Pattern Recognition Letters 74 (2016) 38–45

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Combining diversity measures for ensemble pruning☆

George D.C. Cavalcanti[a,*], Luiz S. Oliveira[a,b], Thiago J.M. Moura[a,c], Guilherme V. Carvalho[a]

[a] Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Av. Jornalista Anibal Fernandes s/n, Recife, Brazil
[b] Universidade Federal do Paraná (UFPR), Rua Cel. Francisco Heraclito dos Santos, 100, Curitiba, Brazil
[c] Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), Av. Primeiro de Maio, 720, João Pessoa, Brazil

ARTICLE INFO

Article history:
Received 17 August 2015
Available online 11 February 2016

Keywords:
Ensemble pruning
Diversity measure
Graphs
Multiple classifier systems

ABSTRACT

Multiple Classifier Systems (MCSs) have been widely used in the area of pattern recognition due to the difficult task that is to find a single classifier that has a good performance on a great variety of problems. Studies have shown that MCSs generate a large quantity of classifiers and that those classifiers have redundancy between each other. Various methods proposed to decrease the number of classifiers without worsening the performance of the ensemble succeeded when using diversity to drive the pruning process. In this work we propose a pruning method that combines different pairwise diversity matrices through a genetic algorithm. The combined diversity matrix is then used to group similar classifiers, i.e., those with low diversity, that should not belong to the same ensemble. In order to generate candidate ensembles, we transform the combined diversity matrix into one or more graphs and then apply a graph coloring method. The proposed method was assessed on 21 datasets from the UCI Machine Learning Repository and its results were compared with five state-of-the-art techniques in ensemble pruning. Results have shown that the proposed pruning method obtains smaller ensembles than the state-of-the-art techniques while improving the recognition rates.

1. Introduction

Ensemble methods began gathering attention of the pattern recognition community after Wolpert's no free lunch theorem [1] stated that given enough problems and two classifiers, the number of problems in which a classifier outperforms the other is roughly equal. This means that searching for a single classifier model that had good performance at a wide array of problems is unproductive. Multiple classifier systems, another name for ensembles of classifiers, avoid the problem stated by Wolpert by combining the output of various classifiers. The combination softens the differences between problems in which the classifiers of the ensemble have different performances. Besides this softening effect ensembles use weaker classifiers which are easier to train.

The main problem with ensemble methods, such as Bagging or AdaBoost, is that the final ensemble has a large number of classifiers. In the late 1990s it had been shown that some of the classifiers in the ensemble could be removed without impairing the ensembles ability to generalize [2,3]. These findings led to more research being done on the area of ensemble pruning since searching exhaustively for the best subset of an ensemble can become intractable for relatively small ensemble sizes.

The seminal work in this field was published by Margineantu and Dietterich [3] where the authors compared five different pruning algorithms on ten datasets and concluded that in most of the experiments the ensemble of decision trees produced by AdaBoost could be pruned substantially with no considerable impacts of the performance. Tamon and Xiang [4] proposed an improvement to one of the methods described by Margineantu and Dietterich [3], the Kappa pruning, and also addressed the boosting pruning problem from a theoretical perspective.

Zhou et al. [5] introduced the GASEN (Genetic Algorithm based Selective ENsemble) method, which selects the classifiers to constitute an ensemble according to some evolved weights that could characterize the fitness of including the classifiers in the ensemble. In their empirical study they used neural networks as classifiers, genetic algorithms, and 20 different datasets. They show that the pruned ensemble generated by the GASEN method was able to outperform the popular ensemble approaches such as Bagging and Boosting. Other examples of methods using global search to prune the ensembles can be found in [6,7].

A different approach, based on a greed local search, was proposed by Martínez-Muñoz and Suárez [8,9], Martínez-Muñoz et al. [10]. In these works they explored the idea that the order in which classifiers are aggregated in ensemble methods can be an

☆ This paper has been recommended for acceptance by Egon L. van den Broek.

* Corresponding author. Tel.: +55 81 2126 8430.
E-mail address: gdcc@cin.ufpe.br, darmiton@gmail.com (G.D.C. Cavalcanti).

http://dx.doi.org/10.1016/j.patrec.2016.01.029
0167-8655/

important tool to prune ensembles. Their algorithm is based on ordering the predictors in the ensemble according to a number of rules that exploit the complementariness of the individual classifiers. Experiments on several UCI repository datasets show that ordered ensembles produced a generalization error lower than the full ensembles created by Bagging.

