Neurocomputing 99 (2013) 581–591

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Letters

Dynamic classifier ensemble using classification confidence

Leijun Li, Bo Zou, Qinghua Hu*, Xiangqian Wu, Daren Yu

Harbin Institute of Technology, Harbin 150001, PR China

ARTICLE INFO

Article history:
Received 16 February 2012
Received in revised form
26 July 2012
Accepted 31 July 2012
Communicated by Zhouchen Lin
Available online 24 August 2012

Keywords:
Dynamic classifier ensemble
Classification confidence
Margin distribution

ABSTRACT

How to combine the outputs from base classifiers is a key issue in ensemble learning. This paper presents a dynamic classifier ensemble method termed as DCE-CC. It dynamically selects a subset of classifiers for test samples according to classification confidence. The weights of base classifiers are learned by optimization of margin distribution on the training set, and the ordered aggregation technique is exploited to estimate the size of an appropriate subset. We examine the proposed fusion method on some benchmark classification tasks, where the stable nearest-neighbor rule and the unstable C4.5 decision tree algorithm are used for generating base classifiers, respectively. Compared with some other multiple classifier fusion algorithms, the experimental results show the effectiveness of our approach. Then we explain the experimental results from the view point of margin distribution.

1. Introduction

Ensemble learning has been a hot topic in pattern recognition and machine learning domains for more than 20 years due to good generalization ability [1,25,26,36]. It means training a group of base learners which jointly solve a given classification or regression task with a fusion strategy. It has been theoretically and empirically demonstrated that combining multiple classifiers can substantially improve the classification performance of its constituent members [2,17,27,34].

How to effectively combine the outputs of the base classifiers is a key issue in ensemble learning. So far a number of fusion strategies have been proposed. In general, there are two basic fusion schemes to follow: one is to use the fixed base classifiers combination for all the test samples. The fixed combination can be constructed with all the base classifiers [4,10,18] or only a subset of them [6,15,20–23,33,35,37]. The other scheme is called dynamic classifier selection, which selects only one classifier to classify a given sample and the selected classifier is thought most likely to be correct for the given sample [12–14,30]. Inspired by the idea of dynamic classifier selection, we propose a dynamic classifier ensemble method in this paper based on the classification confidence of the test sample (termed as DCE-CC). However different from the dynamic classifier selection, DCE-CC dynamically selects a subset of classifiers for a given sample.

The fusion algorithms of using all the base classifiers include simple voting (SV) rule [18], linear weighted voting [4,10], and so on. These algorithms aim at combining all the outputs of the base classifiers in some way to improve the performance of the base classifiers. However it results in a large memory requirement and a slow classification speed [20].

In order to alleviate the drawbacks, selective ensemble algorithms, which select a fraction of the classifiers from the original ensemble and then combine them with simple or weighted voting, were proposed. The key problem is how to find the optimal subset of the base classifiers [20]. In [35], based on the evolved weights, GASEN was designed to select some neural networks to constitute the ensemble. Then in [15], the genetic algorithm was applied to find an approximate solution to the boosting pruning problem. In [33] the subset selection problem was viewed as a quadratic integer programming problem to search the classifiers subsets that have the optimal accuracy-diversity trade-off and semi-definite programming was used to get a good approximate solution. More recently, a new weighted combination method based on the linear programming was constructed for sparse ensemble [37]. However GASEN and semi-definite programming are all global optimization methods to search the appropriate classifiers subset and their computational costs are very high. To overcome this drawbacks, some suboptimal ensemble pruning methods were proposed, such as expectation propagation [6], margin distance minimization (MDM) [21], orientation ordering [22], boosting-based ordering [23], and so on.

These above fusion methods are based on the assumption that the classifiers are independent and equally reliable [8]. However, it is difficult to satisfy such an assumption in real applications. In the scheme of dynamic classifier selection [12–14,30], for each test sample, only one classifier is selected to classify it. The selected classifier for the given test sample is thought to most likely classify it correctly. Therefore it can avoid the error-independence assumption. These dynamic classifier selection algorithms include dynamic classifier selection based on classifier's local accuracy proposed in

* Corresponding author.
E-mail address: huqinghua@hit.edu.cn (Q. Hu).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.07.026

[30], dynamic classifier selection based on multiple classifier behavior [12], and so on. In [30], in order to classify an unknown test sample, the ℓ-nearest neighbors surrounding the sample were firstly estimated and then the classifier with the highest accuracy in the local regions was selected to classify the test sample. Since this algorithm is devised based on the ℓ nearest neighbors, its performance is affected by the choice of ℓ.

Margin distribution is thought as an important factor to improve the generalization performance of classifiers [3,28] and the effectiveness of the ensemble learning methods, especially the boosting method, has to be explained from the improvement of the margin distribution on training sets [29,32]. Therefore improving the margin distribution on the training sets is an effective way to boost the generalization capability of ensemble learning. In this paper a dynamic classifier ensemble method called DCE-CC is proposed based on the classification confidence and the optimization of margin distribution on the training sets. It dynamically selects a subset of classifiers to classify a test sample with the weighted voting and the classification confidence of the test sample on the selected classifiers are the first K largest. In order to estimate the size K, we exploit the optimization of margin distribution based on the ordered aggregation technique [20]. Then the test sample is classified by the selected classifiers using the weighted voting and the weight is the corresponding classification confidence. It is worth remarking that since the classification confidence order for different samples are usually different, the selected classifiers for different samples is usually different.

