RESEARCH ARTICLE OPEN ACCESS

Adaptive Process For Bagging Based Ensembles

Rajni David Bagul [1], Prof. Dr. B. D. Phulpagar [2]
PES Modern College of Engineering
Savitribai Phule Pune University
Shivajinagar, Pune
India

ABSTRACT

Ensemble of classifiers is the approaches of multiple classifiers are learned from same dataset and this multiple trained classifiers are used to predict the unlabelled data. The Performance of using Ensemble of classifiers is best than using single classifier. The Bagging stands for Bootstrap aggregation are the approaches of ensemble. In bagging number of bags(n) are formed where each bag contains number of instances (m) from training examples. Formation of bags is referred as bootstrap sampling in which m instances are selected randomly from training data and instanced can be repeated. Once n bags are formed, no of classifiers(c) are trained using n bags. c classifiers are used for further prediction. It has some drawbacks such as all classifiers are considered as of equal importance; there is no method optimal bag creation. Same problem was faced by Random subspace classifier based ensemble (RSCE) and solved by applying Hybrid Adaptive ensemble learning framework (HAEL).In proposed work, to solve issues in the Bagging HAEL framework is applied.From experimental results, Bagging with HAEL improves classification accuracy than simple bagging.

Keywords:- Ensemble learning ,Bagging, Random Subspace.

I. INTRODUCTION

There are many ensemble approaches described as follows.

1. Boosting

In boosting weight is assign to the training instances that determines how the instances was classified in the previous iteration. The subspace is badly classified are inserted in the training dataset for the next iteration. The instances which are hard to classify on which boosting pays more attention. Boosting does not provide any mechanism to improve the learning of base classifiers.

2. Bagging

It is introduced by Breiman and it is sampling based ensemble classifier approach. From the entire training set bagging generate the multiple base classifiers by training them on the subsets randomly dawn. The predictions of the base classifiers are combined into a final prediction by majority voting. The bagging only suitable for small dataset. The sampling procedure of bagging creates many subsets by bootstrap sampling which results variations in base classifiers.

3. Random Subspace

Random subspace is an ensemble creation method in which use of feature sub sets to create the different data subsets to train the base classifiers. One of the important ensemble approaches is Random subspace classifier ensemble. In RSCE feature/attribute set is sub spaced randomly and for each sub space, classifier is constructed using any learning algorithms. These constructed classifiers are used to classify the test instance with voting majority approach. RSCE method has two major limitations.

- Classifiers are distributed equally without depending on which classifier is constructed of which subspace. For example s1, s2 are two subspace and c1, c2 are classifiers constructed using s1, s2 respectively. s1 contains important attributes and s2 does not contains any important attributes then also s1,s2 are treated equally.

- Which subspace should be selected so that it will increase the accuracy i.e. Sub space selection is completely random. Sometimes due to some irrelevant subspace selection, irrelevant classifiers are constructed and hence results are irrelevant.

To overcome above limitation the drawbacks and proposed Hybrid adaptive ensemble learning (HAEL) framework and applied to RSCE. HAEL includes the two adaptive process 1) Base classifier competition adaptive process (BCCAP) 2) classifier ensemble interaction adaptive process (CEIAP). First adaptive process gives weightage to the classifiers according there importance. To select the optimized subspace is describe in second adaptive process.

II. LITERATURE SURVEY

In literature ensemble of classifiers have many approaches such as Boosting[1],Bagging[2],Random forest[3],Random subspace[4].To improve the classification accuracy ensemble of classifiers integrates multiple classifiers to classify the instance.

A) Boosting

To improve the classifiers accuracy in which subsequent classifier models are trained on misclassified instances of previous classifier model. The set of developed classifiers is used for predicting the labels of the instances using voting majority.

B) Bagging

In bagging n bags are formed where each bag contains m instances from training examples. Formation of bags is referred as bootstrap sampling in which m instances are selected randomly from training data and instanced can be repeated. Once n bags are formed, n classifiers are trained using n bags. C classifiers are used for further prediction.

