FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Ensemble classification based on generalized additive models

Koen W. De Bock[1]

Kristof Coussement[2]

Dirk Van den Poel[3]

December 2009
2009/625

[1] PhD Candidate, Ghent University
[2] Assistant Professor at IESEG School of Management, Université Catholique de Lille
[3] Corresponding author: Prof. Dr. Dirk Van den Poel, Professor of Marketing Modeling/analytical Customer Relationship Management, Faculty of Economics and Business Administration, dirk.vandenpoel@ugent.be; more papers about customer relationship management can be obtained from the website: www.crm.UGent.be

D/2009/7012/77

Ensemble classification based on generalized additive models

Koen W. De Bock*, Kristof Coussement °+, Dirk Van den Poel*

*Ghent University, Faculty of Economics and Business Administration, Department of Marketing,

Tweekerkenstraat 2, B-9000 Ghent, Belgium

° IESEG School of Management, Université Catholique de Lille (LEM, UMR CNRS 8179),

Department of Marketing,

3 Rue de la Digue, F-59000 Lille, France

+University College HUBrussel, Faculty of Economics and Management,

Stormstraat 2, B-1000 Brussels, Belgium

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 comparable performance 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.

Keywords: Data mining, Classification, Ensemble learning, GAM, UCI

Koen W. De Bock: Koen.DeBock@UGent.be

Kristof Coussement: K.Coussement@Ieseg.fr

Corresponding Author: Dirk Van den Poel: Dirk.VandenPoel@UGent.be ; Tel.: + 32 9 264 89 80; Fax: + 32 9 264 42 79

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 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, $X = \{ X_1,...,X_p \}$ 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; $C = \{C_1, C_2, C_3,..., C_m\}$.

2.1. Generalized additive models

Generalized additive models are used as base classifiers in the proposed ensemble algorithms. GAMs were proposed in Hastie and Tibshirani (1986) and have been strongly accepted in several domains as a flexible modeling technique, suited for capturing non-linear, unspecified relationships between predictor variables and a response variable (Berg, 2007; Clements et al., 2005; Kawakita et al., 2005). GAMs generalize the family of generalized linear models (GLMs), by replacing the linear functional form by a sum of smooth functions (Hastie and Tibshirani, 1986, 1987, 1990), enabling the discovery of a nonlinear fit between a variable and a response. In order

to formalize the relationship between a binary response Y and independent variables $\{X_1,...,X_p\}$, the response variable is assumed to follow a binomial distribution and the logistic link function is applied. As many data sets contain discrete variables, a linear parametric part is introduced into the GAM model to allow the inclusion of these categorical variables. As such, the GAM specification that is used in the proposed ensemble classifiers is a logistic, semi-parametric additive model of the following form:

$$\text{logit}(P(Y=1|X)) \equiv \log\left\{\frac{P(Y=1|X)}{1-P(Y=1|X)}\right\} = \sum_{j=1}^{p_c} \mathrm{s_j}(\mathrm{X_j}) + \sum_{k=1}^{p_b} \beta_k \mathrm{X_k} \qquad (1)$$

with $X_j, j=1,...,p_c$ as continuous variables and $X_k, k=1,...,p_b$ as dummy-coded components of categorical variables. In this study, the functions $s_1(X_1), s_2(X_2),...,s_{p_c}(X_{p_c})$ that estimate the nonparametric trend for the dependence of the logit on $X_1, X_2,...,X_{p_c}$ are smoothing splines. A smoothing spline for variable X solves the following optimization problem: amongst all functions $\eta(x)$ with continuous second order derivatives, find the function that minimizes the penalized residual sum of squares via $\sum_{i=1}^{n}(y_i - \eta(x_i))^2 + \lambda\int_a^b(\eta''(t))^2\,dt$ (2)

where λ is a fixed constant and $a \le x_1 \le x_2 \le ... \le x_n \le b$. The goodness-of-fit is measured by the first part of equation (2), while the second term is a penalty term that penalizes curvature in the function, where the degree of penalization is determined by the smoothing parameter λ. The complexity of $\eta(x)$ is measured by λ which is inversely related to the degrees of freedom (*df*). If λ is small (i.e. the *df* are large), $\eta(x)$ is any function that approaches an interpolation to the data. When λ is large (i.e. the *df* are small), $\eta(x)$ is closely related to a simple least squares fit. It is shown that an explicit and unique minimizer for equation (2) exists, i.e. a natural cubic spline with knots at the unique values of x_i (Hastie and Tibshirani, 1990). This study specifies λ

corresponding to a small number of *df* as applied in several GAM-related papers (e.g., Baccini et al., 2007; Marx and Eilers, 1998; Zwane and van der Heijden, 2004) and the examples provided by Hastie and Tibshirani (1990). In order to optimize the GAM, the local scoring algorithm (Hastie and Tibshirani, 1986) is applied.

2.2. GAM Ensemble Classifiers

Based on the previous GAM specification, three ensemble classifier algorithms based on GAMs are proposed. *GAMbag* implements Bagging, *GAMrsm* implements the Random Subspace Method, and *GAMens* implements both. The pseudo code for the algorithms is presented in Figure 1.

[INSERT FIGURE 1 HERE]

The GAM ensemble algorithms require specification of a number of input parameters. A first set of parameters specifies the ensemble strategy. *GAMbag* incorporates Bagging, which requires parameter *b* to be true. *GAMrsm* only implements the Random Subspace Method and requires parameter *s* to be true. For *GAMens*, both parameters are set to true. Secondly, the *m* parameter designates the number of desired GAM base classifiers to be included in the ensemble classifier. Thirdly, the desired number of variables to be selected as random feature subspaces is required (*r* parameter). Finally, specification of the number of degrees of freedom to be used in the smoothing spline estimation is required (*df* parameter).

In the prediction phase, outputs of the ensemble member GAMs are combined into an ensemble prediction $C(x)$ by means of average aggregation (or mean combination rule), which is used in many well-known ensemble classifiers (e.g. RSM and Rotation Forest). In the GAM-based

ensembles, an ensemble prediction for a given observation takes the average of the posterior class membership probabilities produced by the individual ensemble members.

