Neurocomputing 150 (2015) 501–512

Contents lists available at ScienceDirect

Neurocomputing

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

Dynamic ensemble pruning based on multi-label classification

Fotini Markatopoulou [b], Grigorios Tsoumakas [a,*], Ioannis Vlahavas [a]

[a] Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[b] Information Technologies Institute, CERTH, 57001 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 25 December 2013
Received in revised form
19 July 2014
Accepted 21 July 2014
Available online 6 October 2014

Keywords:
Ensemble pruning
Ensemble selection
Multi-label classification
Dynamic classifier fusion

ABSTRACT

Dynamic (also known as instance-based) ensemble pruning selects a (potentially) different subset of models from an ensemble during prediction based on the given unknown instance with the goal of maximizing prediction accuracy. This paper models dynamic ensemble pruning as a multi-label classification task, by considering the members of the ensemble as labels. Multi-label training examples are constructed by evaluating whether ensemble members are accurate or not on the original training set via cross-validation. We show that classification accuracy is maximized when learning algorithms that optimize example-based precision are used in the multi-label classification task. Results comparing the proposed framework against state-of-the-art dynamic ensemble pruning approaches in a variety of datasets using a heterogeneous ensemble of 200 classifiers show that it leads to significantly improved accuracy.

1. Introduction

Supervised ensemble methods are concerned with the production and the combination of multiple predictive models. One dimension along which we could categorize such methods is based on the number of models that affect the final decision. Usually all models are taken into consideration. When models are classifiers, this is called *classifier fusion*. Some methods, however, select just one model from the ensemble. When models are classifiers, this is called *classifier selection*. A third option, standing in between of these two, is to select a subset of the ensemble's models. This is mainly called *ensemble pruning* or *ensemble selection* [26].

Ensemble pruning methods can be either *static*, meaning that they select a fixed subset of the original ensemble for all test instances, or *dynamic*, also called *instance-based*, where a different subset of the original ensemble may be selected for each different test instance. The rationale of using dynamic ensemble pruning approaches is that different models have different areas of expertise in the instance space. Therefore, static approaches that are forced to select a fixed subset prior to seeing an unclassified instance may have a theoretical disadvantage compared to dynamic ones. On the other hand, static approaches lead to improved space complexity as they typically retain a small percentage of the original ensemble, in contrast to dynamic approaches that need to retain the complete original ensemble.

We propose a new approach to the instance-based ensemble pruning problem by modeling it as a multi-label learning task [24]. Labels correspond to classifiers and multi-label training examples are formed based on the ability of each classifier to correctly classify each original training example. This way we can take advantage of recent advances in the area of multi-label learning and attack collectively, instead of separately, the problems of predicting whether each classifier will classify correctly a given unclassified instance. This paper builds upon our previous work [18] and extends it in the following main directions: (a) it approximately doubles the number of datasets of the empirical comparison, providing further evidence of the effectiveness of the proposed algorithm and (b) it employs a thresholding strategy that automatically computes the threshold that optimizes precision, leading to a fairer comparison against the state-of-the-art.

The rest of this paper is organized as follows. Section 2 reviews related work on dynamic ensemble pruning and Section 3 discusses the proposed approach. Section 4 presents the experimental setup. Section 5 presents the empirical study and Section 6 discusses the conclusions of this work.

2. Related work

Many approaches deal directly with instance-based ensemble pruning [27,4,13,10]. These are presented in Section 2.2. However, there are also some that deal with dynamic approaches to classifier selection and fusion [30,7,21,19], which can be considered as extreme cases of ensemble pruning. In addition, some

* Corresponding author.
E-mail addresses: markatopoulou@iti.gr (F. Markatopoulou), greg@csd.auth.gr (G. Tsoumakas), vlahavas@csd.auth.gr (I. Vlahavas).

http://dx.doi.org/10.1016/j.neucom.2014.07.063
0925-2312/

Table 1
Summary of dynamic classifier selection, fusion and pruning methods.

Classifier selection		DS, OLA, LCA, MCB, [19,7]
Ensemble pruning	k-NN-based	DVS, KNORA, [31]
	Clustering-based	[14]
	Ordering-based	[16,32,10,4]
	Other	[17]
Classifier fusion		DV

dynamic classifier selection approaches, may in some cases (e.g. ties) take into consideration more than one model. We therefore discuss in Section 2.1 such methods too. Section 2.3 discusses the issues of time complexity improvement and diversity in the context of dynamic ensemble pruning methods. For ease of reference and navigation of this section, Table 1 shows the category (i.e. selection, fusion, pruning) of each method that is discussed, using either its acronym where available (e.g. KNORA) or its citation.

2.1. Dynamic classifier selection and fusion

The approach in Woods et al. [30] starts with retrieving the k nearest neighbors of a given test instance from the training set. It then classifies this test instance using the most competent classifier within this local region. In case of ties, majority voting is applied. The local performance of classifiers is assessed using two different metrics. The first one, called *overall local accuracy* (OLA), measures the percentage of correct classifications of a model for the examples that exist in the local region. The second one, called *local class accuracy* (LCA), measures the percentage of correct classifications of a model within the local region too, but only for those examples where the model had given the same output as the one it gives for the current unlabeled instance being considered. A very similar approach to this one was proposed independently at the same time [7], also taking the distance of the k nearest neighbors into account.

