VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0045

Performance of Resampling Methods Based on Decision Trees, Parametric and Nonparametric Bayesian Classifiers for Three Medical Datasets

Małgorzata M. Ćwiklińska-Jurkowska[1]

[1] Department of Theoretical Foundations of Biomedical Sciences and Medical Computer Science, Collegium Medicum, Nicolaus Copernicus University, Poland

Abstract. The figures visualizing single and combined classifiers coming from decision trees group and Bayesian parametric and nonparametric discriminant functions show the importance of diversity of bagging or boosting combined models and confirm some theoretical outcomes suggested by other authors. For the three medical sets examined, decision trees, as well as linear and quadratic discriminant functions are useful for bagging and boosting. Classifiers, which do not show an increasing tendency for resubstitution errors in subsequent boosting deterministic procedures loops, are not useful for fusion, e.g. kernel discriminant function. For the success of resampling classifiers' fusion, the compromise between accuracy and diversity is needed. Diversity important in the success of boosting and bagging may be assessed by concordance of base classifiers with the learning vector.

Introduction

Combining classifiers with very close discriminant properties is not useful in the models' fusion. Diversity is suspected to be important for the success of merging classifiers (Banfield et al., 2005; Bi, 2011; Brown et al., 2010; Kuncheva et al., 2000; Kuncheva, 2003; Melville et al., 2005), as well for homogenous combining (with the same kind of constituent classifiers) as for heterogonous classifiers (Kuncheva et al., 2002; Shipp et al., 2002). A question arises: how is this diversity realized during the process of resampling datasets according to the most popular fusion procedures: bagging and boosting (Breiman, 1996, 1998; Freund et al., 1997). It is known that for unstable classifiers resampling methods are useful to decrease generalization error in comparison to the single classifier. Popular unstable classifiers are decision trees or neural networks; however, trees are characterized by smaller computational and memory complexity than neural networks. Thus, for resampling ensemble methods such as boosting or bag-

ISBN 978–83–7431–392–6 ISSN 0860-150X

ging (bootstrap aggregating), the most common constituent classifiers are trees. Therefore, most examinations concerning bagging and boosting fusion are connected with decision trees. Few authors consider linear classifiers joined with bagging (Skurichina et al., 2002, Vu et al., 2009) and boosting (Skurichina et al., 2002). Conclusions concerning the usefulness of bagging linear discriminant classifiers are not concordant in the literature. For large data sets, much bigger than the number of variables, Breiman (1996) concluded that bagging LDC is not useful. Also, further works were based on large data sets (Breiman, 1998; Dietterich, 2000). Skurichina et al. (2002) stated that for critical training sample sizes (when the number of training objects is comparable with data dimensionality) a bagging ensemble is useful for LDC, because then LDC is an unstable classifier. Vu et al. (2009) concluded for small sets of microarray data that bagging is useful for unstable trees and neural networks, but not for LDCs. For boosting, however, fusion might be useful for large training sample sizes (Skurichina et al., 2002).

The aim of this work is to examine the usefulness of fusion of trees and other constituent classifiers like Bayesian parametric and nonparametric discriminant functions in the context of diversity, depending on the number of loops, the size and character of the set, the type of constituent classifiers and kind of merging. The visualization of the single and combined Bayesian classifiers are elaborated to examine the trends of learning curves and to aid the exploration of the effectiveness of bagging and boosting fusion based on such base discriminant functions.

Methods

The practical aims of the discriminant analysis pertaining to medical problems are to find variables with the biggest discriminant power, which is useful for differentiation, and next to support medical decisions according to the chosen classification model based on those variables. High performance of classification models confirms the correctness of the selected set of variables. For classifiers based on single or merged trees, selection of variables is incorporated in the modeling step, so is not necessary as the first step, though it may be beneficial. The theoretical aim is to define which of the potential discriminant methods has the lowest misclassification rate. An application of classification methods to three real medical decision problems of which, the most essential information, hopefully, was included into the data sets, was performed (Table 1). Modeling on selected variables sets may support medical decisions for those problems.

