Pattern Recognition 47 (2014) 3120–3131

Contents lists available at ScienceDirect

Pattern Recognition

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

Exploration of classification confidence in ensemble learning

Leijun Li [a], Qinghua Hu [a,*], Xiangqian Wu [a], Daren Yu [b]

[a] Biometric Computing Research Centre, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b] School of Energy Science and Engineering, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 13 February 2013
Received in revised form
20 March 2014
Accepted 23 March 2014
Available online 1 April 2014

Keywords:
Ensemble learning
Ordered aggregation
Ensemble margin
Classification confidence

ABSTRACT

Ensemble learning has attracted considerable attention owing to its good generalization performance. The main issues in constructing a powerful ensemble include training a set of diverse and accurate base classifiers, and effectively combining them. Ensemble margin, computed as the difference of the vote numbers received by the correct class and the another class received with the most votes, is widely used to explain the success of ensemble learning. This definition of the ensemble margin does not consider the classification confidence of base classifiers. In this work, we explore the influence of the classification confidence of the base classifiers in ensemble learning and obtain some interesting conclusions. First, we extend the definition of ensemble margin based on the classification confidence of the base classifiers. Then, an optimization objective is designed to compute the weights of the base classifiers by minimizing the margin induced classification loss. Several strategies are tried to utilize the classification confidences and the weights. It is observed that weighted voting based on classification confidence is better than simple voting if all the base classifiers are used. In addition, ensemble pruning can further improve the performance of a weighted voting ensemble. We also compare the proposed fusion technique with some classical algorithms. The experimental results also show the effectiveness of weighted voting with classification confidence.

1. Introduction

One of the main aims in the machine learning domain has always been to improve the generalization performance. Ensemble learning has gained considerable research attention for more than twenty years [3,8,12,29,33] owing to its good generalization capability. This technique trains a set of base classifiers, instead of a single one, and then combines their outputs with a fusion strategy. Numerous empirical studies and applications show that the combination of multiple classifiers usually improves the generalization performance with respect to its members [1,28,31,47,51].

There are two key issues in constructing an ensemble system: (1) learning a collection of base classifiers and (2) combining them with an effective technique. Various algorithms have been developed for learning base classifiers by perturbing training samples, parameters, or structures of base classifiers [5,6,12,31,50]. For example, Bagging [5] generates different training sets by bootstrap sampling [11], whereas Zhou and Yu proposed a technique of multi-modal perturbation to learn diverse base classifiers [50]. In 2005, a review on the techniques of learning the diverse members was reported in [6]. The fusion strategy refers to effectively combining the outputs of the base classifiers. Currently available fusion algorithms can be roughly categorized into two schemes: one is to combine all the base classifiers with a certain strategy, such as simple voting [17] and weighted voting [3,13]. However, the investigation in [24] showed that combining part of the base classifiers, instead of all, may lead to better performance. Selective ensembles produced significantly higher accuracies than the original ensembles [41,47,49].

It is well known that a set of diverse and accurate base classifiers is the prerequisite for a successful ensemble. Indeed, effective exploitation of these base classifiers is also an important factor for designing a powerful ensemble. We will focus on the second issue in this work. An ensemble margin is considered an important factor, which has an impact on the performance of an ensemble and is utilized to interpret the success of Boosting [2,34,36,43]. Different boosting algorithms have been developed by constructing distinct loss functions based on the margin [10,12,22,33]. However, the margin defined in [36] just uses the classification decision of the base classifiers and their classification confidences are overlooked. In fact, classification confidence was theoretically proved to be a key factor on the generalization performance [35].

In this work, we want to identify the role of the classification confidence of a base classifier in ensemble learning. We generalize

* Corresponding author.
E-mail addresses: lileijun1985@163.com (L. Li), huqinghua@hit.edu.cn (Q. Hu), xqwu@hit.edu.cn (X. Wu), yudaren@hit.edu.cn (D. Yu).

http://dx.doi.org/10.1016/j.patcog.2014.03.021

the definition of the margin based on the classification confidence. The weights of the base classifiers are trained by optimizing the margin distribution. This strategy is similar to learning a classification function in a new feature space (meta-learning), just like the stacking technique [39,44]. In stacking, the outputs of the base classifiers are viewed as new features to train a combining function. Here, we show the difference of optimizing different margins from the viewpoint of the stacked generalization, and explain the necessity of incorporating the classification confidence into the margin.

Then, we explore how to utilize the weights and the classification confidences in combining the base classifiers. Four strategies are considered in this work. The weights are used to select the base classifiers. In a selective ensemble, a function should be developed to evaluate the quality of the base classifiers [25]. Similar to feature selection, classifier selection is also a combinational optimization problem. Assume that an ensemble consists of L base classifiers. Then, there are 2^L-1 nonempty sub-ensembles. Therefore, it is unfeasible to search the optimal solution via exhaustive search. In order to address this problem, several suboptimal ensemble pruning methods were proposed [1,8,16,47,49,51]. In the ordered aggregation technique, the base classifiers are selected based on the order [24–29]. The base classifiers are sorted by a specified rule, and then, they are added into the ensemble sequentially. A fraction of the base classifiers in the ordered ensemble are selected.

