IJCTA, 9(10), 2016, pp. 4787-4794

Application of Boosting Technique to improve the performance of Decision Tree Classifiers for Parkinson disease

***P. Suganya **C.P. Sumathi**

Abstract : The applicationofmachinelearningalgorithms with boostingtechniqueis applied to Parkinson's data for improving the performance of various weak classifiers is the primary objective of this work.Basically Parkinson's Data from UCI repository is an imbalanced dataset has the voice recording of healthy and nonhealthy persons consisting of 195instances.In this article the classifiers such asID3,Random Forest and J48 algorithm are applied with AdaBoost, anensemble boosting techniquethatdeterminesthe performance oftheabovealgorithms. AdaBoost is an ensemble erudition model where the weights are adjusted for weak classifiers and the iterations make the minority class to be correctly classified.Thevarious classifiers and boosting functionalities are providedbytheWEKAexplorerwhich helps to controlthe efficacy ofa classifier model.The validated results are analyzed inaconfusion matrix which computes the accuracy,precision, recall, kappa andF-score.It is found that there is an increase in the performance of three algorithms up to 5% with the application of AdaBoost technique,the major beneficiary being ID3.

Keywords : Imbalance dataset,Machine Learning, ID3, Random Forest, J48, AdaBoost, Weka.

1. INTRODUCTION

To determine the Parkinson Diseasedataset prediction using SMO -SVM classifier with AdaBoost algorithm attained an accuracy of 84% and 81% [1].In predicting the BCI (Brain computed Interface)data the best results were produced in SVM, AdaBoost,Random Forest,RBF , Bagging, Stacking and Boosting[2].Common spatial patterns have been used in raw data to convert the signals to new space andthe Instance based classifier is compared with different classifiers for EEG (Electroencephalography) measure human brain activity for different activities and its accuracy was termed to be 94.5% which predicted the eye state is opened or closed [3].Data replication method transforms an ordinal problem into a larger binary classification by transforming the dataset to learn the ordering relation using AdaBoost.We have presented a new variant of the well-knownAdaBoost classifierintended for ordinal classificationfor (k-1) strong binary classifiers are combined to yield the multiclass model [4].

The accuracy predictionfor type 2 diabetes patient information to analyze the performance, execution time and error rate using weka usingMeta learning algorithms classifiers and turn them into more powerful learners. From the algorithm one parameter specifies the base classifier and the other specify the number of iterations for outlines such as bagging and boosting.Bagging stacks a classifier to reduce the variance and the other parameterto calculate the out of bag error specifying the threshold for weight pruning and resamples if the base classifier cannot handle weighted instances [5].The paper contains trialoutcomes obtained after classifying 10% of the KDD CUP '99 dataset using ensemble methods like bagging and boosting which associates the presentation with the standard J-48 classification algorithm.For statistical classification and regression problems bagging is used as an ensemble

* Ph.D Scholar, Bharthiar University, Coimbatore-641046, Tamil Nadu, India, Associate

** Prof. & Head, Dept. of Computer Sci, S.D.N.B Vaishnav College for Women, Chromepet, Chennai -44.

algorithm which improves the accuracy. Most importantly, it lessens variances and resolves the problem over fitting the data. Booting algorithm is used for supervised learning is to reduce the preference in the dissimilarorganization techniques used.A weak learner is a process with improved accuracy than a randominference.The weak learner algorithms are boosted to increase the performance using boosting algorithm[6]. To examine the ability of ensemble methods to improve the efficiency of basic J48 machine learning algorithm.The SONAR dataset showed a better perception between sonal signals combined off a roughly cylindrical rock with the application of various algorithms such as bagging, boosting and blending. The ranking and standard deviance functionalities provided by the weka experimenter helps to improve the effectiveness of a classifier model. Boosting is an ensemble methodwhich increases the effectiveness on the training data. A second classifier is concentrated on the second training data where accuracy limit is retained. The AdaBoost M1 is used as a boosting ensemble classifier and the inferencetoensemble is more effective than the individual algorithm[7]. Using weka tool the classification techniques were applied to medical database with the help of bagging and Adaboost. A 10 fold cross validation was utilised which showed that FT tree shows good result with classification in medical diagnosis. [15].DataBoost.IM and SVM algorithm shows a good G-mean and F-Measure metrics in imbalanced liver dataset. The target class is divided into minority and major class for performance evaluation using DataBoost.IM and SVM[16].

