Neurocomputing 139 (2014) 289–297

Contents lists available at ScienceDirect

Neurocomputing

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

A spectral clustering based ensemble pruning approach

Huaxiang Zhang [a,b,*], Linlin Cao [a,b]

[a] Department of Computer Science, Shandong Normal University, Jinan 250014, Shandong, China
[b] Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, China

ARTICLE INFO

Article history:
Received 16 March 2013
Received in revised form
1 January 2014
Accepted 23 February 2014
Communicated by A.M. Alimi
Available online 13 April 2014

Keywords:
Ensemble pruning
Classifier similarity
Spectral clustering

ABSTRACT

This paper introduces a novel bagging ensemble classifier pruning approach. Most investigated pruning approaches employ heuristic functions to rank classifiers in the ensemble, and select part of them from the ranked ensemble, so redundancy may exist in the selected classifiers. Based on the idea that the selected classifiers should be accurate and diverse, we define classifier similarity according to the predictive accuracy and the diversity, and introduce a Spectral Clustering based classifier selection approach (SC). SC groups the classifiers into two clusters based on the classifier similarity, and retains one cluster of classifiers in the ensemble. Experimental results show that SC is competitive in terms of classification accuracy.

1. Introduction

Classification techniques are commonly used to reveal data patterns hidden in large datasets, and have been extensively studied in the field of machine learning. Various algorithms have been developed for constructing classifiers [1], but researche shows that no single algorithm outperforms the others theoretically or empirically in all scenarios [2]. Sometimes we are confused with which algorithm to utilize when facing a practical classification problem. In order to deal with this issue, ensemble classifiers have been proposed. An ensemble consists of a group of classifiers, and classifies instances based on the decisions of all members. Research shows that an ensemble of simple classifiers may achieve better classification performance than any one sophisticated classifier [3,4], and many ensemble approaches have been proposed [5–9]. Both accuracy and diversity play important roles in constructing ensemble classifiers, and many works focus on obtaining a group of accurate classifiers. Among the ensemble approaches, bagging [5] and boosting [6,10–12] are effective and have been extensively studied. Bagging adopts different bootstrap samples to generate diverse classifiers, and boosting constructs ensemble classifiers by using the original training data with weights updated for each classifier.

Ensemble classifiers can achieve remarkable performance, but redundancy may exist in them, and implementing a large number of classifiers requires large memory and slows down the classification. If only parts of the ensemble classifiers are implemented when classifying newly coming samples, the computational cost can be reduced. Many research works have been done to select a subset of ensemble classifiers without sacrificing the performance. Zhou et al. [13] proved the "many-could-be-better-than-all" theorem, and studies show that it is possible to achieve a small yet strong ensemble [14–20].

It is difficult to select an optimal classifier subset, since it needs a combinational search with exponential time complexity. In this paper, we propose a method to choose part of the generated weak classifiers while simultaneously considering both the classifier accuracies and the diversities among a pair of classifiers in one model. As we know learning systems are very common on the internet, and learning from different sources or teachers has been existent for a long time. But the behaviors or teaching styles of the teachers are quite different, and some may have negative effects on learning, thus, it is necessary to distinguish the positive ones from the negative ones. If we consider each classifier as a teacher, then a classifier ensemble can be regarded as a multiple teacher system, and each classifier is responsible for labeling newly coming samples. The teachers are categorized into responsible and irresponsible ones based on their influence on students. All responsible teachers are very similar, since they try to convey the unique right semantic concept, thus forming a cluster. Similarly, the classifiers can be partitioned into similar and dissimilar ones, and similar ones make positive contributions to the ensemble with high probability, thus they should form a cluster. Based on this assumption, we consider accuracy and diversity in one model for evaluating classifier similarity, and adopt clustering techniques to partition ensemble classifiers.

* Corresponding author at: Department of Computer Science, Shandong Normal University, Jinan 250014, Shandong, China.
E-mail address: huaxzhang@hotmail.com (H. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.02.030
0925-2312/

This paper is organized as follows. Section 2 details the related works in the literature. Section 3 defines the classifier similarity, and introduces how the spectral clustering approach is used to cluster classifiers. Section 4 explains the experimental results, and Section 5 concludes the study.

2. Related works

Since a large number of weak classifiers in an ensemble incur computational and storage cost, and the "many-could-be-better-than-all" theorem has been proved [13], many approaches have been proposed for selecting an optimal classifier subset [21–23].

Ensemble classifier selection approaches can be categorized as static and dynamic ones based on whether the selected classifier subset changes or not when different patterns are classified. The approaches that keep the subset unchanged are static, and the approaches that employ different classifier subsets to classify different patterns are dynamic [24,25]. Tsoumakas et al. [26] categorized the ensemble pruning approaches differently into four types: search based methods, clustering based methods, optimization based methods and other methods. Clustering based algorithms are based on the notion of distance to cluster the constructed classifiers.