Table 1. Characterization of applied medical data sets

Data Set	Medical decision problem, coming from:		Number of of cases	Variables number	Number of groups
WDBC	Malignant or benign tumor of the breast	University of Wisconsin Hospitals	569	30	2
Breast cancer	Relapse of breast cancer	Institute of Oncology University Medical Center Ljubljana	286	9	2
Schizophrenia	Discrimination between schizophrenic and control group based on EEG parameters	Department of Psychiatry Nicolaus Copernicus University	80	36	2

The classification methods applied, single and combined, are presented in Duda et al. (2001), Kotsiantis et al., 2006, Rokach (2009, 2010a, 2010b), Webb (2002). For linear discriminant classifiers (LDC) and quadratic discriminant classifiers (QDC) variables were selected by Wilks statistics (measuring variability between groups in relation to total variability). For the nonparametric kernel classifier, the choice was made according to minimization of the 1-Nearest Neighbor leave-one-out error, because 1-Nearest Neighbor is not complex, so is a quick classifier. Performance of various discriminant methods was assessed using apparent (Resubstitution), cross-validation (CV) and leaving-one-out method errors (LOOUT). Exploring performance of classifiers, single and combined, is based on "learning curves", where apparent, CV or LOOUT errors are plotted versus the number of resampling loops. Figures were obtained by use of independent programs based on the PRTOOLS package for Matlab (Duin et al., 2007).

The classical methodological technique (supplying parametric discrimination functions) assumes a jointly normal distribution of the predictive variables for optimality. However, in many problems, this assumption (or assumption of equal covariance matrices in differentiated groups for quadratic discrimination) can be doubtful. Various procedures have been elaborated as alternatives to classical discriminant analysis. One of them is the Parzen classifier, based on kernel estimation of density in discriminated populations. This discriminant Bayesian procedure, in opposite to Bayesian linear and quadratics discriminant functions is nonparametric, i.e. no assumption on distribution in discriminated populations is made. Another discriminant

procedure, coming from quite different methodology, which does not assume anything of distributions, so is also nonparametric, is the creation of trees (Rokach et al., 2005; Quinlan, 1987).

Currently, classifier researchers tend to combine procedures, based on similar type or different base classifier (Kotsiantis et al., 2006; Rokach, 2009, 2010a, 2010b). Especially big attention is focused on families of classifiers coming from two ideas: bootstrap aggregations and boosting (Breiman, 1996, 1998, Freund et al., 1997). Because these combining procedures are time consuming, the constituent classifiers with great complexity may cause computational problems. Combining simple and not optimal classifiers may to some extent bypass the drawbacks of such classifiers. Additionally, relaxed assumptions connected with resampling of the whole training data set may reduce deficiencies of base classifiers built on training sets.

Datasets

Three data sets, characterized by different difficulties, were examined (Table 1). The material, used in the discriminant analysis, comes from a few medical centers. The patients were divided into two groups. Classification was performed by clinicians. The discriminant problems included in the data sets are described in Table 1. Besides group classification, each patient was described, using clinical variables of a number not bigger than both group sizes. Two first sets from Table 1 come from the UCI Machine Learning Repository (Bache et al., 2013). Those sets are of various difficulties in making decisions. The biggest Wisconsin Diagnostic Breast Cancer data set(WDBC) is relatively simple (212 malignant patients from 569), the Breast Cancer data set is more difficult (85 relapses among 286 patients). Schizophrenic data sets consist of 50 schizophrenic patients among 80.

The data sets presented in Table 1 do not have strict multidimensional normal distribution in discriminated groups. However, the Breast Cancer data set shows the largest deviation from normality. The Schizophrenia data set has 36 variables that are linear combinations of disjoint subsets of 96 primary variables coming from computerized EEG equipment. Before summation, those elementary variables were standardized and transformed by logarithm to approach to one-dimensional gaussianity. The process of summing variables was done to reduce the large number of EEG parameters according to the medical knowledge, i.e. the summation of EEG parameters was done with six brain regions, three on the left and three on the right side of the brain.

Classifiers' Visualization for WDBC Data Set

At the beginning of the analysis let's look at the behavior of tree errors during the boosting procedure, because trees are the most common constituent classifiers applied in this ensemble. On the consecutive figures, there are overlaid learning curves of apparent (resubstitution), cross-validation and leave-one-out errors. Each elaborated learning curve represents the dependence of a chosen kind of errors assessment on the number of resampling loops. The figure with overlaid learning curves contains errors of individual classification models for increasing the number of loops (i.e. constituent classifiers number), represented by horizontal axis ($L = 1, \ldots, 100$: "Number of loops in resampling") with the estimated averages of apparent errors for first x loops ("Err. in conseq.loops" and "Mean err. of const.classif", respectively). Fusion errors after merging results of L loops ($L = 1, \ldots, 100$): apparent, cross validated with ten folds and leave-one-out – are also drawn ("Ensemble err", "Ensemble CV10err","Ensemble LOOUTerr", respectively).

Two following plots (Figures 1 and 6) represent overlaid learning curves connected with the diagnosis of breast cancer based on 30 discriminating variables in the WDBC data set (Table 1).