2. DIFFERENT CLASSIFIERS

ID3 ALGORITHM

Iterative Dichotomiser 3 is a decision tree algorithm which was invented by Ross Quinlan who was a researcher in data mining and decision theory. The decision tree contains nodes and leaves which are homogeneous. The nodes are produced from the attribute values and the ID3 algorithm follows Occam's razor principle[8]. It creates smallest decision tree possible with calculation of entropies i.e. the attribute with lowest entropy. The algorithm is as follows

1. If all instancesare belonging to healthy status,return the rootnode with label C1.
2. If all examplesareunhealthy, returnthe root-nodetreewithlabelC0.
3. Check the predicting attributes then return the single node tree Root,with labelequal to majority valueofthetarget attribute inthe instances of the dataset.
4. A = the field that best classifies the target field.
5. DecisionTreeattribute for Root = A.
6. Foreachlikelihood value, *vi*, of A,
7. Adda subdivision for the root from the test condition
8. A = *vi*
9. Let Examples(*vi*), be the subsetof examples that have the value *vi* for A
10. If Examples (*vi*) is empty
11. Then add a new subdivision where majority class values are present.
12. Else below this new branch enhance the sub tree ID3 (*vi*), Target_Attribute,Attributes- {A})
13. End
14. Return Root

RANDOM FOREST

Leo Breiman developed the Random Forest algorithm which is a strong predictive model visualizing for high dimensional data. Its base of CART where there is general rule of tree growing, combining the tree, testing and post processing of the tree. Each note is split into binary or more split of children and the tree grows randomly [9]. Using different random subsampling technique the tree gradually is split its subsequent nodes.

1. Select a small subset of available variables at random actually bootstrap
2. Select square root (K) when there are k the total number of predictors

3. If we have 500 columns of predictors choose only 23 and the split with this 23 and not with 500
4. Radically speed up with this process
5. The splitter may be best, fair or poor then we end with 2 children
6. A new list of eligible predictors set will quite different from node to node.
7. Stop at terminal node with one data record.

Bootstrap sample is used in Random Forest where Breiman followed the following steps to generate single tree with N records with replacement of original data with sampling technique and the tree consists of $\{h(x, \Theta k)$ $k = 1, 2, 3...\}$ where $\{\Theta 1k\}$are independent random vectors. The tested sample tree will be training for the growing tree where the best split is based on the input variable. In original data using bootstrap method the original tuples are drawn from where 1/3 original tuples are left out which is out of bag data. Out of bag has error estimation for each tree generation.The Generalization error of Random Forest algorithm is given as the margin function is given in equ(1),

$$mg(X, Y) = av_k\, I(h_k(X) = Y) - \max_{j \neq y} avk\, I(hk(X) = j) \quad \ldots. (1)$$

Average number of votes denoted using X,Y and the strength of Random forest is used in expected value of margin function which is calculated using equ (2)

$$S = EX, Y(mg(x, Y)) \quad \ldots. (2)$$

J48ALGORITHM

J48 is an algorithm to process the data for mining in an efficient way in classification problems. It handles noisy data and uses classification function for discrete features which can be ranged between a threshold value and formulate a rule. This rule [10] is used for handling real valued features into two ranges based on a threshold which split the decision tree output. The features returns a tree with leaf node which categories label for classification which is discussed below with the following steps,

1. DTree (features, instances) returns a tree
2. If all featured data are in one category, return a leaf node with that class label.Else
3. if the set of featured data is empty, return a leaf node with the class labelthat is the majority class in data set else
4. Choose a feature F and make a node R for it
5. For each possible value vi of F:
6. Let featuresibe the subset of features that have value vi for F
7. Add an out-going edge E to node R labeled with the value vi.
8. Iffeaturesiis emptythen attach a leaf node to edge E labeled with the class label that is the most common in data set else.
9. Call DTree(instancesi, features – {F}) and allocate the resulting tree as the subtree under edge E.

   Return the subtree rooted at R.

3. WEKA INTRINSIC ENSEMBLERS

WEKA

Weka introduced by Waikato University isopen source software which has graphical user interface for performing the classification,clustering and association rule mining algorithms for data mining tasks. For data evaluation and exploration the software is enriched with various interfaces to visualize the various algorithms. Itis mainly used for research and as an educational tool for the students in mining various data[14].

