Classifiers combination with observational Learning

YU Fan,† YANG Li-Ying††, and QIN Zheng†††

†,†††Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
††School of Computer Science and Technology, Xidian University, Xi'an, China

Summary
Ensemble method has shown the potential to increase classification accuracy beyond the level reached by an individual classifier alone. Observational Learning Algorithm (OLA) is an ensemble method based on social learning theory. Previous work focused on OLA for homogeneous ensembles, such as neural networks ensembles. In this paper, OLA for heterogeneous ensembles was proposed, which is a process with three steps: training, observing, and retraining. Experiments on five datasets from the UCI repository show that, OLA outperforms the individual base learner and majority voting when base learners are not capable enough for the given task. Bias-variance decomposition of the error indicates that OLA can reduce both bias and variance.

Key words:
Observational Learning, Social Learning, Classifiers Ensemble, Heterogeneous Ensemble

1. Introduction

The traditional approach to supervised learning problem is "evaluation and selection", which evaluates a set of different learning algorithms against a representative validation set and selects the best one. It works well when a large and representative data set is available, so that estimated errors allow selecting the best classifier. However, in many small sample-size real cases, the validation set provides just apparent errors that differ from true errors. Thus it is impossible to select the optimal learning algorithm, and the worst one could be selected in the worst case. It is quite intuitive that the above case can be avoided by, for example, averaging over the individual classifiers. This is the basic background that ensemble method comes into being. Ensemble method is a learning paradigm where a collection of a finite number of learners is trained for the same task, then all the predictions or just a subset of it are combined to obtain the final result [1][2]. Besides avoiding the selection of the worst learning algorithm, ensemble method can improve the performance of the best individual learner if individual learners are "different" enough. Ensemble method represents one of the main research directions in machine learning [3]. In an ensemble, components can be same learners as well as different learners. We refer to ensembles that composed of same learners as homogeneous ensembles, and analogously, ensembles that composed of different learners as heterogeneous ensembles. A learner traditionally learns by direct experience, which is known as "learning by doing". For a collection of learners, each one can also learn from the others, i.e. learning is done through observing the others besides the direct experience. This is the essence of social learning theory. Learning algorithm based on the observational learning mechanism is called Observational Learning Algorithm (OLA). Previous research mainly concentrated on OLA in homogeneous ensembles [4][5][6]. In this paper, we will focus on the observational learning behavior in heterogeneous ensembles, that is, OLA for heterogeneous ensembles.

The rest of this paper is organized as follows. In section 2, OLA is introduced in the context of heterogeneous ensembles. Bias-variance decomposition for theoretical investigation of classification problems is presented in section 3. Experiments and discussion are given in section 4. Finally, conclusion is drawn in section 5.

2 OLA for Heterogeneous Ensembles

Suppose a group of learners, the training set for each learner is the direct experience. If the information in a learner's training set is not enough for the learner to learn a good model of the task, indirect experience can be obtained by observing other learners, in the way that adding virtual training data to the original ones. How to generate high-quality virtual training data is an important issue in OLA. In this research, we obtained the feature vector of a virtual example by adding Gaussian noise to the feature vector of an example in original training set, and the output of the virtual example by taking the response of the best learner in ensemble. Since OLA is not sensitive to variance, the mean of Gaussian noise was set as zero and variance $\sigma = 2/n$ with n the size of training set [6]. This guaranteed that the observation range was inversely proportional to the number of training set. As for the size of virtual training data, it was set equal to that of the original data in order to avoid underestimating or overestimating the effect of observational learning. So each example in original training set was used just once to

Manuscript received February 25, 2006.
Manuscript revised February 30 , 2006.

generate virtual data, and the virtual data set for each learner was the same one. Then virtual data were added to original ones to train all learners in the ensemble over again. The "observing-retraining" process can be repeated if it has the potential to improve performance.

The OLA for heterogeneous ensembles is a process with three steps: training, observing, and retraining. The algorithm used in this paper is described as follows:

Step 1. Let $\{L_1,...,L_K\}$ be an ensemble of K learners and $D=\{(x_i,y_i)\,|\,x_i\in R^d, d\in N, y_i\in C\}$ a training set, C is the set of class labels and $i=1,...,n$. n is the size of training set.

Step 2. Obtain K classifiers $\{E_1,...,E_K\}$, where E_i denotes the classifier generated by training learner L_i on dataset $D\,(i=1,2,...,K)$.

Step 3. FOR t=1 to T (T denotes epochs)

3.1 Observing the others: FOR each example (x_i,y_i) in the training set D, x_i' is obtained by adding a Gaussian noise with a zero mean and a variance $2/n$ to x_i, y_i' is the output of the best classifiers in $\{E_1,...,E_K\}$. $D'=\{(x_i',y_i')\,|\,i=1,...,n\}$.

3.2 Learning by retraining: obtain K classifiers $\{E_1,...,E_K\}$ by retraining the learners on dataset $D+D'$.

ENDFOR

Step 4. The final ensemble output is computed as the majority voting result of $\{E_1,...,E_K\}$.