Because the performance of elaborated learning curves for a number of loops extending $L = 100$ was not meaningfully changed, the number of loops on the horizontal axes with many overlaid lines was cut to 100 in order to make clear overlaying representation of single classifiers loops possible (if it is helpful, some results for bigger numbers of combined constituent classifiers to 200 may be quoted as numerical, not graphical results). In this way, on one plot we can observe the behavior of individual classifiers constructed on subsamples and the combined classifiers built on all numbers of loops from one till the current at the same time. Namely, on those learning curves, the gray lines without marks denote apparent classification errors made by constituent classifiers, a line with diamonds represents the mean of apparent constituent classification errors. A line marked by triangles represents ensemble apparent classification errors (bagging or boosting), a line with stars shows ensemble CV errors for the increasing number of combined classifiers and, similarly, a line with circles displays leave-one-out ensemble errors.

The instability of the base classifier is expressed in the diversity of errors after resampling of the data set (oscillating line). For the decision tree committee (Figure 1) we can observe useful diversity. The smallest apparent error is achieved for the first resampling loop. For boosting trees, a very

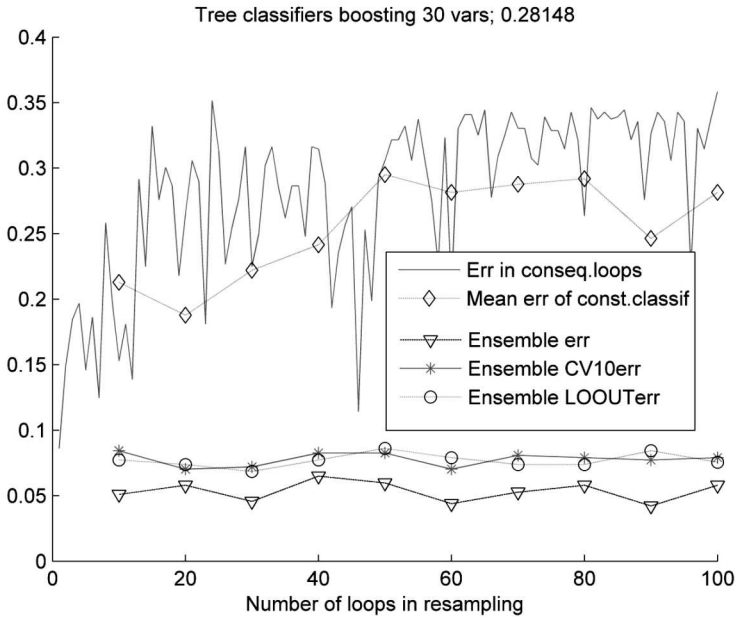

Figure 1. Dependence of classification errors on number of tree boosting loops, for WDBC set recognition based on 30 discriminating variables

high increase for the first 4 loops can be noticed (gray lines severely increasing from 0.09 to 0.2, Figure 1). For decision trees, a boosting aggregation procedure does not have the constant trend of increasing average apparent error over the whole range of the examined number of loops, e.g. significant reduction of mean constituent classifier errors is obtained for 80 resampling loops. The mean of single trees apparent errors have a general increasing tendency till 80 loops. For 50 loops, the maximum level of mean apparent individual errors, equal to 0.29, is obtained and for higher numbers of loops, the levels of individual apparent classifiers are not considerably increasing. The apparent error average of trees constructed on bootstrap subsamples with 100 loops is equal to 0.281 (Figure 1) and for a bigger number of loops, it still generally increases to 0.32 in 200 loops. According to CV and LOOUT assessment of generalization properties, for a relatively easily classified WDBC set, joining thirty loops of boosting trees is sufficient to minimize generalization errors. The smallest CV and LOOUT errors for whole training sets are equal to 0.07 and 0.04 (for 30 loops), respectively. For 100 (Figure 1) and 200 loops CV errors are 0.078 and 0.084, respectively. Similarly, concerning LOOUT errors: for 100 (Figure 1) and 200 loops, LOOUT errors for the whole training set are 0.075 and 0.08, respectively.

