An Efficient Ensemble Based Classification Techniques for Medical Diagnosis

G. Ravi Kumar[1] , Venkata Sheshanna Kongara[2] and Dr. G.A Ramachandra[3]

1. Research Scholar, Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapur-515 003, Andhra Pradesh, India
grkondaravi@gmail.com

2. Research Scholar, Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapur-515 003, Andhra Pradesh, India
kv.sheshu@gmail.com

3. Associate Professor , Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapur-515 003, Andhra Pradesh, India

ABSTRACT

Building accurate and efficient classifiers for Medical databases is one of the essential tasks of data mining and machine learning research. Building effective classification systems is one of the central tasks of data mining. One of the most active areas of research in supervised machine learning has been to study methods for constructing good ensembles of learners. This paper aims to establish an accurate ensemble classification model for Medical prediction, in order to make full use of the invaluable information in clinical data, especially which is usually ignored by most of the existing methods when they aim for high prediction accuracies. This paper presents a comparison among the different ensemble classifiers on the database of Wisconsin Breast Cancer (WBC) and Diabetes data sets. In this experiment, we compare ensemble classification techniques in Weka software and comparison results show that Random forest has higher prediction accuracy than those methods. Different methods for breast cancer detection are explored and their accuracies are compared with these results, we infer that the Random forest are more suitable in handling the ensemble classification problem of Medical prediction, and we recommend the use of these approaches in similar ensemble classification problems.

Keywords: ***Classification, Ensemble, Decision tree, Bagging, Boosting, AdaBoost, Random forest and Weka***

I. Introduction

Classification is one of the most studied problems in machine learning and data mining [10], [18]. Building accurate and efficient classifiers for Medical databases is one of the essential tasks of data mining and machine learning research. Building effective classification systems is one of the central tasks of data mining.

A supervised machine learning task involves constructing a mapping from input data (normally described by several features) to the appropriate outputs. In a classification learning task, each output is one or more classes to which the input belongs. The goal of classification learning is to develop a model that separates the data into the different classes, with the aim of classifying new examples in the future.

Ensemble learning methods instead generate multiple models. Given a new example, the ensemble passes it to each of its multiple base models, obtains their predictions, and then combines them in some appropriate manner (e.g., averaging or voting). The majority of ensemble learning methods are generic, applicable across broad classes of model types and learning tasks. Ensemble learning is an effective technique that has increasingly been adopted to combine multiple learning algorithms to improve overall prediction accuracy [15].

One of the most active areas of research in supervised machine learning has been to study methods for constructing good ensembles of learners. The main discovery is that ensembles are often much more accurate than the individual learners that make them up [2]. When designing an ensemble learning method, in addition to choosing the method by which to bring about diversity in the base models and choosing the combining method, one has to choose the type of base model and base model learning algorithm to use. The combining method may restrict the types of base models that can be used.

The rest of the paper is organized as follows. In section II, we review the related work for ensemble classification. Details of the pre processing and feature selection methods are described in Section III. We describe how each ensemble classifier is selected and how the results are combined to boost up the performance of the ensembled classification techniques and types are presented in section IV. The details of experimental results are discussed in section V and conclude the paper in Section VI.

II. Related Works

Robert E. Banfield et al[13] experimentally evaluate bagging and seven other randomization based approaches to creating an ensemble of decision tree classifiers. They find that boosting, random forests, and randomized trees are statistically significantly better than bagging. Their algorithm uses the out-of-bag error estimate, and is shown to result in an accurate ensemble for those methods that incorporate bagging into the construction of the ensemble.

Ching Wei Wang [5] proposed a new ensemble machine learning algorithm for classification and prediction on gene expression data. The algorithm is tested and compared with three popular adopted ensembles, i.e. bagging, boosting and arcing. The results show that the proposed algorithm greatly outperforms existing methods, achieving high accuracy over 12 gene expression datasets

Gulisong Nasierding et al [8] presents a triple-random ensemble learning method for handling multi-label classification problems. The proposed method integrates and develops the concepts of random subspace, bagging and random k-labelsets ensemble learning methods to form an approach to classify multi-label data. The proposed method is implemented and its performance compared against that of popular multi-label classification methods. The experimental results reveal that the proposed method outperforms the examined counterparts in most occasions when tested on six small to larger multi-label datasets from different domains.