3. Experimental Validation

In order to assess the performance of the proposed algorithms, an experimental validation is performed on 12 two-class classification data sets from the UCI repository (Asuncion and Newman, 2007) that are often used to compare classifier performance (e.g., Kuncheva and Rodriguez, 2007; Rodriguez et al., 2006; Zhang and Zhang, 2008). All categorical variables are dummy coded. The characteristics of the data sets are found in Table 1.

[INSERT TABLE 1 HERE]

The validation of the proposed GAM ensembles is threefold. Firstly, the predictive performances of the GAM-based ensemble classifiers GAMbag, GAMrsm and GAMens are compared to a single GAM model. Secondly, the results of the proposed ensembles are compared against each other. Thirdly, the GAM ensembles are compared against their corresponding decision tree counterparts: RSM, Bagging, Random Forests and Rotation Forest. All these algorithms are well-known and often used in classifier benchmark studies (e.g. Bauer and Kohavi, 1999; Rodriguez et al., 2006). Decision tree Bagging and Random Forest implementations originate from the adabag (Alfaro et al., 2006) and randomForest (Liaw and Wiener, 2002) packages in R (R Development Core Team, 2009). To allow for a fair comparison between GAMbag and Bagging, the fusion rule of Bagging is changed to average aggregation. Rotation Forest is implemented in MATLAB based on the implementation as described in Rodriguez et al. (2006) and the single GAM classifiers are implemented using the gam package in R (Hastie, 2008). The GAMens variations are also implemented in R and made publicly available in the new GAMens package (De Bock et al., 2009) accessible via http://cran.r-project.org. Default settings are used for all classifier

parameters. All decision tree based algorithms use unpruned CART decision trees. The size of random feature subspaces for Random Forests is set equal to the square root of the total number of features, as suggested by Breiman (2001). This setting is also used for RSM, GAMrsm and GAMens. The number of disjoint feature subsets of the Rotation Forests is chosen in order to obtain a fixed number of features per feature subset of three, as suggested by Rodriguez et al. (2006). Moreover, the GAM-based algorithms are trained using four degrees of freedom per smoothing spline. All ensemble-based algorithms are constructed using 100 constituent members.

In order to methodologically benchmark the performance between the algorithms correctly, a 5 times 2-fold cross-validation is performed. Within a 2-fold cross-validation, the training set is randomly split into two parts; the first part is used for model training, while the second part is used for model validation and vice versa. The performance of the classification methods is assessed in terms of Area Under the Receiver Operating Characteristics curve (AUC or AUROC) as argued by several authors like Provost et al. (2000) or Langley (2000) to be an objective performance criterion, well-suited for the comparison of classifier performance. For the detection of significant differences in classifier performance, Demšar (2006) suggests the use of the non-parametric Friedman test (Friedman, 1937, 1940) with the Bonferroni-Dunn post-hoc test (Dunn, 1961) for comparing a control classifier with the proposed benchmarks over multiple datasets .

For every data set and per algorithm, Table 2 provides average AUC values with standard deviations for the 5 times 2-fold cross-validation. The highest average AUC per data set is indicated in bold.

[INSERT TABLE 2 HERE]

Table 3 and Table 4 provide the results of the corresponding Friedman tests with the Bonferroni-Dunn post-hoc tests. The figures in both tables represent the average rank differences, i.e. the

difference between the average rank of the control classifier (CC) and that of the benchmark algorithm (BA). The lower the average rank, the better the algorithm. This implies that a negative average rank difference means that the control classifier has a lower (better) average rank than the benchmark algorithm and vice versa.

[INSERT TABLE 3 HERE]

[INSERT TABLE 4 HERE]

The following conclusions emerge from Table 3 and Table 4. Firstly, a comparison among the GAM ensemble variations uncovers that GAMens and GAMrsm significantly outperform GAMbag at a 5% significance level, while GAMens and GAMrsm appear to have no considerable difference in classification performance.

Secondly, the results reveal that building an ensemble of GAMs is a viable strategy to increase classification performance over the single GAM classifier. This holds for each of the three proposed GAM ensembles; GAMbag, GAMrsm and GAMens.

A third consideration involves a comparison between the newly-proposed ensembles of GAMs and the ensembles of decision trees.

It appears that GAMs as base classifiers in Bagging (GAMbag) perform equally well as using standard decision trees (Bagging). Moreover, GAMbag performs comparably to RSM, Random Forest and Rotation Forest.

Furthermore, GAMrsm demonstrates superior performance over its counterpart RSM and Bagging. The strong performance of GAMrsm is also demonstrated when compared to Random Forest and Rotation Forest. In these cases, GAMrsm performs equally well with respect to these high-performing benchmarks.

GAMens also exhibits good classification performance when compared to the other benchmark ensemble algorithms. GAMens performs significantly better than RSM and Bagging, and there is no significant difference in classification performance when compared to Random Forest and Rotation Forest. While the differences are not significant, GAMrsm and GAMens both exhibit lower average ranks compared to Random Forest and Rotation Forest.

In the following Section, additional experiments are performed to investigate the impact on classifier performance of varying three ensemble parameters: ensemble size (*m*), random feature subset size (*r*) and number of degrees of freedom for smoothing spline estimation (*df)*.

4. Algorithm Parameter Sensitivity Analyses

The described sensitivity analyses are based on average results of a 5 times 2-fold cross-validation using the 12 binary UCI data sets as described in Table 1. All algorithm parameters not under consideration are chosen as in the previous section.

Appendix 1 demonstrates the effect of increasing ensemble size on the cross-validated AUC performance of GAMbag, GAMrsm, GAMens and the other ensemble benchmarks. Three relevant insights are summarized based on Appendix 1. Firstly, it is clear that the gain in AUC performance is high for small ensemble sizes (i.e., less than 25 base classifiers) and rapidly decreases as the forests continue to grow. This trend is confirmed by numerous other studies (e.g. Ho, 1998; Prinzie and Van den Poel, 2008). Secondly, it appears that the order in which algorithms perform remains rather stable over the range of ensemble sizes, while only a few shifts occur in small ensemble size regions. For example, trends from the GAMrsm algorithm demonstrate a faster increase than GAMens in a majority of data sets (i.e. *German, Hepatitis, Ionosphere, Mammo, Sonar and Wisconsin breast*), resulting in few shifts in the small ensemble size regions. A third observation is that the performance of the GAM-based ensembles is more

sensitive to ensemble size for smaller ensemble sizes, i.e. ensemble sizes with less than 10 members. For larger ensemble sizes, the GAM ensembles are less sensitive when compared to the benchmark algorithms. Finally, the plots in Appendix 1 demonstrate for each of the data sets that the classification performance has stabilized at an ensemble size of 100 members, which makes this setting a safe choice.

