Performance Evaluation of Ensemble Classifiers on Benchmark Datasets

[1]H. Benjamin Fredrick David, [2]A. Suruliandi
[12] Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli

Abstract: In supervised machine learning, highest possible performance is achievable and identifying such a perfect model is a crucial research task. Although a single machine learning algorithm performs well in predicting the target variable, it is not up to the highest possible accuracy that is achievable when used with diverse real data. Hence, the native solution is to come up with ideas to combine the models namely ensemble methods that use multiple supervised machine learning algorithms in order to obtain highest predictive performance than those obtained from any of the contributing algorithms alone. These ensemble methods conceal the disadvantage of those single models and improve the performance to provide the best prediction possible. When the validation process is undertaken, it will describe the quality and characteristics of the ensemble models and predict their performance against real data. In this research work, the performance of the two ensemble classifiers namely Bagging and AdaBoost is experimented and evaluated on various folds of cross validation with different experiments on two different benchmark datasets namely credits and diabetes. The results provide better understanding of the accuracy, reliability and usefulness of the models.

Keywords: Data Mining, Classification, Prediction, Diabetes Classification, Credits Classification, Ensemble Classifier.

Introduction

The supervised classification is typically a function which determines the prediction result based on a sample of historical data collected from the existing training samples. The training sample is required for training the classification algorithm, the testing sample is then used to validate the trained classification model built based on the data taken from training samples. This is known as supervised learning. One of the key objectives in classification is to achieve a high accuracy with a small number of training samples to make the classification process as useful and efficient. Although the classifier used for the classification provides high accuracy, it may not be up to the mark for some context sensitive data mining tasks.

In such a case, combining one or more classifiers to create an ensemble of classifiers is useful for getting a classifier with highest predictive accuracy. These ensemble modeling are a powerful way to improve the performance of the classification model. These types of classifiers are commonly called as Ensemble classifiers. Two such attractive and most familiar ensemble learning algorithms are AdaBoost and Bagging that produce more accurate predictions than the widely used alternatives. Ensemble is the art of combining various individual models to improvise on the model's functionality. This may improve the stability or the predictive power or even sometimes both.

Two such familiar ensemble classifiers are AdaBoost and Bagging which is are standard and are performance centric. The AdaBoost [1], short for Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire. AdaBoost when utilized along with decision trees as the weak learners is referred to as the best out-of-the-box classifier [2]. Bagging[3] provides a common way of generalizing the classification accuracy based on the bags of the data.

The above said two algorithms are used in majority of the applications and there has not been a detailed study of these two algorithms being generalized to new datasets. The overall accuracy of the model can be analyzed and evaluated through the help of several evaluation strategies. One such strategy is the k-fold cross-validation which is utilized in this work. In this research work, the above mentioned two meta-classifiers namely AdaBoost and Bagging is used for evaluating the predictive performance on two different benchmark datasets namely credits and diabetes. The research experiments involve taking multiple sets of results for different folds of cross validation and the algorithms are evaluated with standard performance metrics. This study also describes and evaluates the performance of the algorithms with respect to the folds used for cross validation.

Literature Review

In recent years the research intense in ensemble classification is constantly increasing through various means such as improving various learning methods and introduction of new techniques for evaluation. Some of the formidable applications, researches and the ideas of ensemble classification in recent years are listed below.

Nan-Chen Hsieh et. al[4] have proposed an ensemble classifier for credit scoring system by utilizing several data mining techniques with binning to discretize the continuous values through the use of optimal associate binning. The ensemble includes neural network, support vector machine, and Bayesian network. The knowledge obtained from the classifier is represented in multiple forms such as causal diagram and constrained association rules.

Eibe Frank et. al [5] have done a research work which contains an extensive collection of machine learning algorithms and data pre-processing methods that is processed by a GUI for data exploration and the experimental comparison of different machine learning techniques on the same problem. By using Weka, they were able to easily identify a suitable ensemble algorithm for generating an accurate predictive model from it. Frank et. al [6] have described a workbench with convenient interactive graphical

user interfaces are provided for data exploration, for setting up large-scale experiments on distributed computing platforms, and for designing configurations for streamed data processing. These interfaces constitute an advanced environment for experimental data mining. Bühlmann et. al [7] have done a study in which they proved that the hard decisions create instability and bagging is shown to smooth such hard decisions. This provides much smaller variance and mean squared error compared to other. Leo Breiman [8] has written an article on Bagging predictors. His work describes the method for generating multiple versions of a predictor for getting an aggregated predictor. When predicting the class as numerical outcome, the aggregation is averaged. When it's a nominal outcome, the plurality vote is considered when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. Tests on real and simulated data sets using classification and regression trees and subset selection in linear regression show that bagging can give substantial gains in accuracy

Methodology

The methodology of this research work includes the construction of the model using the two state of the art ensemble classifiers namely AdaBoost and Boosting. These two algorithms are used in the various experiments in order to identify the best ensemble algorithm. The proposed model serves the purpose of determining the best classification algorithm for prediction using two benchmark datasets. This provides the and thus achieving maximum accuracy in classification.

