DIVERSITY-INDUCED WEIGHTED CLASSIFIER ENSEMBLE LEARNING

Y. Dong, X.-J. Shen [*], *L.-J. Wang, D. Wornyo*

JiangSu University
ZhenJiang, Jiangsu, 212013, China

Z.-J. Zha

USTC, Hefei,
Anhui, 230022, China

ABSTRACT

Ensemble Learning is widely accepted as an effective technique to improve accuracy and stability of a single classifier. Classifier ensemble generally should combine diverse component classifiers. Accuracy and diversity are two key factors to decide the ensemble generalization error. In this paper, we propose a novel approach to make the tradeoff between accuracy and diversity by maximizing accuracy and diversity simultaneously in an ensemble classifier. Our proposed method is a minimization convex optimization problem. Experimental results on a variety of UCI and artificial datasets have shown that, our proposed method has advantage of superior performances in keeping higher classification results than other ensemble methods, such as Random Forest, AdaBoost, EnsembleSVM and Weighted Classifier Ensemble method Based on Quadratic Forms(QFWEC).

Index Terms— ensemble learning, weighted classifier, classifier diversity

1. INTRODUCTION

Ensemble learning is a technique that combines multiple classifier results in supervised learning to improve the accuracy and stability of an individual classifier. Several studies conducted on its real-world applications proves or shows that [1, 2, 3], classifier ensemble can achieve better classification performance than a single classifier, supported by theoretical or experimental results. Consequently, a variety of ensemble based approaches have been investigated and proposed, this include Bagging [4], Boosting [5] and Random Forest [6].

The most important concern in classifier ensemble learning research is how to improve the accuracy of classifier ensembles. And there are many methods [7, 8, 9] which use different strategies to achieve accuracy of classifier ensembles. These strategies are roughly divided into two categories. One emphasizes the construction of individual classifiers, and the other emphasizes the combination of individual classifiers. The first category focuses on generating different training subsets for individual classifiers. For example, Bagging is a classic ensemble algorithm, which produces different classifiers by selecting subsets from the original data set repetitiously [4]. AdaBoost is another common ensemble and iterative algorithm [5] that allows a new classifier to be generated from the training dataset in each iteration; it further classifies all samples to assess the importance of each sample. The weight of the sample will be higher in the next training. The whole process will not end until the error rate is small enough or up to a certain iteration number. For example, Fazakis et al. proposed Self-trained Rotation Forest for semi-supervised learning [10]. The second category focuses on how to combine the outputs of individual classifiers. The key factor of this category is the value of classifiers coefficients. Recent research on this technique shows that there are three main categories about the value of classifiers coefficients: (a) Simple vote strategy: All values of classifiers coefficients are the same. Especially, it is equivalent to majority vote. Kuncheva proposed how to combine pattern classifiers [11]. Among these strategies proposed, simple vote strategy is the most popularly used rule. (b) Weighted classifier ensemble: All values of classifiers coefficients are different and positive. In this method, each individual classifier has a different contribution for the ensemble classifier. Merz proposed to use correspondence analysis to combine classifiers [12]. The proposed method makes base classifiers have different contributions. Ueda proposed optimal linear combination of neural networks [13], and this method improves classification performance in neural networks. (c) Selective or pruning classifier ensemble: The value of each classifier-s coefficient can be zero or negative. It indicates that some individual classifiers have insignificant or negative effects on improving the performance of the ensemble classifier. Mao et al. proposed a weighted classifier ensemble method based on quadratic forms (QFWEC) [14]. In QFWEC, they introduce an error term with a weight vector, and subtract this error with the quadratic form to obtain approximated error. This subtraction makes minimizing the approximation form equivalent to maximizing the original quadratic form. Zhang and Zhou proposed a framework of sparse ensembles [15] that deals with new linear weighted combination methods for sparse ensembles.

Furthermore, another unique factor which has a classifier diversity that can be considered in designing classifier ensemble is classifier diversity. The accuracy of classifier-

