Towards a Fully Interpretable EEG-based BCI System

Fabien Lotte, Anatole Lécuyer, Cuntai Guan

Towards a Fully Interpretable EEG-based BCI System

F Lotte[1], A Lécuyer[2] and C Guan[1]

[1] Institute for Infocomm Research (I^2R), 1 Fusionopolis Way, #21-01 Connexis, 138632, Singapore

[2] INRIA Rennes-Bretagne Atlantique, Campus Universitaire de Beaulieu, 35042, Rennes Cedex, France

E-mail: fprlotte@i2r.a-star.edu.sg

Abstract. Most Brain-Computer Interfaces (BCI) are based on machine learning and behave like black boxes, i.e., they cannot be interpreted. However, designing interpretable BCI would enable to discuss, verify or improve what the BCI has automatically learnt from brain signals, or possibly gain new insights about the brain. In this paper, we present an algorithm to design a fully interpretable BCI. It can explain what power in which brain regions and frequency bands corresponds to which mental state, using "if-then" rules expressed with simple words. Evaluations showed that this algorithm led to a truly interpretable BCI as the automatically derived rules were consistent with the literature. They also showed that we can actually verify and correct what an interpretable BCI has learnt so as to further improve it.

1. Introduction

Brain-Computer Interfaces (BCI) are communication systems which enable users to send commands to computers by using brain activity only, this activity being generally measured by ElectroEncephaloGraphy (EEG) [1]. Since most BCI are based on machine learning, proposing a BCI design from which we could interpret what has been learnt would have several advantages. Indeed, it could be used to 1) check and improve the BCI design or protocol, 2) discuss what the BCI has learnt with neurophysiologists or other non-technical people, or 3) possibly improve our understanding of the brain [2]. Unfortunately, most BCI are black boxes, i.e., we cannot interpret what they automatically learn from the data [3]. Therefore, the BCI community has stressed the need for signal processing techniques from which humans could gain insights about the brain processes used by the BCI [2, 3].

Surprisingly, few papers have explored the design of interpretable BCI. For the sake of interpretability, some groups have studied the weights of classifiers [2, 4, 5] and spatial filters [6], or feature selection techniques [7, 8]. These methods give insights on what the most relevant features, channels and/or frequency bands are. Although this information is interesting, such methods cannot explain which values for the features correspond to which mental state. They cannot inform about which brain regions (in the whole brain volume, not only the surface) are relevant either, which could be a precious information. Finally, they do not provide a concise and simple explanation of what the BCI has learnt. This makes such BCI virtually impossible to interpret by non-technical people.

In this paper, we propose a method to design a fully interpretable EEG-based BCI. This method can report on what power in relevant brain regions and frequency bands corresponds to which mental state, *by using simple words*. The paper is organized as follows: Section 2 describes the proposed algorithm. Then, Section 3 reports on its evaluation. Finally, Section 4 concludes the paper.

2. Method

Our method to design an interpretable BCI can be broken down into three steps: 1) feature extraction based on inverse solutions, 2) classification based on fuzzy inference systems and 3) interpretability improvement based on linguistic approximation. They are detailed hereafter.

2.1. Feature extraction based on inverse solutions

To design an interpretable BCI, we first need interpretable features. Thus, we use features based on inverse solutions, i.e., methods that can estimate the activity in the whole brain volume from scalp EEG signals only. More precisely, we use FuRIA features [9] which correspond to the power in a small number of brain regions and associated frequency bands, estimated using an inverse solution. With FuRIA, the frequency bands

and brain regions whose power is relevant for classification are automatically identified from training EEG data using an appropriate learning algorithm (see [9] for details). In order to estimate the power in the whole brain volume from the scalp EEG signals, we used the sLORETA inverse solution [10] within FuRIA. As FuRIA features correspond to physiological information, they provide useful insights about the brain. However, they cannot report on what feature values correspond to which mental state on their own, hence the need for an interpretable classifier.

2.2. Classification: Fuzzy Inference Systems

The interpretable classifier we selected to classify FuRIA features is a Fuzzy Inference System (FIS) [11]. A FIS can automatically extract fuzzy "if-then" rules from data, these rules describing which input feature values correspond to which output class, i.e., mental state. When using the selected FIS (see [11] for details), the j^{th} fuzzy rule is as follows:

$$\text{If } X_1 \text{ is } A_{j1} \text{ and } \ldots \text{and } X_i \text{ is } A_{ji} \text{ and } \ldots \text{and } X_N \text{ is } A_{jN} \text{ Then Class is } C_j$$

where X_i is the i^{th} feature (here, a FuRIA feature, i.e., the power in a given brain region and frequency band) and A_{ji} is a fuzzy set, i.e., a function that describes the distribution of values taken by feature i for the mental state of class C_j. In other words, "X_i is A_{ji}" means "the value of feature X_i belongs to the fuzzy set (or distribution) A_{ji}".

