Ensemble learning for brain computer-interface using uncooperative democratic echo state communities

Cédric Gouy-Pailler, Michèle Sebag, Antoine Souloumiac, Anthony Larue

Ensemble Learning for Non-Invasive Brain Computer-Interfaces Using Uncooperative Democratic Echo State Communities

Cédric GOUY-PAILLER, Michèle SEBAG
TAO Team - INRIA Saclay
LRI - Paris-Sud XI University
91405 Orsay Cedex, France

Antoine SOULOUMIAC, Anthony LARUE
CEA, LIST
Laboratoire d'Outils pour l'Analyse de Données
F-91191 Gif-sur-Yvette, France

***Abstract*—This paper deals with the issue of features construction and selection for signals acquired during non-invasive Brain-Computer Interface (BCI) experiments. The Echo State Network (ESN) architecture, a reservoir computing approach proposed by H. Jaeger in 2001, is first adapted to the specific issue of EEG signals classification. In order to predict the performed task at a relatively low computational cost, a basic ESN architecture is combined with regularized logistic regression trained following aggressive subsampling principles. The resulting method is shown to significantly outperform classification rates obtained using raw EEG signals. Basic single ESNs are then integrated to take advantage of the fruitful combination between ensemble learning techniques and aggressive subsampling principles. The resulting novel architecture, constituting an Uncooperative Democratic Echo State Community (UDESC), yields one of the first attempt to provide an efficient subject-independent features construction algorithm. Based on the generative power of individual ESNs as well as the discriminative abilities of ensemble learning combined with aggressive subsampling, UDESC is shown to advantageously integrate the knowledge acquired by each single ESN. The results shown along this paper make an extensive use of a real dataset made available to the BCI community during BCI Competition 2008. This dataset consists of four subjects involved in a two-class motor-imagery BCI experiment.**

I. INTRODUCTION

Brain-Computer Interfaces (BCIs) aim at establishing a direct communication pathway between users will and electronic devices. Initially designed to provide people suffering from severe motor diseases with a tool to restore communication and movement [1], their applications nowadays range from medical and rehabilitation purposes to video games industry. Systems usually rely on the identification of predefined mental tasks within the ongoing brain activity. A typical example is Motor Imagery (MI) and its resulting somatotopical and frequency-specific signals, which allows, *e.g.*, the control of two-dimensional cursors [2]. Such signals are typically measured using electroencephalography (EEG). Over the past twenty years, signal processing methods have been extensively developed and used for the on-line extraction of relevant information from electrical EEG measurements. Usual methodologies entail the training of subject-dependent linear spatial filters, which form a task-specific projection basis [3], [4]. Features such as spectral power or autoregressive coefficients are then calculated in this basis to train a subject-specific classifier able to elucidate the performed task from unseen data.

Recent advances in the machine learning community showed that good generalization of complex dynamical signals can be achieved with non-linear architectures called Echo State Network (ESN) [5]. The idea of ESN is to use a fixed, randomly and sparsely connected Recurrent Neural Network (RNN) of simple units. These units receive a time-invariant mixture of input signals, obey simple update equations responsible for a fading memory effect and constitute a dictionary of complex signals, which can generate any kind of output through learned readout functions. ESNs are a special case of Reservoir Computing (RC) architectures, which subsume the idea of combining dynamical systems with memoryless readout functions as computational devices [6], [7]. Although ESNs might be a perfect alternative when subject-independent features construction techniques are needed, they have been seldom used in the BCI context [8]. The present paper addresses issues raised by the use of ESNs in BCI contexts. Following ideas firstly proposed in [9], proposed solutions will make an extensive use of ensemble learning techniques to increase robustness [10] and aggressive subsampling to maintain a low computational cost. A critical factor of difficulty is indeed the labeling noise [11], [12]. During BCI experiments, subjects are instructed to perform MI tasks during 3 to 4 seconds. Two issues arise from this methodology, first experimenters cannot assess whether the task is actually performed, second subjects are unlikely to perform the task during 3 seconds. The contribution of this work is to provide a subject-independent features construction architecture called UDESC (Uncooperative Democratic Echo State Community), which shows good generalization performances. Results are demonstrated on dataset 1 of BCI Competition IV.