[*] Email: xjshen@ujs.edu.cn

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

s must be of high quality since several poor or weak classifiers can suppress correct predictions of good classifiers. To achieve diversity among classifiers, individual classifiers must be set to uncorrelated error, since these uncorrelated errors among classifiers can help an ensemble rectify errors in some classifiers; this process makes the ensemble achieve better classification performance than individual classifiers. Several diversity- based classifiers were proposed based on the above. Zhang et al. proposed a Corrective Classification (C2) [16], which incorporates error detection, data cleansing and Bootstrap sampling to construct diverse base classifiers that constitute the classifier ensemble. Yin et al. argued that diversity, not direct diversity on samples but adaptive diversity with data, is highly correlated to ensemble accuracy [17]. They proposed a technology for classifier ensemble, learning to diversify, which learns to adaptively combine classifiers. Emilie Morvant et al. paid attention to Majority Vote of Diverse Classifiers [18]. They improve Mean Averaged Precision measure for late fusion while considering the diversity of the voters. Qi et al. presented an Ex-Adaboost learning strategy [19], and then proposed a new Deep Support Vector Machine (DeepSVM) with highest diversity for classification. The Ex-Adaboost strategy improves DeepSVMs performance. Xiao et al. proposed an Ensemble classification based on supervised clustering (ECSC) [20]. In ECSC, supervised clustering is employed to form a number of training subsets, diverse base classifiers are generated. The proposed ensemble method can effectively improve the accuracy of credit scoring.

The above discussions motivate our idea on building diversity-induced weighted Classifier ensembles. We propose a new diversity-based ensemble approach. Our contributions are as follows: Firstly, the proposed method makes the tradeoff between accuracy and diversity in ensemble classifier. We propose to learn classifier weights by optimizing direct but simple criteria: maximizing accuracy and diversity simultaneously [17]. In addition, we formulate this procedure in a convex minimization problem. Secondly, extensive experiments are carried out on a variety of UCI and artificial datasets to have advantage of superior performances in keeping higher classification results than other ensemble methods, such as Random Forest, AdaBoost, EnsembleSVM [21] and QFWEC.

This paper is structured as follows. In Sect. 2 the proposed method to balance accuracy and diversity is elaborated. In Sect. 3 the experimental results are presented and discussed. Finally Sect. 4 concludes this paper.

2. PROPOSED METHOD

In this section, we introduce a new weighted ensemble method to balance accuracy and diversity. Accuracy and diversity are the two key factors that are used to evaluate generalization errors of ensemble classifier [14]. In our proposed method, the optimal weight vector of base classifiers is obtained by maximizing the ensemble accuracy and diversity simultaneously. It is modeled to weight each classifier in a convex minimization way with balancing between accuracy and diversity.

Firstly, the accuracy of an ensemble system is equal to the probability of samples whose predictive labels are the same as their true labels. It is an important factor in ensemble system. Suppose a two-class classification problem (classes:[-1,+1]) has a training set $\mathbf{X}$ with class labels ($\mathbf{X} = \{(x_1, y_1), ..., (x_N, y_N)\}$) and a testing set $\mathbf{X}_t$ without labels ($\mathbf{X}_t$=$\{x_1, ..., x_M\}$), where $x_n(x_n \in R^d, n = 1, ..., N)$ expresses a training sample, y_n is the true class label of x_n, and $x_m(x_m \in R^d, m = 1, ..., M)$ expresses a testing sample. Then L individual classifiers are produced from the other ensemble classifiers (e.g, Random Forest), shown by a set $\mathbf{H} = \{h_1, ..., h_L\}$. In general, the error rate of an ensemble is calculated based on the difference between the final predictive labels and the true labels. The goal is to minimize the difference between the predictive and true labels for the accuracy of an ensemble classifier.

$$\begin{aligned} &\underset{w}{min} : \|\mathbf{Kw} - \mathbf{y}\|^2 \\ &s.t. \sum_{i=1}^{L} w_i = 1, -1 < w_i < 1 \end{aligned} \tag{1}$$

where $\mathbf{w}$ denotes a weight vector of individual classifiers, $\mathbf{w} = [w_1, ..., w_L]^T$, $\mathbf{y}$ is a vector of the true labels, $\mathbf{y} = [y_1, ..., y_L]^T$, $\mathbf{K}$ expresses a $N \times L$ matrix, the elements of each row in the matrix are the predicted labels by the different classifiers for the same sample, the elements of each column in the matrix are the predicted labels by the same classifier for the different samples. $\mathbf{K} = [\mathbf{k}_1, ..., \mathbf{k}_L]$, where $\mathbf{k}_j$ expresses a vector of predictive labels gained by h_j, $\mathbf{k}_j = [k_{1j}, ..., k_{Nj}]^T, j = 1, ..., L$. $\mathbf{Kw}$ denotes N predictive labels by the entire ensemble classifier. Meanwhile, two constraints($\sum_{i=1}^{L} w_i = 1$ and $-1 < w_i < 1$) should be added, because $\mathbf{Kw}$ is equivalent to an approximation of $sgn(\mathbf{Kw})$ based on two constraints.