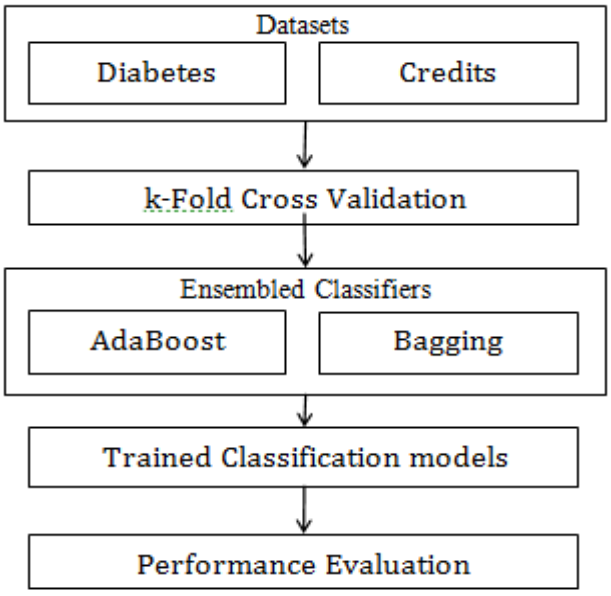

Fig.1. Research Methodology

The methodology shown in the above figure 1 describes the proposed model being used to analyze the two ensemble classification algorithms. The benchmark datasets are first validated using the cross validation using k-folds with the help of two mentioned classification algorithms. The experiments produce the classification models trained. This further is used to make predictions using the testing data. The Performance of the AdaBoost and Bagging ensemble method is evaluated at the end and the comparison is made.

A. K-Fold Cross Validation

Cross-validation is also called as rotation estimation. The main purpose of this cross validation is to estimate how accurately a predictive model will perform in practice and test the classification model's ability to predict new data that was not used in estimating it. This also gives an insight on how the model will generalize to an unknown dataset for a real problem. In general, the k-fold cross-validation process is an evaluation strategy in which the sample represents the original sample of the dataset which is randomly partitioned into k equally partitioned subsamples. A single subsample out of the k samples is kept as the data to be used for testing the model built, and the remaining k-1 are used as training data for training the classifier. The cross-validation process is repeated k times with each of the k subsamples used exactly once as the testing data. This results in k results from each fold which are then averaged to produce a single estimation.

B. Ensemble Classifiers

1) AdaBoost

AdaBoost[1] which is shortened from Adaptive Boosting, is a machine learning meta-algorithm which could be used alone with many different types of learning algorithms in order to improve the performance. The output of constituting learning algorithms is then combined as a weighted sum that represents the final output of the boosted classifier. The word adaptive refers to the algorithm's nature of tweaking the subsequent weak learners in the favor of those instances that are misclassified by previous classifiers in the list. Hence, the AdaBoost is sensitive to noisy data and outliers. A generalized algorithm for the AdaBoost is given below.

• Sample - S1, S2, S3…… Sn

• Desired output – Predicted Class {-1,+1}

• Initial weights – w1,w2,w3,……. wn

• Error Function – f

• Weak learners – c1, c2,c3,…..,cn

1. For each Classifier x in Total Classifiers (T)
2. Choose h1(x)
3. Find weak learners that minimizes the weighted sum error for misclassified points
   - Add to Ensemble
   - Update Weights and coefficients
4. End for
5. Based on coefficients with biased weights, the final desired output is produced.

2) Bagging

Bagging [3] termed as Bootstrap aggregating is a meta-learning machine learning algorithm that is designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. Although bagging is known to be applied with the decision tree methods, it can also be used with any type of classification methods. Bagging follows a model averaging approach and it also decreases the variance amount in order to avoid over fitting of the data. The generalized algorithm for Bagging is given below.