This paper is organized as follows. The first section is devoted to the description of the basic Echo State Network architecture, the learning algorithm and the proposed extension called Uncooperative Democratic Echo State Community (UDESC). The real dataset is then described in the next section. Results are shown in a third section. They are lastly discussed in a final section.

II. OVERVIEW OF THE METHOD

A. Echo State Network Architecture

Echo State Networks (ESNs) have been introduced in 2001 by H. Jaeger [5]. We first fix the notations and present the basic principles of ESNs. The global structure

of ESNs is depicted in figure 1. ESNs consist of K input units, typically the EEG signals or frequency-filtered measurements. K thus corresponds to the number of EEG sensors. Activations of input units are real-valued, $\mathbf{u}(t) = [u_1(t), \cdots, u_K(t)]^T \in \mathbb{R}^K$. The internal units form the reservoir of the ESN, it consists of N single neurons. The activations of internal units are also real-valued numbers $\mathbf{x}(t) = [x_1(t), \cdots, x_N(t)]^T \in \mathbb{R}^N$. $\mathbf{p}(t) = [p_1(t), \cdots, p_L(t)]^T \in [0,1]^L$ denotes the output layer. Although general formulation of ESNs considered real-valued output units, BCI contexts involve discrimination. Thus output units represent probabilities and for all t, $\sum_{l=1}^{L} p_l(t) = 1$. Input units are linearly projected onto the reservoir units using fixed weights that are gathered in a matrix $W_{\text{in}} \in \mathbb{R}^{N \times K}$. The temporal evolution of the reservoir activations is governed by the function f_{act} and interconnection weights gathered into $W \in \mathbb{R}^{N \times N}$. The activation of internal units is updated according to

$$\mathbf{x}(n+1) = f_{\text{act}}\left(W^{\text{in}}\mathbf{u}(n+1) + W\mathbf{x}(n)\right) \quad , \qquad (1)$$

where $f_{\text{act}} = \tanh(.)$. It should be noted that all previously mentioned weights are fixed. W_{in} is a random dense matrix such that

$$[W_{\text{in}}]_{i,j} = \begin{cases} -\frac{1}{K} & \text{with prob } 1/2 \\ \frac{1}{K} & \text{with prob } 1/2 \end{cases} \quad , \qquad (2)$$

while W is a sparse random matrix characterized by the density of non-zero coefficients and the maximum absolute eigenvalue. In [5], a sufficient condition such that the ESN has the interesting "echo state" property is that the maximum absolute eigenvalue is 1. Many authors yet observed interesting properties for matrices W with maximum absolute eigenvalue greater than 1. In the following the ESNs will have a maximum eigenvalues of 0.8 (obtained by scaling a randomly chosen matrix) and a density of connections of 0.5%. Coefficients of W were chosen among the vector $[-1, 0, 1]$ with probabilities $[0.0025, 0.995, 0.0025]$. The influence of these values on the final result is negligible as far as the density of connection remains small [7].

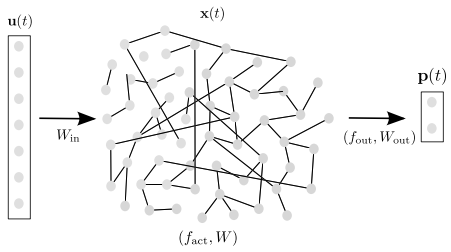

Figure 1. A simple Echo State Network (ESN) architecture.