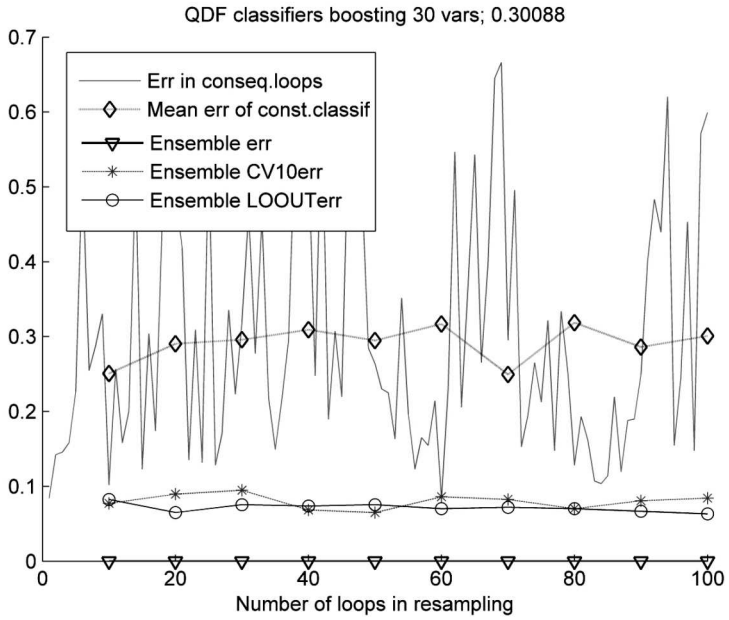

Figure 2. Dependence of classification errors on number of QDC boosting loops for WDBC set

Single classical trees are classifiers characterized by identification regions with boundaries consisting of linear parts parallel (perpendicular) to axes. Constituent decision trees depend strongly on the subset of the training set drawn by resampling, because of trees' instability. A single unstable tree classifier on the whole data set is characterized by a CV error equal to 0.074 and LOOUT error equal to 0.076. The resampling method improves the stability of a decision tree. The diversity of classifiers in consecutive loops is observable in error diversity.

Single quadratic classifier regions have boundaries that are multidimensional quadrics. The kind of quadric depends on the relationship between covariance matrices within discriminated groups. Each loop results in a different boundary. For boosting fusion of the quadratic discriminant function (QDC), the smallest apparent error is achieved for the fist resampling loop. A very high increase for the first 6 loops can be noticed (gray line severely increasing from 0.09 to 0.53, solid black line without additional marks in Figure 2). Looking at the lines with diamonds in Figure 2, we can observe that for QDC, the boosting aggregation procedure does not have the constant trend of increasing average apparent error over the whole range of the examined number of loops. In the QDC boosting combiner, the mean apparent errors of constituent classifiers have a general tendency to increase until

the first 60 loops (Figure 2), but a significant reduction of mean constituent classifier errors is obtained for 70 combined classifiers. The average apparent error follows a strict increasing trend until 40 loops. The mean of apparent errors of trees constructed on subsequent subsamples with 100 loops is equal to 0.281. The smallest LOOUT and CV errors are equal to 0.065, obtained by CV for 50 loops and by LOOUT on the 20^{th} loop. According to CV and LOOUT error assessment, the results of boosting trees are comparable with QDC (Figures 1 and 2).

Classifiers' Visualization for Breast Cancer Data Set

The analysis for the next set will begin by examining the constituent Parzen classifier in detail. The approximated value of optimal radius (r) is chosen with the usage of CV error. In boosting the Parzen classifier, the mean apparent constituent classifier does not have any tendency (Figure 3). Average reclassification errors substantially vary across the whole range of examined loops till 100. In contrast, for linear discriminant function (LDC) on the same Breast Cancer Data Set (Figure 4) there is an observable increasing trend, which confirms the intuition that consecutive loops of boosting procedure work on samples more difficult for classification (boosting is focused on objects problematic for identification). For strong and flexible Parzen classifiers there are fewer difficult patterns to identify than for other Bayesian classifiers (smaller CV and LOOUT errors of the constituent Parzen kernel classifier based on the whole data set, equal to 0.25 and 0.26, respectively). Modeling on a selected set of variables may support medical decisions for those problems. In the context of Parzen boosting results for the Breast Cancer data set, it should be noticed that the Parzen classifier has the beneficial property of tuning the parameter of smoothing (r) on the basis of CV error, so it has better discriminant properties than other single examined Bayesian methods and obtains the smallest generalization error of them – CV 10 equal to 0.21, LOOUT error level of 0.22.

Quite a different situation concerning Breast Cancer data was observed for LDC boosting classifier (Figure 4). For LDC boosting, the mean of apparent single classifier errors, which is interpretable by lines with diamonds, has a tendency to increase till 90 loops and after that to not grow until the number of 200 loops is reached. The single constituent LDC discriminant function obtains classification errors close to errors of the ensemble – about 0.25 – after resampling of the dataset. Few classifiers reach the level

