proceedings

Proceedings

Semi-Automated Sleep EEG Scoring with Active Learning and HMM-Based Deletion of Ambiguous Instances †

Martin Macaš *, Nela Grimová, Václav Gerla and Lenka Lhotská

Czech Institute of Informatics, Robotics and Cybernetics; Czech Technical University in Prague, 16000 Prague, Czech Republic; nela.grimova@cvut.cz (N.G.); vaclav.gerla@cvut.cz (V.G.); lenka.lhotska@cvut.cz (L.L.)

* Correspondence: martin.macas@cvut.cz; Tel.: +420-737-853-299

† Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.

Published: 20 November 2019

Abstract: Sleep scoring is an important tool for physicians. Assigning of segments of long biomedical signal into sleep stages is, however, a very time consuming, tedious and expensive task which is performed by an expert. Automatic sleep scoring is not well accepted in clinical practice because of low interactivity and unacceptable error, which is often caused by inter-patient variability. This is solved by proposing a semi-automatic approach, where parts of the signal are selected for manual labeling by active learning and the resulting classifier is used for automatic labeling of the remaining signal. The active learning is disturbed by noisy ambiguous data instances caused by continuous character of the sleep stage transitions and a removal of such transitional instances from the training set prior to active learning can improve the efficiency of the method. This paper proposes to use the hidden Markov model for the detection of the transitional instances. It shows experimentally on 35 sleep EEG recordings that such a method significantly improves the semi-automatic method. A complete methodology for semi-automatic sleep scoring is proposed and evaluated, which can be better accepted as a decision support tool for sleep scoring experts.

Keywords: sleep scoring; EEG; pattern classification; active learning; hidden Markov models; expert-in-the-loop; ambiguity

1. Introduction

Generally, sleep scoring can be defined as an assignment of segments of some biomedical signal into sleep stages. Most typically, the signal is electroencephalogram (EEG) or polysomnogram (PSG) and the sleep scoring is performed manually by an expert. A sleep scoring process produces a sequence of sleep stage labels which can be very useful in evaluation of various disorders like sleep disorders, neurological and cardiological diseases or diabetes. An example of such a sequence, called hypnogram, is depicted in the upper subfigure of Figure 1. The sleep scoring expert must have sufficient skills and practice and thus the sleep scoring task is very time-consuming, tedious, and expensive. Moreover, the expert's fatigue causes labeling errors, subjectivity and higher inter-expert variability. An automatic computer-based sleep scoring system could help to avoid and reduce many of mentioned drawbacks. It can significantly reduce the time requirements, personal costs and errors. After decades of research devoted to automatic sleep scoring, automatic systems are, however, still not widely accepted by physicians and electrophysiologists. This is caused by insufficient accuracy, which is partly caused by high inter-personal variability, where recordings come from different measurement devices and training data are labeled by different experts. Another reason for limited acceptance of fully automated systems is lack of interpretability and interactivity, where the physicians cannot make serious decisions on the basis of non-interactive computer systems not well supported by an interpretable evidence.

This paper combines a manual and fully automated system, which makes a trade-off between savings of manual labeling effort and accuracy of the scoring system. By training and operating the automatic classifier on the data from the same patient, the approach removes the problem of inter-personal variability at the expense of an increased manual labeling effort. While there is no manual labeling effort in conventional automatic systems, in the proposed semi-automatic system, part of the signal is labeled manually and used as a training data set for a pattern classifier, which is finally used for automatic labeling of the rest of the signal.

The approach is depicted in Figure 2. It starts with an initial data set consisting of segments of EEG signal that are not scored (are unlabeled). The first step is an initial selection of segments that should be manually labeled by the expert. This initialization is out of scope of this paper and one data segment from each class is randomly sampled at the beginning. Next, an iterative loop runs, where unlabeled segments are selected one by one and manually labeled until a certain sufficiency criterion is met and sufficient training data are formed consisting of a subset of signal segments and their corresponding labels. The sufficiency criterion is irrelevant here because we compare the methods during the first 150 iterations. Nevertheless, it can be an achievement of some validation error threshold. Finally, a classifier is trained and used for classification of remaining segments. Since only a subset of the signal segments is labeled, this approach can significantly reduce the labeling effort and labeling costs, and improve the final scoring result by eliminating mistakes caused by expert's fatigue.