In order to learn a matching between the reservoir activation and the task performed by the subjects, a logistic regression is used. Yet the activation values are not directly used to produce the probabilities, activations are rather integrated on small temporal windows. This step is mathematically written

$$[x']_n(t) = \log\left(\frac{1}{T}\sum_{\tau=t-T+1}^{t}[x]_n(\tau)^2\right) \quad . \qquad (3)$$

In the remaining of this paper T is set to 80, which corresponds to 0.8 s as the frequency rate considered further will be 100 Hz. A logistic regression is then applied to produce the outputs. Recall that the logistic regression model is written in the case of two-class outputs

$$\ln\frac{\mathbb{P}(y(t) = c_1|\mathbf{x}'(t))}{\mathbb{P}(y(t) = c_{-1}|\mathbf{x}'(t))} = W_{\text{out}_0} + \sum_{n=1}^{N} W_{\text{out}}\mathbf{x}'(t) \quad , \quad (4)$$

where $W_{\text{out}} \in \mathbb{R}^N$ and $c_{\pm 1}$ represent the available classes. Output probabilities can thus be deduced from this model by writing

$$\mathbb{P}(y(t) = c_1|\mathbf{x}'(t)) = \frac{\exp(W_{\text{out}_0} + \sum_{n=1}^{N} W_{\text{out}}\mathbf{x}'(t))}{1 + \exp(W_{\text{out}_0} + \sum_{n=1}^{N} W_{\text{out}}\mathbf{x}'(t))}$$

and using the fact that $\mathbb{P}(y(t) = c_1|\mathbf{x}'(t)) + \mathbb{P}(y(t) = c_{-1}|\mathbf{x}'(t)) = 1$. W_{out} gathers the only coefficients that are learned in the architecture.

Note that, strictly speaking, ESNs comprise the definition of an architecture and a supervised learning principle based on linear regression. The formulation proposed thus slightly differs from the initial one in the sense that the reservoir-to-output function has been adapted for BCI purposes by introducing the logistic regression output function. General ESN formulations mention potential output-to-reservoir back-propagations as well as input-to-output direct projections. These possibilities have been discarded in this work for the sake of simplicity and computational cost.

B. Aggressive Subsampling Learning

Training the architecture depicted in figure 1 is done in a supervised fashion. As the number of reservoir units can be huge, the choice of W_{out} (equivalent to the selection of dynamics within the reservoir dictionary) is crucial. Obviously most of the reservoir units will be useless and regularizations techniques based on ℓ_1/ℓ_2 norms have to be employed to select the most useful ones. The *elasticnet* algorithm is used [13], [14] to obtain sparse W_{out} solutions. Regularizations in *elasticnet* can be weighted through the coefficients α and λ. In the remaining of this paper, α will take the commonly used value of 0.9 and the influence of λ will be studied. Note that increasing the parameter λ results in selecting fewer dimensions.

Given a training set consisting of trials lasting a few seconds, different learning strategies can then be adopted. Note that at each time t, a feature is computed by the ESN, thus the amount of data is huge. Two substantial drawbacks would thus result from learning features computed at each time point t:

- the computational learning cost, namely the *elasticnet* algorithm that finds W_{out} coefficients along the λ regularization path, would be huge;

- as training features are likely to contain a high proportion of noisy labels due to tasks performance uncertainty, resulting logistic regression coefficients are not reliable.

Following principles explained in [9], an aggressive subsampling technique is therefore used to train the regularized logistic regression. This drastically decreases the computational load of the learning step because approximately only 2 % of the features are used to train a single classifier. Although robustness is of course not achieved in the case of one single ESN compared to using all time points, the aggressive subsampling has been shown to greatly improve robustness when ensembles of classifiers are used. This idea leads to the Uncooperative Democratic Echo State Community (UDESC).

C. Echo State Community

This section describes the core of the article. The ESN architecture and the aggressive learning principle have been described in previous sections. We will now explain how ESNs can be combined to take advantage of ensemble learning principles [10]. The basic idea is depicted in figure 2. Let M denote the number of ESNs in the community. EEG signals acquired using K electrodes are band-passed filtered around the frequency $\{f_c\}_{[1,M]}$. This constitutes a multi-dimensional filter bank from which the output is provided to the corresponding ESN. The single ESN components are used to predict the set of probabilities $\{\mathbb{P}_m(c_1)\}_{m\in[1..M]}$ and $\{\mathbb{P}_m(c_{-1})\}_{m\in[1..M]}$, which are combined to yield the probabilities