How to rank the base classifiers in the aggregation process is the key issue for this technique. In 1997, Reduce-Error pruning and Kappa pruning were proposed [24]. For Reduce-Error pruning, the first classifier is the one with the lowest classification error and the remaining classifiers are sequentially selected to minimize the classification error. Then, in 2004, Reduce-Error pruning without backfitting, Complementarity Measure, and Margin Distance Minimization were proposed to decide the order of the base classifiers [26], respectively. Based on the Complementarity Measure, the classifier incorporated into a sub-ensemble is the one whose performance is most complementary to this sub-ensemble. Recently, ensemble pruning via individual contribution ordering (EPIC) and uncertainty weighted accuracy (UWA) were proposed [23,29]. Moreover, in [25], the performances of some ordered aggregation-pruning algorithms have been extensively analyzed. For the proposed method, the base classifiers are sorted based on their weights in the descending order, which is similar to the method, MAD-Bagging, proposed in [46]. However, MAD-Bagging does not consider the classification confidence. While the major objective of this work is to analyze the influence of the classification confidence in ensemble learning. We try some ordered aggregation techniques to combine the base classifiers. Both the weighted and simple voting strategies are tested after pruning. The objective is to elucidate how to use the weights and the classification confidences of the base classifiers in ensemble optimization.

The main contributions of the work are listed as follows. First, we introduce the classification confidence in defining the ensemble margin and design a margin-induced loss function to compute the weights of the base classifiers. Second, we test several strategies to utilize the weights and the classification confidences in combining the base classifiers. Finally, extensive experiments are conducted to test and compare different techniques, and some guidelines for constructing a powerful ensemble are given.

The rest of the paper is organized as follows. In Section 2, we present some main notations and review the related works. In Section 3, we show how to learn the weights of the base classifiers and reveal the difference of optimizing different margins. In Section 4, we explore how to utilize these weights and the classification confidences to combine the base classifiers and propose a new ordered aggregation ensemble pruning method. Then, we analyze the proposed method in Section 5. Further, we test our algorithm on open classification tasks and study its mechanism for improving the classification performance in Section 6. Finally, Section 7 presents the conclusions.

2. Notations and related works

The main notations used in this paper are summarized as follows:

$h_j (j=1,2,\dots,L)$: the base classifiers
L: the total number of the base classifiers
$X=\{(x_i,y_i),\ i=1,2,\dots,n\}$: the pruning set
y_i: the true class label of the sample x_i
$\hat{y}_{ij}$: the classification decision of x_i estimated by the classifier h_j
r_{ij}: the classification confidence of x_i estimated by the classifier h_j

In the following, first, we introduce some works related to the classification confidence, margin, and stacked generalization, and then, we present some ordered aggregation pruning methods used in our experiments.

Classification confidence is used in this paper. A classifier h_j assigns a classification confidence r_{ij} to its decision $\hat{y}_{ij}$. For example, considering a linear real-valued classifier $h(x)=\psi\cdot x-b$, the classification decision of the sample x is 1 if $h(x)\ge 0$ and -1 otherwise. Then, the value $|h(x)|$ can be deemed as the classification confidence for its decision. In [35], the bound on the generalization error for this linear classifier was given, and it indicated that the classification confidence was an important factor for generalization. The linear classifier was also generalized to non-linear function and the detailed information can be obtained in [35]. Moreover, the classification confidence has been utilized in certain ensemble learning algorithms [12,30,32,45].

The margin is also considered as an important factor for the generalization performance of ensemble learning [36,43]. In [36], the margin of a sample with respect to an ensemble was introduced. Given a sample $x_i\in X$, its margin with respect to $\{h_1,\dots,h_L\}$ is defined as

$$m(x_i)=\sum_{j=1}^{L} w_j\Lambda_{ij},\qquad \text{s.t. } w_j\ge 0,\quad \sum_{j=1}^{L} w_j=1, \tag{1}$$

where w_j is the weight of the classifier h_j and

$$\Lambda_{ij}=\begin{cases}1 & \text{if } y_i=\hat{y}_{ij}\\ -1 & \text{if } y_i\ne\hat{y}_{ij}\end{cases} \tag{2}$$

In this work, a generalized definition of the margin is proposed based on the classification confidence and the weights of the base classifiers are learned through the optimization of the margin distribution. We will discuss the difference of optimizing different margins from the viewpoint of stacked generalization [39,44]. The stacking algorithms learn the weights of the base classifiers by training a function in a new feature space. In [44], the classification decision of the base classifier was used as the input feature. Then, the classification confidence was introduced and the stacking performance was improved [39].

In the classifiers ensemble, we are generally given a set of base classifiers $\{h_1,\dots,h_L\}$, which are obtained by certain learning algorithms [5,6,12,31,50]. Then, they are combined with some strategies such as the simple voting or the weighted voting. The simple voting implies that the class that receives the most votes is considered as the final decision. In the weighted voting, the votes are weighted and the final ensemble decision is the class that receives the largest weight coefficients sum of votes.

It was shown that selectively combining some of the base classifiers may lead to a better performance than combining all of