A C T A   U N I V E R S I T A T I S   L O D Z I E N S I S
FOLIA OECONOMICA 235, 2010

Małgorzata Misztal[*]

ON THE APPLICATION OF BUNDLING IN MEDICAL DIAGNOSIS

Abstract. *Bundling* was proposed by Hothorn (2003a) as modification of *bagging* (Breiman 1996). The main idea is to use the *out-of-bag* (OOB) observations of a bootstrap sample to build classifiers of arbitrary type (i.e. LDA or NN). The predictions of those classifiers are computed for the observations in the bootstrap sample and are used as predictors offered to a classification trees in addition to the original predictors (Hothorn 2003a).

In the study *bundling* was applied to improve prediction accuracy in some classification tasks in medical diagnosis.

Key words: classification trees, aggregated models, *bagging*, *bundling*, medical diagnosis.

I. INTRODUCTION

The main disadvantage of classification trees is the lack of stability. In other words, small changes in the learning set can lead to a completely different model. To improve the stability and prediction accuracy of classification trees we can use ensembles of classifiers or hybrid models.

Ensemble of classifiers can be defined as the set of single classifiers (usually homogenous) which predictions are aggregated using majority voting:

$$\hat{D}^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{v=1}^{V} I(\hat{D}_v(\mathbf{x}) = y) \right\}. \tag{1}$$

The most popular methods of aggregation are *bagging* (Breiman 1996), *boosting* (Freund & Schapire 1997) and *random forests* (Breiman 2001).

Hybrid (or composite) model is the combination of two different models – recursive partitioning method and one of some others algorithms (i.e.: linear discriminant functions – as in CRUISE (Kim & Loh 2003), logistic regression as in LOTUS (Chan & Loh 2004) and PLUS (Lim 2000) or distance-based algorithms as in k-NNTree (Buttrey & Karo 2002).

[*] Ph.D., Chair of Statistical Methods, University of Łódź.

In 2003 Hothorn proposed *bundling* as a new method for combining classifiers. *Bundling* is the modification of *bagging* and its main idea is to use the *out-of-bag* (OOB) observations of a bootstrap sample to build classifiers of arbitrary type (i.e. LDA or NN). The predictions of those classifiers are computed for the observations in the bootstrap sample and then are used as predictors offered to a classification trees in addition to the original predictors (Hothorn 2003).

In the paper *bundling* is applied in some medical diagnosis tasks. The goal of the study was to compare the prediction accuracy of *bundling* and some other classification algorithms (logistic regression, discriminant models, distance-based algorithms and aggregated models).

II. BUNDLING

Bundling was proposed by Hothorn (2003a) as modification of *bagging* (Breiman 1996). In *bagging* (*bootstrap aggregation*) V bootstrap samples, $U_1,\ldots, U_v$, each consisting of N cases, are drawn at random but with replacement from the learning set. Each bootstrap sample of size N covers about 63,2% of the observations of the learning sample. The observations which are not in the bootstrap sample are called the OOB (*out-of-bag*) sample. According to Breiman (1996) it can be used for estimating the misclassification error or for improving class probability estimates.

Hothorn & Lausen (2003) proposed to use *out-of-bag* sample for estimation of the coefficients of linear discriminant functions. The corresponding linear discriminant variables are then computed for the bootstrap sample and used as additional predictors for the classification tree. Procedure is repeated V times and the outcome prediction is made using majority voting. The combination of LDA (linear discriminant analysis) and classification tree is called *double – bagging*.

Hothorn (2003a) extended the described procedure to the combination of classifiers of different types. For each bootstrap sample a number of classifiers (logistic regression, LDA, k-NN, etc.) can be trained using the OOB sample. The predictions of those classifiers – predicted classes, estimated conditional class probabilities, linear discriminant values – are in the next step computed for the observations in the bootstrap sample and added to the set of original predictors for *bagging* of classification trees (Hothorn 2003a). This extension is called *bundling*.