$$\mathbb{P}(c_{\pm 1}) = \sum_{m=1}^{M} \mathbb{P}_m(c_{\pm 1}) \tag{5}$$

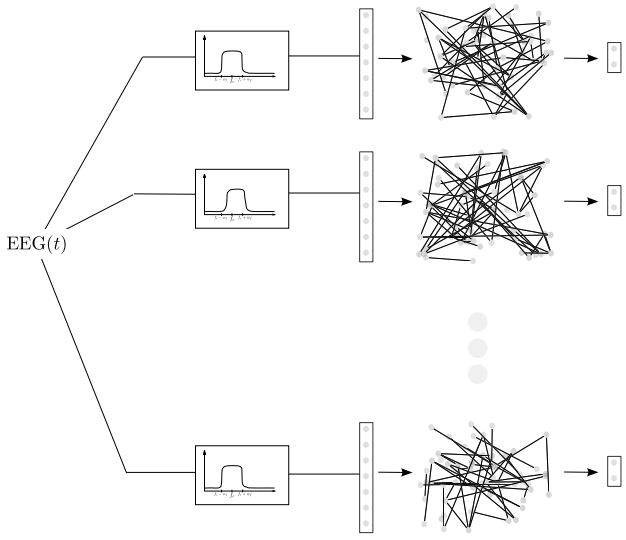

Figure 2. Uncooperative Democratic Echo State Community (UDESC)

The resulting architecture is called Uncooperative Democratic Echo State Community (UDESC). It is composed of independent single ESNs (hence Uncooperative), from which outputs are combined in a fashion resembling to a voting procedure (hence Democratic). Note that each ESN has a fixed random structure.

This architecture is trained using the aggressive subsampling technique. It means that each single ESN independently choose its own small training dataset among the set of available training examples (about 2 % of the training set).

III. Paradigm and Dataset

BCI Competition 2008 provided a dataset untitled "motor imagery, uncued classifier application", in which participants had to continuously identify the mental state of the subject [15]. The data were recorded from four healthy subjects. The experiment consisted in two motor-imagery-based sessions without feedback. Only the training session is used in this paper. The four subjects had to choose two classes of MI among *left hand*, *right hand* and *foot*. Instructions about the mental task to be realized were presented as visual cues on a computer screen. Cues were displayed for a period of 4 s during which the subject had to perform the cued MI. Two hundred trials were performed (one hundred of each task). The datasets were recorded at 1000 Hz, using 59 EEG sensors. Data were downsampled at 100 Hz for analysis purposes.

IV. Results

A. Single Echo State Network

Single Echo State Networks (ESN) are first compared to using raw EEG signals (RAW). A common band-pass filtering is first used in both cases. As motor imagery tasks are known to mainly involve frequency bands around 10 Hz, cut-off frequencies are set to 8 and 12 Hz. The number of units in the reservoir is set to various values ranging from 100 to 700. Various regularizations values λ are used, ranging from $10^{-3.5}$ to $10^{-0.8}$. Note that an efficient coordinate descent is used to compute the weights W_{out} along the whole regularization path, hence drastically reducing the computational load. While the task was performed during about 4 seconds, 3 seconds were actually used and the remaining data were discarded from the analysis. The generalization performances are evaluated using a 1×5 cross-validation procedure. The global dataset is equally split between the training and the test sets. Whereas only 2 % of the training set is used to train the classifier, prediction is done using the whole test set. This yield a classification rate corresponding to the proportion of correctly classified time points. This procedure is applied five times and the average is depicted on figure 3.