• Number of models – N

• Sample with replacement – X

• Classification Algorithm – C

• Trained Models – m1, m2, m3, ……., mN

• Desired output – (E) i.e. Predicted Class {-1,+1}

1. For each models untrained
   - Train a Classification algorithm C on this sample
   - Save the model mi
2. End For
3. Estimate from each models m
4. Average the estimates to produce E.

Experimental Results and Discussion

The dataset is given in section A, the metrics utilized for comparison in these experiments are given in section B, the experimental results of the Bagging and AdaBoost on the two databases are given in section B and C respectively.

A. Dataset

This research work includes two dataset. These two dataset utilized for this experiment is given in table 1.

Table 1. Dataset description

S.No	Dataset	Type	Attributes	Instances
1	German Credit	Nominal	20	1000
2	Pima Indians Diabetes Database	Numeric	8	768

B. Metrics

The performance metrics used for the comparison of the two algorithms are described in table 2 given below.

1) Accuracy

The classification accuracy is the number of correct predictions divided by the total number of predictions. It is given in equation (1)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2) Error rate

Error rate is the 1 minus the accuracy that produces the error made in prediction of the target variable. It is given in equation (2)

$$Errorrate = 1 - Accuracy \quad (2)$$

3) Precision

Precision which is also called positive predictive value is the fraction of retrieved instances that are relevant. It is given in equation (3)

$$\frac{TP}{TP + FP} \quad (3)$$

4) Recall

Recall is the fraction of relevant instances that are retrieved. The equation is given in equation (4)

$$\frac{TP}{TP + FN} \quad (4)$$

5) F-Measure

The F measure is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. It is given in equation (5).

$$\frac{2}{(\frac{1}{P} + \frac{1}{R})} \quad (5)$$

6) Mean Absolute Error

It is the mean absolute error is an average of the absolute errors. The equation is given in equation (6)

$$MAE = \frac{1}{N} \sum_{x \in S} | f(x) - h(x) | \quad (6)$$

7) Root Mean Square Error

It denotes how much error value resembles the two datasets. It compares the predicted value and the observed value. The equation is given in equation (7)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - P_o)^2}{n}} \quad (7)$$

8) Relative Absolute Error

It is the Relative Error It is the ratio of the absolute error of a measurement to the measurement being taken. The equation is given in equation (8)

$$RAE = \frac{AE}{M_{AE}} \quad (8)$$

9) Root Relative Squared Error

This simple predictor is just the average of the actual values. The equation is given in equation (9)

C. Experimental result of AdaBoost and Bagging in German Credits database

The Bagging and AdaBoost classification algorithm are evaluated with different folds of k-fold cross validation on the German Credits database. The experimental results are given in table 2 given below.

Table 2. Performance of the Bagging and AdaBoost algorithm in German Credits Database

Classifier	Folds	Accuracy	Error	Precision	Recall	F-Measure
Bagging	10	74.3	25.7	0.738	0.743	0.694
	20	73.6	26.4	0.724	0.736	0.686
	30	74.6	25.4	0.735	0.746	0.706
	40	74.8	25.2	0.743	0.748	0.703
	50	74.2	25.8	0.732	0.742	0.697
	60	74.6	25.4	0.741	0.746	0.7
	70	**75.6**	**24.4**	**0.758**	**0.756**	**0.713**
	80	74.4	25.6	0.738	0.744	0.697
	90	75.3	24.7	0.746	0.753	0.714
Adaboost	10	72.3	27.2	0.702	0.723	0.705
	20	71.8	28.2	0.697	0.718	0.7
	30	72.4	27.6	0.702	0.724	0.703
	40	72.5	27.5	0.705	0.725	0.707
	50	71	29	0.69	0.71	0.695
	60	72.3	27.7	0.704	0.723	0.708
	70	70.6	29.4	0.682	0.706	0.686
	80	**72.9**	**27.1**	**0.712**	**0.729**	**0.716**
	90	71.3	28.7	0.693	0.713	0.698

As given in the above mentioned table, the two classifiers performs well and at 70 folds cross-validation, the Bagging performs best with 75.6% accuracy and at 80 folds, the AdaBoost performs best with 72.9% accuracy and the error rates, precision, recall and f-measure at the respective folds are also the highest resulting in better quality in the target variable prediction.

Similarly, the Bagging and AdaBoost classification algorithm error rates are validated on the German Credits database. The results for the experiments are given in table 3 given below.

