Information Sciences 422 (2018) 144–160

Contents lists available at ScienceDirect

Information Sciences

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

Heterogeneous classifier ensemble with fuzzy rule-based meta learner

Tien Thanh Nguyen[a,*], Mai Phuong Nguyen[b], Xuan Cuong Pham[c], Alan Wee-Chung Liew[a]

[a] School of Information and Communication Technology, Griffith University, Australia
[b] FPT Company, Hanoi, Vietnam
[c] Department of Computer Science, Water Resources University, Hanoi, Vietnam

ARTICLE INFO

Article history:
Received 14 January 2017
Revised 17 July 2017
Accepted 2 September 2017
Available online 6 September 2017

ABSTRACT

In heterogeneous ensemble systems, each learning algorithm learns a classifier on a given training set to describe the relationship between a feature vector and its class label. As each classifier outputs different result on an observation, uncertainty is introduced. In this paper, we introduce a heterogeneous ensemble system with a fuzzy IF-THEN rule inference engine as the combiner to capture the uncertainty in the outputs of the base classifiers. In our method, fuzzy rules are generated on the outputs of an ensemble of base classifiers, which can be viewed as the class posterior probability of the observations. The performance of our method was evaluated on thirty datasets and in comparison with nine ensemble methods (AdaBoost, Decision Template, Decision Tree on meta-data, and six fixed combiners) and two single learning algorithms (SVM with L2-loss function and Decision Tree), and was shown to significantly outperforms these algorithms in terms of classification accuracy.

1. Introduction

It is widely known that the best classifier for a given problem is often problem dependent. A natural question that arise is: can we combine multiple classifiers to achieve higher classification accuracy than a single classifier? This is the idea behind a class of methods called ensemble method. Ensemble method is defined as the combination of multiple classifiers with the aim of achieving lower classification error than using a single classifier [46].

Ensemble methods can be divided into two categories [31,44]:

- Homogeneous: The same learning algorithm is used to learn a set of classifiers based on different training sets obtained from an original one. The outputs of these classifiers are then combined to give the final decision. Several state-of-the-art ensemble methods in this category are Bagging [5], Random Forest [6], and AdaBoost [10].
- Heterogeneous: A set of diverse learning algorithms is used to generate the base classifiers from a single training set. The outputs of these heterogeneous base classifiers (called Level1 data or meta-data) are then combined to give the final decision [23,26,28–31,37,38].

* Corresponding author.
E-mail address: tienthanh.nguyen2@griffithuni.edu.au (T.T. Nguyen).

http://dx.doi.org/10.1016/j.ins.2017.09.009
0020-0255/

In this study, we focus on the heterogeneous ensemble system where each learning algorithm uses a different learning methodology to learn the base classifier on a given training set. The output of the base classifiers therefore reflects the agreements and disagreements between the different base classifiers for an observation. A combiner based on fuzzy IF-THEN rules then operates on the output of the base classifiers to produce the final discriminative decision model. Fuzzy inference systems are well suited to deal with imprecise data [25,35] and to represent knowledge with uncertainty [40,41]. Fuzzy IF-THEN rule based classifiers have found wide application in problems involving pattern recognition, decision making, and image processing [3], and extensive research have been done on areas such as fuzzy membership function design [7,18,35,40], fuzzy rules selection [12,16,19,24,25], and fuzzy rules refinement [41]. Whereas features in the original data often differ in scale and type, meta-data, which can be viewed as a transformation from feature domain to posterior probability domain, has all values between [0, 1]. Therefore, there is no need to build feature-specific antecedent fuzzy set as is the case for the original data [13]. Moreover, for a D-dimension dataset and $|\mathcal{A}|$ antecedent fuzzy sets, the total number of generated rules is $|\mathcal{A}|^D$, i.e. the number of rules increases exponentially with the dimension of the feature vector [16,35,40]. In contrast, in heterogeneous ensemble system with a small number of base classifiers, meta-data with a small number of classes usually has a much lower dimension than the original data. Therefore, the number of fuzzy rules generated by a fuzzy rule-based combiner from the meta-data is usually much smaller than that generated from the original data. Our ensemble system is able to significantly outperform many state-of-the-art ensemble classifiers as shown by our experiments.