Results corresponding to subjects a, b, g and f are presented. The use of the ESN architecture clearly improves the classification rate for subjects a, g and f for values of N greater than 300. An important point is that the same regularization parameter can be chosen for each subject ($\lambda = 10^{-2}$). It has to be noticed that subject b completely differs from the other subjects. The performances obtained by ESN and RAW methods perform similarly bad for this subject. While the improvement achieved with $N = 500$

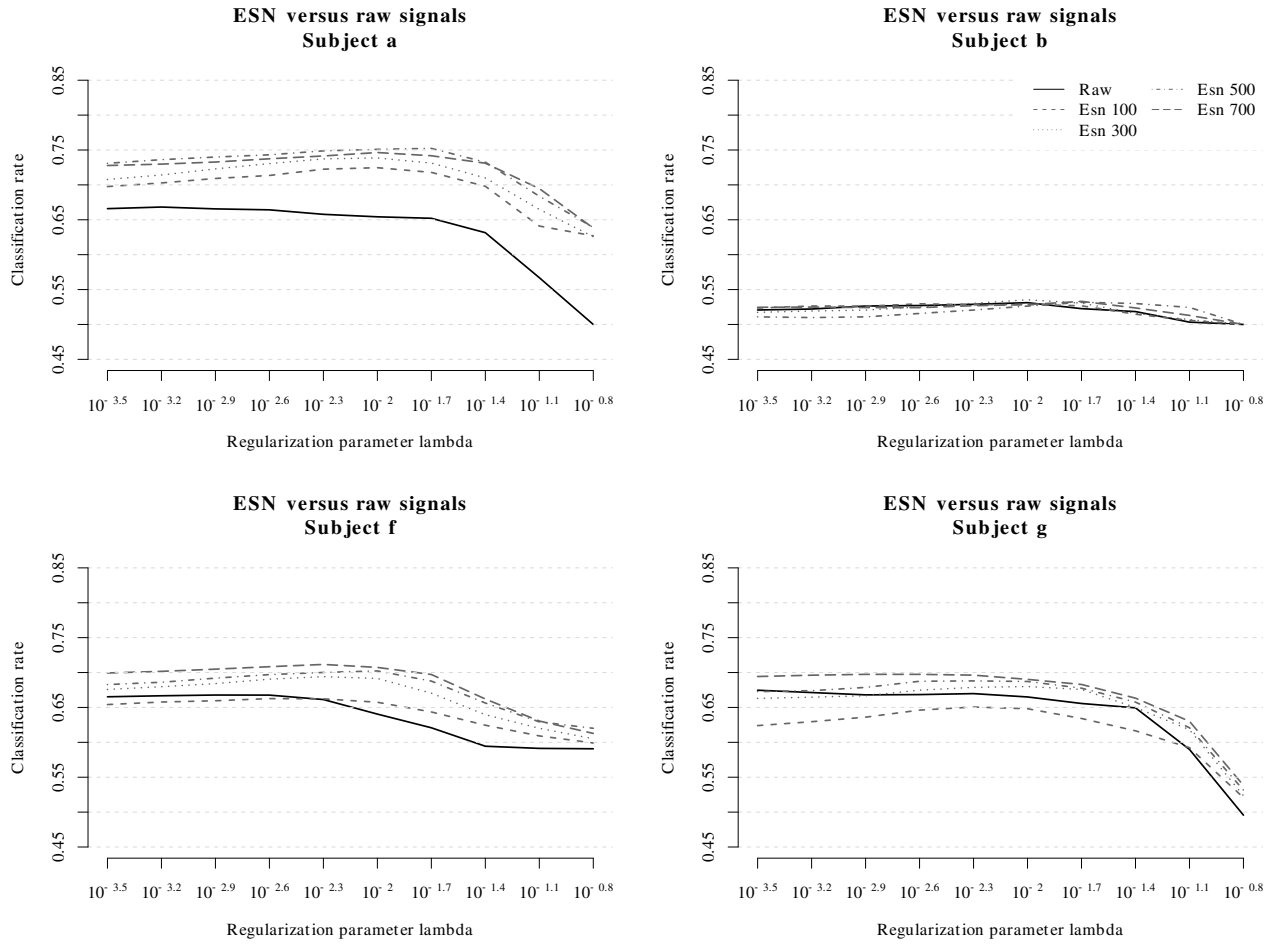

Figure 3. Generalization performances of the ESN as well as the RAW approaches versus the value of the regularization parameter λ.

compared to $N = 300$ is still sensible, results obtained for values of N greater than 500 are not sensibly different. These observations support the parameters chosen for performing simulations with UDESC. In the remaining of this paper, the number of units in the reservoirs is set to $N = 500$ and the regularization parameter is set to $\lambda = 10^{-2}$.