C) Random subspace

Features in the training dataset are sub-spaced randomly. For example there are n features in the training data then to form subspace any m features are selected randomly where m << n. Number of subspace (S) are formed then data is sub-spaced according to feature subspace. This sub-spaced datasets are used to form S classifiers. Work related to ensemble of classifiers can be divided into three categories. First category is about design of the ensemble approaches; second category is about how to improve existing ensemble solutions.

Design of the Ensemble approaches:

The problem of space complexity of ensemble focuses on instance selection to reduce the space complexity. Aim of the instance selection is that selected instances and whole dataset should yield classifiers which will give same results. Instance selection method is combined with boosting and proposed generic ensemble approach with instance selection. From the experimental results it is clear that proposed approach in this paper can form simple and better ensembles than Random subspace based knn ensemble, other traditional ensemble approaches for c4.5 and SVM are describe in N.Garcia-Pedrajas [6].

The drawbacks of using undirected bi-relation graph for prediction describe in G.Yu,C. Domenoconi[7].To overcome this problem paper proposed a system which used directed bi-relation graph with transductive multi-label classifier. Ensemble of TMC is used then to improve the accuracy of prediction. Different from traditional methods that make use of multiple data sources by data integration, TMEC takes advantage of multiple data sources by classifier integration.

In this paper proposed cluster based ensemble of classifiers. Base classifiers give boundaries of clusters then cluster confidences are mapped to class decision is described in B. Verma and A.Rahman[8].Training dataset is categorized into clusters using labels, these categorized data used to train the classifiers. Base classifiers evaluate the cluster boundaries and generate cluster confidence vectors. Proposed system makes learning efficient by modifying domain for classifiers.

Work to improve the ensemble approaches

In this paper method which evaluates how many classifiers in ensemble are needed to be queried for prediction of complete ensemble. In general, unlabeled instance is classified by all classifiers in the ensemble describe in D. Hernandez-Lobato [9]. There is no need to classify instance with all classifiers in the ensemble, only subset of classifier can predict the prediction of all classifiers. Voting process is stopped when probability of prediction change is below threshold. Number of queries to be done which will reduce the probability of change below threshold depends on the instance. Instances which are on classification boundary need more queries. This method is called as instance based pruning and can be combined with any ensemble approach.

Distinct pruning strategies which are designed to improve ensemble of classifiers are analysed is describe in paper [10]. In pruning selects the subset of function in the ensemble which can perform better than whole ensemble. The approach aggregation order is random therefore generalization error decreases as number of functions is increased describe in simple bagging. Many pruning methods are based on modifying order of aggregation of classifiers ensemble. Performance of ensemble of classifiers can be improved by ordering aggregation to minimize the generalization error. Performance of the ensembles with pruning strategies are is better than normal ensembles.

In this paper investigated the effect of the accuracy and the diversity in the design of a multilayer perceptron (MLP)-based classifier ensemble describe in paper [11]. They also considered how to reduce added classification errors in a classifier ensemble based on the Walsh coefficients. Kuncheva et al. [12] studied how to use a kappa-error diagram to analyze the performance of classifier ensemble approaches.

The drawbacks of RSCE (discussed in introduction) and proposed Hybrid adaptive ensemble learning (HAEL) framework and applied to RSCE to overcome above limitation describe in paper [5]. HAEL includes the two adaptive process 1) Base classifier competition adaptive process (BCCAP) 2) classifier ensemble interaction adaptive process (CEIAP). First adaptive process gives weightage to the classifiers according there importance. To select the optimized subspace is describe in second adaptive process.

Applications of the Ensemble
In this paper non-parallel plane proximal classifier ensemble which gives more accuracy than single non-parallel plane proximal classifier. This ensemble is applied to classify unknown tissue samples using known gene expressions as training data are describe in paper [13]. Genetic algorithm based scheme is used to train non-parallel plane proximal classifier. To predict the unlabelled data, classifiers with positive performance are selected. To aggregate the prediction results minimum average proximity-based decision combiner is used. System is compared with SVM proves that it gives comparable accuracy with less average training time. Takemura et al. [14] designed combined heterogeneous classifier ensembles using a kappa statistic diversity measure, and applied it to the electromyography signal datasets. Takemura et al. [14] used the classifier ensemble approach to identify breast tumours in the ultrasonic images. In the area of data mining, Windeatt et al. [15] applied the MLP ensembles to perform feature ranking. Rasheed et al.