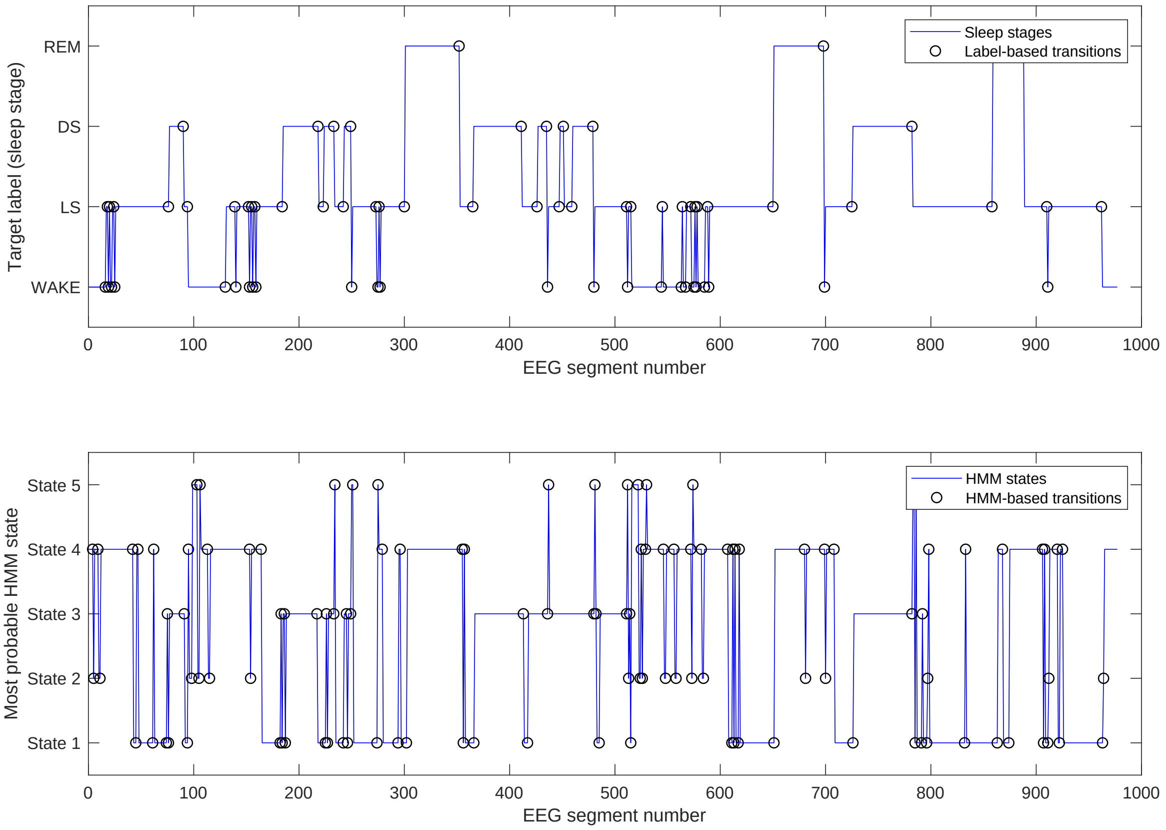


Figure 1. The difference between sleep stage sequence called hypnogram (top) and most probable state sequence of trained hidden Markov model (bottom). The label-based transitions must be be computed from the labels (sleep stages), which are not available in the active learning scenario. On the other hand, HMM-based transitions can be computed from the most probable state sequence of the HMM trained on feature time series.

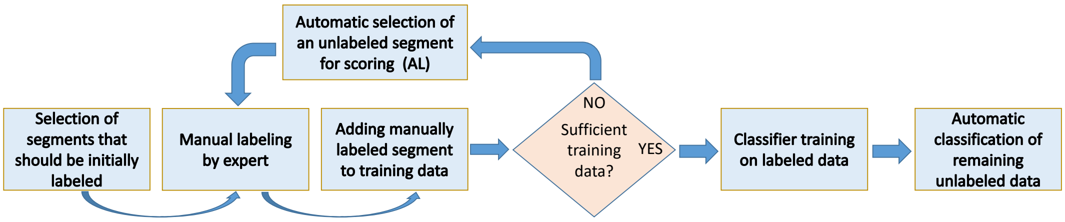

Automatic selection of an unlabeled segment for scoring (AL)
Selection of segments that should be initially labeled
Manual labeling by expert
Adding manually labeled segment to training data
NO
Sufficient training data?
YES
Classifier training on labeled data
Automatic classification of remaining unlabeled data

Figure 2. The flow diagram of semi-automated scoring of EEG data.