Martínez-Muñoz et al. [16] analyzed several ordered aggregation based ensemble pruning techniques, and evaluated their performances on benchmark datasets. They concluded that ordered aggregation techniques sometimes could generate effective pruned bagging ensembles. The investigated pruning approaches employ specific metrics to rank classifiers, or perform a heuristic search in the classifier space while evaluating the collective merit of a candidate classifier subset [19,27–30]. Other ensemble pruning techniques employ genetic algorithms [31], randomized greedy selective strategy and ballot [32] or semi-definite programming [20] to perform classifier pruning.

Rokach [33] took into account the predictive capability of classifiers along with the degree of redundancy among them, and selected a high accuracy and low inter-agreement classifier subset. The approach implements the best first search strategy in a $2n$ huge search space (n: the number of classifiers), and reports that over half of the original classifiers can be pruned.

Aksela and Laaksonen [34] proposed a classifier selection method based on an exponential diversity error measure, and evaluated their approach on handwritten character patterns. Meynet and Thiran [35] took into account the classifier diversity and accuracy in the definition of information theoretic score (ITS), and selected a classifier subset with the optimal ITS. ITS is obtained by selecting one classifier at each iteration to maximize its value. It is not differentiable and its calculation incurs large time complexity.

Different from static approaches, Xiao et al. [36] proposed a dynamic classifier selection approach to noise data classification, and introduced several data handling methods for dynamic classifier selection. Statistical analysis and experimental results show that their approach has stronger noise-immunity ability than several other strategies. Many other dynamic approaches have also been proposed [24,37–39].

Zhu [40] integrated data envelopment analysis and stacking, and described a hybrid approach to classifier selection. Bakker and Heskes [41] proposed a clustering method for ensemble classifier extraction, in which a small collection of representative entities is used to represent a large entity collection. A method was used to extract the small representative model set through clustering the constructed models. Different from the proposed classifier pruning approaches, the small representative models are not part of the original ones.

Based on the analysis of the relationship between the proposed on-line allocation algorithm and the boosting algorithm, Freund and Schapire [42] proposed variants of adaboost, and proved the error bound of each variant. In order to handle the classification of a dataset with overlapping patterns from different classes, Verma and Rahman [43] first clustered the classified data, and used a group of component classifiers to learn the decision boundaries between pairs of clusters. A fusion classifier is responsible for the class decision based on the decisions of the component learners.

Different ensemble construction approaches may lead to quite different performances for a classifier ensemble, and many classifier ensemble approaches have been reported. We are only interested in how to prune a constructed ensemble for performance improvement.

3. The proposed approach

The above-mentioned research works show that, if both the accuracy and the diversity are taken into account in the process of classifier selection, the performance of the pruned ensemble may be improved. In this study, we focus on static classifier selection approaches, and propose a classifier similarity concept. The classifiers are used to construct a graph with weighted edges, and the spectral clustering approach is employed to analyze classifier aggregation based on the assumption that highly similar classifiers will aggregate into one group and lowly similar classifiers will aggregate into another group.

3.1. Classifier similarity

For each classifier, we calculate its classification accuracies on the training datasets used to construct weak classifiers, and use the obtained accuracies to construct an accuracy vector. Let D_1, D_2, …, and D_n denote n training datasets, and classifier h_i is modeled on D_i. Given h_i's classification accuracies a_i^1, a_i^2, …, a_i^n on $D_1, D_2, \ldots, D_n$ correspondingly, we use $\mathbf{a}_i = (a_i^1, a_i^2, \ldots, a_i^n)^T$ to represent the accuracy vector of h_i. Since we evaluate the prediction performances on different samples from the original training data, the resulting estimates may suffer from training bias. This problem does not matter, because each sample has an equal chance to influence the classifier's performance when calculating each classifier's accuracy vector. Hence, the effect of the training bias can be decreased by the vector entries.

For each pair of classifiers h_i and h_j with corresponding accuracy vectors $\mathbf{a}_i$ and $\mathbf{a}_j$, the accuracy similarity between them is defined by

$$S_{ij}(a) = \begin{cases} \sqrt{(\mathbf{a}_i \cdot \mathbf{a}_j)/n}, & (i \neq j) \\ 0, & (i = j) \end{cases} \tag{1}$$

where " $\cdot$ " denotes a scalar product of two vectors. $S_{ij}(a)(\in[0,1])$ is large if both classifiers perform well to some extent on all the sampled datasets, and small otherwise. It is also symmetric for each pair of classifiers, since we have $S_{ij}(a) = S_{ji}(a)$.

Diversity also plays an important role in the success of an ensemble [44], and it can be viewed as a measure of dependence, complement or even orthogonality among classifiers [45]. Diverse classifier ensembles are preferred, and Giacinto and Roli [46] stated that ensemble classifiers should be accurate and diverse. There exist many diversity measures [47,48], and no one has been proved to be the best. We employ the widely used diversity measure Q-statistics [45] to calculate the diversity of pairs of classifiers.

Given a dataset D, Q is calculated as

$$Q = (N_{11} \cdot N_{00} - N_{01} \cdot N_{10})/(N_{11} \cdot N_{00} + N_{01} \cdot N_{10}) \tag{2}$$