III. SYSTEM ARCHITECTURE

A) Bag of instance creation-

In this section N bags of instances created from training dataset Dtr. Training dataset Dtr contain m records. Size m' of the bag is taken as input from user. m' instances are selected randomly to create the bag of the instances. In such way n bags are created.

B) Classifier learning

N classifiers are trained on n bags created in previous step. In our work, system used J48 i.e. weka implementation of C4.5 decision tree algorithm. Specifically,

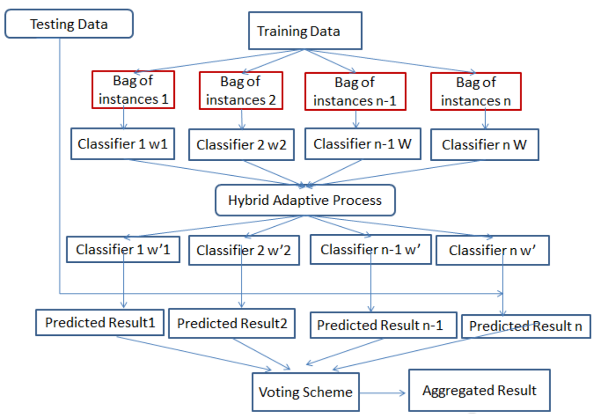
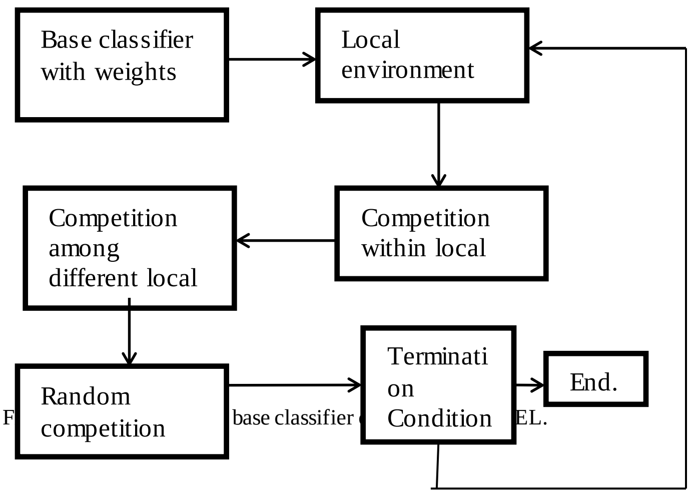


Fig.1.Overview of hybrid adaptive ensemble learning (HAEL) for the random subspace-based classifier ensemble approach.

To represent an attribute the decision tree adopts an internal node, uses an outgoing branch of the node to represent an attribute value, and associates each leaf node with a classification. Most of the decision trees apply the information gain, which is calculated from the entropies in the information theory, for performing the gain, which is calculated from the entropies in the information theory, for performing the Learning process c1, c2, c3, cn classifiers are generated in this step. When instance i is to be classified it is classified by all n classifiers and output of classification of instance I by c1 is denoted by y1.So classification step gives y1, y2, yn. Majority voting is done to get final classification result by ensemble .n classifiers are learned.

C) Hybrid adaptive ensemble learning process

1) Base classifier competition adaptive process
Weight is evaluated for each classifier in the ensemble describe in this step. From previous step we get set of n classifiers.



WI = 1/n is the initial weight for each classifier. First step of this process is to assign the initial weight to the each classifier. In the next step T training instances are selected randomly from main training data(Dtr). Each selected instance is classified by each classifier. Each classifier maintains predicted values of each instance. Indicator vector IV is generated for each classifier is generated whose length is equal to T. ith Value in the indicator vector denotes the whether ith predicted value was correct or not. 0 values indicate correct prediction and 1 indicates error. Instances S which are predicted correctly by some classifiers and predicted wrongly by some classifiers are considered for the further procedure.