The main problem of such a semi-automatic method is how to select the signal segments that should be labeled manually. This problem can be solved by active learning approaches well known in pattern classification area [1]. An active learning method typically selects data instances that are worth to be labeled. For example, it reduced labeling time for a clinical concept extraction in [2]. In [3], active learning was used for an efficient labeling of artifacts in EEG signal. The following advantages of involving active learning can be expected: First, a proper selection of training instances can reduce the generalization error, only a subset of available training data need to be labeled which reduces the labeling effort. Second, if the annotator is human, the reduced labeling effort leads to less fatigue-based errors. Third, a smaller training set can significantly reduces the training time and memory requirements.

Active learning has been applied to EEG data in the area of Brain Computer Interface (BCI) [4], where it typically improved classification performance and reduced training set that lead to reduced time and memory requirements. Typically, the active learning is not used in a semi-automatic manner, where the inter-personal variability issue can be overcame or labeling effort reduced. Those two problems are focused in [5], where transfer learning together with semi-supervised learning was considered to calibrate a pre-trained classifier to a specific subject and thus avoid error related to inter-personal variability. The active learning was also used for regression task in [6], where driver drowsiness was predicted from EEG, but its main purpose was an improvement of training set and not reduction of data acquisition costs.

In [7], we proposed an active learning based semi-automatic sleep scoring approach. It was found that the training data instances whose labels differ from the labels of their adjacent instances tend to be noisy since they are hard to be assigned clearly to one of sleep stages. Such ambiguous instances can disrupt the training procedure and especially the active learning approaches, which have a tendency to prefer them by selecting data close to the decision boundaries. Although it was demonstrated that removal of such instances from the training set improves the performance of the semi-automatic system, a crucial problem related to detection of such instances was pointed out. Since the active learning starts with partly unlabeled data, the complete sequence of target labels is unknown and the transitional instances cannot be found by observing difference of the labels of adjacent segments. The paper thus theoretically demonstrated a potential improvement, but did not provide solution for its practical realization. This point is focused in this paper, which proposes to detect the transitions via hidden Markov models (HMMs). The paper follows the experimental scenario from [7], but describes the removal of HMM based transitions instead of label-based transitions. The difference is shown in Figure 1. The instances corresponding to the label transitions are marked as black circles in the upper subfigure. They are simply detected from the sequence of labels provided by expert in fully automatic approaches. This sequence is, however, not available in our approach and thus we propose to train HMM and label the EEG time series by its most probable states. Such an approach detects data instances that correspond to natural transitions in EEG signals and does not need the labeled training data.

2. Active Learning and Ambiguous Instances

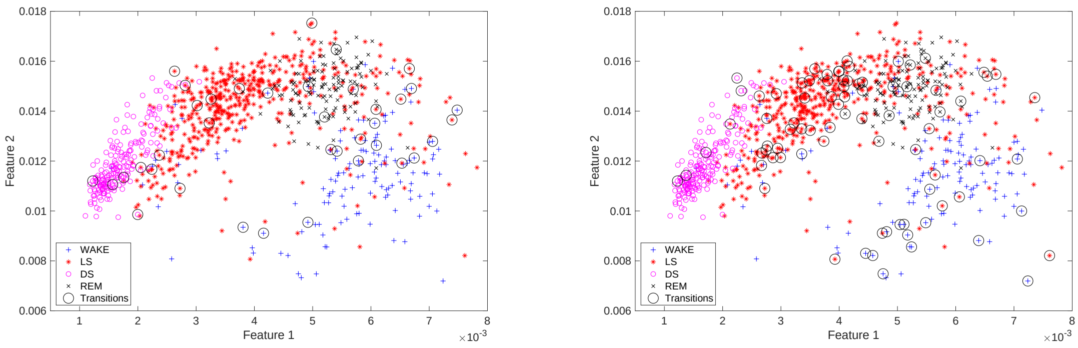
As in [7], this paper focuses on a pool-based sampling scenario of active learning [1], where a pool of unlabeled instances (feature vectors) is given and instances for labeling are sequentially selected. An uncertainty sampling is the most popular and the simplest strategy [8]. It was proposed by Lewis et al. [9]. It is particularly suitable for probabilistic classification models, but it might also be applied to non-probabilistic classifiers [8]. Uncertainty sampling approaches typically use posterior probability $\mathrm{p}(\omega|\mathbf{x})$ estimating probability that object described by feature vector $\mathbf{x}$ comes from the class with label ω. At each iteration of the sampling process, an uncertainty measure is computed from the posterior probabilities. Next, an instance or instances with minimum uncertainty, whose label the classifier is least certain of, is queried. In margin uncertainty sampling proposed by Scheffer et al. [10] the algorithm selects an instance with the smallest difference of the highest and the second highest posterior probability: $\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathrm{p}(\omega_1|\mathbf{x}) - \mathrm{p}(\omega_2|\mathbf{x})$, where ω_1 and ω_2 is the most probable and the second most probable label of the feature vector $\mathbf{x}$, respectively. Many other variants of uncertainty sampling exist [8] based on entropy or posterior probability. Moreover, many other types of query selection strategies exist like query-by-committee strategy, whose higher time complexity can be an important problem for the user of the semi-automated system.