Table 3. Error measures of the Bagging and AdaBoost in German Credits Database

Classifier	Folds	MAE	RMSE	RAE	RRSE
Bagging	10	0.3558	0.4105	84.6808	89.5882
	20	0.3565	0.4103	84.8506	89.5379
	30	0.3589	0.4116	85.4127	89.8103
	40	**0.3547**	**0.4071**	**84.4105**	**88.8296**
	50	0.3575	0.4093	85.0859	89.3121
	60	0.3582	0.4112	85.2571	89.7313
	70	0.3564	0.4086	84.8139	89.1712
	80	0.3578	0.4101	85.1459	89.4836
	90	0.358	0.4109	85.2042	89.6617
Adaboost	10	0.3441	0.4229	81.9037	92.2843
	20	0.3464	0.4236	82.4523	92.4379
	30	0.3425	0.4195	81.5198	91.5442
	40	0.3414	0.4195	81.247	91.5318
	50	0.3431	0.4193	81.66	91.4879
	60	**0.341**	**0.4185**	**81.1696**	**91.3217**
	70	0.3449	0.4237	82.0885	92.4626
	80	0.3414	0.4193	81.2599	91.5035
	90	0.3425	0.4196	81.5098	91.5583

Similar to the above mentioned table, the error measure for the two ensemble classifiers are considerably low at the 70 and 80 folds respectively. Still, from the above table it is also noted that the error values achieve the lowest possible values at the 40 and 60 folds respectively for Bagging and AdaBoost.

D. Experimental result of Bagging and AdaBoost in Pima Indians Diabetes Database

Similar to the previous experiment, this experiment is validated for the Indian Diabetes database. The performance of the Bagging and AdaBoost classification algorithm are evaluated with different folds of k-fold cross validation on the Pima Indians Diabetes Database. The experimental results are given in table 4 given below.

Table 4. Performance of Bagging and AdaBoost in Pima Indians Diabetes Database

Classifier	Folds	Accuracy	Error	Precision	Recall	F-Measure
Bagging	10	76.3021	23.6979	0.758	0.763	0.758
	20	75.7813	24.2188	0.751	0.758	0.75
	30	75.7813	24.2188	0.752	0.758	0.752
	40	**77.6042**	**22.3958**	**0.771**	**0.776**	**0.769**
	50	75.5208	24.4792	0.749	0.755	0.749
	60	75.7813	24.2188	0.751	0.758	0.75
	70	75.5208	24.4792	0.749	0.755	0.749
	80	77.2135	22.7865	0.767	0.772	0.767
	90	75.9115	24.0885	0.753	0.759	0.752
Adaboost	10	75.2604	24.7396	0.746	0.753	0.746
	20	76.0417	23.9583	0.755	0.76	0.755
	30	76.9531	23.0469	0.765	0.77	0.765
	40	76.3021	23.6979	0.758	0.763	0.759
	50	76.0417	23.9583	0.755	0.76	0.756
	60	**77.8646**	**22.1354**	**0.774**	**0.779**	**0.774**
	70	76.0417	23.9583	0.755	0.76	0.755
	80	76.3021	23.6979	0.757	0.763	0.758
	90	76.6927	23.3073	0.761	0.767	0.761

It is noted from the above table that, the performance of the Bagging on the Diabetes dataset is highest in the 40 folds of cross-validation and whereas the AdaBoost classifier performs best in the 60 folds cross-validation.

Similarly, the experimental results for Bagging for the same dataset are given in table 5 given below.

Table 5. Error measures of Bagging and AdaBoost in Pima Indians Diabetes Database

Classifier	Folds	MAE	RMSE	RAE	RRSE
Bagging	10	0.3244	0.3978	71.383	83.4604
	20	0.3213	0.3945	70.6816	82.7714
	30	0.3226	0.3977	70.9796	83.437
	40	0.3229	0.3962	71.0341	83.1215
	50	0.3211	0.3969	70.6495	83.2617
	60	0.3242	0.3981	71.3213	83.5053
	70	0.3191	0.3941	70.2057	82.6726
	80	0.3208	0.3957	70.5736	83.0024
	90	**0.3202**	**0.3954**	**70.4485**	**82.944**
Adaboost	10	0.3057	0.4082	67.266	85.6399
	20	0.3088	0.41	67.9317	86.0244
	30	**0.301**	**0.4001**	**66.2222**	**83.9446**
	40	0.3041	0.4039	66.8988	84.7285
	50	0.3041	0.4046	66.9115	84.8721
	60	0.3024	0.4005	66.5352	84.0213
	70	0.3042	0.404	66.9317	84.7549
	80	0.3022	0.4015	66.4927	84.2357
	90	0.3041	0.405	66.8965	84.9749

The 40 and 60 folds of cross-validation provided highest predictive accuracy and better precision, recall and f-measure for the Bagging and AdaBoost classifiers. But, the error measures are very low in the 90 and 30 folds respectively.