We presented the algorithm used in experiments above. Note that y_i' can also be obtained by other means (such as the majority voting result of $\{E_1,...,E_K\}$), so does the final ensemble output. Experiments on these cases were also carried out, and observations agreed with that of the above algorithm. They are not included in the paper for the sake of space.

3 Bias-variance Decomposition for the 0/1 Loss Function

Many theoretical investigations have been proposed to justify the success of ensemble methods, among which there are two main theoretical threads. One thread considers the ensembles in the framework of large margin classifiers, showing that ensemble method enhances generalization ability by enlarging the margins [7][8][9]. The other research thread follows classical bias-variance decomposition of error, indicating that ensemble method improves performance by reducing bias, variance or both of them [10]. Domingos proved that Schapire's notion of margins could be expressed in terms of bias and variance, and vice versa [11]. So the two explanations are of equivalence. We follow the theoretical thread of bias-variance decomposition to analysis the effectiveness of OLA.

Bias-variance decomposition has been originally developed in the standard regression setting, where the squared error is usually used as loss function [12]. This decomposition cannot be automatically extended to classification problems, where the 0/1 loss function is usually applied. Consider a set $D=\{D_i\,|\,i=1,...,m\}$ with $D_i=\{(x_{ij},y_{ij})\,|\,j=1,...,n\}$ a training set. Given a training set D_i, a learner produces a model f_i. Given a test example *x*, the model produces a prediction $y=f_i(x)$. Let *t* be the true value of the predicted variable for the test example *x*. The goal of learning is to produce a model that minimum the loss *L* over the example. So what we are interested in is the expected loss $E_{D,t}[L(t,y)]$, for which the following decomposition holds [11]:

$$E_{D,t}[L(t,y)]=c_1N(x)+B(x)+c_2V(x) \quad (1)$$

$N(x),B(x),V(x)$ denote noise, bias and variance respectively. c1 and c2 are multiplicative factors that will take on different values for different loss functions. The average loss over all examples can be obtained by averaging Equation (1) over all test examples. The variance of biased examples reduces error in two-class problems, but it is not true in multiclass problems. For multiclass problems, only the variance of part biased examples benefits the performance, and the more the classes, the less of the benefit. Note that noise is the unavoidable component of loss and it is incurred independently of learning algorithm, we suppose $N(x)=0$.

4 Experiments and Discussion

Five learners used in this work are listed as follows: (1) LDC, Linear Discriminant Classifier. (2) QDC, Quadratic Discriminant Classifier. (3) KNNC, K-Nearest Neighbor Classifier with K optimizes leave-one-out error for the training set. (4) TREEC, a decision tree classifier. (5) BPXNC, a neural network classifier based on

MATHWORK's trainbpx with 1 hidden layer and 5 neurons in this hidden layer. OLA based on the above learners were applied to five real world problems from the UCI repository: Iris, Soybean, New Thyroid, Zoo and Wine [13]. The characteristics of these data sets are shown in Table 1.For each dataset, 2/3 examples were used as training data and 1/3 test data.

Table **1** Data sets used in the study

	#Samples	*#Features*	*#Classes*
Iris	150	4	3
Soybean	47	35	4
New Thyroid	215	5	3
Zoo	101	16	7
Wine	178	13	3

Generalization abilities of base learners and their combination by majority voting were compared with OLA, where the "observing-retraining" process was carried out just once. Table 2 shows the errors on test sets. In order to avoid randomicity, all experiments were repeated for 10 runs and averages were computed as the final results.

Table **2** Individual classifiers vs MAJORC and OLA in the term of error on test set

	Iris	*New-Thyroid*	*Soybean*	*Wine*	*Zoo*
LDC	0.0216	0.0889	0.2400	0.0100	0.0941
QDC	0.0314	0.0361	0.8000	0.0100	0.9176
KNNC	0.0471	0.3056	0.5933	0.3817	0.5882
TREEC	0.0902	0.0778	0.2267	0.1150	0.1265
BPXNC	0.0392	0.0292	0.7800	0.0233	0.0706
MAJORC	0.0294	0.0361	0.4600	0.0017	0.0882
OLA	0.0294	0.0417	0.0067	0.0050	0.0676

From the experiments, we can predict that, heterogeneous ensembles composed of weak classifiers will benefit from OLA (such as Soybean and Zoo problems), while heterogeneous ensembles composed of effective classifiers will buy little at the cost of additional computation (such as Iris, New-thyroid and Wine problems). This is in agreement with intuition. In a heterogeneous ensemble, each component performs learning in a distinguished way. If the learning ability of an individual is weak, its performance can be promoted by observing the others. But this is not true when the individual's capability is good enough by itself, since learning from others may just add trivial information, even intervention in this circumstance.

In order to investigate the effect of "observing-retraining" process on OLA, experiments were carried out on five data sets respectively for different epochs T with $T \in \{1,2,3\}$. Results were plotted in Figure 1. The plot shows that iteration of "observing-retraining" process has effect on the accuracy of OLA. Performaces are promoted on three datasets (Iris, New Thyroid, Soybean) with the increase of variable T. This is not true on the other datasets(Wine and Zoo), which is abnormal. More attention were paid to the datasets themselves and some obsevations were drawn to explain the abnormal results. There are more features in Wine and Zoo than in Iris and New Thyroid, accordingly more noise is injected when OLA is carried out. If features size is big enough in the dataset, the effect of injected noise counteract the benefit from OLA. System performance would decrease in the case. From this point of view, OLA suits datasets with less features. As for Soybean, the original sample size is very small, so the benefit obtained from OLA outperforms the injected noise.