Xiao-Dong Zeng et al [19] proposed a novel ensemble model that refines the bagging algorithm with an optimization process. The optimization process mainly emphasizes on how to select the optimal classifiers according to the accuracy and diversity of the base classifiers. The empirical results reveal that the new model does outperform the original method in terms of learning accuracy and complexity.

III. Feature Selection

Feature selection is one of the important and frequently used techniques in data preprocessing for data mining [1], [9]. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications: speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility.

Many irrelevant attributes may be present in data to be mined. So they need to be removed. Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore feature selection techniques needs to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are to avoid over fitting and improve model performance and to provide faster and more cost-effective models [15]. Attribute selection methods can be broadly divided into filter and wrapper approaches.

In the filter approach the attribute selection method is independent of the data mining algorithm to be applied to the selected attributes and assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low scoring features are removed. The subset of features left after feature removal is presented as input to the classification algorithm.

Wrapper methods embed the model hypothesis search within the feature subset search. In the wrapper approach the attribute selection method uses the result of the data mining algorithm to determine how good a given attribute subset is. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The major characteristic of the wrapper approach is that the quality of an attribute subset is directly measured by the performance of the data mining algorithm applied to that attribute subset. The wrapper approach tends to be much slower than the

filter approach, as the data mining algorithm is applied to each attribute subset considered by the search. In addition, if several different data mining algorithms are to be applied to the data, the wrapper approach becomes even more computationally expensive [11].

Another category of feature selection technique was also introduced, termed embedded technique in which search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses. Just like wrapper approaches, embedded approaches are thus specific to a given learning algorithm. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods [20].

IV. Ensemble classification techniques

1 Bagging

Bagging stands for Bootstrap Aggregating(Bagging) which is one of the successful ensemble learning methods [6]. It generates multiple bootstrap training sets from the original training set and uses each of them to generate a classifier for inclusion in the ensemble [5]. It consists in training different classifiers with bootstrapped replicas of the original training data-set. That is, a new data-set is formed to train each classifier by randomly drawing (with replacement) instances from the original data-set (usually, maintaining the original data-set size). Hence, diversity is obtained with the resampling procedure by the usage of different data subsets. Finally, when an unknown instance is presented to each individual classifier, a majority or weighted vote is used to infer the class.

2. Boosting

This model is based on averaging method and is highly applied in combination methods. Boosting is one of the most powerful ideas presented during the last ten years and is aimed to create a strong classifier through the mixture of multiple weak classifiers. Boosting algorithm creates a strong learning algorithm by the combination of three weak algorithms. This is similar to Bootstrap Aggregation method except that the classification is performed within several steps. Samples, according to the classification`s accuracy, are given definite weights until the final model could be created.

3. AdaBoost

AdaBoost stands for Adaptive Boosting, it is a well known, effective technique for increasing the accuracy of learning algorithms. The sequence of base classifiers, produced by AdaBoost from the training set, is applied to the validation set, creating a modified set of weights. The training and validation sets are switched, and a second pass is performed. Re-weighting and re-sampling are two methods implemented in AdaBoost[7]. The fixed training sample size and training examples are re-sampled according to a probability distribution used in each iteration. In term of re-weighting, all training examples with weights assigned to each example are used in each iteration to train the base classifier [21].

4. MultiBoosting

MultiBoosting is an extension to the highly successful AdaBoost technique for forming decision committees and can be viewed as combining AdaBoost with wagging. It is able to harness both AdaBoost's high bias and variance reduction with wagging's superior variance reduction [16].