An issue not to be neglected when building ensembles of classifiers is the diversity, which is the underpinning to successful deployment of classifiers ensemble. Empirical results have shown that there exists positive correlation between performance of the ensemble and diversity among the base classifiers [11,12]. On the other hand, the usefulness of diversity measures to build ensembles of classifiers is questioned by some authors. Kuncheva and Whitaker [13] performed a considerable amount of experiments but could not find a definitive connection between the diversity measures and the improvement of the ensemble accuracy. In other words, designing diverse classifiers is important but the problem of measuring this diversity and so using it effectively for building better ensembles is still an open problem. Ko et al. [14] investigated 10 diversity measures into a pairwise fusion matrix transformation to combine classifiers and concluded that the use of diversity might slightly improve the methods for classifier combination in some problems, but the effect is not significant. Tang et al. [15] evaluated six different measures of diversity and concluded that none of them is suitable for the task of building ensemble of classifiers. According to the authors, if one exploits diversity measures as criteria to select the base classifiers, then the diversity measure is required to be precise, since the choice of diversity measure will directly influence the final ensemble and subsequently the classification result.

As one may notice, understanding how diversity can be used to build ensembles remains an open problem. In spite of that, the literature shows us several cases where the diversity has been successfully applied to build ensembles of classifiers. Tsymbal et al. [16] point out the importance of the diversity measures during the search problem for ensemble feature selection. Oliveira et al. [17] show that diversity is quite useful to build ensembles of classifiers through feature selection since it helps preventing overfitting during the search. Li et al. [18] presented a theoretical study on the effect of diversity in voting. They concluded that by enforcing large diversity, the hypothesis space complexity of voting can be reduced, and then better generalization performance can be expected. These findings were used to build a method called DREP (Diversity Regularized Ensemble Pruning) which explicitly exploit diversity regularization. Experimental results show that with the help of diversity regularization, DREP is able to achieve significantly better generalization performance with smaller ensemble size than the compared methods.

Motivated by the success of Li et al. [18] and also by the findings of Kuncheva [19], which suggests that a single measure of diversity might not be accurate enough to capture all the relevant diversities in the ensemble, in this study we argue that the combination of several diversity measures can be an useful tool to prune an ensemble of classifiers. To support this idea, we propose an ensemble pruning method where the undermining concept is the combination of different pairwise diversity matrices. The weights of this combination are provided by a genetic algorithm. From the combined diversity matrix we are able to group similar classifiers, i.e., those with low diversity, that should not belong to the same ensemble. In order to generate the candidate ensembles, we transform the combined diversity matrix into one or more graphs and then apply a graph coloring method. The fitness of the genetic algorithm is provided by the ensemble that minimizes the error on a validation set.

Through a set of comprehensive experiments on 21 datasets of the UCI repository we show that the proposed method is able to considerably reduce the original size of the ensemble while improving the recognition rates. The results reached by our method compare favorably to other published methods.

The rest of this article is organized as follows: Section 2 reviews the diversity measures used in this work; Section 3 describes the proposed method for pruning a pool of classifiers; Section 4 reviews the methodology and experiments run to validate the proposed method; Section 5 lists the conclusions that can be taken from the experiments.

Table 1
Contingency table for two classifiers d_i and d_j.

	$d_i = +$	$d_i = -$
$d_j = +$	a	c
$d_j = -$	b	d

2. Diversity measures

There is not a widely accepted definition of diversity between classifiers. For that reason there are many definitions used throughout the literature. In the proposed method five pairwise diversity measures are combined to reach a broader definition of diversity. This section describes these five measures and how to calculate them.

The diversity measures are calculated using a contingency table [20] that summarizes the behavior of two classifiers d_i and d_j across a dataset. Table 1 shows an example of a contingency table. The values on the table have the following meaning: a is the number of examples in the dataset correctly classified by both d_i and d_j; b is the number of examples correctly classified by d_i and incorrectly classified by d_j; c is the number of examples incorrectly classified by d_i and correctly classified by d_j; and d is the number of examples incorrectly classified by both classifiers.

Disagreement is the proportion of examples differently classified by d_i and d_j. Its value is calculated by Eq. (1), where $m = a + b + c + d$. Its value ranges from 0 to 1, with higher values indicating more diversity.

$$dis_{ij} = \frac{b + c}{m} \tag{1}$$

The Q-statistic is defined by Eq. (2). Q_{ij} ranges from -1 to 1, where 0 means the two classifiers are independent, 1 both classifiers make similar predictions, and -1 the classifiers make different predictions.

$$Q_{ij} = \frac{ad - bc}{ad + bc} \tag{2}$$

The Correlation Coefficient of two classifiers is calculated by Eq. (3) and the meaning of its value is similar to that of the Q-statistic.

$$\rho_{ij} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} \tag{3}$$

The Kappa-statistic is widely used in statistics and was used to analyze the diversity between classifiers for the first time by Margineantu and Dietterich [3]. κ_p (Eq. (4)) is equal to 1 if the classifiers completely agree, 0 if they randomly agree, and less than 0 is a rare case that happens when they agree less than what is expected by chance.

$$\kappa_p = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2} \tag{4}$$

where

$$\Theta_1 = \frac{a + d}{m}, \tag{5}$$