The *dynamic selection* (DS) and *dynamic voting* (DV) approaches in Puuronen et al. [21,20] are in the same spirit as Woods et al. [30] and Giacinto and Roli [7]. A kNN approach is initially used to find the most similar training instances with the given test instance. DS selects the classifier with the least error within the local area of the neighbors weighted by distance. In fact, DS, is very similar to the weighted version of OLA presented in Giacinto and Roli [7]. DV is different, as it is a classifier fusion approach. It combines all models weighted by their local competence.

Yet another approach along the same lines is Giacinto and Roli [8]. After finding the k nearest neighbors of the test instance, this approach further filters the neighborhood based on the similarity of the predictions of all models for this instance and each neighbor. In this sense, this approach is similar to the LCA variation in Woods et al. [30]. It finally selects the most competent classifier in the reduced neighborhood. The predictions of all models for an instance are in this paper called collectively *multiple classifier behavior* (MCB).

The approach in Ortega et al. [19] estimates whether the ensemble's models will be correct/incorrect with respect to a given test instance, using a learning algorithm, trained from the k-fold cross-validation performance of the models on the training set. It can be considered as a generalization of the approaches we have seen so far in this subsection, where a nearest neighbor approach was specifically used instead. The approach we propose in this paper is based on the same principle, with the difference that multi-label learning algorithms are employed and therefore the binary tasks of predicting correct/incorrect decision for each model are viewed in a collective way.

2.2. Dynamic ensemble selection

Similar with the dynamic Classifier Selection methods, the majority of the Dynamic Ensemble Selection methods start with retrieving the k nearest neighbors of a given test instance from the training set, in order to construct a new set of instances known as local region of competence [31]. The selection algorithms decide for the appropriate subset of the initial ensemble based on different properties (e.g. accuracy, diversity) of the base classifiers in this local region.

Dynamic voting with selection (DVS) [27,28] is an approach that stands in between the DS and DV algorithms that were mentioned in the previous subsection. First, about half of the models in the ensemble, those with local errors that fall into the upper half of the error range of the committee, are discarded. Then, the rest are combined using DV. Since this variation, eventually selects a subset of the original models, we can consider it as an instance-based ensemble pruning approach.

The primary goal of k-nearest-oracles (KNORA) [13] is improving the accuracy compared to the complete ensemble. Four different versions of the basic KNORA algorithm are proposed, all based on an initial stage of identifying the k nearest neighbors of a given unclassified instance. KNORA-ELIMINATE selects those classifiers that correctly classify *all* k neighbors. In case none such exists, the k value is decreased until at least one is found. KNORA-UNION selects those classifiers that correctly classify at least *one* of the k neighbors. KNORA-ELIMINATE-W and KNORA-UNION-W are variations that weight the votes of classifiers according to their Euclidean distance to the unclassified instance.

While the above methods only consider the accuracy of the ensemble within the local region, the method proposed by Xiao et al. [31] simultaneously considers both the accuracy and the diversity of the pruned ensemble. Specifically, this method utilizes the symmetric regularity criterion to measure the accuracy of the ensemble and the double-fault measure to estimate the diversity. Finally, a GMDH-based neural network describes the relationship between the class labels of the local region of competence and the test instance.

We can distinguish two more categories of dynamic ensemble selection methods that do not consider the k-nearest neighbors of test instances: Clustering based methods and Ordering based methods. Clustering based methods use clustering algorithms (k-means, Gaussian Mixture Models etc.) in order to estimate the local regions Kuncheva [14]. In contrast with the k nearest neighbors based methods that generate the local regions of competence during the test phase, clustering based methods estimate them offline during the training phase. Only the selection of a winning local region and the appropriate classifier ensemble is selected during the test phase.

Ordering based methods utilize statistical or probabilistic measures in order to produce a decreasing order of the base classifiers from the most suitable for a given test instance to the less suitable. The method proposed by Li et al. [16] assumes that base classifiers not only make a classification decision but also return a confidence score that shows their belief that their decision is correct. Dynamic ensemble selection is performed by ordering the base classifiers according to the confidence scores and fusion is performed using weighted voting. The method proposed by Yan et al. [32] is a two-step approach. In the first step classifiers are ordered based on their diversity using the Fleiss's statistic, in the second step classifiers in this rank are selected until a confidence threshold is reached.

Recently, a statistical approach has been proposed for instance-based pruning of homogeneous ensembles, with the provided that the models are produced via independent applications of a randomized learning algorithm on the same training data, and that majority voting is used [10]. It is based on the observation that given the decisions made by the classifiers that have already been queried, the probability distribution of the remaining class predictions can be calculated via a Polya urn model. During prediction, it samples the