ADABOOST

AdaBoost technique reduces the error rates which combine weak classifiers to get accuracy in more active way. AdaBoost outstripped breast cancer[11] in accuracy, sensitivity, specificity which uses stratified 10 folds

cross validation.The weak learner algorithm is boosted with AdaBoost that assigns a set of weights over the training dataset. The training set ($x1, y1$) and (xn, yn) where each x_i belongs to some domain or instance space X and each label y_i belongs to $y = \{-1, +1\}$.The weights on the training example i on round t is denoted by D(i) .The same weight will set to concentrate on the hard examples in the training set. The AdaBoost algorithm is presented in the following steps,

1. Assign N example ($x1, y1$) ... (xn, yn) ; Xi $\in$ X, Yi $\in \{-1, +1\}$
2. Initialise the weights of $D_1(i) = 1/N$, $i = 1, \ldots N$
3. For $k = 1 \ldots K$
4. Train weak learner using distribution D_k
5. Get weak hypothesis h_k, $X \rightarrow R$ with its error
6. Choose $\sum_{i = h_k(x_i) \neq y_i} D_k(i)$
7. Update $D_{k+1}(i) = \dfrac{D_k(i) \exp(-a_k y_k k_k(x_k))}{z_k}$

   Where z_k is normalization factor (chosen so that D_{k+1} will be a distribution)
8. Output of the hypothesis $H(x) = \text{sign}\left(\sum_{k=1}^{k} a_k h_k(x)\right)$

   Based on the outcome of the hypothesis there are different versions of AdaBoost algorithm.

4. EXPERIMENTS AND METHODOLOGIES

Representation and quality of data is most important before running an analysis. Irrelevant and redundant data has to be preprocessed with cleaning, normalization, transformation, feature extraction and selection etc. This improves the efficiency of mining techniques when applied and hence the quality of the trained data is more efficient. In this article data preprocessing is done with normalization and discretization methods. There are no outliers in this data. No missing values. The normalization involved over here is z score analysis and discretization used here is binning. The product of data preprocessing is the final trained data.

PREPROCESSING

Normalization : The attributes are scaled to fit a specified range such as -1.0 to 1.0 or 0 to 1.Here we have used 0 to 1 as a scale of range. There are no missing values and outliers in this dataset.

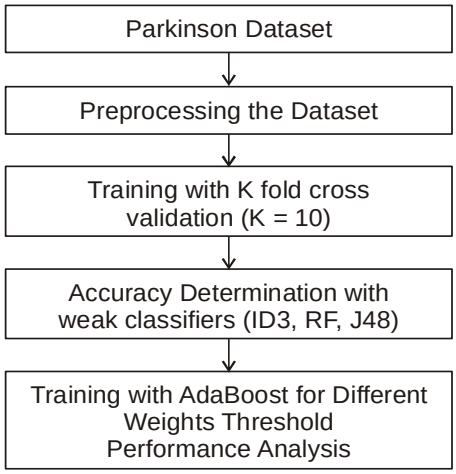

Fig. 1. Graphical Representation of the working process.

Binning method : The data is smooth by consulting the neighborhood data or values around it and the sorted values are distributed into a number of buckets or bins and they perform local smoothing method.After preprocess the data is trained with *k* fold cross validation which set k value to 10. The weak classifiers are trained to obtain the accuracy strategy and the AdaBoost is used to train with different weights initialized to get the performance of various weak classifiers.The following figure 1. Shows how the data is experimented with flow of information from initial test to performance evaluation of the data.

In this article, Parkinson diseases dataset was gathered from UCI repository [12] which has 195 instances of 23 features with 147 as diseased and 48 without disease. The raw data has the frequency measures of biomedical voice measurements from 31 people where each column indicates the voice recording from healthy/non healthy people. The status column indicates the class attribute. The attributes and their descriptions are listed in the following Table 1.

Table 1. Attribute Description

Patient Details	***Field Name***	***Description***
Vocal Frequency	MDVP: Fo(Hz)	Fundamental average frequency
	MDVP: Fhi(Hz)	Fundamental maximum frequency
	MDVP: Flo(Hz)	Fundamental minimum frequency
Fundamental Frequency	MDVP: Jitter(%), MDVP: Jiter(Abs) MDVP: RAP MDVP: PPQ Jitter: DDP	Several measures of variation
Amplitude	MDVP: Shimmer MDVP: Shimmer(dB) Shimmer: APQ3 Shimmer: APQ5 MDVP: APQ Shimmer: DDA	Several measures of variation
Ratio of noise	NHR HNR	Two measures of tonal components in the voice
Class Complexity	Status RPDE D2	Health status of the subject (Healthy-C0 / Parkinson's - C1) Two nonlinear dynamical measures
Signal	DFA	Signal fractal scaling exponent
Nonlinear Measures	spread1 spread2 PPE	Three fundamental frequency variation

