Computational Statistics and Data Analysis 54 (2010) 1535–1546

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Ensemble classification based on generalized additive models

Koen W. De Bock [a], Kristof Coussement [b,c], Dirk Van den Poel [a,*]

[a] Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Ghent, Belgium

[b] IESEG School of Management, Université Catholique de Lille (LEM, UMR CNRS 8179), Department of Marketing, 3 Rue de la Digue, F-59000 Lille, France

[c] University College HUBrussel, Faculty of Economics and Management, Stormstraat 2, B-1000 Brussels, Belgium

ARTICLE INFO

Article history:
Received 1 July 2009
Received in revised form 15 December 2009
Accepted 23 December 2009
Available online 4 January 2010

Keywords:
Data mining
Classification
Ensemble learning
GAM
UCI

ABSTRACT

Generalized additive models (GAMs) are a generalization of generalized linear models (GLMs) and constitute a powerful technique which has successfully proven its ability to capture nonlinear relationships between explanatory variables and a response variable in many domains. In this paper, GAMs are proposed as base classifiers for ensemble learning. Three alternative ensemble strategies for binary classification using GAMs as base classifiers are proposed: (i) *GAMbag* based on Bagging, (ii) *GAMrsm* based on the Random Subspace Method (RSM), and (iii) *GAMens* as a combination of both. In an experimental validation performed on 12 data sets from the UCI repository, the proposed algorithms are benchmarked to a single GAM and to decision tree based ensemble classifiers (i.e. RSM, Bagging, Random Forest, and the recently proposed Rotation Forest). From the results a number of conclusions can be drawn. Firstly, the use of an ensemble of GAMs instead of a single GAM always leads to improved prediction performance. Secondly, GAMrsm and GAMens perform comparably, while both versions outperform GAMbag. Finally, the value of using GAMs as base classifiers in an ensemble instead of standard decision trees is demonstrated. GAMbag demonstrates performance comparable to ordinary Bagging. Moreover, GAMrsm and GAMens outperform RSM and Bagging, while these two GAM ensemble variations perform comparably to Random Forest and Rotation Forest. Sensitivity analyses are included for the number of member classifiers in the ensemble, the number of variables included in a random feature subspace and the number of degrees of freedom for GAM spline estimation.

1. Introduction

Ensemble classifiers or multiple classifier systems (MCS) have received considerable attention in applied statistics (Hastie et al., 2001), machine learning (Dietterich, 2000) and pattern recognition (Kuncheva, 2004) for over a decade. Several studies demonstrate that the practice of combining several base classifier models into one aggregated classifier leads to significant gains in classification performance over its constituent members (Bauer and Kohavi, 1999). Over the years, different ensemble algorithms have been proposed, which differ along three structural dimensions of ensemble design, i.e. (i) the choice of the base or member classifier, (ii) the treatment of the input training data and (iii) the aggregation strategy for the outputs of member classifiers. Firstly, two broad strategies exist for choosing the members of an ensemble (Canuto et al., 2007). In hybrid ensembles, different types of algorithms are combined, whilst in non-hybrid ensembles, one classifier algorithm is chosen as base classifier, and replicated multiple times in order to constitute an ensemble. Secondly, many algorithms differ in terms of the treatment of the training data, used as input for each base classifier. Possibilities include

* Corresponding author. Tel.: +32 9 264 89 80, +32 484617961; fax: +32 9 264 42 79.
E-mail addresses: Koen.DeBock@UGent.be (K.W. De Bock), K.Coussement@ieseg.fr (K. Coussement), Dirk.VandenPoel@UGent.be (D. Van den Poel).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2009.12.013

data sampling schemes (Breiman, 1996), variable selection (Ho, 1998) or more complex data transformations (Kuncheva and Rodriguez, 2007; Rodriguez et al., 2006). A third ensemble design characteristic involves the fusion rule used for the ensemble member outputs, ranging from simple average aggregation to more complex combination rules (Skurichina and Duin, 2000).

The most popular classifier ensemble schemes are non-hybrid and apply a base classification algorithm to differently permutated training sets. A well-known method in this category is *Bagging* (Breiman, 1996), an acronym of bootstrap aggregating. Although numerous variations have been proposed since its introduction (e.g. Bauer and Kohavi, 1999; Bühlmann, 2002; Croux et al., 2007; Hothorn and Lausen, 2005), Breiman's original implementation is still a widely used ensemble classifier. In Bagging, each ensemble member is trained on a bootstrap sample of the training data, i.e. a random sample of observations drawn with replacement and having the same size as the original training data. Ensemble classification is obtained by means of uniform majority voting, where an unlabeled observation is assigned the class with the highest number of votes among the individual classifiers' predictions. Theoretically, bootstrapping can induce large differences in the constructed individual classifiers which substantially improves the accuracy of the ensemble classifier (Breiman, 1996).