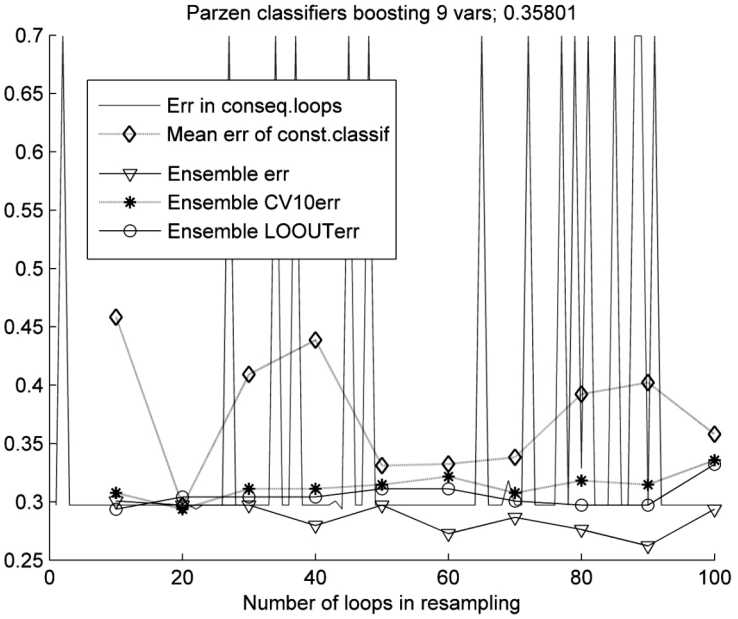

Figure 3. Dependence of classification errors on the number of Parzen boosting loops for breast cancer recognition based on 9 variables

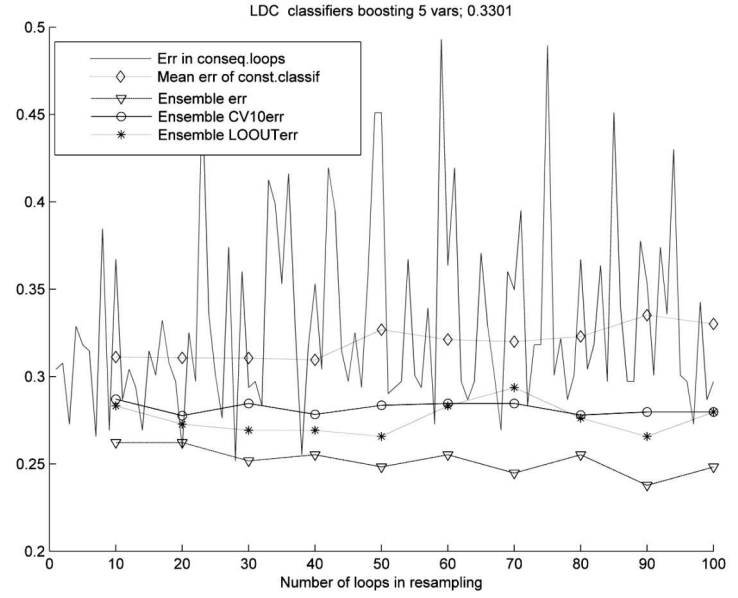

Figure 4. Dependence of classification errors on number of LDC boosting loops for breast cancer recognition based on the 5 best discriminating variables

of error of almost 0.5. The mean of reclassification errors after 100 loops is equal to 0.33, while fusion constructed by 100 loops of bagging obtains a generalization error level of 0.28.

Single linear classifier regions have boundaries that are multidimensional hyperplanes, which depend on the relationship between covariance matrices and centroids of populations. Each loop results in a different boundary. The diversity of discriminant functions in consecutive loops is visible as variability in error levels.

Classifiers' Visualization for Schizophrenia Data Set

To visualize more exactly classification during resampling methods, the schizophrenia data set was reduced to two dimensions, according to the optimization of variables selection criterion. In the data connected with the problem of recognizing schizophrenia, let's analyze the visualization of another resampling method – bootstrap aggregation (bagging).

The QDC classifier base error is 0.24, when estimated by CV, which means, after comparison with CV fusion errors not exceeding 0.14 after fourty loops, that the fusion bagging committee in this classification problem is certainly useful (Figure 5). Assessment of CV errors for bagging QDC shows the greatest decrease till 20 loops, where the smallest level error, 0.10, is reached. Increasing the number of loops above 30 is not useful. In contrast to boosting (e.g. Figure 2), the mean apparent errors of individual classifiers does not have any clear tendency. Because bagging is based on trials that are nondeterministic (as opposed to boosting, in which the draw in subsequent loops is associated with assigning higher weights for the observation poorly classified in the previous step), the average error base classifiers vary and no clear trends are established (line with diamonds on Figure 6). CV errors of the single LDC on the whole data set is equal to 0.216, thus it appears that CV error estimate (not bigger than 0.201) shows substantial reduction in CV error after the LDCs combining. Thus, according to CV error, LDC bagging gives improvement of the generalization properties, though CV errors are very diverse along increasing numbers of loops (Figure 6). For numbers of loops between 100 and 200, the CV is about 0.22, so adding more constituent classifiers is not beneficial (not included in the figure).