Weight of the classifier is increased if it predicts the S correctly otherwise reduced. Predicted error of the classifiers for the i-th training samples is calculated as follows:

$$Ei = \sum_n Wn * IVin / \sum_n Wn \qquad (1)$$

Where IVin indicates value in the indicator vector for ith instance and nth classifier

Cumulative error of samples is calculated as

$$E = \sum_{i=1}^{T'} Ei \qquad (2)$$

Where T' are instances whose prediction is varied across the classifiers. T' < T. Local environments are created as per described in the [5]. Each classifier is assigned to the local environment. Three operators are used for weight calculation of classifier. 1) Competition among classifiers in the same local environment (CSLE) , the competition among the classifiers In different local environments (CDLE), and the random competition (RC), and executes three operators one by one.

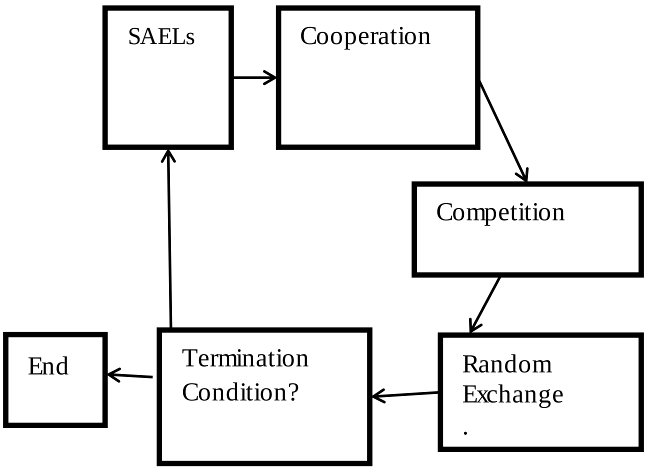


Fig.3System Architecture: Adaptive process for bagging based ensemble

In CDLE classifiers two distinct local environments are compared and weight is updated as shown below. In CSLE classifiers in same local environments are compared and weight is updated as shown below. In RC, any two random classifiers are compared and weight is updated as shown below.

- If b and j are two classifiers and Wb and Wj is prior weight respectively.

- AccB and AccJ is predicted accuracy of b and j classifier then

- Wb' = Wb + a1 Wb If (AccB > AccJ)
- Wb' = Wb if (AccB= AccJ)
- Wb' = Wb – a1 Wb if (AccB < AccJ)
- Wj' = Wj + a1 Wj if (AccJ > AccB)
- Wj' = Wj if (AccB= AccJ)
- Wj' = Wj – a1 Wj if (AccJ < AccB)

– Where Wj' and Wb' is updated weight

Above process is repeated until termination condition is reached for each local environment. Termination condition may be any of the following the number of iterations is larger than the maximum number of iterations pre-specified by the user and 2) the weight vector does not change in several iterations after termination of the iterative process average weight of the classifiers is calculated. Classifiers whose weight is above average are reset to 1 otherwise it is set to 0.

2) Classifier ensemble interaction adaptive process

Set of Ensembles is given as input in this step. Set of ensembles are updated with aim to optimize the bags. All ensembles are with updated weight of classifiers using BCCAP. Ensemble of classifiers is considered as classifier in this step. S = {S1, S2, S3…Sn} is set of ensembles. Selected instances for training of ensemble of classifier (n) are indicated by indicator vector Vpn, 0 denotes that instances is not used by ensemble else 1. Aggregated indicator vector Vp' for pth ensemble defined as follows.

$$Vp' = \sum_{b} Wb \; Vpb \qquad (3)$$

Where Wb is weight values of bth classifier

For calculating the similarity between two ensembles each ensemble is having one aggregated indicator. Knn of ensemble is calculated using aggregated indicator and considered as Local environment of the ensemble.

Three operators are used for interaction in ensemble of classifiers.

- Cooperation operator – Input to this operator will be two ensembles from same local environment. First all the classifiers in both ensembles are sorted according to their weight. Half number of classifiers from both sites is selected to form new ensembles.

- Competition Operator – Input to this operator will be three ensembles which are selected randomly irrespective of local environment. Ensemble which has lowest accuracy on training dataset is removed from set of ensembles in this process.

- Random operator - Input to this operator will be single ensemble. Any bag in this ensemble is replaced by any random bag and its respective classifier. If performance of newly generated ensemble is less than old one, newly generated ensemble is used for further process otherwise change is discarded.