The rest of the paper is organized as followed. Section 2 describes the proposed ensemble system in detail. Experiment results and discussions are reported in Section 3. The conclusion is given in the last section.

2. Literature review

2.1. Heterogeneous ensemble methods

In general, there are two techniques to combine the outputs of the base classifiers in a heterogeneous ensemble system, namely fixed combining methods and trainable combining methods [22,31,44]. The fixed combining methods do not consider the label information in the meta-data of training set when building the combiner. Hence, they are simpler to build and run. Several popular fixed combining methods are: Sum, Product, Vote, Max, Min, Average, Median, and Oracle Rule [22]. The Ordered Weighted Averaging operator (OWA), one of the most well-known operators applied to Decision Making Systems, has also been applied to the combiners in ensemble systems [28]. This operator is used to compute average value based on weight, but instead of focusing on the original data it is linked to the order of data. As a result, elements at specific locations like head, tail or middle can receive more attention than the others.

On the other hand, trainable combining method uses the label information in the meta-data of training observations to train a meta-learner to learn the prediction model. Here, meta-learner is used to refer to combining algorithm that learns on meta-data generated by base classifiers under a Cross Validation (CV)-based procedure. The common feature of trainable combining methods is that the meta-data of training set is used to train a combining classifier to obtain the final prediction model.

Several trainable strategies have been proposed to exploit the label information in the meta-data. First, weight-based classifiers methods are based on the assumption that each classifier puts a different weight on each class and the combining algorithm is then obtained based on the M weighted linear combinations of posterior probabilities for the M classes. There are several approaches to find the weights. Ting et al. [37] proposed the Multi-Response Linear Regression (MLR) method by solving M Linear Regression models corresponding with the M classes based on the meta-data and the labels to find the weights. Zhang and Zhou [45] proposed using linear programming to find the weights. Sen et al. [33] introduced a method that was inspired by MLR which uses hinge loss function and group sparse regularization in the combiner.

Besides, Decision Template is based on computing the distance between each template associated with a class and the meta-data of an unlabeled observation. In Decision Template method [23], the meta-data of the training set is grouped according to the class label of the training observations and the Decision Template for each class is constructed by averaging the meta-data associated with each class label. Kuncheva et al. [23] proposed eleven measures between a Decision Template and the meta-data of an unlabeled observation to predict the class label. The advantage of this method is that it saves time in both training and testing due to its simple computation.

Merz [26] combined Stacking, Correspondence Analysis (CA) and k Nearest Neighbor (kNN) (SCANN) in a single algorithm. The idea of that algorithm is to discover the relationship between the observations and the base classifiers' outputs by applying CA to the indicator matrix formed by the meta-data and their true labels. After that, kNN is used to classify unseen observations in the new scaled space. In real-world application, SCANN sometimes is impractical due to the singularity characteristic of the indicator matrix obtained by CA. Moreover, the classification process of SCANN is more complicated than that of other combining classifier algorithms and this increases the classification time.

Another approach proposed by Todorovski and Džeroski [38] is the Meta Decision Tree, a new Decision Tree on the meta-data where at each node instead of selecting value for splitting an attribute, a classifier is chosen. The authors also proposed an expansion for the meta-data by adding entropy and maximum posterior probability so as to increase the discrimination ability. However, no theoretical basis was provided about the effectiveness of that expansion. Recently, Nguyen et al. [31] proposed a combining method based on Bayesian inference. In the proposed method, prediction is obtained by selecting the class label associated with the maximum posterior probability that an observation belongs to a class. The pos-