Boosting LDC for the schizophrenia dataset with 2 selected variables is not beneficial in comparison to bagging, because CV and LOOUT errors for all numbers of loops till 200 exceed 0.2 (not presented in graphical way).

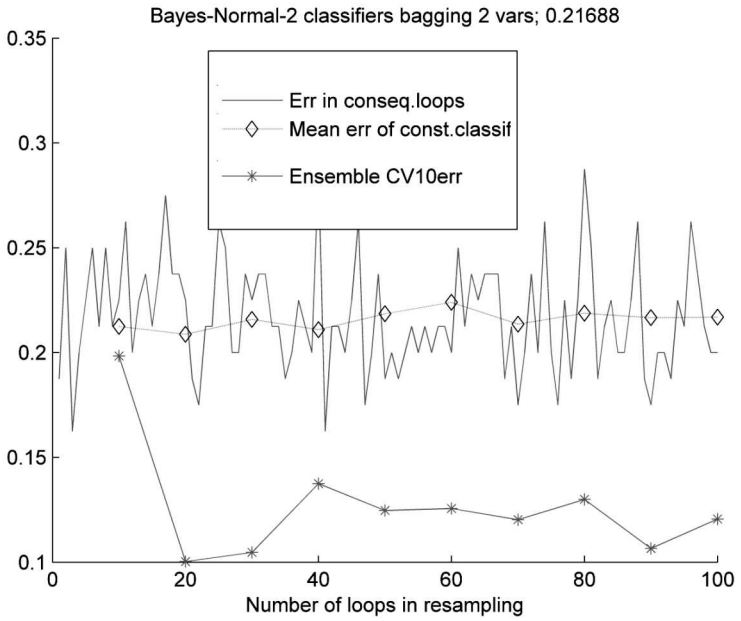

Figure 5. Dependence of classification errors on number of QDC bagging loops for schizophrenia recognition based on 2 best discriminating variables

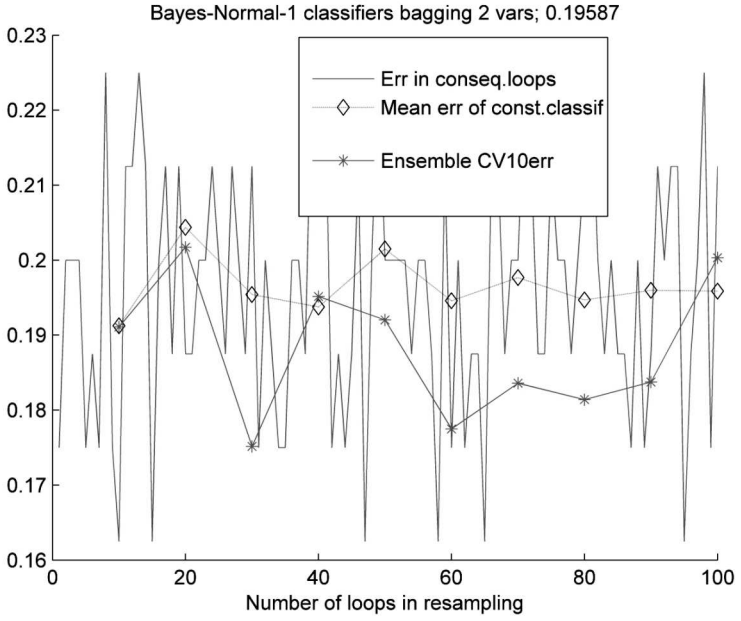

Figure 6. Dependence of classification errors on number of LDC bagging loops for schizophrenia recognition based on 2 best discriminating variables

Bagging is known as appropriate for smaller data sets, while boosting is elaborated for rather bigger data sets (Skurichina, 2001). The Schizophrenia data set is the smallest of the examined sets, though, when this set is considered with the number of chosen two best discriminating variables, it is not very small in comparison to dimensionality.

