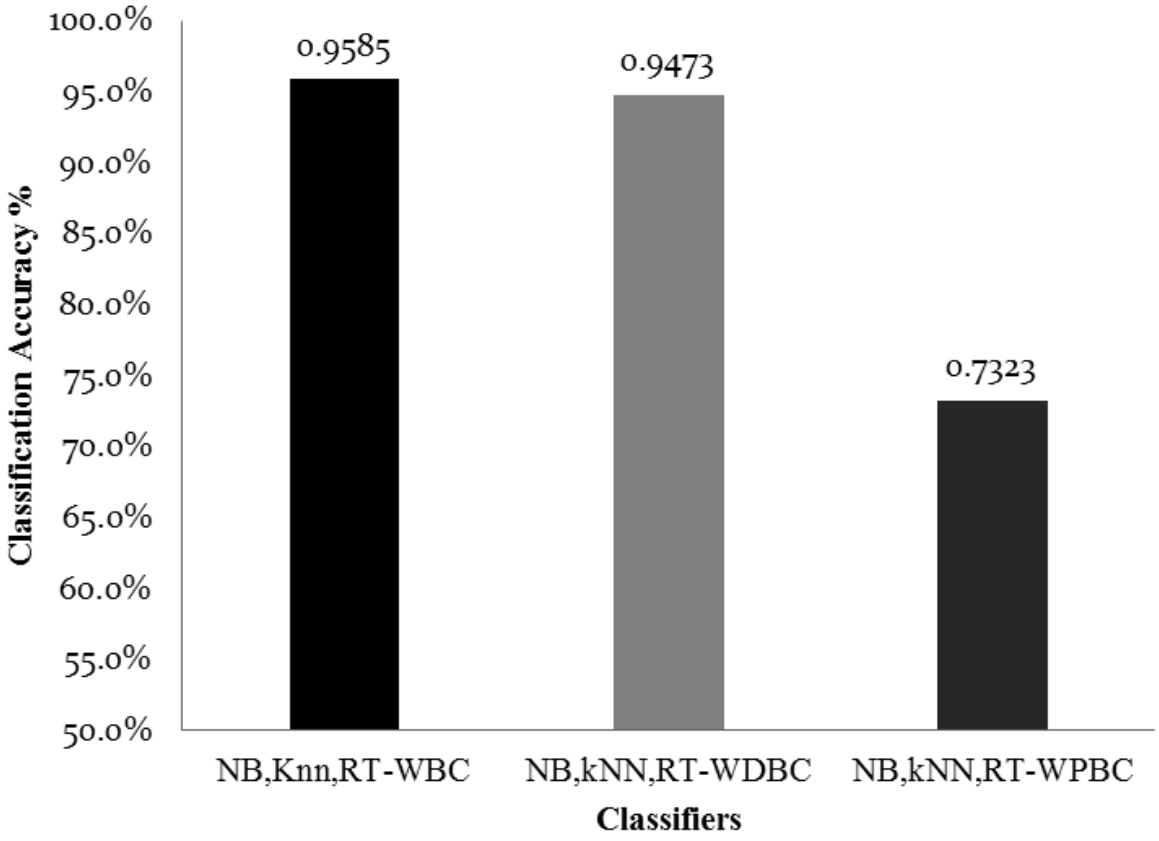

Figure 3: The Fusion of three classifiers on three datasets WBC, WDBC, and WPBC.

Figure 3 shows the result of fusion for three classifiers (Naïve Bayes, *k*-NN, and Random Tree). It shows that combining the three classifiers in three datasets still maintained satisfactory classification accuracy on WBC (0.9585) and WDBC (0.9473) while a significant improvement in classification accuracy on WPBC dataset (0.7323).

Conclusion:

Classification fusion on three well-known machine learning classifiers on breast cancer dataset. This work can confirm the argument that the best combination of a set of classifiers depends on the application and on the classifiers characteristics. In addition, there is no best combination of classifiers that suites all datasets. However in the current experiments, Naïve Bayes and *k*-NN produced better

results when they combined as one classifier with maximum classification accuracy obtained on WBC dataset (0.9642).

References:

1. Tsoumakas, G., L. Angelis, and I. Vlahavas, Selective fusion of heterogeneous classifiers. *Intelligent Data Analysis*, 2005. 9(6): p. 511-525.

2. Džeroski, S. and B. Ženko, Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 2004. 54(3): p. 255-273.

3. Kuncheva, L. and C. Whitaker, Feature Subsets for Classifier Combination: An Enumerative Experiment, in Multiple Classifier Systems, J. Kittler and F. Roli, (Editors). 2001, *Springer Berlin Heidelberg*. p. 228-237.

4. Grabusts, P., The Choice of Metrics for Clustering Algorithms, *in Proceedings of the 8th International Scientific and Practical Conference. Volume I1*. 2011. p. 70-76.

5. Jurman, G., Riccadonna, S., Visintainer, R., & Furlanello, C., Canberra distance on ranked lists. *In Proceedings, Advances in Ranking–NIPS 09 Workshop* , 2009, p. 22-27.