Secondly, classifier diversity denotes the divergence among base classifiers. Through the $\mathbf{K}$ in Eq.1, instead of the original output, the oracle output $\mathbf{O}$ of the ensemble classifier is often used for ensemble optimization, which is a $N \times L$ matrix, and its element is

$$O_{ij} = \begin{cases} +1, & k_j(x_i) = y_i \\ -1, & k_j(x_i) \neq y_i \end{cases} \tag{2}$$

Each column in the matrix $\mathbf{O}$ denotes that whether each base classifier can correctly classify different samples. Generally, different base classifiers have different classification effects, so each column in the matrix $\mathbf{O}$ is also different. The diversity of the ensemble classifier is

$$\mathbf{D} = \frac{1}{2}N(N\mathbf{1}_{L\times L} - \mathbf{O}^T\mathbf{O}) \tag{3}$$

where N is the number of the samples. $\mathbf{D}$ is the diversity matrix of base classifiers, and each element in the matrix represents the difference between two base classifiers. It is worth noting that diagonal element of the matrix $\mathbf{D}$ is zero.

Our approach aims to balance accuracy and diversity for overall generalization performance. We propose the following model:

$$\begin{aligned} &\underset{w}{min}: \|\mathbf{K}\mathbf{w} - \mathbf{y}\|^2 - \alpha\mathbf{w}^T\mathbf{D}\mathbf{w} \\ &s.t. \sum_{i=1}^{L} w_i = 1, -1 < w_i < 1 \end{aligned} \tag{4}$$

where α is a priori balance factor of accuracy and diversity. Eq.4 can be transformed into Eq.5

$$\begin{aligned} min: f(w) = \frac{1}{2}\|\mathbf{K}\mathbf{w} - \mathbf{y}\|^2 - \frac{1}{2}\alpha\mathbf{w}^T\mathbf{D}\mathbf{w} \\ + \frac{1}{2}\lambda(1 - (\mathbf{1}_{L\times 1})^T\mathbf{w})^2 \end{aligned} \tag{5}$$

where λ is a regulation factor parameter that forces $\sum_{i=1}^{L} w_i = 1$. We take the derivative of Eq.5 with respect to w and get Eq.6

$$f'(w) = \mathbf{K}^T\mathbf{K}\mathbf{w} - \mathbf{K}^T\mathbf{y} - \alpha\mathbf{D}\mathbf{w} + \frac{1}{2}\alpha(\mathbf{1}_{L\times L}\mathbf{w} - \mathbf{1}_{L\times 1}) \tag{6}$$

Eq.7 can be obtained when $f'(w)$ equals to zero.

$$\mathbf{w} = (\mathbf{K}^T\mathbf{K} - \alpha\mathbf{D} + \lambda\mathbf{1}_{L\times L})^{-1}(\mathbf{K}^T\mathbf{y} + \lambda\mathbf{1}_{L\times 1}) \tag{7}$$

According to Eq.7, the optimal weight of the ensemble classifier can be obtained. We obtain an ensemble classifier by combining the various base classifiers linearly. To a certain extent, the ensemble method can balance accuracy and diversity well and it has a good generalization performance.

In the following section, experimental results will be shown.

3. EXPERIMENTS

As mentioned in Section 3, several datasets with different characteristics have been tested in order to validate our proposed ensemble method. In the experimental study, we mainly focus on two important aspects: the parameter selection in the proposed ensemble method and the performance of different ensemble classification methods.

Table 1. Descriptive information for the ten datasets

Dataset	Data Points	No. of features
Australian	690	14
Pima	768	8
Liver-disorders	345	5
Breast-cancer	683	10
Sonar	208	60
A1a	28000	14
Heart	270	13
Emotions	593	72
Scene	2407	294
German	1000	24

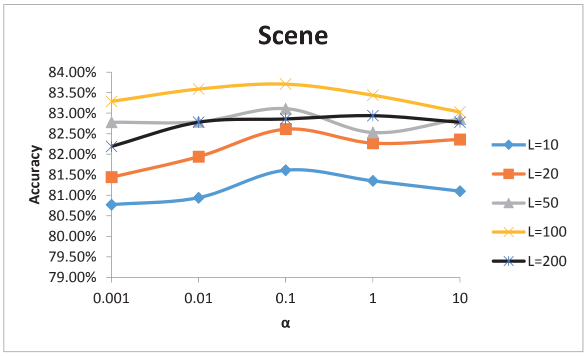

Fig. 1. The average accuracy of our proposed ensemble method on Scene dataset by different parameter selections

3.1. Parameter Selection

To evaluate the effectiveness of the Diversity-Induced weighted classifier ensemble learning, we used ten public datasets from different domains, named Australian, Pima, Liver-disorders, Breast-cancer, Sonar, A1a, Heart, Emotions, Scene, and German. The Australian, Pima, and Heart datasets are extracted from UCI[1]; the Liver-disorders, Breast-cancer, Sonar, A1a, Emotions, Scene, and German datasets are obtained by LIBSVM Data[2]. These datasets are processed into binary data. Basic descriptive information about datasets used in experimental analysis appears in Table 1.