Appendix 2 shows the sensitivity of GAMrsm and GAMens performance depending on the size of random feature subsets (RFS). Plots are also included for RSM and Random Forests, in order to investigate whether these algorithms demonstrate comparable sensitivity to the parameter. The resolution of the RFS size range depends upon the total number of features in a data set, i.e. in steps of 1 for data sets with 20 features or less, in steps of 2 for data sets with features between 20 and 50 features, in steps of 3 for the German data set with 59 features and in steps of 6 for the Horse colic data set. The plots uncover three distinct patterns for GAMrsm and GAMens: an inverted U, a descending and an invariant curve. The AUC performance follows an inverted U-curve in four out of the twelve data sets, meaning that there is an increase in performance until a maximum is reached, followed by a downward trend. Further, the descending trend is observed in four out of the twelve data sets, where larger random feature subsets show a negative impact on AUC performance. In four out of the twelve data sets, the performance is more or less invariant to changes in RFS size. Overall, RSM and Random Forests demonstrate deviating trends and are on average slightly less sensitive to specification of the parameter. The vertical dotted lines represent the random feature subset sizes that are used in the experimental validation in Section 3. In this validation, RFS size is equal to the square root of the total number of features. The plots demonstrate that the default setting of the RFS size as the square root of the number of features as suggested by Breiman (2001) is close to the optimum, i.e. where AUC reaches a maximum value, for a majority of the data sets. This observation confirms the experiments on RFS size in Random

Forests (Bernard et al., 2009) and the related Extremely Randomized Trees Ensemble (Geurts et al., 2006).

The sensitivity analysis on the variation of the *df* parameter did not reveal significant differences in prediction performance. The sensitivity plots are not included in the paper, but they can be obtained by contacting the corresponding author.

5. Conclusions

In this paper generalized additive models (GAMs, Hastie and Tibshirani, 1986) are introduced as base classifiers for binary ensemble classification using Bagging and/or the Random Subspace Method. GAMs constitute a powerful nonparametric technique to model nonlinear relationships between explanatory variables and a response variable. We present and evaluate three algorithms using GAMs as base classifiers: *GAMbag* applying Bagging, *GAMrsm* implementing the Random Subspace Method and *GAMens* combining both previous approaches. The results of the experimental validation on 12 UCI binary data sets show evidence of the advantage of using GAMs as members in an ensemble classifier. Firstly, constructing an ensemble of GAMs increases classification performance over a single GAM classifier. Secondly, both GAMrsm and GAMens perform better than GAMbag, while there are no considerable differences in performance between GAMrsm and GAMens. Thirdly, the results demonstrate that GAMrsm and GAMens significantly improve performance over RSM and Bagging and perform at least as well as Random Forest and Rotation Forest on a majority of data sets, while GAMbag performs comparably well to RSM and Bagging.

Moreover, sensitivity analyses are performed in order to investigate the sensitivity of classification performance to algorithm parameters (i) ensemble size, (ii) number of elements in random feature subsets and (iii) number of degrees of freedom for smoothing spline estimation.

Sensitivity plots in (i) demonstrate that the GAM-based ensembles are overall less sensitive to ensemble size compared to the benchmark algorithms for medium to large ensemble sizes (10 – 100), while for small ensemble sizes (less than 10 ensemble members) AUC performance is generally more sensitive to ensemble size. In (ii), the dependence of classification performance upon random feature subspace size specification is demonstrated. The plots indicated near-optimal performance of the default setting for random feature subspace size, i.e. equal to the square root of the total number of independent variables in the data set. The sensitivity analyses on the number of degrees of freedom (iii) do not show significant differences in classification performance.

