Avoiding overfit by restricted model search in tree-based EEG classification

Klaschka, Jan
Institute of Computer Science, Academy of Sciences, Department of Neural Networks and Nonlinear Modelling
Pod Vodárenskou věží 2
CZ-18207 Prague 8, Czech Republic
E-mail: klaschka@cs.cas.cz

1. Introduction

This paper presents the results of a computational experiment based on real data. It follows up previous works Klaschka (2007, 2008), accomplished in the framework of a broader project aimed at prevention of drivers' microsleeps and the traffic accidents resulting from them. In the papers cited, electroencephalography (EEG) frequency spectra of a group of experimental subjects (persons) were analyzed in order to find accurate enough classifiers discriminating somnolence (sleepiness) from other brain states. The classifiers considered were complex models whose building blocks were classification forests.

The starting point of the present study was a strange and undesirable behavior of the best models from Klaschka (2008), observed when the size of the forests, which was constant in the study cited, was varied. Some of the models deteriorated with the growing forest size. After applying restrictions to some parameters of the model (i.e. when the number of candidate models was reduced), the trend of model deterioration with the growing forest size vanished or, at least, was considerably attenuated.

In Section 2, the data and classification problems are characterized. Section 3 describes the kind of classification models studied and outlines the principles of model search. Section 4 summarizes the results of previous works. In Section 5, the key section of the paper, the problem of model deterioration with increasing forest size is demonstrated, together with the results of more restricted model search. Finally, some brief concluding remarks may be found in Section 7.

2. Experimental data and classification problems

The data dealt with in the present study come from an experiment performed at the Joint Laboratory of System Reliability, Faculty of Transportation Sciences, Czech Technical University, Prague. (For a more detailed description of the experiment, see Faber et al. (2005).) The data set used in the present analyses consists of the frequency spectra of 677 EEG segments from 18 experimental subjects (individuals, persons). The data set $\mathcal{L}$ of 677 cases is the union of sets $\mathcal{L}_1, \ldots, \mathcal{L}_{18}$ (sizes 25–52) of the data contributed by the individual experimental subjects.

Each case is classified *a priori* as corresponding to one of the following brain states:

- *somnolence* (sleepiness) – 293 cases,
- relaxed *wakeful state* – 188 cases,
- *mentation* – mental activity, namely solving a part of the Raven test – 196 cases.

Each case possesses a vector of 62 numerical predictors – spectral powers (extracted from the raw EEG signals using the Burg filter of order 20) for 31 frequencies 0–30 Hz by 1 Hz, from two EEG electrodes (namely T3 and O1 channels – see Jasper (1958)).

There are four classification problems of interest:

- The 3-class problem (Somnolence vs. Wakeful vs. Mentation),
- three partial 2-class problems, namely Wakeful vs. Somnolence, Mentation vs. Somnolence, and Wakeful vs. Mentation.

3. Models and search strategies

The principal classification tool within this study is the Random Forests (RF) method by Breiman (2001). Besides positive experience by other authors, there are additional good reasons for such a choice. First, the method proved competitive in classification study by Štefka and Holeňa (2007) where many different classification methods were applied to the same data as those used in this paper. Second, the classification forest are, due to their ensemble nature, well suited for model combining.

Let us denote the set of all subjects S. Let, further, for $J \subseteq S$, $\mathcal{L}_J = \cup_{j \in J} \mathcal{L}_j$. Finally, let M_J denote a model obtained applying the RF method to $\mathcal{L}_J$.

A straightforward application of the RF method to data set $\mathcal{L}$ yields so called *global model* (M_S in the notation introduced above). When the RF method is applied to the data sets $\mathcal{L}_i$, we obtain *individual models* M_i for subjects $i \in S$. It appeared in an early stage of the work with the given data that the individual models outperformed clearly the global model. (It is not surprising since the EEG signals are highly individual, so that a different classifier is needed for each subject rather than a common one for all the subjects.)

Several previous works were focussed at so called *mixed models*. A mixed model for subject i results from combining an individual model M_i with a model M_J trained on $\mathcal{L}_J$ where J is a set of properly selected subjects (naturally, $i \notin J$).