The λ is a priori value and its value is set to 10000. The values of parameters α and L are decided based on each dataset in our experiments. The parameter α is a coefficient of diversity and it is as a tradeoff among various base classifiers. Each base classifier should have different contribution to an ensemble classifier and their weights should be different. The generalization ability of an ensemble classifier will be bad if the value of the parameter α is quite small. However, the weights of worse base classifiers may be higher when α is quite big. Generally, we give $\alpha \in \{0.001, 0.001, 0.1, 1, 10\}$.

[1] http://archive.ics.uci.edu/ml/datasets.html.

[2] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/.

Table 2. Accuracies (%) of classification on test sets of 10 datasets by 100 classifiers ensemble

Method Dataset	AdaBoost	Random Forest	Simple Vote Rule	QFWEC	EnsembleSVM	Our method
Australian	85.79±0.75	86.54±0.60	86.53±0.61	86.05±0.70	86.85±0.60	**86.95±0.58**
Pima	75.78±1.52	81.10±0.61	81.00±1.18	79.05±1.46	81.25±0.62	**81.75±0.61**
Liver-disorders	51.50±0.65	59.00±0.79	59.25±0.76	58.67±0.76	59.13±1.03	**59.70±0.67**
Breast-cancer	96.72±0.72	99.09±0.31	99.18±0.31	98.22±0.82	99.27±0.31	**99.34±0.24**
Sonar	73.00±1.45	**82.85±1.42**	80.77±1.72	74.7±1.44	82.49±1.25	72.89±0.43
A1a	77.91±0.23	83.72±0.11	83.70±0.43	83.65±0.10	83.70±0.12	**83.79±0.10**
Heart	76.67±0.74	83.75±0.70	83.75±0.70	83.17±0.70	83.96±0.42	**84.00±0.38**
Emotions	73.27±1.12	80.05±0.62	80.20±0.99	76.94±1.20	80.69±0.31	**80.86±0.26**
Scene	76.67±0.33	83.00±0.26	83.08±0.32	83.05±0.22	83.51±0.23	**83.58±0.24**
German	72.50±1.15	75.30±1.26	76.50±1.17	75.14±0.85	79.00±1.22	**79.50±0.87**

The parameter L denotes the number of base classifiers. The generalization ability of an ensemble classifier will be good if there are enough base classifiers. However, excessive base classifiers may consist of many worse base classifiers and result in low classification accuracy. Therefore, we use $L \in \{10, 20, 50, 100, 200\}$. Different combinations of α and L on Scene dataset is used to select best parameters. Fig.1 shows that average accuracy of our proposed ensemble method on different parameter selections.

As described in Fig.1, the average accuracy is up to the maximum when α is given 0.1 and L is 100. It shows that the classification effect of our proposed method has the best result on Scene dataset when α is given 0.1 and L equals 100. To investigate the effect of the selected parameters, we analyze the parameter α and the parameter L respectively. Before discussing these two parameters, the size of the dataset should be specified, because the size has certain influence on experimental results. The Scene dataset contains of 2407 instances with 294 features. On the one hand when the number of base classifiers is fixed, we conclude that the value of the parameter α should be 0.1. Obviously, 0.1 is the appropriate value of the parameter on Scene dataset. On the other hand, when the value of the parameter α is fixed, we summarize that the number L of base classifiers should be 100.

3.2. Performance of Ensemble Classification Methods

In order to demonstrate statistically the performance of ensemble algorithms, we conduct some experiments on ten datasets by using different ensemble methods and make a comparison with the proposed method and others. 100 base classifiers are used for every ensemble method. The results of classification can be seen in Table 2.

In the Table 2,the result is shown by $(A \pm B)$. A and B express the mean of classification accuracy (%) and the standard deviation of classification accuracy (%) on the test set of each dataset, respectively.

Additionally, a result is bolded in each row of the table when it is the highest accuracy in results of six algorithms. Based on this analysis, it is obvious that the accuracy of our proposed method is higher than other techniques on most datasets, which indicates that the proposed method outperforms others. There is an exception that the accuracy of the proposed method on sonar dataset is low compared to other methods. It results in bad accuracy in the proposed method that the number of sonar dataset is too small and the number of base classifiers is quite big. On the whole, we can conclude that our proposed method obtains a better classification performance as compared with other ensemble methods according to all the experimental results of classification performance.