Several variations upon Bagging have been proposed in search for further performance improvements. Two popular strategies involve (i) increasing variation in the training data for base classifiers and (ii) the use of alternative base classifier algorithms.

Firstly, several studies have shown the impact of variations of the input data used for the training of base classifiers. Varying the training data of the members of an ensemble is a strategy to increase diversity amongst member classifiers, which is generally perceived as a key driver of ensemble performance (Kuncheva and Whitaker, 2003). In the *Random Subspace Method* (RSM; (Bryll et al., 2003; Ho, 1998)), variables are randomly sampled to create training data sets for a decision tree ensemble. RSM, also referred to as Attribute Bagging (Bryll et al., 2003), specifies that each ensemble member is trained using a random feature subset (RFS), i.e. a random selection of explanatory variables sampled without replacement and of a predefined size. A related method is the *Random Forest* algorithm by Breiman (2001), which has demonstrated high classification performance in many fields of research (e.g. Archer and Kirnes, 2008; Diaz-Uriate and de Andres, 2006; Gislason et al., 2006; Prasad et al., 2006; Svetnik et al., 2003). A Random Forest combines Bagging and a specific form of RSM where random feature subset selection is performed at each node of a member decision tree. More recently, Rodriguez et al. (2006) proposed *Rotation Forest*, an ensemble classifier based on rotations of the feature space through principal component analysis (PCA). The purpose of Rotation Forest is to increase the individual classifier performance and the diversity within the ensemble. Diversity is achieved for each classifier by applying feature extraction, while one tries to increase the performance by using all principal components and training the model on the whole data set.

A second strategy to increase classification performance is to select an alternative base classifier algorithm. Many studies have proposed ensembles based on alternative base classifiers, such as Artificial Neural Networks (Hansen and Salamon, 1990; Maclin and Shavlik, 1995; Opitz and Shavlik, 1996; Schwenk and Bengio, 2000; Zhou et al., 2002), Support Vector Machines (Kim et al., 2002, 2003), parametric regression techniques (Prinzie and Van den Poel, 2008) and nonparametric regression techniques (Borra and Di Ciaccio, 2002).

This paper introduces generalized additive models (GAMs; (Hastie and Tibshirani, 1986)), a statistical technique for nonparametric or semi-parametric modeling, as ensemble members for ensemble classification. It contributes to the ensemble literature by proposing three GAM ensemble classifiers for binary classification based on Bagging, the Random Subspace Method and a combination of both. In each of the proposed methods, average aggregation is used to combine posterior class membership probabilities, generated by the member GAMs. In an experimental validation using 12 binary classification data sets from the UCI repository, classification performance is compared to single GAM performance, and amongst the three GAM ensemble algorithms. Further, the GAM ensemble approaches are compared to their counterparts based on decision tree base classifiers: RSM, Bagging, and Random Forest, which implements both Bagging and a specific form of RSM. The recently proposed Rotation Forest algorithm is included as an additional high performance benchmark, which also consists of decision trees trained in parallel, and demonstrated superior performance over Random Forest and ordinary Bagging earlier (Rodriguez et al., 2006).

The paper is organized as follows. In Section 2, GAMs are reviewed and three variations of the GAM ensemble algorithm are presented. Section 3 reports the experimental results. Section 4 includes sensitivity analyses of classification performance based on the ensemble size, the number of variables per random feature subspace and the number of degrees of freedom for spline smoothing. In the last section, conclusions and suggestions for further research are given.

2. Methodology

This section briefly presents an overview of generalized additive models and the GAM specification used for ensemble members, and presents details of the proposed ensemble classifiers. Consider the following notations. X is a set of p independent variables, $X = \{X_1, \ldots, X_p\}$ and Y is a binary response variable. Denote a training data set by $D = \{(x_i, y_i)\}_{i=1}^{n}$ consisting of n observations. Each observation (x_i, y_i) is a combination of an input vector x_i and a response y_i with $y_i \in \{0, 1\}$. Training a base classifier C_l involves using the training data to formalize a mapping of the input variable space onto the binary response variable, Y. The prediction of a base classifier C_l is the conditional class membership probability $P(Y = 1|X)$. An ensemble classifier C consists of m base classifiers; $C = \{C_1, C_2, C_3, \ldots, C_m\}$.