In many real world pattern classification problems, the labels are often assigned to objects with a certain level of uncertainty. This can happen if the oracle makes mistakes (noisy oracles) or the classes are ambiguously defined (e.g., there are objects belonging to multiple classes and crisp labeling is not sufficient). Such an erroneous labeling can confuse the training procedure and increases a lower boundary for the classification error. Moreover, it causes a high inter- and intra- expert variability. Consequently, it leads to lower acceptance of automatic pattern classification systems by their potential users. Such issues can be solved by the use of soft labels and fuzzy methods. In active learning scenarios, several approaches relevant to such issues were also proposed [11–15].

A typical example of such an issue is the application of active learning to the sleep EEG processing. It is a well known problem that sleep is a process with continuous transitions. Then, the labeling expert (physician) has problems with assignment of the transitional parts of the EEG signal into one particular sleep stage (i.e., to use crisp labels). Such state transitions correspond to the ambiguous data instances. In [7], we observed that such instances are located on the border between different classes. An example is depicted in Figure 3, where a data set with 2 features and four classes is depicted in a scatter plot. Instances corresponding to label transitions from upper part of Figure 1 are highlighted in Figure 3a as black circles. It can be seen that such instances often appear close to the borders between classes. Unfortunately, the majority of current AL techniques (especially most popular query-by-committee and uncertainty sampling) prioritize the border instances and thus the ambiguous instances are often queried for labeling. This can lead to the failure of AL methods.

In [7], only an unrealistic method for detection of transitions based on apriori known labels was considered. Its purpose was to demonstrate that the transitional instances disturb the iterative learning process and their removal can improve the learning. It was assumed that such noisy segments are those, whose labels differ from one of its adjacent segments. Based on this assumptions, all such training instances can be removed, which occur before or after a transition of class labels. However, the unrealistic nature of such an approach is caused by the fact that in the proposed iterative expert-in-the-loop scenario, one does not dispose of labels in advance and cannot use such a detection rule. In [7], this problem was intentionally postponed for future work. Here we provide and analyse a solution of such a problem, based on detection of real transitions directly from unlabeled time series of feature values. To find the unlabeled training instances that correspond to potential transitions between sleep stages, we propose to use HMMs that are popular in EEG and PSG signals processing [16]. Recently, HMMs were used in different manners for sleep EEG artefact detection [17], sleep stage classification [18] or post-hoc refinement of classification results [19].

In our approach, HMM is trained on multidimensional time-series composed of unlabeled training data. It is trained on extracted features rather then on original EEG signals due to much smaller computational requirements. The main aim is to identify inherent state transitions no matter what is the real interpretation of the states. HMM with five states and Gaussian emission distributions was trained using Expectation-Maximization (EM) algorithm. The number of states was chosen empirically corresponding to original number of sleep stages in Rechtschaffen and Kales scoring manual, although the method does not need a clear correspondence between sleep stages and HMM states. After the model is trained, the most probable state sequence is computed from the feature time series using Viterbi algorithm [16]. Regardless of the interpretation of HMM states, the instances whose states labels differ from state labels of adjacent instances are considered as transitions and removed from the training set. An example of such HMM states is shown in the lower part of Figure 1, where the HMM-based transitions are highlighted by black circles. In Figure 3b, the scatter plot with such transitions is demonstrated. One can see that data in stances corresponding to those HMM-based transitions also exhibit a tendency to appear close to the class borders. On the other hand, when comparing the two subfigures of Figure 1, the HMM-based transitions do not fully correspond to label-based transitions both in number and positions. Nevertheless, some similarities can be visually observed.