4. CONCLUSIONS

In this paper, we investigate the problem of how to combine different base classifiers. In order to avoid the problem generated by constructing an ensemble algorithm based on diversity, we propose diversity-induced weighted classifier ensemble method. The proposed method makes the tradeoff between accuracy and diversity and learns classifier weights by maximizing accuracy and diversity simultaneously. According to the theoretical analysis, we obtain several conclusions: Firstly, the value of parameters need to be adjusted in different datasets. Secondly, our proposed method outperforms other ensemble methods, such as Random Forest, AdaBoost, EnsembleSVM. Future work will be based on how to seek the optimal classifier weights from multi- class problems and extend the proposed method in regression.

Acknowledgement

This work was funded in part by the National Natural Science Foundation of China(No.61572240,61672268), and the Open Project Program of the National Laboratory of Pattern Recognition(NLPR)(No.201600005).

5. REFERENCES

[1] J.Liu, S.Shang , and K.Zheng, "Multi-view ensemble learning for dementia diagnosis from neuroimaging: An artificial neural network approach," *Neurocomputing*, , no. 195, pp. 112–116, 2016.

[2] S.Ghorai, A.Mukherjee, and S.Sengupta, "Cancer classification from gene expression data by nppc ensemble," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 8, no. 3, pp. 659–671, 2011.

[3] L.Liu, L.Shao , and P.Rockett , "Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition," *Pattern Recognition*, vol. 46, no. 7, pp. 1810C1818, 2013.

[4] L.Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[5] Y.Freund , "Experiments with a new boosting algorithm," in *Thirteenth International Conference on Machine Learning*, 1996, pp. 148–156.

[6] T.K.Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[7] Z. N. Li, I.W.Tsang, and Z. H. Zhou, "Efficient optimization of performance measures by classifier adaptation," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1370C1382, 2013.

[8] N.Garcła-Pedrajas , "Constructing ensembles of classifiers by means of weighted instance selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 258–277, 2009.

[9] H.Yuan, M.Fang, and X.Zhu , "Hierarchical Sampling for Multi-Instance Ensemble Learning," *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 12, pp. 2900–2905, 2013.

[10] N.Fazakis, S.Karlos, and S.Kotsiantis , "Self-trained Rotation Forest for semi-supervised learning," *IEEE Transactions on Knowledge & Data Engineering*, pp. 1–12, 2016.

[11] L.I.Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Inc, Hoboken, NJ, 2004.

[12] C.J.Merz, "Using Correspondence Analysis to Combine Classifiers," *Machine Learning*, vol. 36, no. 1, pp. 33–58, 1999.

[13] N.Ueda, "Optimal linear combination of neural networks for improving classification performance," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 22, no. 2, pp. 207–215, 2000.

[14] S.Mao, L.Jiao, and L.Xiong , "Weighted classifier ensemble based on quadratic form," *Pattern Recognition*, vol. 48, no. 5, pp. 1688–1706, 2015.

[15] L.Zhang, W.D.Zhou , "Sparse ensembles using weighted combination methods based on linear programming," *Pattern Recognition*, vol. 44, no. 1, pp. 97–106, 2011.

[16] Y.Zhang, X.Zhu, and X.Wu , "Corrective Classification: Classifier Ensembling with Corrective and Diverse Base Learners," in *Proceedings of the Sixth International Conference on Data Mining*, 2006, pp. 1199–1204.

[17] X.C.Yin, C.Yang, and H.W.Hao, "Learning to Diversify via Weighted Kernels for Classifier Ensemble," *Eprint Arxiv*, 2014.

[18] E.Morvant, A.Habrard, and S.Ayache , " Majority Vote of Diverse Classifiers for Late Fusion," *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 153–162, 2014.

[19] Z.Qi, B.Wang and Y.Tian , " When Ensemble Learning Meets Deep Learning: a New Deep Support Vector Machine for Classification," *Knowledge-Based Systems*, , no. 107, pp. 54–60, 2016.

[20] H.Xiao, Z.Xiao, and Y.Wang , "Ensemble classification based on supervised clustering for credit scoring," *Applied Soft Computing*, , no. 43, pp. 73–86, 2016.

[21] Marc Claesen, Frank De Smet and A.K.Johan , "EnsembleSVM: A Library for Ensemble Learning Using Support Vector Machines," *Journal of Machine Learning Research*, , no. 15, pp. 141–145, 2014.