Pattern Recognition Letters 31 (2010) 125–132

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Forests of nested dichotomies

Juan J. Rodríguez *, César García-Osorio, Jesús Maudes

Escuela Politécnica Superior, Universidad de Burgos, 09006 Burgos, Spain

ARTICLE INFO

Article history:
Received 19 January 2009
Received in revised form 12 August 2009
Available online 19 September 2009

Communicated by T. Breuel

Keywords:
Nested dichotomies
Classifier ensembles
Multiclass classification
Decision trees

ABSTRACT

Ensemble methods are often able to generate more accurate classifiers than the individual classifiers. In multiclass problems, it is possible to obtain an ensemble combining binary classifiers. It is sensible to use a multiclass method for constructing the binary classifiers, because the ensemble of binary classifiers can be more accurate than the individual multiclass classifier.

Ensemble of nested dichotomies (END) is a method for dealing with multiclass classification problems using binary classifiers. A nested dichotomy organizes the classes in a tree, each internal node has a binary classifier. A set of classes can be organized in different ways in a nested dichotomy. An END is formed by several nested dichotomies.

This paper studies the use of this method in conjunction with ensembles of decision trees (forests). Although forests methods are able to deal directly with several classes, their accuracies can be improved if they are used as base classifiers for ensembles of nested dichotomies. Moreover, the accuracies can be improved even more using forests of nested dichotomies, that is, ensemble methods that use as base classifiers a nested dichotomy of decision trees. The improvements over forests methods can be explained by the increased diversity of the base classifiers. The best overall results were obtained using MultiBoost with resampling.

1. Introduction

Some methods for constructing classifiers are inherently binary (e.g., support vector machines). Nevertheless, other methods were first devised for binary problems, although later were extended for the multiclass case (e.g., decision trees, logistic regression, some neural networks). Hence, several approaches have been proposed for using binary methods with multiclass problems. Interestingly, these methods for binarizing multiclass problems can be useful even for methods that are able to construct multiclass classifiers, because they can improve the accuracy of the classifiers (Anand et al., 1995; Fürnkranz, 2002; Frank and Kramer, 2004). Therefore, they can be considered as ensemble methods, because the obtained classifiers are formed by several classifiers.

There are two basic approaches for combining binary classifiers for multiclass problems. The first one constructs a classifier for each class (Anand et al., 1995; Rifkin and Klautau, 2004). Each classifier discriminates between one class and the others. This approach is called one vs. all and one vs. the rest. The second approach is to construct a classifier for each pair of classes, that discriminates between them (Hastie and Tibshirani, 1998; Fürnkranz, 2002; Quost et al., 2007). This approach is called one vs. one, pairwise and round robin classification.

There are more complex approaches. A method that combines the previous ones is presented in (García-Pedrajas and Ortiz-Boyer, 2006). In error-correcting output codes (ECOC) (Dietterich and Bakiri, 1995) each binary classifier discriminates between two non-empty, disjoint, subsets of the set of classes. The union of the two subsets is the set of all the classes. That is, for all the binary classifiers each original class has to be in one of the subsets. In (Allwein et al., 2000) a generalized approach is presented, the binary classifiers are trained to discriminate between two subsets of classes, but not all the classes have to appear in one of the subsets.

Ensemble of nested dichotomies (END) is a recent approach for this problem (Frank and Kramer, 2004). A nested dichotomy (ND) is a binary tree, each node has a set of classes associated. In the internal nodes, the classes are split using a binary classifier to the two children. END combines several nested dichotomies, where each tree is generated randomly. In this case the word "ensemble" is not used to indicate a family of methods, but a specific one. The ensemble method used in END is based only on the randomness of the base classifier (in this case nested dichotomies). From the same training set, different classifiers can be obtained because the base method has an intrinsic source or randomness.

Another approach for using nested dichotomies is presented in (Pujol et al., 2006; Escalera et al., 2007). In this case only one tree is constructed, but instead of generating it randomly, an optimization

* Corresponding author. Tel.: +34 947258988; fax: +34 947258910.
E-mail addresses: jjrodriguez@ubu.es (J.J. Rodríguez), cgosorio@ubu.es (C. García-Osorio), jmaudes@ubu.es (J. Maudes).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.09.015

criterion is used. The tree is not used directly, but to generate an ECOC code.

ENDs have been studied with decision trees and logistic regression as binary classifiers. This paper studies their use with ensembles (e.g., bagging, boosting) of decision trees as binary classifiers. This approach improves the results of ENDs of decision trees and forests of multiclass trees. Moreover, another way of combining classical ensemble methods with nested dichotomies is considered. It can be seen as replacing the ensemble method used in END with other ensemble method. NDs of decision trees are used as the base classifiers for these ensemble methods. This approach gives even better accuracies.

Although the presented method could be used with ensembles of classifiers obtained using any method, this paper will consider decision trees. They are very commonly used as base classifiers in ensemble methods: they can be used with mixed type variables, are fast and sensitive to changes of the training data. The last property is relevant in ensemble methods because the diversity of the base classifiers is desirable for classifier ensembles.