(**a**) Label-based transitions (**b**) HMM-based transitions

Figure 3. Scatter plot of two features with label transition instances (**a**) and HMM states transition instances (**b**) marked as black circles.

3. Experimental Setting

This section first describes the two classification methods that were used - one very simple and one very popular classifier. Further, the EEG data, their acquisition and their processing including feature extraction are described. The methodology of evaluation and comparison of methods is finally described in detail.

3.1. Classifiers

Two classifiers were assumed - linear Bayes classifier (LB) and support vector machine (SVM). Observing a feature vector $\mathbf{x}$, Bayes classifiers use Bayes theorem to compute a posteriori probability $\mathrm{p}(\omega|\mathbf{x})$ of each class from likelihood $\mathrm{p}(\mathbf{x}|\omega)$ and class prior probability $\mathrm{p}(\omega)$ as $\mathrm{p}(\omega|\mathbf{x}) = \mathrm{p}(\mathbf{x}|\omega)\mathrm{p}(\omega)/\mathrm{p}(\mathbf{x})$. The feature vector $\mathbf{x}$ is further assigned into the class that maximizes the a posteriori probability, i.e., the result of classification by classifier γ is $\gamma(\mathbf{x}) = \arg\max_{\omega\in\Omega} \mathrm{p}(\omega|\mathbf{x})$. Uniform prior probabilities and Gaussian likelihoods with equal covariance matrices are assumed that lead to linear decision boundary.

SVM classifier with linear kernel was tested, because of its popularity and reasonable computational time. The training procedure uses quadratic programming to maximize margin, which is a width of the largest 'tube' not containing samples that can be drawn around the decision boundary. For linearly separable data, it can be proven that this particular solution has the highest generalization ability [20] among all linear separating boundaries.

3.2. Sleep EEG Data

The recorded data set contains 36 EEG recordings acquired in the National Institute of Mental Health of Czech Republic (NIMH-CZ). Recordings on 18 healthy subjects and 18 insomniac patients are available, while the membership of the patients within those two groups is ignored, but it corresponds to realistic assumption that the semi-automatic labeling will be performed within a clinical practice, where both types of people can occur. Some details about the recorded data can be found in [7].

All data were measured using standard 10-20 montage referenced to mastoids (M1, M2). 19 EEG channels were used for further feature extraction and are also summarized in [7]. All EEG recordings were completely and manually labeled, i.e., sleep classes were assigned to all samples of the signals, by accredited sleep neurologist according to the American Association of Sleep Medicine Scoring Manual (AASM 2012) [21] (classes Wake, N1, N2, N3 and REM). N1 and N2 were merged into one class.

Within the feature extraction, the signals were split to 30 s long segments and 21 features commonly used in EEG sleep staging were extracted for each segment of each channel. This gives $21 \times 19 = 399$ features in total. The complete list of features and their more detailed description can be found in [7,22]. To increase a robustness and resistance to artifacts we used a median operation over each feature across all EEG channels. Through this process we reduced the number of features from 399 to 21.