Preprocessed data is fed to various weak classifiers and itsperformance correctnessis found. Then applied AdaBoost with its threshold weights are changed. The effectiveness is visualized either through graph or confusion matrix. The objective of this article is to boost the various classifiers such as Id3, Random Forest,J48 with its accuracy,precision,recall and F-Score.Performance of the algorithms are evaluated using 10 fold cross validation and results are analyzed in a confusion matrix. The general structure of the confusion matrix is given in Table 2.The accuracy of class1(C0) and class2(C1) is listed with the help of weka experimenter in Table3.

Table 2. Structure of confusion matrix

	Predicted Class	
Actual Class	***Yes***	***No***
Yes	True Positive(TP)	False Positive(FP)
No	False Negative(FN)	True Negative(TN)

Table 3. Confusion Matrix for Each Classifier

Algorithm		***Confusion Matrix***		
ID3		***C0***	***C1***	***Sum***
	C0	126	21	147
	C1	9	39	48
Random Forest		***C0***	***C1***	***Sum***
	C0	135	12	147
	C1	11	37	48
J48		***C0***	***C1***	***Sum***
	C0	133	14	147
	C1	6	42	48

The various performance measure calculations are listed below,

Accuracy : Percentage of testing set examples correctly classified by the classifier.

$$TP = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision : It is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved in the database.

$$TP = \frac{TP}{TP + FP}$$

Recall : It is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is expressed in percentage.

$$TPR = \frac{TP}{TP + FN}$$

Fmeasure : Fmeasure is defined as the harmonic mean for precision and recall

$$Fmeasure = \frac{2*(precision* recall)}{(precision + recall)}$$

Kappa : Kappa is a measure of true agreement. It indicates the proportion of agreement beyond that expected by chance [13].

$$Kappa = \frac{Total\ Accuracy - Random\ Accuracy}{(1 - Random\ Accuracy)}$$

Before applying the AdaBoost algorithm the various performance measures of ID3, Random Forest and J48 are shown in Table 4.

Table 4. Performance Measure of various classifiers

Performance Measure	*ID3*	*Random Forest*	*J48*
Accuracy (%)	84.61	88.20	89.74
Precision	0.87	0.88	0.90
Recall	0.86	0.88	0.89
Fmeasure	0.86	0.88	0.90
Kappa	0.65	0.68	0.73

Among the three classifiers J48 produced higher performance with respect to performance metrics considered in this work. The training of weak classifiers with different weights Threshold are given to AdaBoost which enhances the accuracy performance of various classifiers and are tabulated in Table 5, 6and 7 using weka tool.

Table 5. Performance Measure of ID3 with AdaBoost

Performance Measure	*ID3 Different Weights Threshold*					
	100	*90*	*80*	*70*	*60*	*50*
Accuracy (%)	87.69	88.71	89.23	88.20	88.20	88.20
Precision	0.93	0.94	0.94	0.89	0.89	0.89
Recall	0.89	0.90	0.91	0.88	0.88	0.88
FMeasure	0.91	0.92	0.92	0.88	0.88	0.88
Kappa	0.68	0.70	0.71	0.69	0.69	0.69

Table 6. Performance Measure of Random Forest with AdaBoost

Performance Measure	*Random Forest Different Weights Threshold*					
	100	*90*	*80*	*70*	*60*	*50*
Accuracy (%)	89.23	88.71	88.71	88.20	88.20	88.20
Precision	0.89	0.89	0.89	0.88	0.88	0.88
Recall	0.89	0.88	0.88	0.88	0.88	0.88
FMeasure	0.89	0.88	0.88	0.88	0.88	0.88
Kappa	0.71	0.70	0.70	0.69	0.69	0.69

Table 7. Performance Measure of J48 with AdaBoost

Performance Measure	*J48 Different Weights Threshold*					
	100	*90*	*80*	*70*	*60*	*50*
Accuracy (%)	90.76	89.23	89.23	88.20	88.20	88.20
Precision	0.91	0.90	0.90	0.89	0.89	0.89
Recall	0.90	0.89	0.89	0.88	0.88	0.88
FMeasure	0.91	0.89	0.89	0.88	0.88	0.88
Kappa	0.76	0.72	0.72	0.70	0.70	0.70

From Table 5, it is implicit that the accuracy of ID3 is significantlyincreased with threshold weight of 80.