The models are combined through linear combinations of their votes. The votes $\mathbf{v}(\mathbf{x}, k)$ of forest consisting of trees $T_1, T_2, \ldots, T_m$ are given as the proportions of those trees in the forest that classify predictor vector $\mathbf{x}$ into classes $k = 1, \ldots, K$. The votes of a mixed model are defined as a linear combination

$$\mathbf{v}(\cdot,\cdot) = (1-\alpha)\mathbf{v}_i(\cdot,\cdot) + \alpha \mathbf{v}_J(\cdot,\cdot) \tag{1}$$

of the votes of forests M_i and M_J ($0 \leq \alpha \leq 1$) for $J \neq \emptyset$.[1] (For $J = \emptyset$, $\mathbf{v}(\cdot,\cdot) = \mathbf{v}_i(\cdot,\cdot)$.) The class predicted by the mixed model is then given, as in the case of "ordinary" classification forest, by majority voting, i.e. as $\text{argmax}_k\, \mathbf{v}(\cdot, k)$.

Note that a mixed model is very similar to a classification forest: It is composed of a number of trees whose votes determine the predicted class. The only difference in comparison with classification forests is the fact that different subsets of the trees are trained on (bootstrap samples from) different data sets and their votes have different weights.

A mixed model for subject i is given by a specific choice of set J and constant α. Various strategies of the search for the "best" set J were developed and tested experimentally. While in Klaschka (2007) only the maximal set $J = S \setminus \{i\}$ is considered, work Klaschka (2008) studies more sophisticated strategies of choice of J from within a (possibly big) set of candidates, e.g. stepwise forward search, either unrestricted, or with some restrictions. As regards the choice of α, an exhaustive search within a discrete grid (most often from 0 to 1 by 0.1) is performed.

During the search for the "best" model among candidate models (combinations of J and α), models are mutually compared using misclassification error estimates calculated in the following way.

[1] In the previous works, the mixed models based on weights (1) were referred to as the Type 2 models. There were Type 1 models, too, with different votes. The original Type 1 models, however, yielded worse results, so that they are not dealt with here any more.

Votes (1) where $\mathbf{v}_i$ is replaced with out-of-bag votes[2] $\mathbf{v}_i^{OOB}$ are calculated for cases from $\mathcal{L}_i$. Predicted classes for cases from $\mathcal{L}_i$ are then determined by majority voting. Misclassification error estimate is the proportion of those cases in $\mathcal{L}_i$ whose predicted class does not match the true class. Of two models, the one with lower error estimate is considered better.

The generalization error of the model chosen by a specific model search strategy as the "best" one is estimated as the proportion of the misclassified cases by leave-one-out cross-validation (jackknife).

4. Previous results

In the study Klaschka (2008), several model search strategies were examined. Each strategy was applied 100 times (fewer times for the most computationally expensive strategies) with different random seeds. For the sake of simplicity, evaluation of strategy performance was based on the "overall error" of the models yielded by the strategy, i.e. on the proportion of misclassified cases in $\mathcal{L}$ (or, equivalently, on the weighted average of 18 errors of the mixed models for all the subjects $i \in S$, the weights being proportional to the sizes of $\mathcal{L}_i$, $i \in S$).

Some of the strategies failed and some succeeded in reducing the misclassification error in comparison with the individual models. For instance, the simplest strategy with a single candidate set $S \setminus \{i\}$ for subject i was among those that have failed (for two of the four classification tasks, the means of 100 overall errors were greater than the analogous results for the individual models).

One of those only partially successful strategies was the unrestricted stepwise forward search: The best candidate set is found as the optimal element of a sequence o nested sets consisting of the best singleton $J_1 = \{j\}$ $(j \neq i)$, the best 2-element superset J_2 of J_1, the best 3-element superset J_3 of J_2, and so forth. The mean overall errors were lower than those of the individual models in all 4 classification tasks, but the improvements were small.

In comparison with the unrestricted stepwise forward search, the "winning" strategy of the study performs a much less extensive model search (much fewer candidate sets are examined). It will be referred to here, for the sake of briefness, by its code in the original work, as the *F-strategy*, and the models resulting from it will be called *F-models*. In order to describe the strategy, the following notation will be introduced. For $i, j \in S$, $i \neq j$, e_{ij} denotes the misclassification error of model M_i on $\mathcal{L}_j$, and e_{ii} is the out-of-bag estimate (Breiman (1996)) of the error of M_i. Further, r_{ij} will denote the rank of e_{ij} in $\{e_{ij}; i \in S\}$ sorted in the ascending order. Each subject $i \in S$ is assigned score $s_i = \sum_{j \in S} \lambda^{r_{ij}}$ where λ (the *base*) is a proper constant $(0 < \lambda < 1)$. In the computational experiments, λ was set to $(\sqrt{5} - 1)/2 \approx 0.62$ (the well known golden section).