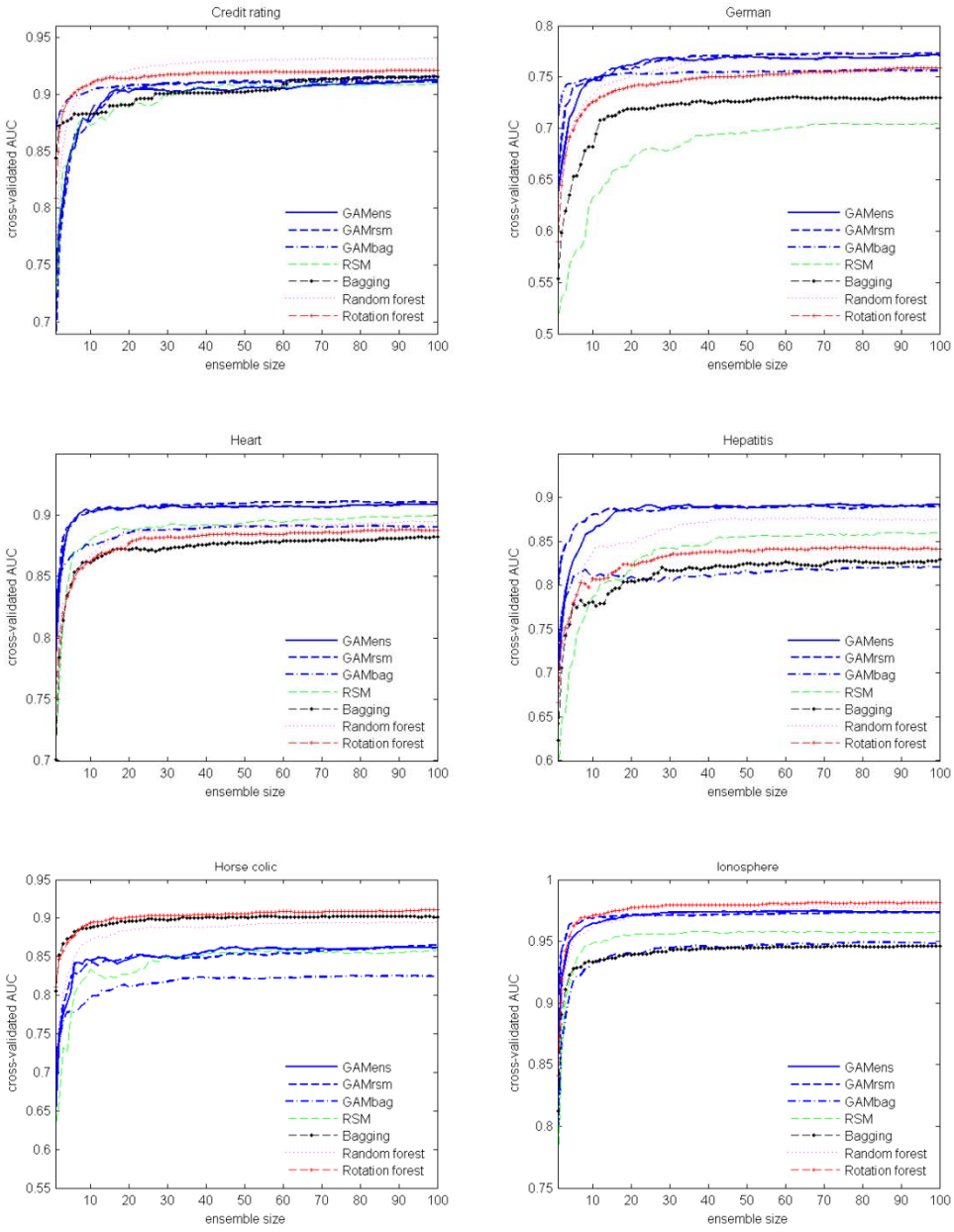
Whilst we are confident that our study adds significant value to the current ensemble learning literature, a number of limitations and directions for future research are identified. Firstly, the proposed GAM ensemble algorithms are validated in a binary classification context, based on the original specification of generalized additive models by Hastie and Tibshirani. Future work can extend the proposed GAM-based ensemble classifiers to multiclass classification based on an extension of the GAM framework to multi-class problems, as for example proposed by Abe (1999). Secondly, a number of well-known benchmark algorithms are selected based on the frequently-used Bagging and RSM based on decision trees. One can of course argue about using (i) other base classifiers in the ensemble (e.g. Support Vector Machines (Kim et al., 2002) or Neural Networks (Opitz and Shavlik, 1996)) or (ii) other ensemble strategies (e.g. Boosting (Friedman et al., 2000), Bragging (Bühlmann, 2002) or Trimmed Bagging (Croux et al., 2007)) to compare the proposed GAM ensemble algorithms to. Thirdly, the main ensemble components in GAMens, GAMrsm and GAMbag are based on the manipulation of the training data for the member classifiers via Bagging and RSM, and average aggregation as a fusion rule for ensemble member output. While it is not feasible to take into account all variations from those components

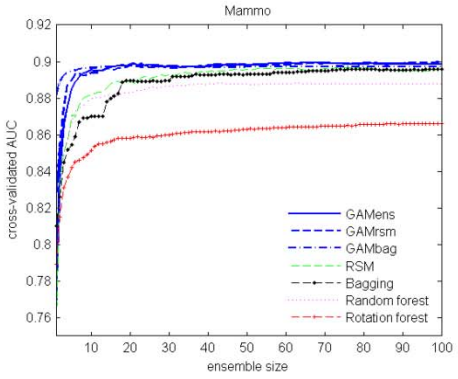
in this study, further investigation may be conducted to analyze alternative approaches and their impact on classification performance.

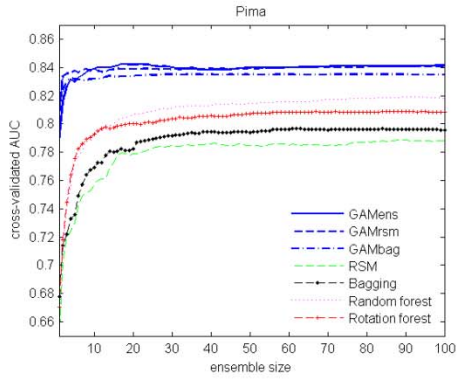
Acknowledgments

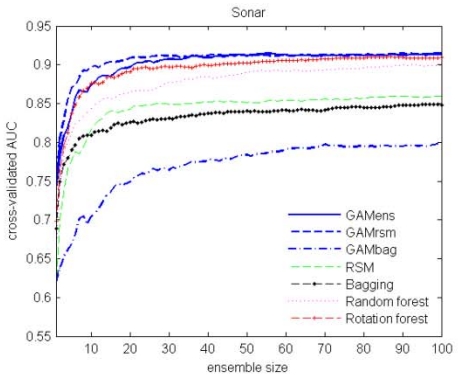
The authors thank Stefan Van Aelst for friendly-reviewing this research paper and Ghent University for funding the PhD project of Koen W. De Bock.

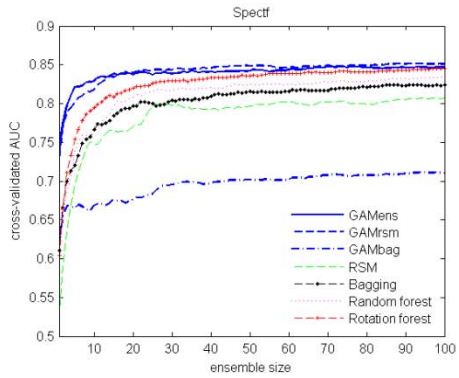
Appendix 1: Sensitivity plots for number of members in the ensemble classifier