Both – *bagging* and *bundling* – are implemented in ipred package in R system. The performance of *bundling* was tested in Hothorn (2003b) using several data sets from the UCI machine learning repository (Blake & Merz 1998). The results are shown in Table 1. The smallest misclassification errors for each data set are bolded.

Table 1. Misclassification errors for UCI data sets

Data set	Classifier							
	Single tree	sLDA	5-NN	10-NN	Logistic regression	Bagging	Bundling	Random Forests
Twonorm	24.7	2.5	3.9	3.4	5.1	6.9	2.8	4.8
Threenorm	33.1	17.4	18.4	16.9	17.9	19.6	16.6	17.8
Ringnorm	23.0	38.9	47.4	49.3	38.1	10.0	6.5	7.3
Breast Cancer	5.5	3.4	6.7	8.3	7.3	4.0	2.9	3.1
Ionosphere	12.3	13.9	15.7	16.6	12.5	7.8	6.0	6.4
Diabetes	25.8	26.9	28.6	26.5	22.4	24.3	24.2	23.7
Glass	30.2	42.4	32.7	38.1	35.2	23.0	24.2	21.3
Satellite	16.2	19.3	8.7	9.6	19.2	8.4	7.2	7.6
Shuttle	0.1	8.2	0.4	0.6	2.9	0.1	0.1	0.1
DNA	9.1	8.1	18.6	16.4	10.4	4.6	2.7	5.5

Source: Hothorn (2003b).

As we can see, the smallest error rates are achieved most of all for *bundling* of sLDA, k-NN and logistic regression.

In the next section we try to use *bundling* to improve prediction accuracy in some classification problems from medical research.

III. APPLICATION OF BUNDLING

In the paper *bundling* was applied to classify patients from the following data sets:

1. AF – 300 patients undergoing aortic valve replacement (AVR) due to aortic valve defect. The dependent variable y was the binary variable with two possible values: 0 for patients with no atrial fibrillation complications (non-AF) and 1 for patients with atrial fibrillation complications (AF).

2. ICU – 337 patients undergoing CABG (Coronary Artery Bypass Grafting). The dependent variable y was the binary variable describing the length of stay in the Intensive Care Unit: 0 for patients with ICU-stay ≤ 2 days and 1 for patients with ICU-stay longer than 2 days.

3. OP_RISK – 4634 patients treated surgically due to Coronary Artery Disease. The dependent variable y was the binary variable connected with the outcome after the operation: 0 for patients with good outcome and 1 for patients with cardiac complications and / or deaths.

All data sets were randomly divided into the learning sample and the test sample. Short description of data sets used in the study is presented in Table 2.

Table 2. Characteristics of data sets.

Data set	Number of predictors	Number of cases in the learning set	Number of cases in the test set
AF	9	150	150
ICU	17	200	137
OP_RISK	19	2634	2000

Source: own elaboration.

All calculations were performed using the R system (packages rpart, ipred and e1071). The following classification algorithms were used:

- Single classification tree (rpart);
- Logistic regression (LR);
- Linear discriminant analysis (LDA) and stabilized linear discriminant analysis (sLDA);
- k-*Nearest Neighbors* algorithm (k-NN);
- Support Vector Machines (SVM);
- *Bagging*;
- *Boosting* (Adaboost);
- *Random Forests*;
- *Bundling*.

All the classifiers were trained using the learning sample and their performance was assessed using the test sample. The results (misclassification error rates for the test sample) are presented in Table 3. The smallest misclassification errors are bolded.

Table 3. Misclassification error rates for the test sample

Classifier	% of incorrect classifications		
	AF	ICU	OP_RISK
1	2	3	4
rpart	46.00	42.34	18.05
LR	39.33	35.04	17.95
LDA	39.33	35.77	18.05
sLDA	36.67	35.77	28.40
1-NN	29.33	45.26	37.00
5-NN	48.00	40.15	33.00
7-NN	48.00	40.15	33.00
10-NN	44.67	40.88	30.25
15-NN	x	x	29.35
SVM	43.33	35.04	18.00

Table 3 (cont.)