The rest of the paper is organised as follows. Section 2 gives a brief introduction to ensembles of nested dichotomies and describes how to use them with decision forest. The experimental study, using 44 datasets and 51 variants of methods is presented in Section 3. In Section 4 kappa-error diagrams are used to analyze the relationship between ensemble methods when decision trees and nested dichotomies of decision trees are used as base classifiers. Finally, Section 5 presents some concluding remarks.

2. Nested dichotomies and decision forests

A nested dichotomy (Frank and Kramer, 2004) is a tree with the following properties:

- Each node has associated a non-empty set of classes.
- The root node includes all the classes, while the leaf nodes include only one class.
- The tree is strictly binary, that is, all the non-leaf nodes have two children.
- The classes in two siblings form a partition of the classes in the parent node. That is, their intersection is empty and their union is the set of all the classes in the parent.
- Each internal node has associated a binary classifier, that discriminates between the two set of classes in the children.

Fig. 1 shows two different dichotomies for a six class problem. Nested dichotomies are trees, but they are not decision trees. In the latter, internal nodes have an attribute that is used to split the examples, while in the former there is a classifier. In fact, the classifier associated to each internal node can be a decision tree (Frank and Kramer, 2004; Dong et al., 2005). Another difference is that in a nested dichotomy each class is assigned to only one leaf, while in a decision tree several leaves can predict the same class.

In the training phase, the binary classifiers are trained using all the examples available in the original training data for the classes in the node. Note that a training example can be misclassified by the binary classifier associated to a node. This can happen if the method includes some approach to avoid overfitting, for instance, pruning in decision trees. These misclassifications are not taken into account when constructing the nested dichotomy classifier. That is, when a binary classifier for a node is constructed, it is trained with all the training examples of the corresponding classes, regardless if the classifiers in the nodes (in the path from the root) would bring the example to that node.

In the classification phase, nested dichotomies could be used in the typical top-down traversal. The problem with this approach is

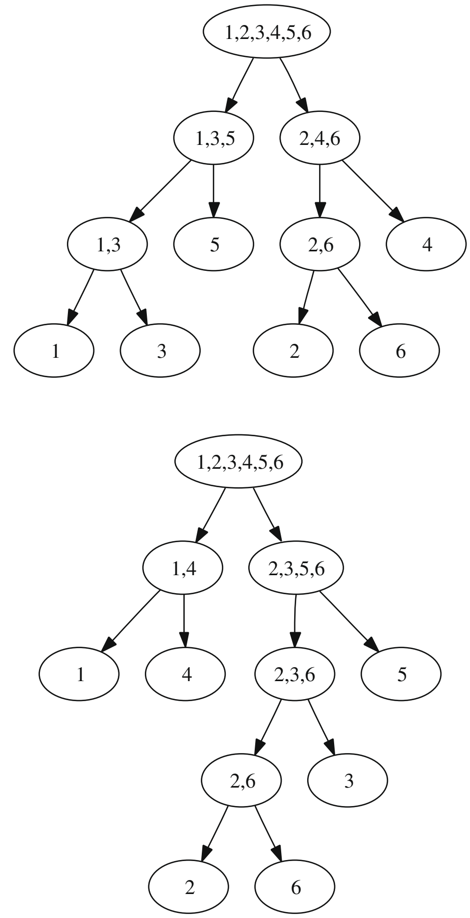

Fig. 1. Two nested dichotomies for six classes.

that the binary classifiers can make mistakes, and they cannot be corrected further down in the tree. Hence, all the tree is used. Let consider that the output of the binary classifiers are probabilities. Then, the probability of a class (leaf) can be calculated as the product of the probabilities given by the classifiers in the path from the root to the leaf. If the output of all the binary classifiers could only be zero or one, then one of the leaves would have a probability of one while the probability of the others would be zero. In this case, the result would be equivalent to a top-down traversal. In general, each class will have an associated probability. As usual, the class with the greatest probability will be predicted.

Fig. 2 shows a possible classification of an example using a nested dichotomy. The example is classified by all the binary classifiers, they will assign a probability to each subset of classes. These probabilities are shown in the arrows. The probability of a class is the product of the probabilities assigned to all the subsets that contain that class. In the figure, the class with the greatest probability is the class six, although in a top-down traversal the predicted class would be class three.

Given a set of classes, it is possible to construct different nested dichotomies, as shown in Fig. 1. Ensembles of nested dichotomies (ENDs) are formed by several nested dichotomies (Frank and Kramer, 2004; Dong et al., 2005). The structure of each nested dichotomy is obtained randomly. Predictions are obtained by averaging the output probabilities of each nested dichotomy.

It must be noted that the word "ensembles" as used in ENDs did not refer to the generic term but to an specific ensemble method. This method is based only in the randomness of the base classifier, different classifiers are obtained from the same training data be-