Such fuzzy rules enable us to 1) classify EEG signals [11] and 2) interpret what the FIS has learnt. Indeed, they can report on which power in each relevant brain region and frequency band corresponds to which mental state. Although they can be interpreted by researchers familiar with fuzzy sets, they are not naturally easy to understand. Thus, to further ease their interpretability, we finally employ a method known as linguistic approximation to express fuzzy sets with simple words.

2.3. Improving interpretability: linguistic approximation

Fuzzy logic being a methodology for "computing with words" [12], it can be used to express fuzzy sets with words [12, 13]. This appears as useful since humans understand and manipulate words routinely. Linguistic approximation consists in replacing a set of rules based on fuzzy sets (i.e., mathematical functions) by a set of rules based on linguistic terms, i.e., words associated to fuzzy sets [12]. The fuzzy sets learnt by our FIS are not linguistic terms. Therefore, using linguistic approximation enables us to express our FIS with words.

Based on Yager's framework [13], the first step of linguistic approximation consists in defining a vocabulary V, i.e., a collection of linguistic terms L_k ($k \in [1..N_t]$). Thus, each L_k is a fuzzy set which represents and describes the word W_k. As our FIS uses two-sided Gaussian fuzzy sets (i.e., Gaussians with a plateau and potentially different standard deviations on each side), we use the same kind of fuzzy sets for linguistic terms. We define our vocabulary as a collection of N_t (an odd number) such fuzzy sets regularly

spaced in the [-1:1] interval, all with the same standard deviations. The k^{th} fuzzy set L_k of our vocabulary is defined as having the following left mean μ_{L_k}, right mean μ_{R_k} and standard deviation σ ($\sigma_{L_k} = \sigma_{R_k} = \sigma$):

$$\mu_{L_k} = -1 + \frac{2k}{(N_t - 1)} - \frac{1}{2(N_t - 1)} \tag{1}$$

$$\mu_{R_k} = -1 + \frac{2k}{(N_t - 1)} + \frac{1}{2(N_t - 1)} \tag{2}$$

$$\sigma = \frac{1}{2(N_t - 1)\sqrt{2ln(2)}} \tag{3}$$

Then, each fuzzy set L_k is assigned to a word W_k. For instance, for $N_t = 3$, we use the words "Low", "Medium" and "High" to describe the value of a feature. Figure 1 shows an example of vocabulary with $N_t = 5$ linguistic terms.

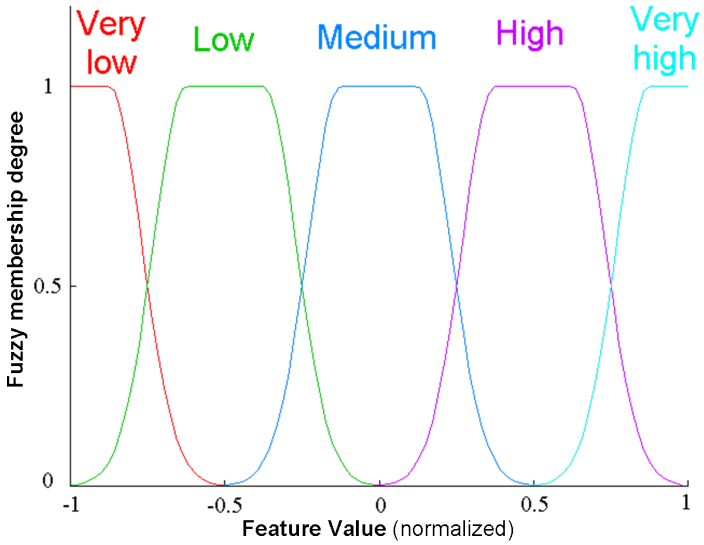

Figure 1. A vocabulary with $N_t = 5$ linguistic terms (Very low, Low, Medium, High, Very high).