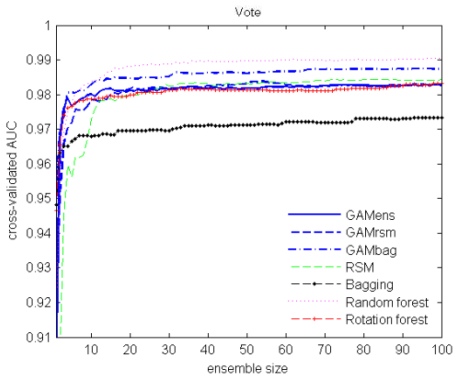


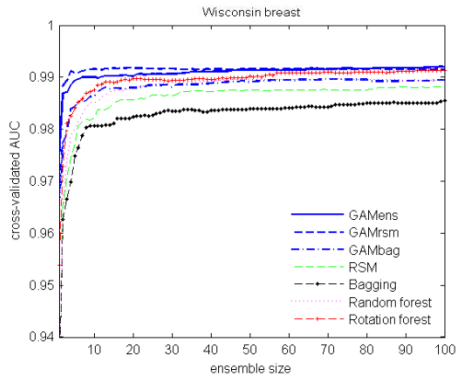

Mammo
cross-validated AUC
ensemble size
GAMens
GAMrsm
GAMbag
RSM
Bagging
Random forest
Rotation forest


Pima
cross-validated AUC
ensemble size
GAMens
GAMrsm
GAMbag
RSM
Bagging
Random forest
Rotation forest

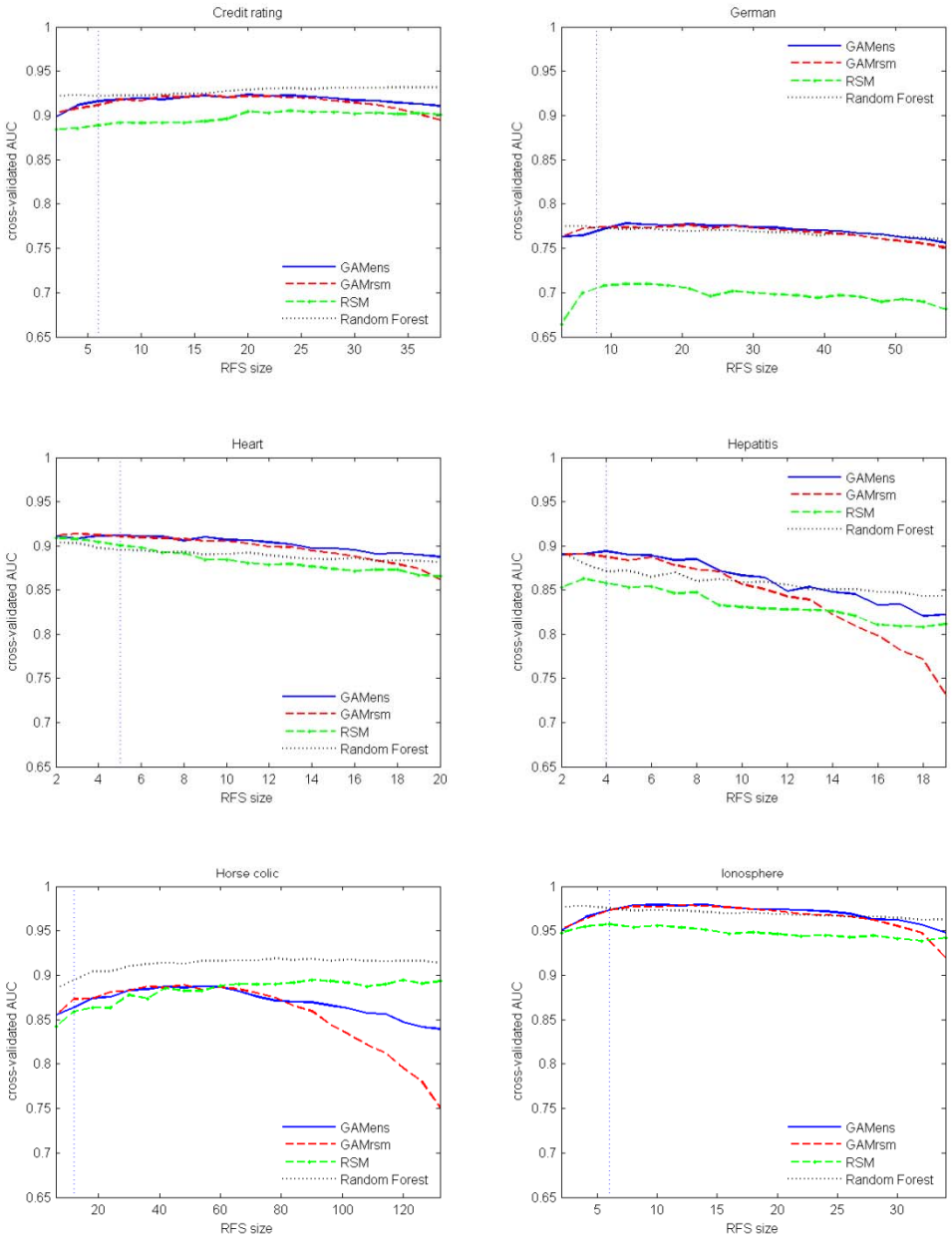

Sonar
cross-validated AUC
ensemble size
GAMens
GAMrsm
GAMbag
RSM
Bagging
Random forest
Rotation forest

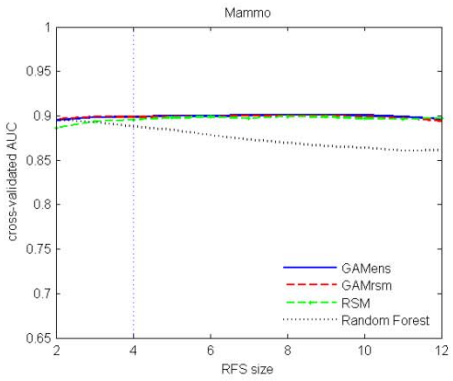

Spectf
cross-validated AUC
ensemble size
GAMens
GAMrsm
GAMbag
RSM
Bagging
Random forest
Rotation forest

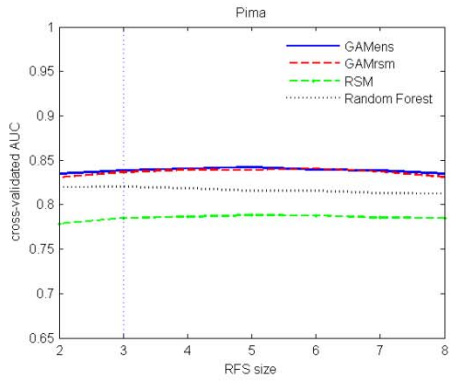

Vote
cross-validated AUC
ensemble size
GAMens
GAMrsm
GAMbag
RSM
Bagging
Random forest
Rotation forest