When building an F-model for subject i, there are nested candidate sets $J_1 \subset J_2 \subset J_3 \subset \ldots$, where J_k is the set of subjects with k smallest s_j values in $\{s_j; j \in (S \setminus \{i\})\}$. Concerning the candidate sets, note that their number is small, and they are "almost independent" of i. (The only differences between the candidate sets for different subjects i follow from $i \notin J$.)

The F-models reduced, in comparison with the individual models, the mean overall errors by at least the factor of approx. 0.9 consistently for all 4 classification tasks.

For the detailed results of the computational experiment, see the source paper Klaschka (2008).

5. New problem and experimental results

In the study Klaschka (2008), the forest size (number of trees per forest) was set uniformly to 500, the default of the R extension package randomForest (R Development Core Team (2006), Liaw and Wiener (2002)). Later on, the part of the experiment related to the individual models and F-models

[2] The trees in the forest are trained on different bootstrap samples from $\mathcal{L}_i$. The out-of-bag votes related to a case are based on only those trees whose training set did not contain the case.

was extended and various forest sizes – namely 100, 200, 500, 1 000, 2 000, 5 000, 10 000 and 20 000 – tried out. All the calculations were repeated 50 times with different random seeds. (The reason why the number of repetitions was smaller than in the earlier study was an extremely high computational cost of the jackknife error estimates of the F-models for the biggest forest sizes: With 20 000 trees per forest, a single overall error for the 3-class problem required approx. 36 hours on a computational cluster node.)

The original aim of forest size varying was a fine tuning of the method (the F-strategy): The improvement might have proved, in comparison with the individual models, either bigger for larger forests, or, conversely, essentially the same for a smaller forest size.

Computational experimenting, however, lead to surprising and undesirable results: Fig. 1 shows the mean overall errors ($\pm 2SE$) of the individual models (dotted lines) and of the F-models (dashed lines) for 8 forest sizes ranging from 100 to 20 000. (The solid lines will be explained below.) We can see that the mean F-model errors, starting from the forest size of 1 000, increase for the 3-class problem (Fig. 1a), and an even more dramatic deterioration takes place concerning the Mentation vs. Wakefulness problem (Fig. 1b). That has changed the direction of the research and initiated the present study.

Deterioration of the generalization properties with growing model complexity results often from an overfit due to a too extensive model search. The tendency to overfit, however, is not, as a rule, a property of the components of the mixed models, i.e. of the forests (Breiman (2001)). (Note that the individual models, being "ordinary" forests, adhere to the rule, since their mean error curves in Fig. 1 are free of any deterioration.)

When building a mixed model, multiple pairs (J, α) of a candidate set and a constant are tried out before a final one is found, and that is where overfit might take place. There is, seemingly, no reason for the tendency to overfit to strengthen with the growing forest size, since the number of the candidate models (i.e. of (J, α) pairs examined) does not change.

A possible explanation of the observed tendency might, however, be the following hypothesis: *Combining of bigger forests is more prone to overfit than that of the smaller ones.* If so, a stronger restriction of the model search (i.e. decreasing the number of the candidate models) might be a remedy.

The above considerations led to an experimental examination of a more restricted version of the F-strategy: While the J sets of the mixed models corresponding to the dashed lines in Fig. 1 could contain up to 9 subjects, the solid lines were obtained when the size limit for J was set to 3 and, moreover, the grid of allowed α values was reduced from 0–1 by 0.5 to 0–0.5 by 0.1.

The search restriction was fully successful in the 3-class problem where the model deterioration vanished, and the the resulting models appear uniformly better than the original ones (Fig. 1a). As regards the Mentation vs. Wakefulness problem (Fig. 1b), the trend of model deterioration was attenuated markedly, but is still present. In the two remaining classification problems, where the original F-models did not call much for corrections, the effect of search restriction was negative (Somnolence vs. Wakefulness problem, Fig. 1c), or negligible (Somnolence vs. Mentation problem, Fig. 1d).

Fig. 2 shows additional experimental results for the Mentation vs. Wakefulness problem. Further reduction of the size limit for set J from 3 to 2 and 1 (while keeping the limitations concerning α) has led to a further attenuation of the increase of mean errors with growing forest size but, at the same time, to less accurate classifiers. In the earlier study, one model search strategy, namely the strategy based on a single "well-guessed" candidate set J (for details, see Klaschka (2008)), did markedly better than the F-strategy for the Mentation vs. Wakefulness problem. (Nothing like that happened for the other 3 problems.) The same remains true even when the forest size is varied – see the lowest line in Fig. 2. The corresponding curve is uniformly satisfactorily low, and does not increase with the growing forest size.