B. Uncooperative Democratic Echo State Community

A similar evaluation procedure is employed for UDESC. But contrary to the previous section, regularization parameter is fixed. The influence of the number of ESN in the community is studied. Note also that the cut-off frequencies of the filters are also tightly randomized in order to increase the heterogeneity of each single ESN. Five cross-validations are performed on equally sized train and test sets using a UDESC community of 50 ESNs. The training is done by each ESN using an independent subsample of the training set, while the test phase is performed on each time point of the test set. For computational reasons, we do not perform the simulation for each separate value of M but rather radomly choose among the 50 ESN a subset to be combined to obtained the predicted class. After applying this procedure 100 times for each value of M and each cross-validation, a classification rate is obtained for each value of M. The results are shown in figure 4.

The influence of ensemble learning is clearly seen on this figure. The median results using a community size of 40 outperform the performances using one single ESN. Classification rates clearly grow for community sizes between 2 and 20 whereas improvements are small for community sizes greater than 20.

V. Discussion and Conclusion

Echo State Networks have been proved to outperform the raw EEG classifcation scores for 3 out of 4 subjects. These results confirm the good generalization abilities of the structure proposed by H. Jaeger in 2001 [5]. It should be noted that the ESN architecture used in this paper has been greatly simplified over the general structure presented initially by H. Jaeger. Notably the output-to-reservoir back-propagation has been removed. This obviously reduce the memory ability of the ESN but also substantially simplify the learning. As computational cost becomes crucial when dealing with cross-validation procedures, back-propagation have thus been deliberately removed. Including such weights will be considered in