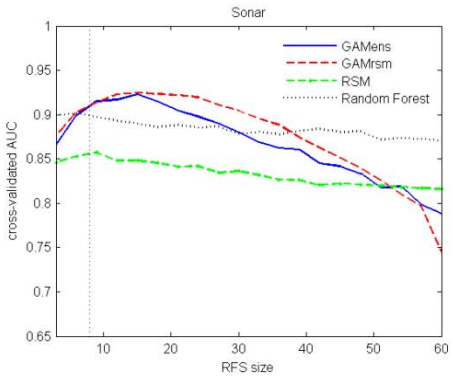

Wisconsin breast
cross-validated AUC
ensemble size
GAMens
GAMrsm
GAMbag
RSM
Bagging
Random forest
Rotation forest

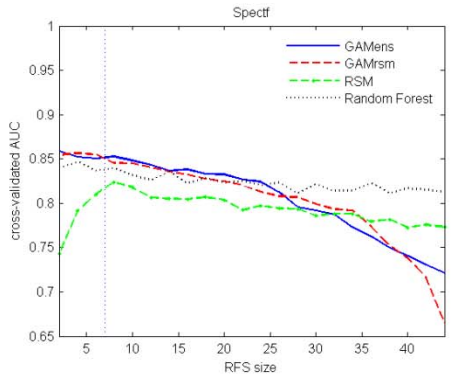
Appendix 2: Sensitivity plots for random feature subspace (RFS) size

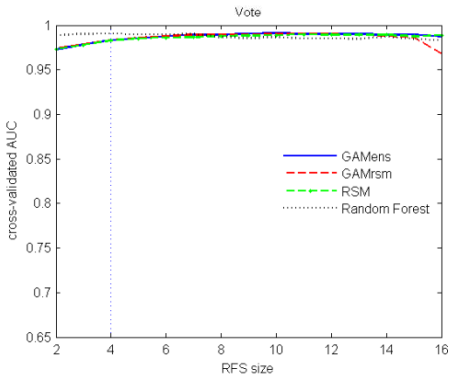


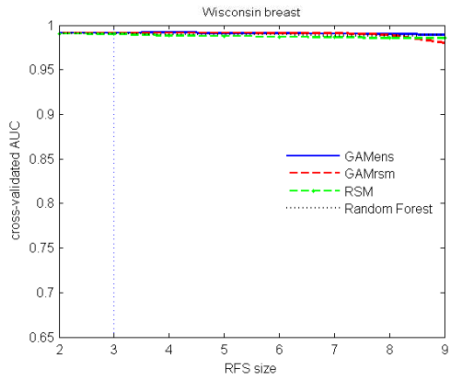

Mammo
cross-validated AUC
RFS size
GAMens
GAMrsm
RSM
Random Forest


Pima
cross-validated AUC
RFS size
GAMens
GAMrsm
RSM
Random Forest


Sonar
cross-validated AUC
RFS size
GAMens
GAMrsm
RSM
Random Forest


Spectf
cross-validated AUC
RFS size
GAMens
GAMrsm
RSM
Random Forest


Vote
cross-validated AUC
RFS size
GAMens
GAMrsm
RSM
Random Forest


Wisconsin breast
cross-validated AUC
RFS size
GAMens
GAMrsm
RSM
Random Forest

References

[1] Abe, M., 1999. A generalized additive model for discrete-choice data. Journal of Business & Economic Statistics, 17 (3) 271-84.

[2] Alfaro, E., Gámez, M., and García, N., 2006. adabag: Applies Adaboost.M1 and Bagging, R Package version 1.1.

[3] Archer, K. J. and Kirnes, R. V., 2008. Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 52 (4) 2249-60.

[4] Asuncion, A. and Newman, D. J., 2007. UCI Machine Learning Repository. Irvine, CA., University of California, School of Information and Computer Science.

[5] Baccini, M., Biggeri, A., Lagazio, C., Lertxundi, A., and Saez, M., 2007. Parametric and semi-parametric approaches in the analysis of short-term effects of air pollution on health. Computational Statistics & Data Analysis, 51 (9) 4324-36.

[6] Bauer, E. and Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36 (1-2) 105-39.

[7] Berg, D., 2007. Bankruptcy prediction by generalized additive models. Applied Stochastic Models in Business and Industry, 23 (2) 129-43.

[8] Bernard, S., Heutte, L., and Adam, S., 2009. Influence of hyperparameters on Random Forest accuracy. In: Benediktsson, J. A., Kittler, J., and Roli, F. (Eds.), Proc. of 8th International Workshop on Multiple Classifier Systems (MCS 2009), Springer-Verlag, Berlin / Heidelberg.

[9] Borra, S. and Di Ciaccio, A., 2002. Improving nonparametric regression methods by bagging and boosting. Computational Statistics & Data Analysis, 38 (4) 407-20.

[10] Breiman, L., 1996. Bagging predictors. Machine Learning, 24 (2) 123-40.

[11] Breiman, L., 2001. Random forests. Machine Learning, 45 (1) 5-32.

[12] Bryll, R., Gutierrez-Osuna, R., and Quek, F., 2003. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition, 36 (6) 1291-302.

[13] Bühlmann, P., 2002. Bagging, subagging and Bragging for improving some prediction algorithms, in: Akritas, M. G. and Politis, D. N. (Eds.), Recent Advances and Trends in NonParametric Statistics. Elsevier, Amsterdam.

[14] Canuto, A. M. P., Abreu, M. C. C., Oliveira, L. D., Xavier, J. C., and Santos, A. D., 2007. Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensembles. Pattern Recognition Letters, 28 (4) 472-86.

[15] Clements, M. S., Armstrong, B. K., and Moolgavkar, S. H., 2005. Lung cancer rate predictions using generalized additive models. Biostatistics, 6 (4) 576-89.