In case of Random Forest it is evident from Table 6 that the accuracy increases with a threshold weight of 100 and from Table 7 the J48 classifier has an accuracy value of 90.76% with 100 as threshold value. Hence the threshold weights of AdaBoost algorithm emphasize a strong change in case of ID3 when compared to J48 and Random Forest.

5. CONCLUSION

The AdaBoost is a boosting technique which strengthens the accuracy of various classifiers with pruned weighted techniques. The weak classifiers are combined with AdaBoost which make the classification more powerful. ID3 showed an accuracy of 89.23% from 84.61% with the help of boosting technique which make it to be trained effectively. In case of Random Forest and J48 the accuracy of were increased from 88%to 89% and 89%to 90%. The performance measures such as precision,recall,fmeasure and kappa are also evenly increased with the experimental results. The kappa statistic for j48 remains a good agreement for the closeness of the positive class.Thus the AdaBoost makes the classifiers to give a good performance measure with different weights threshold. The Parkinson healthcare is an imbalanced dataset can be balanced using different techniques for future work and the cost effectiveness could be found formisclassification instances in medical domain area. Also the stability of the framework can be validated by testing the model or framework with number of datasets in the medical domain. Further the application of neuro fuzzy classification can be explored.Various cost sensitive methods like AdaC1, AdaC2 and AdaC3 can be used in future to estimate the cost sensitive learning methods to update the associationin boosting algorithms.

6. REFERENCES

1. Billah, Muhtasim, RubelBiswas, and MdZahangirAlam. “Symptom Analysis of Parkinson Disease using SVM-SMO and Ada-Boost Classifiers.” BRAC University, Dhaka, Bangladesh (2014).
2. AlZoubi, Omar, Irena Koprinska, and Rafael A. Calvo.“"Classification of brain-computer interface data”, Proceedings of the 7th Australasian Data Mining Conference-Vol. 87, Australian Computer Society, Inc., 2008.
3. MriduSahu, N.K. Nagwani, ShrishVerma, and SaranshShirke,Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal, International Journal of Knowledge Engineering, Vol.1(2), September 2015.
4. Costa, Joao, and Jaime S. Cardoso. “An ADABOOST variant for Ordinal Classification.”, pages 68-75.
5. Ahmed, Kawsar, and TasnubaJesmin. “Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach.” International Journal of Science and Engineering 7.2 (2014): 155-160.
6. Kulkarni, Rohan D. “Using Ensemble Methods for Improving Classification of the KDD CUP'99 Data Set.”Journal of Computer Engineering,vol.16(5):57-61.
7. Tiwari, Aakash, and AdityaPrakash, “Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using Weka”, International Journal of Engineering and Technical Research, ISSN: 2321-0869, Vol. (6), September 2014.
8. Balamurugan M and K.Viji. “Predicting Eligible Educator Category for Disability Student Welfare using Decision Tree Method.” International Journal of Computer Applications 119.12 (2015).
9. Liaw, Andy, and Matthew Wiener. “Classification and regression by Random Forest.” R news 2.3 (2002): 18-22.
10. Tina R.Patil, Mrs. S.Sherekar, Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification, International Journal of Computer Science and Applications , Vol. 6(2), April 2013.
11. Thongkam, Jaree, et al. “Breast cancer survivability via AdaBoost algorithms.” Proceedings of the second Australian workshop on Health data and knowledge management-Vol.80. Australian Computer Society, Inc., 2008.
12. https://archive.ics.uci.edu/ml/datasets/Parkinsons
13. Julius Sim, chris C Wright,The Kappa Statistic in Reliability Studies: Use, Interpretation and Sample Size Requirements, Journal of the American Physical Therapy Association, March 2005, vol. 85(3).
14. https://weka.waikato.ac.nz/explorer
15. Dhakate, Payal, K. Rajeswari, and DeepaAbin. “Analysis of Different Classifiers for Medical Dataset using Various Measures.” International Journal of Computer Applications 111, no. 5 (2015).
16. K.Lokanayaki and Dr. A. Malathi, “A Prediction for classification of Highly Imbalanced Medical Dataset using Databoost.IM with SVM”, International Journal of Advanced Research in Computer Science and Software Engineering,April 2014,Vol. 4(4).