1	2	3	4
Bagging	26.67	43.07	18.35
Bundling of LR	27.33	38.69	18.20
Bundling of LDA	26.00	40.88	18.15
Bundling of sLDA	26.67	42.34	18.05
Bundling of 1-NN	27.33	41.61	18.40
Bundling of k-NN	28.67	37.96	17.90
Bundling of SVM	30.00	42.34	17.90
Bundling of more classifiers	27.33	37.96	17.95
Adaboost	33.33	40.15	18.00
Random Forests	32.00	37.23	18.15

Source: author's calculations.

As we can see, the results for *bundling* are always better than for single classification tree and for *bagging*. For 2 data sets: AF and OP_RISK we obtained the smallest error rates using *bundling* with one additional classifier. The best predictions for ICU-stay were obtained for logistic regression and SVM method.

IV. CONCLUSIONS

There are some problems connected with *bundling*:

1) there is no instruction about the number and the best type of additional classifier; all the choices should be made experimentally;

2) the algorithm is not fast (in the sense of the time of calculations) when large data sets are analysed;

3) the results from *bundling* are some kind of a "black box", there is no possibility to get decision rules and to gain insight and understanding into the predictive structure of the data.

Bundling can outperform many popular classifiers but we should remember, according to *No-Free-Lunch* (NFL) Theorem (Wolpert & Macready 1997) that there is no algorithm, better than all the others, for all classification tasks and all classifiers perform – on average – equally well as they are applied to different problems.

REFERENCES

Blake C., Merz C. J. (1998), *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, http://ics.uci.edu/~mlearn/MLRepository.html.

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

Breiman L. (2001), Random forests, *Machine Learning* 45, 5–32.

Buttrey S. E., Karo C. (2002), Using k-nearest-neighbor classification in the leaves of a tree, *Computational Statistics & Data Analysis* 40 (2002), 27–37.

Chan K.-Y., Loh W.-Y. (2004), LOTUS: An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees, *Journal of Computational and Graphical Statistics*, Vol. 13, Issue 4, 826–852.

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, 119–139.

Hothorn T., Lausen B. (2003), *Double-bagging: Combining classifiers by bootstrap aggregation*, Pattern Recognition, 36, 1303–1309.

Hothorn T. (2003a), *Bundling classifiers with an application to glaucoma diagnosis*, Dissertation, Department of Statistics, University of Dortmund, Germany, 2003. http://eldorado.uni-dortmund.de:8080/bitstream/2003/2790/1/hothornunt.pdf

Hothorn T. (2003b), *Bundling Predictors in R*, (in:) Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), K. Hornik, F. Leisch, A. Zeileis (eds.), Vienna, Austria.

Kim H., Loh W.-Y. (2003), Classification Trees with Bivariate Linear Discriminant Node Models, *Journal of Computational and Graphical Statistics*, 12, 512–530.

Lim T.-S. (2000), *Polytomous Logistic Regression Trees*, PhD Thesis, Department of Statistics, University of Wisconsin, Madison.

Wolpert D. H., Macready W. G. (1997), No Free Lunch Theorems for Optimization, *IEEE Transactions on Evolutionary Computation*, 1 (1), 62–68.

Małgorzata Misztal

**O ZASTOSOWANIU METODY WIĄZANIA MODELI (*BUNDLING*)
W DIAGNOSTYCE MEDYCZNEJ**

Metoda wiązania modeli (*bundling*) została zaproponowana przez Hothorna (2003a) jako modyfikacja metody *bagging* (Breiman 1996). Polega ona na wykorzystaniu dodatkowych modeli, innych klas niż drzewa klasyfikacyjne, budowanych na podstawie zbioru OOB (*out-of-bag*), zawierającego obserwacje spoza aktualnej próby bootstrapowej. Na podstawie tych modeli dokonuje się predykcji dla obserwacji w próbie bootstrapowej a następnie wyniki predykcji traktuje się jako dodatkowe zmienne objaśniające przy budowie drzewa klasyfikacyjnego (Hothorn 2003a).

W referacie przedstawiono wyniki wykorzystania metody wiązania modeli do poprawy dokładności predykcji w wybranych problemach diagnostyki medycznej.