[16] Croux, C., Joossens, K., and Lemmens, A., 2007. Trimmed bagging. Computational Statistics & Data Analysis, 52 (1) 362-68.

[17] De Bock, K. W., Coussement, K., and Van den Poel, D., 2009. GAMens: Applies GAMens, GAMrsm and GAMbag ensemble classifiers, R Package version 1.0.

[18] Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7 1-30.

[19] Diaz-Uriate, R. and de Andres, S. A., 2006. Gene selection and classification of microarray data using random forest. Bmc Bioinformatics, 7.

[20] Dietterich, T. G., 2000. Ensemble methods in machine learning. In: Kittler, J. and Roli, F. (Eds.), Proc. of 1st International Workshop on Multiple Classifier Systems (MCS 2001), Springer-Verlag, Berlin / Heidelberg.

[21] Dunn, O. J., 1961. Multiple comparisons among means. Journal of the American Statistical Association, 56 (293) 52-64.

[22] Friedman, J., Hastie, T., and Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. Annals of Statistics, 28 (2) 337-74.

[23] Friedman, M., 1940. A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, 11 (1) 86-92.

[24] Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32 (200) 675-701.

[25] Geurts, P., Ernst, D., and Wehenkel, L., 2006. Extremely randomized trees. Machine Learning, 63 (1) 3-42.

[26] Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R., 2006. Random Forests for land cover classification. Pattern Recognition Letters, 27 (4) 294-300.

[27] Hansen, L. K. and Salamon, P., 1990. Neural Network Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (10) 993-1001.

[28] Hastie, T., 2008. gam: Generalized Additive Models, R package version 1.0.

[29] Hastie, T. and Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.

[30] Hastie, T. and Tibshirani, R., 1986. Generalized additive models. Statistical Science, 1 (3) 297-318.

[31] Hastie, T. and Tibshirani, R., 1987. Generalized Additive Models: Some applications. Journal of the American Statistical Association, 82 (398) 371-86.

[32] Hastie, T., Tibshirani, R., and Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York.

[33] Ho, T. K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (8) 832-44.

[34] Hothorn, T. and Lausen, B., 2005. Bundling classifiers by bagging trees. Computational Statistics & Data Analysis, 49 (4) 1068-78.

[35] Kawakita, M., Minami, M., Eguchi, S., and Lennert-Cody, C. E., 2005. An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. Fisheries Research, 76 (3) 328-43.

[36] Kim, H. C., Pang, S., Je, H. M., Kim, D., and Bang, S. Y., 2003. Constructing support vector machine ensemble. Pattern Recognition, 36 (12) 2757-67.

[37] Kim, H. C., Pang, S., Je, H. M., Kim, D., and Bang, S. Y., 2002. Support vector machine ensemble with bagging. In: Lee, S. E. Verri A. (Eds.), Proc. of 1st International Workshop on Pattern Recognition with Support Vector Machines, Springer-Verlag, Berlin / Heidelberg.

[38] Kuncheva, L. I., 2004. Combining pattern classifiers: methods and algorithms. John Wiley & Sons, Hoboken, New Jersey.

[39] Kuncheva, L. I. and Rodriguez, J. J., 2007. Classifier ensembles with a random linear oracle. IEEE Transactions on Knowledge and Data Engineering, 19 (4) 500-08.

[40] Kuncheva, L. I. and Whitaker, C. J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51 (2) 181-207.

[41] Langley, P., 2000. Crafting papers on Machine Learning. In: Langley, P. (Eds.), Proc. of 17th International Conference on Machine Learning (ICML-2000), Stanford University, Stanford.

[42] Liaw, A. and Wiener, M., 2002. Classification and Regression by randomForest. R News, 2 (3) 18-22.

[43] Maclin, R. and Shavlik, J. W., 1995. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. In: Mellish, C. S. (Eds.), Proc. of 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Morgan-Kauffman, San Francisco, CA.

[44] Marx, B. D. and Eilers, P. H. C., 1998. Direct generalized additive modeling with penalized likelihood. Computational Statistics & Data Analysis, 28 (2) 193-209.

[45] Opitz, D. W. and Shavlik, J. W., 1996. Generating accurate and diverse members of a neural-network ensemble. Advances in Neural Information Processing Systems, 8 535-41.

[46] Prasad, A. M., Iverson, L. R., and Liaw, A., 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems, 9 (2) 181-99.

[47] Prinzie, A. and Van den Poel, D., 2008. Random forests for multiclass classification: Random MultiNomial Logit. Expert Systems with Applications, 34 (3) 1721-32.

[48] Provost, F., Fawcett, T., and Kohavi, R. The Case against Accuracy Estimation for Comparing Induction Algorithms. In: Shavlik, J. (Eds.), Proc. of 15th International Conference on Machine Learning (ICML-1998), Morgan Kaufman, San Francisco, CA.

[49] R Development Core Team, 2009. R: A Language and Environment for Statistical Computing, Vienna, Austria.

[50] Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J., 2006. Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (10) 1619-30.

[51] Schwenk, H. and Bengio, Y., 2000. Boosting neural networks. Neural Computation, 12 (8) 1869-87.

[52] Skurichina, M. and Duin, R. P. W., 2000. The role of combining rules in bagging and boosting. In: Ferri, F. J., Inesta, J. M., Amin, A., and Pudil, P. (Eds.), Proc. of Joint International Workshops SSPR 2000 and SPR 2001, Springer-Verlag, Berlin / Heidelberg.

[53] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P., 2003. Random forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences, 43 (6) 1947-58.

[54] Zhang, C. X. and Zhang, J. S., 2008. RotBoost: A technique for combining Rotation Forest and AdaBoost. Pattern Recognition Letters, 29 (10) 1524-36.

[55] Zhou, Z. H., Wu, J. X., and Tang, W., 2002. Ensembling neural networks: Many could be better than all. Artificial Intelligence, 137 (1-2) 239-63.

[56] Zwane, E. N. and van der Heijden, P. G. M., 2004. Semiparametric models for capture-recapture studies with covariates. Computational Statistics & Data Analysis, 47 (4) 729-43.