Above three operators are executed until termination condition is reached which is given by user. Output of this step will be S with an updated ensembles. Ensemble of classifier with minimum cumulative error is selected as optimal ensemble.

4) Testing

In this step instances are tested using ensemble of classifier obtained in above process.

IV. EXPERIMENTAL RESULT

Dataset Name	HBBE	BBE
Australian	85.50%	84.04%
Bupa	66.81%	60.68%

TABLE 1: Comparison Of Accuracy (in %) Of HBBE And BBE

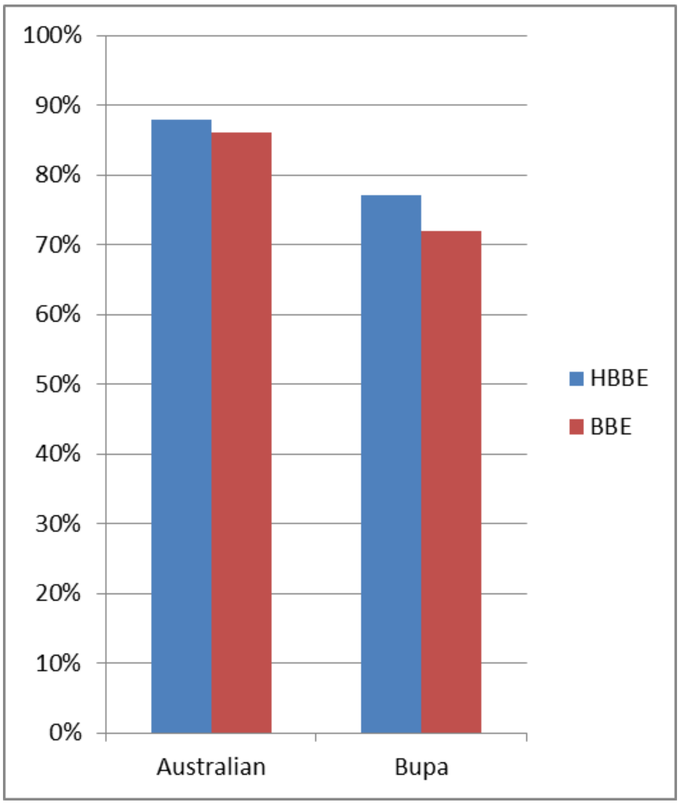


TABLE 2: Comparison Of Accuracy Of HBBE And BBE

A) Experimental Setup

Goal of the experimental evaluation is to check the effectiveness of the proposed approach. In proposed work HAEL is applied on Bagging based ensemble. Accuracy, time and memory requirements of the Bagging based ensemble with HAEL (HBBE) and existing bagging based ensemble (BBE) are compared. Random subspace based ensemble with HAEL (HRSBE) and normal random subspace based ensemble (RSBE) is also implemented. Comparison of results of HRSBE and RSBE is done in [5]. HRSBE gives more accuracy than RSBE but it takes more time and memory than RSBE. From those results [5] it was expected that HBBE will give more accuracy than BBE. For experiment, number of SAEL was set 5, number of classifiers for each SAEL was set 5, termination condition for CEIAP was 5 and termination condition for BCCAP was 5. In all cases J48 will be used as base classifier. Bagging size was set equal to 50%. Table 1 and Graph 1 show the obtained results. Australian and Bupa datasets from KEEL [17] repository were used for experiments.66% data was used for training purpose and 34 % data was used for testing purpose.

B) Results

BBE gives 60.68% accuracy on bupa dataset and HBBE gives 66.81% accuracy on bupa dataset. For Australian dataset BBE gives 84.04C% accuracy and HBBE gives 85.50% accuracy. From result table 1 and Graph 1 it is clear that HBBE gives more accuracy than simple BBE. Time requirement and memory requirement were observed for both systems. Obliviously HBBE takes much more time and memory than simple BBE.

V. CONCLUSIONS

This paper describes the effectiveness of ensemble learning, various approaches for ensemble learning, Problems in Random subspace ensemble learning and method to overcome this method. We observed bagging based ensembles also faces same problems as RSCE and proposed method HBBE to solve the problem in bagging based ensembles. Bagging Based ensemble by using HAEL approach gives results into increase in prediction accuracy.