In this paper, the ordered aggregation technique is utilized to find an appropriate classifier subset for each sample, where the weights of base classifiers are learned by minimization of margin loss on the training sets. This strategy has been used in the selective ensembles such as Complementarity Measure [21], margin distance minimization (MDM) [21], orientation ordering [22] and boosting-based ordering [23]. Then the performance of these algorithms has been analyzed in [20]. The key problem for the ordered aggregation technique is how to reorder the classifiers in the ensemble process. In DCE-CC, the order of aggregation of the classifiers is estimated according to the classification confidence of the sample.

The major contributions in this work are listed as follows. First, based on the classification confidence, DCE-CC and a new margin are proposed. Second, the optimization of margin distribution and the ordered aggregation technique are utilized for the estimation of the size of an appropriate subset. Besides, the weighted voting based on the classification confidence is proposed to combine the selected classifiers for an unseen sample. Third, we use the stable nearest-neighbor rule and the unstable C4.5 decision tree algorithm to train base classifiers, a set of experiments are presented to test the rationality and the effectiveness of the proposed algorithm. DCE-CC is competent compared with the single classifier, a dynamic classifier selection algorithm DCS-LA and a selective ensemble algorithm called MDM [21].

The rest of the paper is organized as follows. Related work and a margin based on the classification confidence are introduced in Section 2. DCE-CC algorithm and the generation algorithm of the base classifiers are presented in Section 3. Then we discuss the rationality of DCE-CC and present our experimental results in Section 4. Finally, Section 5 offers the conclusions and future work.

2. Related works

Denote by $X=[x_1,\ldots,x_n]$ the training set which contain n samples and $D_1,\ldots,D_L$ the classifiers in the ensemble. Let $Y=[y_1,\ldots,y_n]$ be the true class labels of training set and $H_i=[h_{1i},\ldots,h_{ni}]$ be class labels of training set estimated by the classifier D_i. Besides, every classifier D_i provides for the training set the classification confidence $R_i=[r_{1i},\ldots,r_{ni}](r_{ij}\in[0,1])$. Intuitively, the higher the confidence provided by the classifier, the higher the probability that the classifier has correctly classified the sample.

Since DCE-CC algorithm proposed in this paper utilizes the optimization of margin distribution, the definition of margin is first given. In [29], the margin of a sample is defined as the difference between the number of correct votes and the maximum number of votes received by any incorrect label.

Definition 1 (*Schapire et al. [29]*). For $x_i\in X(i=1,2,\ldots,n)$, let $\omega=\{\omega_1,\ldots,\omega_c\}$ be the set of class labels, $H=\{h_{ij}|h_{ij}\in\omega\}$ be the classification decision of x_i by the classifier D_j $(j=1,2,\ldots,L)$. The margin of the sample x_i is denoted by

$$M_1(x_i)=\frac{N(\omega_i)-\max\{N(\omega_j)|i\neq j\}}{L} \tag{1}$$

where L is the number of the classifiers, $N(\omega_i)$ means the number of ω_i in H and ω_i is the true label of x_i.

From Definition 1, we can see that the margin is a number in the range $[-1, 1]$ and a sample x_i is classified correctly if and only if $M_1(x_i)>0$. A large positive margin can be interpreted as a "confident" correct classification, so the larger the margin on the test samples, the better the classification accuracy on the test samples. When the outputs of the classifiers are given, we expect the margin of each sample is as large as possible.

The margin distribution on the training sets is an important factor for the generalization performance of the ensemble learning methods. In [29], the generalization error of voting classifiers is bounded by the margin distribution, the number of training examples and the complexity of the set from which the base classifiers are chosen.

Theorem 1 (*Schapire et al. [29]*). *Let S be a sample of m examples chosen independently at random according to D. Assume that the base hypothesis space H is finite, and let $\delta>0$. Then with probability at least $1-\delta$ over the random choice of the training set S, every weighted average function f satisfies the following bound for all $\theta>0$:*

$$P_D[yf(x)\le 0]\le P_S[yf(x)\le\theta]+O(1/\sqrt{m}(\log m\log|H|/\theta^2+\log(1/\delta))^{1/2})$$

More generally, for finite or infinite H with VC-dimension d, the following bound holds as well:

$$P_D[yf(x)\le 0]\le P_S[yf(x)\le\theta]+O(1/\sqrt{m}(d\log^2(m/d)/\theta^2+\log(1/\delta))^{1/2})$$

In the theorem, H is the base classifier set, d is the VC dimension of H and θ is a threshold for the margin of an example (x,y), $P_D(yf(x)\le 0)$ denotes the probability of $yf(x)\le 0$ when an example (x,y) is chosen randomly according to the distribution D and $P_S(yf(x)\le\theta)$ denotes the probability with respect to choosing an example (x,y) uniformly at random from the training set S. This theorem states that with high probability $1-\delta$ the generalization error of any majority vote hypothesis can be bounded in terms of the number of training examples with margin below a threshold θ, the number of training examples S and the complexity measure of the base classifier set H.

Theorem 1 shows that a small generalization error for a voting classifier can be obtained by a good margin distribution on the training set. A good margin distribution refers to most training examples have large margins so that $P_S[yf(x)\le\theta]$ is small for not too small θ.