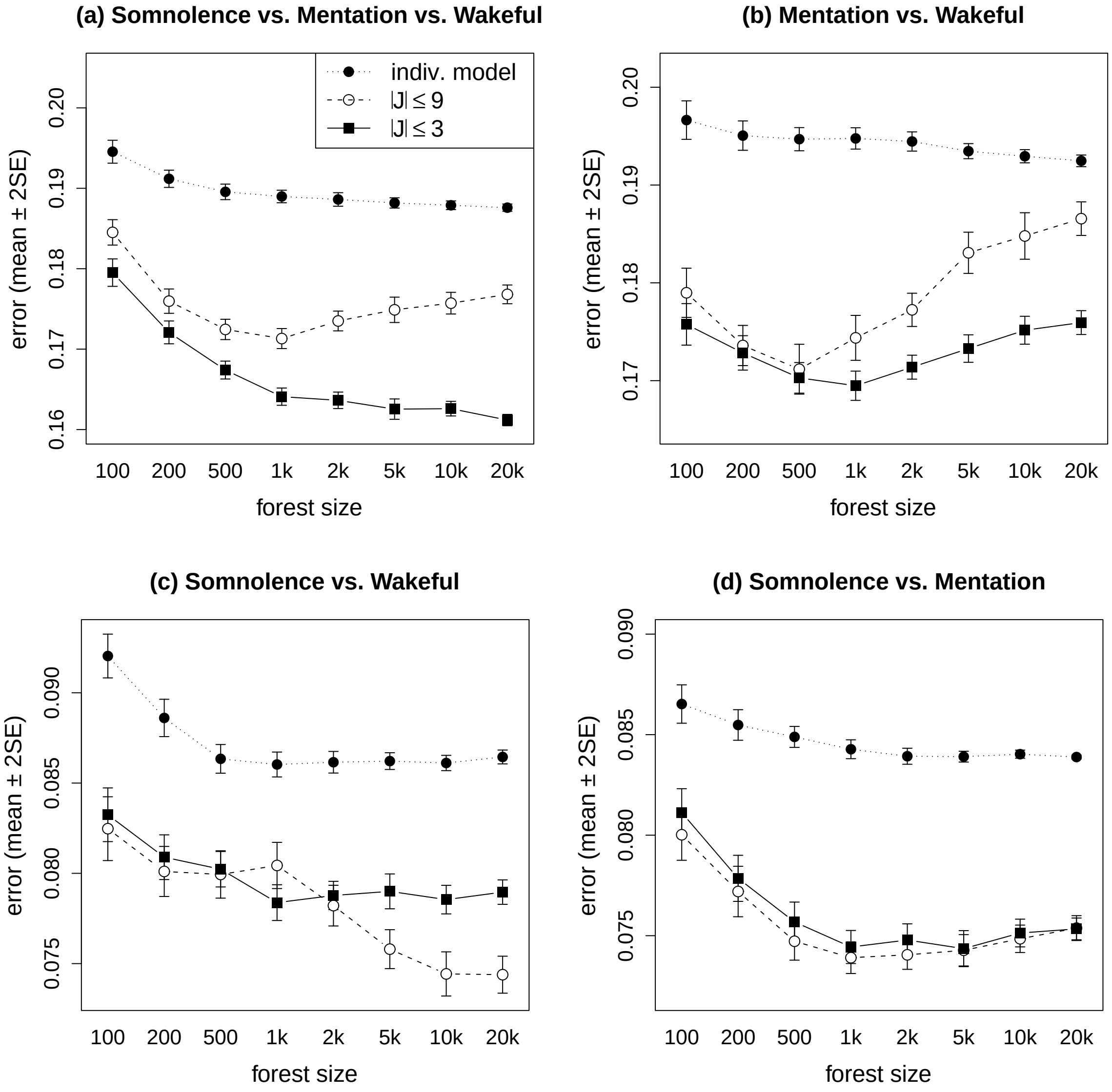

Figure 1. Overall errors (means $\pm 2SE$ ***of 50 repetitions) of the individual models, original F-models (legend*** $|J| \leq 9$***) and F-models resulting from restricted search (***$|J| \leq 3$***) for 8 different forest sizes (1k = 1 000 etc.).***

6. Conclusions

The hypothesis that combining forests into mixed models is the more prone to overfit, the bigger are the forests, remains still just one of possible explanations of the phenomena studied in Section 5. The experimental results, nevertheless, seem to support it. Moreover, even if it is true, it is still unclear, whether its validity is a specific of the given domain, or more general.