Discussion

According to all learning curves, apparent constituent error averages are substantially higher and considerably more diverse than ensemble errors. Assessment of misclassification rates for ensemble methods may also be regarded in the context of the diversity of a set of constituent classifiers. From learning curves we can observe that the average apparent error of individual loops is much higher than the error of the weighted voting for boosting loops (Figures 1–4). This corresponds to a linear formula for combining individual models, where the committee error is the sum of individual model error averages and of components related to the measure of the heterogeneity of a set of models. For example, error-ambiguity decomposition was proposed by Krogh et al. (1995) for regression tasks. Krogh et al. (1995) proved that for a single observation (x), the squared error of the combined estimator, obtained by weighing the base linear classifiers results, is expressed as the difference of weighted average base classifiers' squared error and component specifying ambiguity (indicating the diversity of base classifiers). If, instead of considering the arithmetic mean, the geometric mean is chosen as the fusion procedure, then as s a measure of the accuracy of the combined classifier, the mean squared error can no longer be applied, but the Kulback-Leibler (D_{KL}) directed divergence for two distributions can. On the base of Heskes (1998) work, Brown et al. (2010) gave the formula for D_{KL}:

$$D_{KL} = (y \| \bar{f}) = \frac{1}{L} \sum_{l=1}^{L} D_{KL}(y \| f_l) - \frac{1}{L} \sum_{l=1}^{L} (\bar{f} \| f_l)$$

where f_l is the l-th discriminant function ($l = 1, \ldots, L$) and y is the learning vector.

Thus the results coming from figures representing the learning curves that the ensemble errors benefit mean apparent errors can be explained by attached theoretical reasons.

From attached theoretical background we can see that in order to achieve a small error classifier fusion, a compromise is needed between diversity and the average error of the base classifiers. However, those decom-

position models take into account the estimated mean square error of the linear combination function or Kullback-Leibler divergence, while in bagging and boosting methods we are dealing with a merger constructed by a vote, weighted or unweighted. Brown et al. (2010) presented an analogous decomposition of errors in the case of classifiers combined by voting. Decomposition for the majority vote error and *"0–1"* loss function for L base classifiers with labels $\{-1, 1\}$ is the following:

$$e_{MV} = \int e_{AvgInd} - \int\limits_{x+} \frac{1}{L} \sum_{l=1}^{L} d_l(x) + \int\limits_{x-} \frac{1}{L} \sum_{l=1}^{L} d_l(x)$$

where

- e_{MV} is Majority Vote (MV) error
- e_{AvgInd} is individual classifiers error average
- d_l – binary variable denoting the mismatch of l-th base classifier ($l = 1, \ldots, L$) with the vote, subspaces of the data set:
- x_+ where the vote fusion is correct
- x_- where the vote fusion incorrect.

The last decomposition indicates that the mean of individual errors and the diversity both have an impact on the classifier committee error. The second component is beneficial diversity and the third component is unbeneficial diversity. By considering the third component, we can explain the phenomenon that combining diverse but inaccurate classifiers is not beneficial and that the compromise between diversity and accuracy of component classifiers is needed.

The success of boosting and bagging is connected with the diversity of the ensemble. Different sets in resampling cause different classifiers (with different classification regions) and they have different performance. Diversity of classifiers reflects diversity of classification errors. However, they are not the same, because constituent classifiers with the same errors may be different, e.g. may have quite dissimilar boundaries, and therefore different sensitivity and specificity. The component classifiers' diversity, visualized in the presented learning curves, reflects the concordance of the constituent classifier with the learning vector. Another, though to some extent analogous, measure of diversity used for the active enforcement of base classifiers differentiation, to build the accurate ensembles, was applied by Melville et al. (2005); this is the average binary incompatibility of individual classifiers' results with the aggregated classifier.

The tendency of increase of the average resubstitution errors was found for boosting methods. For bagging, such a clear trend cannot be seen. In particular, the fastest increase in apparent errors, compared with fusion meth-

ods of Bayesian classification, was found in the trees. They are known as unstable classifiers. Although in applications of bagging and boosting methods usually the number of loops used is at least 100, for the data sets which do not have the problem of small size relative to the size, exceeding the number of a few tens of loops is not necessary. It may be an essential observation in the context of the complexity of the ensemble procedure. Contrary to the opinions of some authors, linear discriminant functions may also be useful in resampling combining. This is concordant with the results of Skurichina (2001). Additionally, for quadratic discrimination, combining by bagging or boosting committee may also be beneficial. Constituent learners, which do not hold an increasing trend for resubstitution errors in subsequent boosting procedures loops, are not useful for the ensemble. This fact may mean that they little correct errors for patterns close to the classification boundaries. An example of such a classifier is the Parzen classifier. The kernel classifier is known as strong and flexible discriminant procedure. Thus, boosting is beneficial for nonparametric decision tree classifiers, but may not be useful for nonparametric Bayesian classifiers.