The performance of the two ensemble classification algorithms on two datasets namely Credits and Diabetes are considered and the mean accuracy is plotted in Fig 2.

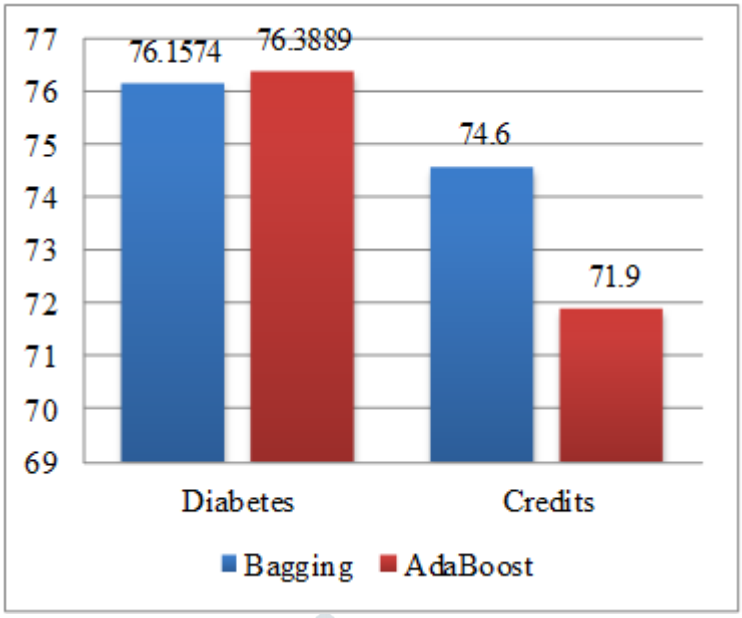

Fig 2. Accuracy of the AdaBoost and Bagging on two datasets

Based on the results and conclusions drawn from the above figure, it is observed that AdaBoost outperforms the Bagging in Pima Indian Diabetes dataset and whereas in German Credits dataset, the Bagging performs best when compared to AdaBoost ensemble classification algorithm with higher Accuracy. Thus the models generated depend on the learning algorithms that are utilized in ensemble classification and this largely depends on the dataset for making the predictions. This research work compares the significant performance improvement and identifies the best suited ensemble classifier for real world problems.

Conclusion

The overall objective of the work is to study and compare the two ensemble classification algorithms namely AdaBoost and Bagging. This study is performed on two benchmark datasets to prove the stability of the algorithms. The algorithms are evaluated using different folds of cross validation for comparing the significant performance improvements among the two algorithms. The performance evaluation has been completed and the study shows promise in ensemble learning when compared to the single learning model.

The Future work of this research work can be made to include various other ensemble classification algorithms. The algorithms can be fine-tuned to perform better involving the Meta learners from different classification algorithms. Feature selection methods can be used to reduce the number of features or attributes that are relatively important and this might improve the performance of the classification algorithms on the whole.

References

[1] Hastie, Trevor, et al. "Multi-class adaboost." Statistics and its Interface 2.3 (2009): 349-360.

[2] Kégl B. The return of AdaBoost. MH: multi-class Hamming trees. arXiv preprint arXiv:1312.6086. Dec 2013.

[3] Quinlan, J. Ross. "Bagging, boosting, and C4. 5." AAAI/IAAI, Vol. 1. 1996.

[4] Hsieh, Nan-Chen, and Lun-Ping Hung. "A data driven ensemble classifier for credit scoring analysis." Expert systems with Applications 37.1 (2010): 534-545.

[5] Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, Ian H. Witten; Data mining in bioinformatics using Weka, Bioinformatics, Volume 20, Issue 15, 12 October 2004, Pages 2479–2481, https://doi.org/10.1093/bioinformatics/bth261

[6] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2009). Weka-a machine learning workbench for data mining. In Data mining and knowledge discovery handbook (pp. 1269-1277). Springer, Boston, MA.

[7] Bühlmann, Peter, and Bin Yu. "Analyzing bagging." The Annals of Statistics 30.4 (2002): 927-961.

[8] Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140..