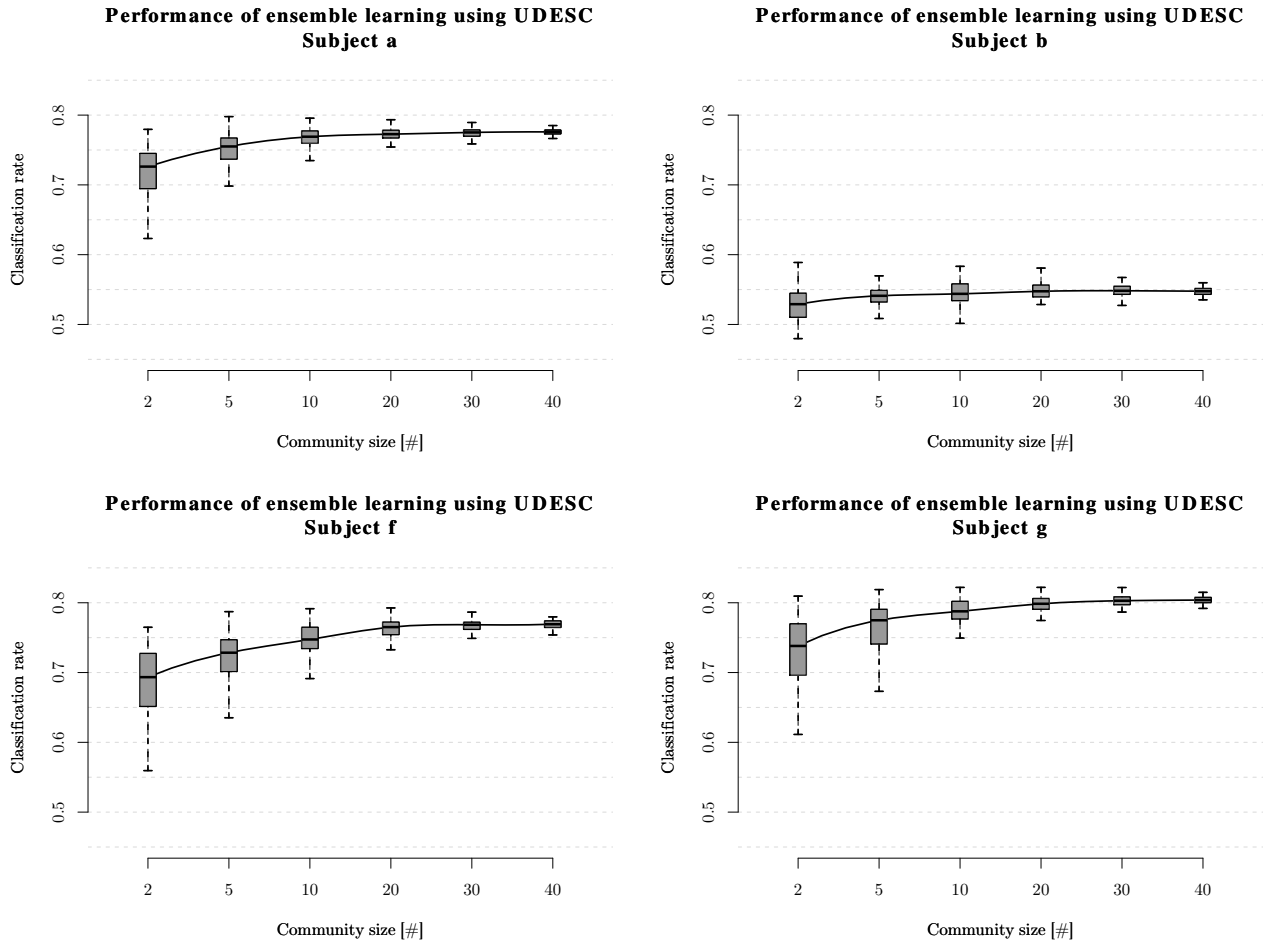

Figure 4. Performances obtained using various community sizes are depicted. Each boxplot summarizes the classification rates obtained by the choice of 100 sets of M ESN among 50 available in 5 cross-validations. Note that learning and testing sets differ between each cross-validation whereas the 100 scores obtained by randomizing the choice of the M voting ESNs concern the same learning and test sets. Each boxplot hence summarizes 500 classification scores by showing the median (straight line inside the box), lower and upper quartile (edges of the box).

future extensions of this work. The absolute classification scores obtained in this work, ranging from 53 % for subject b to 75 % for subject a, are relatively low compared to those usually reported in classical 2-class motor-imagery BCIs. This is explained by the fact that the problem considered in this work is much more complicated than the usual ones. Whereas classical approaches aim at labeling each trials corresponding to 4 seconds of recorded EEG data, each time point was classified in the present work. Temporal decisions across the trials should thus be integrated in order to be able to compare the two approaches. The ESN approach should rather be interpreted in the context of asynchronous BCIs [3]. Asynchronous BCIs relate to contexts where systems are not aware of tasks timing information and the EEG signals have to be continuously decoded to decide whether subjects are sending commands. Nevertheless a substantial difference between this work and asynchronous BCIs lies in the identification of idle states. Future work will consider the use of the architectures proposed in this work for asynchronous BCI paradigms.

Computational costs are crucial in BCI. Algorithms should be able to be applied in real-time, which mean that the time needed to process a time sample do not exceed the time between two acquisition (sample rate). Basically, in case of single ESNs, the operations needed to process one time sample consist in matrix-vector multiplications and non-linear functions. The global cost remains reasonable as far as the sparse structure of the ESN reservoir is used. Indeed, the matrix-vector multiplications implying dense W would be time-consuming. Fortunately, the sparse structure of W results in efficient implementations that are linear in the number of non-zero coefficients in W. The same principle should be used in the implementation to produce the probability output from the sparse vector W_{out}. The proposed extension consisting in using a community of ESN obviously lead to an increase of the computational cost by a factor of M. As we have seen, the number of ESN needed to obtain excellent performances is quite small and do not jeopardize the real-time specifications. Real-time is not necessary during training but computational load needed by the UDESC

architecture during training step is negligible compared to the classifier training algorithm, the use of aggressive subsampling is mandatory to achieve reasonable training times.