As shown in Fig. 2, the search restriction, though successful in some cases, should be applied with care, and by no means is a universal way to the optimal classification results. It cannot compensate for good candidate models "overlooked" by a given type of search strategy (see the case of J fixed), and it may rather, when overdone, do harm by eliminating most of promising candidates (case $|J| \leq 1$).

As regards fine tuning of the F-strategy, search restriction is advisable for 3 of the 4 tasks, but not for the Somnolence vs. Wakefulness problem. Forest size up to 1 000 is enough for the Mentation vs. Wakefulness and Somnolence vs. Mentation problems. For the other two problems, it remains open whether the error decrease with the forest size growth is worth the increased computational cost.

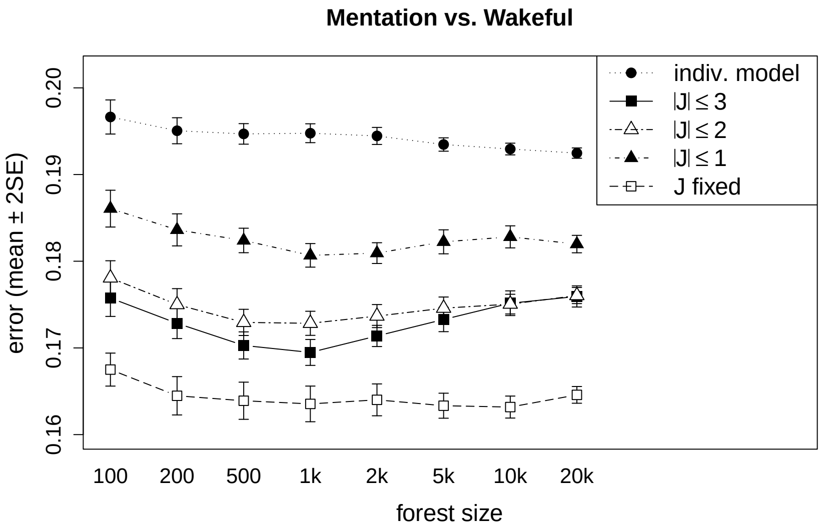

Figure 2. Further results for the Mentation vs. Wakefulness problem. Overall errors (means ±2SE of 50 repetitions) of the individual models, F-models resulting from restricted search with size limit for set J reduced to 3, 2, and 1, and a mixed model with a fixed set J, all for 8 different forest sizes (1k = 1 000 etc.).

ACKNOWLEDGEMENT

This work was supported by grant ME 949 from the Ministry of Education, Youth and Sports of the Czech Republic, and by the Institutional Research Plan AV0Z10300504.

REFERENCES

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

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

Faber J., Novák M., Tichý T., Svoboda P. and Tatarinov V. (2005). *Driver psychic state analysis based on EEG signals.* Novák M. (ed.), Neurodynamic and Neuroinformatics Studies (Second Book on Micro-Sleeps), 33–48, Czech Tech. Univ. in Prague, Faculty of Transport. Sci.

Jasper H. (1958). *Report of the committee on methods of clinical examination in electroencephalography.* Electroencephalography and Clinical Neurophysiology **10**, 371–375.

Klaschka J. (2007). *Combining Individual and Global Tree-Based Models in EEG Classification.* Bulletin of the ISI 56th Session. Proceedings, 1–4, International Statistical Institute, Lisboa.

Klaschka J. (2008). *Classification of Heterogeneous EEG Data by Combining Random Forests.* Mizuta M. and Nakano J. (eds.), Proceedings of IASC 2008, 888-896, Japanese Society of Computational Statistics, Tokyo.

Liaw A. and Wiener M. (2002). *Classification and Regression by randomForest.* R News **2**, 18–22.

R Development Core Team (2006). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Štefka D. and Holeňa M. (2007). *Using fuzzy k-NN ensembles in EEG data classification.* Novák M. (ed.), Neuroinformatic Databases and Mining of Knowledge of them (Third Book on Micro-Sleeps), 200-211, Czech Tech. Univ. in Prague, Faculty of Transport. Sci.