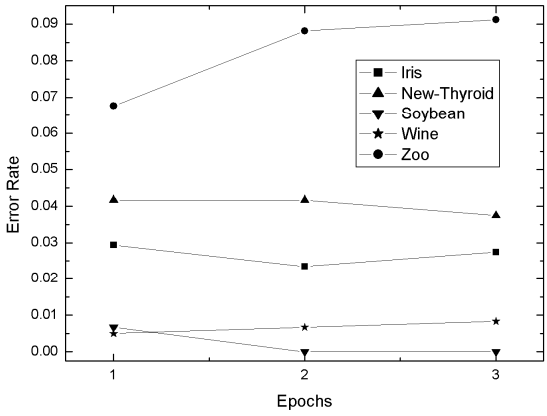

Fig. **1.** The effect of "observing-retraining" iteration on OLA

Bias-variance decomposition provides a powerful tool to investigate the effectiveness of ensemble method. Here it was used to analyze OLA. For the training set, 50 bootstrapping replicates were obtained to compute the bias-variance decomposition. Decomposition results on dataset Soybean were given in Table 3. We can see that,

bias and variance are both reduced in OLA compared with base learners and MAJORC.

Table **3** Bias-variance decomposition of error on Soybean

	Error	*Bias*	*Variance*	*Unbiased Variance*	*Biased Variance*
	0.2400	0	0.2400	0.2400	0
LDC					
	0.8000	0.8000	0	0	0
QDC					
	0.5973	0.6000	-0.0027	0	0.0027
KNNC					
	0.1707	0	0.1707	0.1707	0
TREEC					
	0.7760	0.8000	-0.0240	0.1280	0.1520
BPXNC					
	0.3387	0.3333	0.0053	0.1547	0.1493
MAJORC					
	0.0107	0	0.0107	0.0107	0
OLA					

5 Conclusion

OLA for heterogeneous ensembles was proposed in this paper. Experiments were carried out on five real world problems, which show that OLA performs better than the best base learner when the ensemble is composed of weak classifiers for the given task. The results also show that system performance can be improved further by repeat the "observing-retraining" process, and this effect is remarkble for datasets with less features. Bias-variance decomposition on Soybean showed that bias and variance were both reduced in OLA. Besides avoiding data insufficiency, the superiority of OLA may come from that virtual data added to the training set can also prevent learners from being overfitted to the original data, and the virtual output data are observational results from all learners rather than pure noise.

References

[1]. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions On Pattern Analysis and Machine Intelligence, 3 (1998) 226-239
[2]. Sharkey, A. J. C., Sharkey, N. E., Gerecke, U., Chandroth, G. O.: The "test and select" approach to ensemble combination. LNCS 1857, Springer-Verlag (2000) 30-44
[3]. Dietterich, T.G.: Ensemble methods in machine learning. LNCS 1857, Springer-Verlag (2000) 1-15
[4]. Cho, S., Cha, K.:Evoulution of neural network training set through addition of virtual samples. International conference on evolutionary computation, Nagoya, Japan (1996) 685-688
[5]. Cho, S., Jang, M., Chang, S. J.: Virtual sample generation using a population of networks. Neural processing letters, 2 (1997) 83-89
[6]. Jang, M., Cho, S.: Observational learning algorithm for an ensemble of neural networks. Pattern analysis & applications, 5 (2002) 154-167
[7]. Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics, 5 (1998) 1651-1686
[8]. Mason, L., Bartlett, P., Baxter J.: Improved generalization through explicit optimization of margins. Machine Learning, 38 (2000) 243-255
[9]. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: a unifying approach for margin classifiers. Journal of Machine Learning Research, 1(2000) 113-141
[10]. Breiman, L.: Bias, variance and arcing classifiers. Technical Report TR 460, Statistics Department, University of California, Berkeley, CA (1996)
[11]. Domingos, P.: A Unified Bias-Variance Decomposition for Zero-One and Squared Loss. In Proceedings of the Seventeenth National Conference on Artificial Intelligence, Austin, TX, AAAI Press, (2000) 564-569
[12]. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural Computation, 1 (1992) 1-58
[13]. Blake, C., Keogh, E., Merz, C. J.: UCI Repository of Machine Learning Databases (1998) www.ics.uci.edu/~mlearn/MLRepository.html

YU Fan received the B.S. and M.S. degrees in Aerospace Engineering from Northwest Polytech University in China, in 1998 and 2001, respectively. He is a now a Ph.D student in Xi'an Jiaotong University, his major is software theory.

Yang Liying received the Ph.D. degree in 2005 in computer science from Xi'an Jiaotong University, P. R. China. Her research interests are in pattern recognition, machine learning and multi-strategy learning. She is currently a lecturer in Xidian University, P. R. China