ACKNOWLEDGMENT

Rajani Bagul received the B.E in Information Technology from Savitribai Phule Pune University, Pune. From 2007 – 2013. and M.E.(Appear) Degree in Computer Engineering from PES Modern College of Engineering. Pune University 2014.

REFERENCES

[1] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput Syst Sci.,vol. 55, no. 1, pp.119–139, 1997

[2] L. Breiman, "Bagging predictors," Mach. Learn., *vol*.24, no. 2, pp. 123–140, 1996.

[3] L. Breiman, "Random forests," Mach. Learn., vol. *45*,no. 1, pp. 5–32, 2001.

[4] T. K. Ho, "The random subspace method forconstructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell., vol. 20*, no. 8, pp. 832–844, Aug.1998.

[5] Zhiwen Yu,"Hybrid Adaptive ClassifierEnsemble"Senior Member, *IEEE, Le Li, Jiming Liu,Fellow, IEEE, and Guoqiang Han,vol 45*,NO.2,February 2015.

[6] N. García-Pedrajas, "Constructing ensembles of classifiers by means of weighted instance selection," *IEEE Trans. Neural Netw., vol. 20*, no. 2, pp. 258–277, Feb. 2009.

[7] G. Yu, C. Domeniconi, H. Rangwala, G. Zhang, and Z. Yu, "Transductive multi-label ensemble classification for protein function prediction," in Proc.18th ACM SIGKDD KDD, New York, NY, USA,2012, pp. 1077–1085.- archive/macros/latex/contrib/supported/IEEEtranc.

[8] B. Verma and A. Rahman, "Cluster-orientedensemble classifier: Impact of multiclustercharacterization on ensemble classifier learning,"*IEEE Trans. Knowl. Data Eng., vol*. 24, no. 4, pp.605–618, Apr. 2012 "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.

[9] D. Hernandez-Lobato, G. Martinez-Muoz, and A.Suarez, "Statistical instance-based pruning inensembles of independent classifiers," *IEEE Trans.Pattern Anal. Mach. Intell., vol*. 31, no. 2, pp. 364–369, Feb. 2009.

[10] G. Martinez-Muoz, D. Hernandez-Lobato, and A.Suarez, "An analysis of ensemble pruning techniques based on ordered aggregation," *IEEE Trans. Pattern Anal. Mach. Intell., vol. 31*, no. 2, pp. 245–259, Feb.2009.

[11] T. Windeatt and C. Zor, "Minimising added classification error using Walsh coefficients," *IEEE Trans. Neural Netw., vol. 22*, no. 8, pp. 1334–1339, Aug. 2011.

[12] L. I. Kuncheva, "A bound on kappa-error diagramsfor analysis of classifier ensembles," *IEEE Trans.Knowl. Data Eng., vol. 25*, no. 3, pp. 494–501, Mar.2013.

[13] S. Ghorai, A. Mukherjee, S. Sengupta, and P. K.Dutta, "Cancer classification from gene expression data by NPPC ensemble," *IEEE/ACM Trans.Comput. Biol. Bioinf., vol. 8*, no. 3, pp. 659–671,May–Jun. 2011.

[14] A. Takemura, A. Shimizu, and K. Hamamoto,"Discrimination of breast tumors in ultrasonic images using an ensemble classifier based on the adaboost algorithm with feature selection," *IEEE Trans.Medical Imaging, vol. 29*, no. 3, pp. 598–609, Mar.2010.

[15] T. Windeatt, R. Duangsoithong, and R. Smith,"Embedded feature ranking for ensemble MLP classifiers," *IEEE Trans. Neural Netw., vol. 22*, no. 6,pp. 988–994, Jun. 2011.

[16] S. Rasheed, D. W. Stashuk, and M. S. Kamel,"Integrating heterogeneous classifier ensembles forEMG signal decomposition based on classifieragreement," *IEEE Trans. Inf. Tech. Biomed*., vol. 14,no. 3, pp. 866–882,May 2010.

[17] http://sci2s.ugr.es/keel/datasets.php