Some diversity measures based on oracle outputs, examined by Kuncheva et al. (2001, 2003), use only the information about concordance of pairs for constituent classifier decisions. Additional methods of diversity valuation may be the changes in constituent classifier error assessment, presented in the current work, connected with elaborated learning curves. We can observe the relationships between such differentiation and classification errors. Further development may be the assessment of the correlations between the new measure of diversity suggested by the current examination, variance or standard errors of consecutive constituent classification errors, and ensemble errors. It would also be interesting to study how this diversity correlates with the stability of constituent and combined classifiers.

Conclusions

The usefulness of bagging and boosting methods comes from diversity. Discriminant functions, which do not have an increasing trend for base resubstitution errors in subsequent boosting deterministic procedure loops, are not advantageous for the fusion, like the kernel Parzen discriminant function is. For the resampling success of classifier fusion, a compromise between accuracy and diversity is necessary. Diversity important in the success of boosting and bagging may be evaluated by concordance of component classifier with the learning vector.

Acknowledgments

The author is grateful to Prof. Wiktor Dróżdż from Department of Psychiatry at Nicolaus Copernicus University for the schizophrenic patients data set.

REFERENCES

Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from http://archive.ics.uci.edu/ml.

Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2005). Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1), 49–62.

Bi, Y. (2011). Analyzing the Relationship between Diversity and Evidential Fusion Accuracy. In C. Sansone, J. Kittler, F. Roli (Eds.), *Multiple Classifier Systems*, LNCS 6713, 249–258.

Breiman, L. (1996). Bagging predictions. *Machine Learning*, 24(2), 123–140.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, 26(3), 801–849.

Brown, G., & Kuncheva, L. I. (2010). Good and Bad Diversity in majority vote ensembles. In N. El Gayar, J. Kittler, & F. Roli (Eds.), *Multiple Classifiers Systems*, LNCS 5997, 124–133.

Dietterich, T. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40(2), 139–157.

Duin, R. P. W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D. M. J., & Verzakov, S. (2007). *PRTools4.1, A Matlab Toolbox for Pattern Recognition*. Delft University of Technology.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1),119–139.

Heskes, T. (1998). Bias/variance decomposition for likelihood-based estimators. *Neural Computations*, 10(6), 1425–1433.

Kotsiantis, S. B, Zaharakis, I. D., & Pintelas, P. E. (2006). Supervised machine learning. *A review of classification and combining techniques*, 26(3), 159–190.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation and active learning. *Advances in Neural Information Processing Systems*, 7, 231–238.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2), 181–207.

Kuncheva, L. I., Skurichina, M., & Duin, R. P. W. (2002). An Experimental Study on Diversity for Bagging and Boosting with Linear Classifiers. *Information Fusion*, 3(4), 245–258.

Kuncheva, L. I., & Whitaker, C. J. (2001). Ten Measures of Diversity in Classifier Ensembles: Limits for Two Classifiers. Proceeding IEEE Workshop on Intelligent Sensor Processing, 14 February 2001.

Kuncheva, L. I., Whitaker, C. J., Ship, C. A., & Duin, R. P. W. (2000). Is independence good for combining classifiers? In International Conference on Pattern Recognition (ICPR'00), 3–8 September 2000 (Volume 2, 168–171). Barcelona, Spain.

Kuncheva, L. I. (2003). That elusive diversity in classifier ensembles. In F. J. Perales, A. J. C. Campilho, N. P. de la Blanca, & A. Sanfeliu (Eds.), *Pattern Recognition and Image Analysis*, LNCS 2652, 1126–1138.

Melville, P., & Mooney, R. J. (2005). Creating diversity in ensembles using artificial data. Diversity in Multiple Classifier Systems. *Information Fusion*, 6(1), 99–111.

Quinlan, J. R. (1987). Simplifying Decision Trees. *Int. J. Man – Machine Studies*, 27(3), 221–234.

Rokach, L. (2010a). *Pattern Classification Using Ensemble Methods*. Series in Machine Perception and Artificial Intelligence, World Scientific Publishing.

Rokach, L. (2010b). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1–2), 1–39.

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics and Data Analysis*, 53(12), 4046–4072.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers – a survey. *IEEE Transactions on Systems, Man and Cybernetics*, Part C, 35(4), 476–487.

Shipp, C. A., & Kuncheva, L. I. (2002). Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 3(2), 135–148.

Skurichina, M. (2001). *Stabilizing weak classifiers* (Doctoral dissertation). Delft University of Technology.

Skurichina, M., & Duin, R. P. W. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis and Applications*, 5(2), 121–135.

Vu, T. T., Braga-Neto, U., & Dougherty, E. R. (2009). Bagging degrades the performance of linear discriminant classifiers. In IEEE International Workshop on Genomic Signal Processing and Statistics, GENSiPS, 17–21 May 2009. Minneapolis, MN, USA.