3.3. Evaluation of Classification Performance

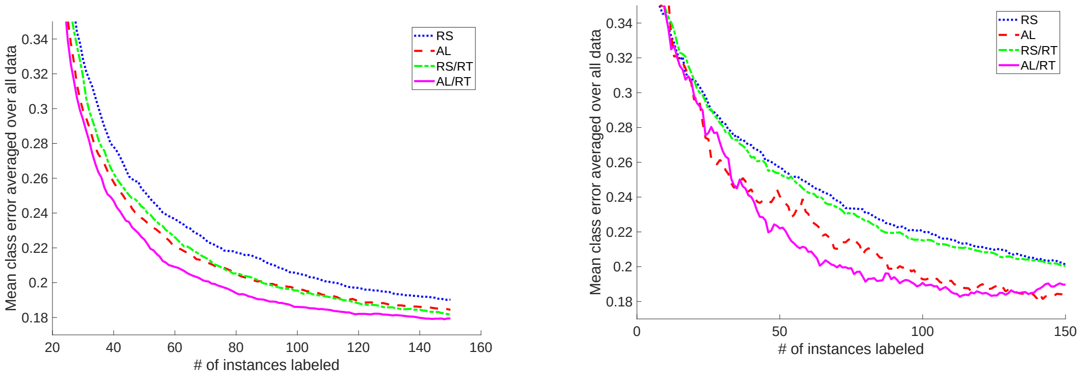
To compare active learning approaches to random sampling strategy (RS) and evaluate the influence of the proposed removal of transitions (RT), we used a mean class error defined as the ratio of incorrectly classified instances from particular class averaged over all four classes. This is actually analogical to $1-$sensitivity in binary classification cases. The estimation of true mean class error is performed via hold-out estimate, where all the data instances are randomly split to two disjunctive subsets of the same size that are called training and testing subset. Since our classifier will be used in an intra-personal manner—trained on labeled segments and applied to classification of the rest—the testing is performed on each of 36 patient records separately which gives 36 values of the mean class error for each of the evaluated methods. The methods are further compared using average computed over all patient records. Let e_{ij} is the relative missclassification error obtained for patient i on class j. The mean class error averaged over all 36 patients is defined as: $E = 1/36 \sum_{i=1}^{36} 1/4 \sum_{j=1}^{4} e_{ij}$.

For a statistical comparison of four active learning strategies on multiple data sets, we use the Friedman test. The test and its application on comparison of classifiers is described in [23]. It is a non-parametric equivalent of the "repeated-measures ANOVA". It can be used for comparison of our 4 algorithms on 36 data sets. The null hypothesis is that there are no differences between the algorithms. For each data set, the algorithms are ranked according to the performance. The best performing algorithm has rank 4 and the worst performing algorithm has rank 1. The Friedman statistics is computed from the average ranks and corresponding p-value is found. If the p-value is lower then a significance level α, the null hypothesis can be rejected.

Once the Friedman test rejects its null hypothesis, one can use a post-hoc procedure to find the particular pairwise differences [24]. According to [25], the Bergmann-Hommel procedure [26] is the most powerful one for pairwise comparisons of classification methods, although it requires more intensive computation. In our experiments, we used an implementation from [25].

4. Experimental Results

The average learning curves that show the dependence of the performance criterion defined above on the number of labeled instances are depicted in Figure 4. The curve is plotted only for the first 150 sampling iterations. The first reason is that it does not make a sense to use the semiautomatic scenario if the user must label too many segments since the reduction of labeling effort would be too small. The second reason is a better visualization of differences. The following conclusions can be made about the average behavior. First, the random sampling without the removal of transitions (RS) is the worst method during the first 150 sampling iterations for both tested classifiers. On the other hand, the active learning with the removal of transitions (AL/RT) is the best among all tested methods for all tested sampling iterations. For LB classifier, there seems to be nearly no difference between active learning without transitions removal (AL) and random sampling with transitions removal (RS/RT). This means that the removal of transitions makes the random sampling comparable to active learning. Such results show that both the active learning strategy and transitions removal bring some additive benefits. The reduction of labeling effort depends on the user's tolerance of an error. If the user can tolerate 18% error (average error on completely labeled dataset is about 17%), one can see that such an error is reached approximately after 100 iterations. Since the actual number of segments in each record is about 700, the user could label only 15% of segments and 85% of the labeling effort would be saved.



(**a**) Linear Bayes classifier (**b**) Support Vector Machine

Figure 4. Learning curves for linear Bayes classifier (**a**) and support vector machine (**b**) for random sampling (RS), random sampling with removal of transitions (RS/RT), active learning (AL) and active learning with removal of transitions (AL/RT).

We should point out that all the mentioned observations are the same as those made in our previous paper [7]. Although there is no clear correspondence between HMM-based transitions and label-based transitions (as can be seen in Figure 1), the newly proposed method keeps the main advantages while being practically implementable.

For a better clearness, there were no standard deviations or confidence intervals depicted in Figure 4. To make more credible statistical evaluation, the Friedman test accompanied by Bergmann-Hommel posthoc procedure was used. The average rankings of the particular methods and particular iteration of the sampling process are summarized in Table 1 for both classifiers. It can be clearly observed for LB classifier that the combination of active learning and transitions removal leads to the best average rank for all three tested moments of the sampling process and both classifiers.