5. Random Forest

A random forest [6] is an ensemble of classifiers, where diversity among the predictors is obtained by using bagging implemented by growing many classification trees and having them "vote" for a final decision according to a majority role. However, due to its capability to estimate the importance of the features, it can also be applied as a feature quality estimator and selector (by applying a threshold to the feature quality estimates).
Random forest classifier works by building a set of decision trees where a single tree node growing is done using a limited set of randomly chosen features. Since it includes many trees, this set is called a forest

V. Result and discussion

Our experiments on two databases in UCI machine learning database repository as shown in Table 1. In order to validate the prediction results of the comparison of the various popular ensemble classification techniques and the 10-fold crossover validation is used. The k-fold crossover validation is usually used to reduce the error resulted from random sampling in the comparison of the accuracies of a number of prediction models. The present study divided the data into 10 folds where 1 fold was for testing and 9 folds were for training for the 10-fold crossover validation.

Table 1 provides the attribute information of two datasets.

Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Breast Cancer (WBC)	11	699	2
Diabetes	9	768	2

.

1. Evaluation Methods

We have used the Weka toolkit to experiment with these three data mining algorithms [17]. The Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA version 3.6.9 was utilized as a data mining tool to evaluate the performance and effectiveness of the 6-breast cancer prediction models built from several techniques. This is because the WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models.

The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. The computation time is noted for each classifier is taken in to account. The evaluation parameters are the specificity, sensitivity, and overall accuracy.

The sensitivity or the true positive rate (TPR) is defined by $\frac{TP}{(TP+FN)}$ while the specificity or the true negative rate (TNR) is defined by $\frac{TN}{TN+FP}$ and the accuracy is defined by $\frac{(TP+TN)}{(TP + FP + TN + FN)}$

- True positive (TP) = number of positive samples correctly predicted.

- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

These values are often displayed in a confusion matrix as be presented in Table 2. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values with the predicted values in the trained model.

Table 2: Confusion Matrix

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

2. Result.

The confusion matrix of each Classification method is presented in Table 3; the values to measure the performance of the methods (i.e. accuracy, sensitivity, specificity, error rate and time) are derived from the confusion matrix and showed in Table 4.

An experiment was set up to compare Decision Tree with Bagging, AdaBoost, MultiBoosting and Random forest. In all ensemble methods, Decision Tree was used as the base classifiers. In implementing of the experiments, we used the WEKA software to gain access to different classifiers. Decision Tree construction method was the J48 algorithm from the WEKA.

In this experiment, the accuracy of ensemble classification techniques is based on the selected classifier algorithm. The accuracy for each of the classifier algorithm for full attributes for two datasets is shown in Table 3 and 4. The results for full attribute present the highest accuracy of model is Random forest is 94.49% for WBC data and MultiBoostAB is 76.31% for Diabetes data, which is the results could be considered as an indicator to the potential ensemble classification algorithm for human talent data. Based on table 3 and 4 we derived the accuracies, errors and execution times are shown in figure 1, figure 2 and figure 3.

Table 3: Confusion Matrix of Two data sets.

Algorithm	Breast cancer Data (699)			Diabetes Data (768)		
	Desired Result	Output Result		Desired Result	Output Result	
		Benign	Malignant		Negative	Positive
J48	Benign	437	21	Negative	407	93
	Malignant	17	224	Positive	108	160
Bagging	Benign	438	20	Negative	418	82
	Malignant	15	226	Positive	107	161
AdaBoostM1	Benign	441	17	Negative	423	77
	Malignant	22	219	Positive	120	148
MultiBoostAB	Benign	439	19	Negative	422	78
	Malignant	23	218	Positive	104	164
RandomForest	Benign	443	15	Negative	421	79
	Malignant	16	225	Positive	121	147

Table 4: Performance of Ensemble Classification.