GAMbag, GAMrsm and GAMens algorithms

Input

- D : a training set, $D = \{(x_i, y_i)\}_{i=1}^{n}$; $x_i \in X \subset R^p$; $y_i \in Y = \{0,1\}$
- m: number of GAMs in the ensemble
- r: number of randomly selected variables, $r \leq p$
- df: number of degrees of freedom used to estimate smoothing splines $s_1(.), s_2(.), \ldots, s_p(.)$
- b: true for Bagging (*GAMens* and *GAMbag)*
- s: true for Random Subspace Method (*GAMens* and *GAMrsm)*

Training phase

For $l = 1,2,\ldots,m$

- If s then $R_l = D$ with subset of r randomly selected variables from X
  Else $R_l = D$
- If b then D_l = bootstrap sample of R_l
  Else $D_l = R_l$
- In D_l, identify continuous variable set $R_{l,c}$ and binary variable set $R_{l,b}$. The numbers of elements of both sets are indicated by $p_{l,c}$ and $p_{l,b}$.
- Estimate l-th base classifier C_l as a semi-parametric GAM with logistic link function and df degrees of freedom for smoothing splines $s_1(.), s_2(.), \ldots, s_p(.)$

$$C_l : P_l(Y = 1 | X) = 1 / \left\{ 1 + \exp\left(-\left(\sum_{j=1}^{p_{l,c}} s_{l,j}(X_j) + \sum_{k=1}^{p_{l,b}} \beta_{l,k} X_k \right) \right) \right\}$$

with $X_j \in R_{l,c}$ and $X_k \in R_{l,b}$

Prediction phase

- The probability for observation x to belong to class 1, predicted by ensemble classifier C, is

$$C(x) = \frac{1}{m} \sum_{l=1}^{m} C_l(x)$$

Figure 1: GAM ensemble algorithms pseudo code

Data set				
	Observations	*Discrete variables*	*Cont. Variables*	*Variables after dummy coding*
Credit rating	690	9	6	38
German	1000	13	7	59
Heart	270	0	13	20
Hepatitis	155	13	6	19
Horse colic	368	16	7	133
Ionosphere	351	0	34	34
Mammo	961	2	3	12
Pima	768	0	8	8
Sonar	208	0	60	60
Spectf	267	0	44	44
Vote	435	16	0	16
Wisconsin breast	699	0	9	9

Table 1: UCI data set characteristics

Dataset	Algorithm							
	RSM	*Bagging*	*Random Forest*	*Rotation Forest*	*GAM*	*GAMens*	*GAMrsm*	*GAMbag*
Credit rating	0.9093 (0.019)	0.9154 (0.013)	**0.9314** (0.008)	0.9207 (0.008)	0.892 (0.015)	0.9126 (0.015)	0.9154 (0.014)	0.9107 (0.011)
German	0.7041 (0.026)	0.73 (0.025)	0.7727 (0.019)	0.7591 (0.018)	0.749 (0.018)	0.7713 (0.024)	**0.7729** (0.024)	0.7563 (0.017)
Heart	0.8994 (0.019)	0.8822 (0.026)	0.8947 (0.02)	0.8875 (0.023)	0.862 (0.036)	0.9088 (0.019)	**0.9106** (0.019)	0.8901 (0.022)
Hepatitis	0.8606 (0.036)	0.8293 (0.053)	0.8748 (0.035)	0.8417 (0.051)	0.722 (0.08)	**0.892** (0.038)	0.8897 (0.036)	0.8207 (0.052)
Horse colic	0.8582 (0.043)	0.9019 (0.035)	0.8947 (0.031)	**0.9108** (0.034)	0.714 (0.047)	0.8618 (0.029)	0.8652 (0.026)	0.8249 (0.03)
Ionosphere	0.9576 (0.011)	0.9465 (0.017)	0.9764 (0.008)	**0.9812** (0.008)	0.831 (0.026)	0.9737 (0.012)	0.9729 (0.01)	0.9482 (0.011)
Mammo	0.8957 (0.008)	0.8958 (0.009)	0.8877 (0.009)	0.8662 (0.01)	0.894 (0.008)	0.8985 (0.008)	**0.8996** (0.005)	0.8972 (0.007)
Pima	0.7881 (0.012)	0.7958 (0.031)	0.8182 (0.016)	0.8083 (0.017)	0.831 (0.016)	**0.8413** (0.015)	0.8409 (0.017)	0.8349 (0.015)
Sonar	0.8596 (0.044)	0.8482 (0.039)	0.8995 (0.037)	0.9092 (0.028)	0.733 (0.045)	0.9136 (0.033)	**0.9153** (0.027)	0.7976 (0.025)
Spectf	0.8076 (0.026)	0.8237 (0.026)	0.834 (0.02)	0.8447 (0.019)	0.625 (0.062)	0.847 (0.013)	**0.8513** (0.015)	0.7104 (0.053)
Vote	0.9844 (0.005)	0.9734 (0.012)	**0.9905** (0.004)	0.9832 (0.006)	0.968 (0.019)	0.9827 (0.007)	0.9827 (0.007)	0.9874 (0.005)
Wisconsin breast	0.9882 (0.004)	0.9854 (0.005)	0.9895 (0.005)	0.9913 (0.004)	0.98 (0.012)	**0.992** (0.003)	0.9915 (0.003)	0.9895 (0.006)

Table 2: Classification performance in AUC: average (standard errors)

		Benchmark Algorithms (BA)		
		GAMbag	*GAMrsm*	*GAMens*
Control Classifier (CC)	*GAMbag*	x	1.42**	1.08**
	GAMrsm	-1.42**	x	-0.33
	GAMens	-1.08**	0.33	x

*= p < 0.10, ** = p< 0.05

Table 3: Average rank differences (CC-BA) among GAM ensembles

		Benchmark Algorithms (BA)				
		GAM	*RSM*	*Bagging*	*Random Forest*	*Rotation Forest*
Control Classifier (CC)	*GAMbag*	-1.92*	-0.25	-0.58	1.33	0.92
	GAMrsm	-3.50**	-2.25**	-2.50**	-0.75	-1.00
	GAMens	-3.33**	-2.08**	-2.25**	-0.50	-0.83

*= $p < 0.10$, ** = $p < 0.05$

Table 4: Average rank differences (CC-BA) between GAM ensembles and benchmark algorithms