For all columns of the table, the p-value computed by the Friedman test suggests that at least one of the four tested methods gives significantly different median than the others at level of significance $\alpha = 0.01$. When performing pairwise comparisons using the Bergmann-Hommel procedure several statistically significant conclusions can be drawn at level of significance 0.1. First, both AL methods were significantly better than simple random sampling at all three moments of the sampling procedure. Second, AL/RT method outperformed RS/RT method for all moments. Although for SVM there is no sufficient experimental data to reach statistically significant conclusion that AL was improved by transitions removal, for LB classifier AL/RT was significantly better than AL after 40 and 60 iterations.

This statistically supports our original hypothesis that the removal of instances that correspond to HMM-based transitions can improve the semi-automatic sleep scoring. This is statistically stronger outcome than in [7] and thus the removal of HMM-based transitional instances seems to be even more promising than the unrealistic removal of label-based transitions.

Table 1. Average rankings of the sampling strategies and Friedman P-values

	LB			SVM		
Iteration	**40**	**60**	**100**	**40**	**60**	**100**
RS	1.71	1.74	1.86	1.74	1.69	1.74
AL	2.57	2.66	2.63	2.97	2.80	3.06
RS/RT	2.43	2.23	2.34	2.20	2.14	1.89
AL/RT	3.29	3.37	3.18	3.09	3.37	3.31
Friedman p-value	8.9×10^{-6}	1.35×10^{-6}	2.7×10^{-4}	1.05×10^{-5}	1.6×10^{-7}	8.8×10^{-9}

5. Conclusions

The paper confirms one conclusion from [7], that the active learning outperforms the random sampling in semi-automatic EEG-based sleep scoring in terms of mean class error. Its main conclusion is, however, a solution of problem of detection of potentially ambiguous data instances that should be not queried for labeling. It was shown that the method based on most probable state sequence of HMM can find data instances whose deletion from training set can statistically significantly improve both the random sampling and the active learning procedure.

An interesting finding is that the HMM-based method supports the active learning better than label-based method. It means that even if all labels would be theoretically known (which does not make a sense), the deletion of label transitions improves the semi-automatic scoring less than the deletion of HMM state transitions. This can be caused by the fact that the transitions provided by human expert do not correspond to real transitions that exist in EEG signal. Such transitions are better detected by HMM method.

Although we demonstrated the use of our method in the sleep staging application, it can be relevant to many other fields. First, if a time series is classified and classes change continuously in transitional manner, the problem with ambiguity can be avoided by the proposed detection and removal of transitions. Second, if it is needed to minimize an interaction with human oracle during training data acquisition, active learning can help to choose data to be labeled efficiently. The combination of those two properties is common in ambient intelligence applications, where adaptive models of time series coming from sensors are often required that avoid an excessive disturbance of human users.

Author Contributions: Conceptualization, M.M., N.G.; Methodology, M.M.; Software, M.M. and N.G.; Validation, M.M.; Investigation, M.M.; Writing—Original Draft Preparation, M.M.; Writing—Review & Editing, M.M.; Visualization, M.M.; Data Preparation, V.G.; Funding Acquisition and Project Leader, L.L.

Funding: This research was funded by Czech Science Foundation grant number 17-20480S: Temporal context in analysis of long-term non-stationary multidimensional signal.

Acknowledgments: We would like to acknowledge Martin Brunovsky, Jana Koprivova, Daniela Dudysova and Alice Heuschneiderova from National Institute of Mental Health Czech Republic for their assistance, discussions, and contribution to the acquisition and evaluation of the EEG data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Settles, B. Active learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2012**, *6*, 1–114.
2. Kholghi, M.; Sitbon, L.; Zuccon, G.; Nguyen, A. Active learning reduces annotation time for clinical concept extraction. *Int. J. Med. Inform.* **2017**, *106*, 25–31.