While this paper only dealt with two-class classification problems, the basic ESN architecture and its combination with a multi-class logistic regression is natural. The weights vector W_{out} then become a matrix containing weights to be applied to obtain the probability of each class.

The UDESC architecture efficiently combines the generalization ability of individual Echo State Networks, the low computational load resulting from aggressive subsampling, the good performances of regularized logistic regression and the robustness of ensemble learning. UDESC was inspired by the well-known Random Forests [16] proposed by L. Breiman, which efficiently combine ensemble learning principles with discrimination based on decision trees. UDESC provides a subject-independent features construction method. Such methods might be of great interest to build subject-independent BCI systems. On the one hand it might allow a better control of the amount of data needed to calibrate a BCI. In classical methods the amount of data needed to train the system depends on the one needed to train the features construction method (*e.g. linear spatial filters*) and the one needed to train the classifier. The global amount of time thus equals the maximum of these two quantities. When subject-independent features construction methods are used, the amount of data needed simplifies to the amount of data needed to train the classifier. Using such algorithms might thus improves controlling the number of training trials. On the other hand the proposed approach might constitutes a first step toward the design of universal BCI systems. Subject-independent features construction methods indeed provide an invariant features representation. Future works might consider using a unique UDESC representation for each subject and combining the output linear functions as a database to be used by new subjects. Training a BCI system might then consist in selecting the adapted output vectors.

VI. Acknowledgments

This work is part of the DIGIBRAIN project, funded by DIGITEO (région Île-de-France). The third and fourth authors are also funded by ANR (Agence Nationale pour la Recherche) within the Open-ViBE2 framework.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, Jun 2002.

[2] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, July 2001.

[3] C. Gouy-Pailler, *Interfaces Cerveau-Machines : Modèles dynamiques de l'activité cérébrale pour la conception de systèmes asynchrones*. Editions Universitaires Européennes, May 2010, no. ISBN: 978-6131503764.

[4] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Jan 2009. [Online]. Available: http://dx.doi.org/10.1109/TBME.2009.2012869

[5] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks," Fraunhofer Institute for Autonomous Intelligent Systems, Tech. Rep., 2001.

[6] B. Schrauwen, L. Büsing, and R. Legenstein, "On computational power and the order-chaos phase transition in reservoir computing," in *Proceedings of Neural Information Processing Systems, NIPS 2008*, Vancouver, Canada, December 2008.

[7] M. Lukosevicius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, Aug 2009.

[8] Y. N. Rao, S.-P. Kim, J. C. Sanchez, D. Erdogmus, J. C. Principe, J. M. Carmena, M. A. Lebedev, and M. A. Nicolelis, "Learning mappings in brain machine interfaces with echo state networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, 2005.

[9] C. Gouy-Pailler, M. Sebag, A. Larue, and A. Souloumiac, "Sabin: a resampling-based learning algorithm for idle state identification in asynchronous brain-computer interfaces," in *ICPR Workshop on Brain Decoding*, Istanbul, Turkey, Aug 2010.

[10] B. V. Dasarathy and B. V. Sheela, "Composite classifier system design: concepts and methodology," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 708–713, 1979.

[11] M. Kearns and M. Li, "Learning in the presence of malicious errors," *SIAM Journal on Computing*, vol. 2, no. 4, pp. 807–837, Aug 1993.

[12] D. Angluin and P. Laird, "Learning from noisy examples," *Machine Learning*, vol. 2, pp. 343–370, 1988.

[13] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. New-York, USA: Springer, 2009.

[14] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Soft.*, vol. 33, no. 1, pp. 1–22, Feb 2010.

[15] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive berlin brain-computer interface: fast acquisition of effective performance in untrained subjects," *Neuroimage*, vol. 37, no. 2, pp. 539–550, Aug 2007.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.