Algorithm	Breast cancer Data (699)					Diabetes Data (768)				
	Acc	Senst	Spec	Err	Time	Acc	Senst	Spec	Err	Time
J48	93.79	0.96	0.91	6.21	0.09	73.83	0.81	0.60	26.17	0.20
Bagging	93.96	0.95	0.93	6.04	0.17	75.39	0.84	0.60	24.61	0.42
AdaBoostM1	93.87	0.96	0.90	6.13	0.12	74.34	0.85	0.55	25.65	0.22
MultiBoostAB	93.69	0.95	0.91	6.31	0.06	76.31	0.85	0.61	23.69	0.05
RandomForest	94.49	0.96	0.91	5.51	0.14	73.95	0.84	0.55	26.05	0.50

Note:- Acc - Accuracy, Senst - Sensitivity, Spec - Specificity, Err - Errors and Time - Execution time

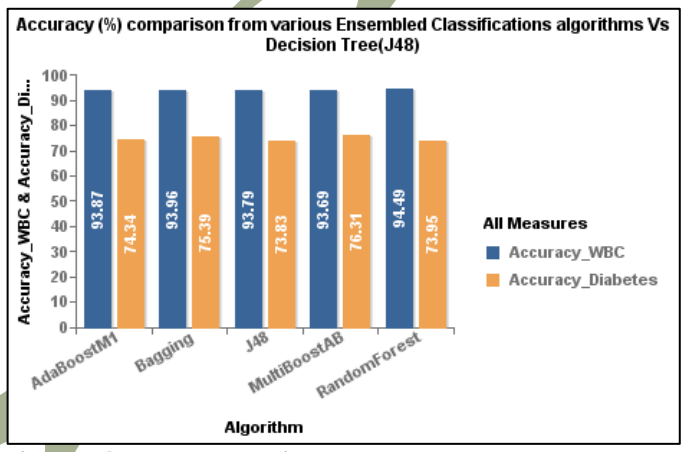

Figure-1:Accuracy comparison

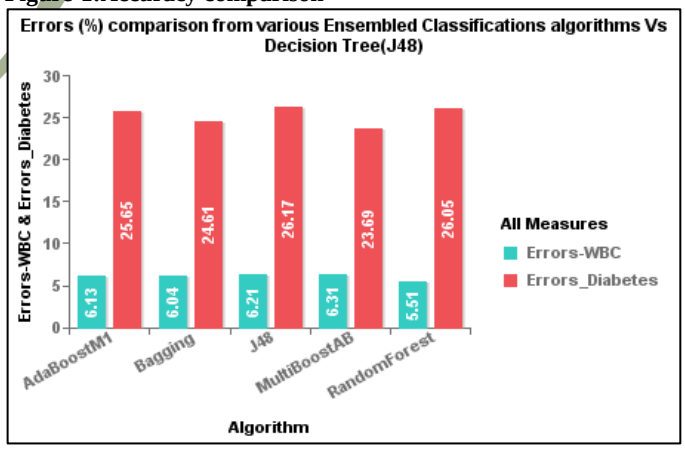

Figure-2:P Errors comparison

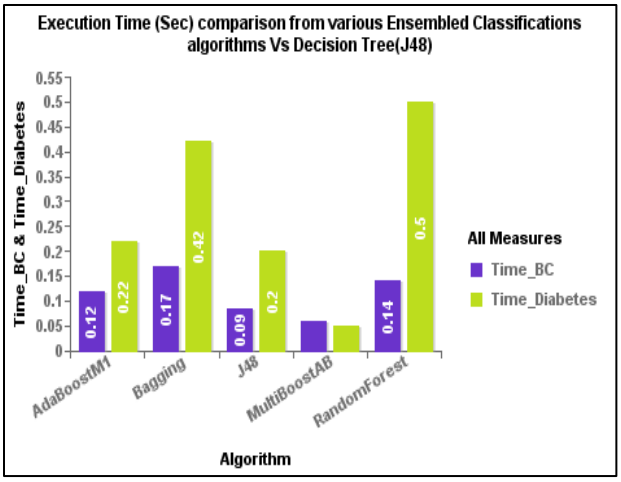

Figure-3:Execution time comparison

VI. Conclusions

In this paper, the accuracy of Ensemble classification techniques is evaluated based on the selected classifier algorithm. An important challenge in data mining and machine learning areas is to build precise and computationally efficient ensemble classifiers for Medical applications. The performance of Random forest shows the high level compare with other ensemble classifiers. Hence Random forest and MultiBoostAB shows the concrete results with Breast Cancer and Diabetes disease of patient records. Therefore Random forest and MultiBoostAB classifier is suggested for diagnosis of breast cancer and Diabetes diseases based ensemble classification to get better results with accuracy, low error rate and performance.