3. Lawhern, V.; Slayback, D.; Wu, D.; Lance, B.J. Efficient labeling of EEG signal artifacts using active learning. In Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon, China, 9–12 October 2015; pp. 3217–3222.
4. Hossain, I.; Khosravi, A.; Hettiarachchi, I.; Nahavandi, S. Batch Mode Query by Committee for Motor Imagery-Based BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *27*, 13–21.
5. Wu, D. Active semi-supervised transfer learning (ASTL) for offline BCI calibration. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 246–251.
6. Wu, D.; Lawhern, V.J.; Gordon, S.; Lance, B.J.; Lin, C.T. Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning (EBMAL) for regression. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016; pp. 000730–000736.
7. Macaš, M.; Grimová, N.; Gerla, V.; Lhotská, L.; Saifutdinova, E. Active Learning for Semiautomatic Sleep Staging and Transitional EEG Segments. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018; pp. 2621–2627.
8. Settles, B. *Active Learning Literature Survey*; Computer Sciences Technical Report 1648; University of Wisconsin-Madison: Madison, WI, USA, 2009.
9. Lewis, D.D.; Catlett, J. Heterogeneous Uncertainty Sampling for Supervised Learning. In Proceedings of the Eleventh International Conference on Machine Learning, New Brunswick, NJ, USA, 10–13 July 1994; pp. 148–156.
10. Scheffer, T.; Decomain, C.; Wrobel, S. Active hidden markov models for information extraction. In Proceedings of the International Symposium on Intelligent Data Analysis, Cascais, Portugal, 13–15 September 2001; pp. 309–318.
11. Zhao, L.; Sukthankar, G.; Sukthankar, R. Robust Active Learning Using Crowdsourced Annotations for Activity Recognition. In Proceedings of the 2011 AAAI Workshop, San Francisco, CA, USA, 8 August 2011; Volume 32.
12. Sheng, V.S.; Provost, F.; Ipeirotis, P.G. Get another label? improving data quality and data mining using multiple, noisy labelers. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 614–622.
13. Zheng, Y.; Scott, S.; Deng, K. Active learning from multiple noisy labelers with varied costs. In Proceedings of the 2010 IEEE 10th International Conference on Data Mining (ICDM), Sydney, Australia, 13–17 December 2010; pp. 639–648.
14. Zhang, C.J.; Chen, L.; Jagadish, H.; Cao, C.C. Reducing uncertainty of schema matching via crowdsourcing. *Proc. VLDB Endow.* **2013**, *6*, 757–768.
15. Bouguelia, M.R.; Nowaczyk, S.; Santosh, K.; Verikas, A. Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *Int. J. Mach. Learn. Cybern.* **2018**, *9*, 1307–1319.
16. Rabiner, L.R.; Juang, B.H. An introduction to hidden Markov models. *IEEE Assp Mag.* **1986**, *3*, 4–16.
17. Malafeev, A.; Omlin, X.; Wierzbicka, A.; Wichniak, A.; Jernajczyk, W.; Riener, R.; Achermann, P. Automatic artefact detection in single-channel sleep EEG recordings. *J. Sleep Res.* **2018**, e12679, doi:10.1111/jsr.12679.
18. Fonseca, P.; den Teuling, N.; Long, X.; Aarts, R.M. A comparison of probabilistic classifiers for sleep stage classification. *Physiol. Meas.* **2018**, *39*, 055001.
19. Jiang, D.; Lu, Y.N.; Yu, M.; Yuanyuan, W. Robust sleep stage classification with single-channel EEG signals using multimodal decomposition and HMM-based refinement. *Expert Syst. Appl.* **2019**, *121*, 188–203.
20. van der Heijden, F.; Duin, R.; de Ridder, D.; Tax, D. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*; John Wiley and Sons: Hoboken, NJ, USA, 2004.
21. Berry, R.B.; Brooks, R.; Gamaldo, C.E.; Harding, S.M.; Marcus, C.; Vaughn, B. *The AASM Manual for the Scoring of Sleep and Associated Events*; American Academy of Sleep Medicine: Darien, IL, USA, 2012.
22. Gerla, V.; Djordjevic, V.; Lhotska, L.; Krajca, V. PSGLab Matlab toolbox for polysomnographic data processing: Development and practical application. In Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine, Corfu, Greece, 3–5 November 2010; Volume 20, pp. 47–50.
23. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.

24. Hommel, G.; Bretz, F.; Maurer, W. Multiple Hypotheses Testing Based on Ordered p Values—A Historical Survey with Applications to Medical Research. *J. Biopharm. Stat.* **2011**, *21*, 595–609.
25. Garcia, S.; Herrera, F. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *J. Mach. Learn. Res.* **2008**, *9*, 2677–2694.
26. Bergmann, G.; Hommel, G. Improvements of general multiple test procedures for redundant systems of hypotheses. In *Multiple Hypotheses Testing*; Springer: Berlin, Germany, 1988; pp. 100–115.