References

[1]. A.L. Blum and P. Langley, "Selection of Relevant Features and Examples in Machine Learning," Artificial Intelligence, vol. 97, pp. 245-271, 1997.

[2]. Bauer E, Kohavi R (1999) "An empirical comparison of voting classification algorithms: bagging, boosting, and variants", Mach Learn 36:105–139

[3]. Breiman, L., "Bagging predictors", Mach. Learn. 24(2):123–140, 1996.

[4]. Breiman, L., "Random forests", Mach. Learn. 45(1):5–32, 2001

[5]. Ching Wei Wang, "New Ensemble Machine Learning Method for Classification and Prediction on Gene Expression Data", Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006, pp 3487-3481

[6]. Dietterich TG," Ensemble methods in machine learning. In: Proceedings of Multiple Classifier System", vol. 1857. Springer; 2000. pp. 1–15.

[7]. Freund, Y., and Schapire, R. E., "A decision-theoretic generalization of on-line learning and an application to Boosting", J. Comput. Syst. Sci. 55(1):119–139, 1997

[8]. Gulisong Nasierding et al, "A Triple-Random Ensemble Classification Method for Mining Multi-label Data", 2010 IEEE International Conference on Data Mining Workshops, IEEE Computer Society, pp: 49-56

[9]. H. Liu and H. Motoda, eds. Boston,"Feature Extraction, Construction and Selection", A Data Mining Perspective, Kluwer Academic, 1998, second printing, 2001.

[10]. J. Han and M. Kamber, "Data Mining—Concepts and Technique" (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. San Mateo, CA: Morgan Kaufmann, 2006.

[11]. M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn and A.K. Jain,"Dimensionality Reduction Using Genetic Algorithms", IEEE Transactions On Evolutionary Computation, Vol. 4, No. 2, 2000

[12]. P.N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining", Reading, MA: Addison-Wesley, 2005

[13]. Robert E. Banfield et al, " A Comparison of Decision Tree Ensemble Creation Techniques", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No.1,January 2007, pp 173-180

[14]. Valentini, G., and Dietterich, T. G., "Low bias bagged support vector machines", In: Fawcett, T., and Mishra, N. (Eds.), International conference on machine learning. AAAI press, California, 2003.

[15]. W. Daelemans, V. Hoste, F.D. Meulder and B. Naudts, "Combined Optimization of Feature Selection and Algorithm Parameter Interaction in Machine Learning of Language", Proceedings of the 14th European Conference on Machine Learning (ECML-2003), Lecture Notes in Computer Science 2837, Springer-Verlag, Cavtat-Dubrovnik, Croatia, 2003, pp. 84-95

[16]. Webb, G. I., "MultiBoosting: a technique for combining Boosting and wagging", Mach. Learn. 40(2):159–197, 2000.

[17]. Weka "Data Mining Software in Java", http://www.cs.waikato.ac.nz/ml/weka/

[18]. Witten H.I., Frank E., "Data Mining: Practical Machine Learning Tools and Techniques", Second edition, Morgan Kaufmann Publishers, 2005.

[19]. Xiao-Dong Zeng et al, "Optimization of bagging classifiers based on SBCB algorithm", Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010, pp: 262-267

[20]. Y. Saeys, I. Inza and P. Larranaga, "A review of feature selection techniques in bioinformatics", Bioinformatics-19, 2007, pp. 2507–17.

[21]. Zhang, C. X., Zhang, J. S., and Zhang, G. Y., "An efficient modified Boosting method for solving classification problems". J. Comput. Appl. Math. 214(2):381– 392, 2008