The vocabulary defined, linguistic approximation then consists in selecting a linguistic term to replace each fuzzy set used in the FIS. Among the possible selection criteria [13], we used the most intuitive one, namely the "closeness", which reflects how close two fuzzy sets are from each other. We defined closeness with respect to the distance $dist(A, B)$ between fuzzy sets A and B:

$$closeness(A, B) = \frac{1}{1 + dist(A, B)} \tag{4}$$

In our vocabulary, all fuzzy sets have the same standard deviation σ, which enables us to define $dist(A, B)$ independently of σ:

$$dist(A, B) = |\frac{(\mu_{L_A} + \mu_{R_A})}{2} - \frac{(\mu_{L_B} + \mu_{R_B})}{2}| \tag{5}$$

This distance and vocabulary require that the means of the fuzzy sets learnt by the FIS be normalized to [-1:1] for each feature. We can then replace each "X_i is A_{ji}" by "X_i is L_k" where L_k is the linguistic term from the vocabulary for which $closeness(A_{ji}, L_k)$ is the highest. We can finally express each "X_i is A_{ji}" by "X_i is W_k" where W_k is the word described by L_k.

The linguistic approximation completed, what the BCI has learnt automatically can be described by simple and easy-to-understand "if-then" rules. Indeed, these rules report on what power in relevant brain regions and frequency bands corresponds to which mental state, by using words only. Figure 2 summarizes the whole process to design an interpretable BCI. It should be mentioned that this whole approach can deal equally well with binary or multiclass problems, as both FuRIA and FIS are multiclass algorithms [9, 11].

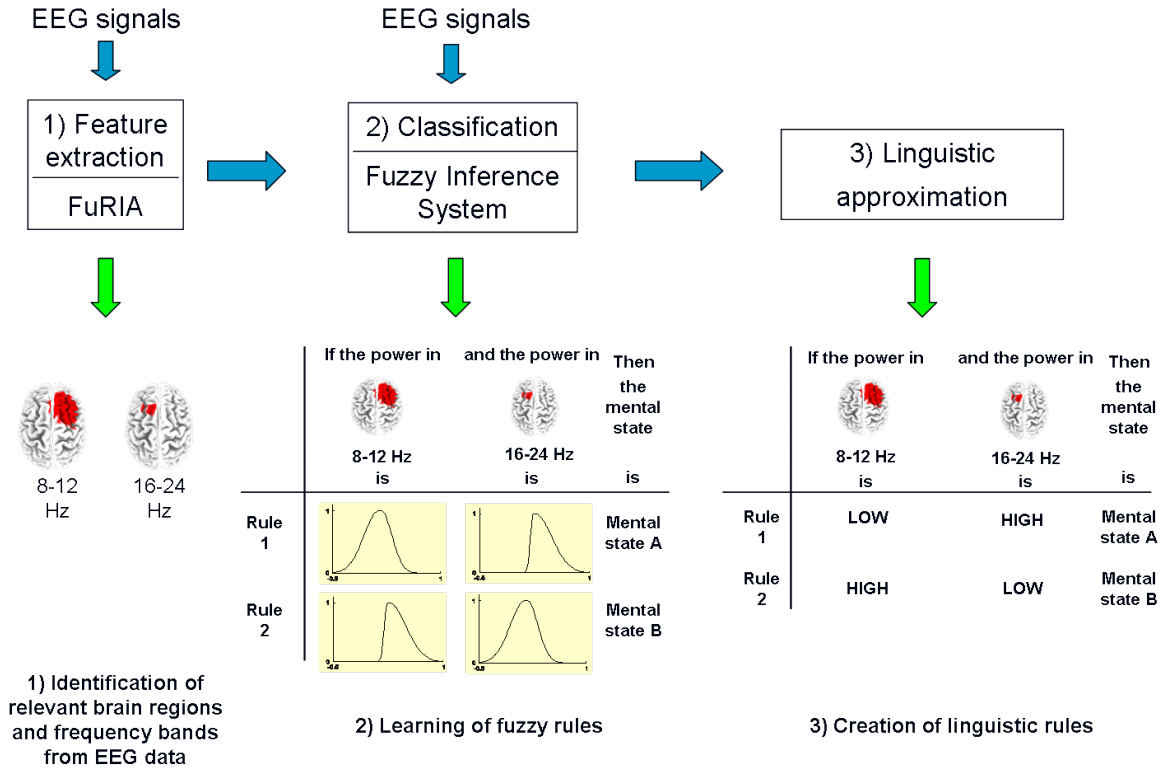

Figure 2. Schematic representation of the proposed algorithm to design an interpretable BCI. An artificial example is provided to ease understanding. In the tables provided, each row corresponds to an "if-then" rule, and each column to a feature. These rules describe the feature values for the mental state they infer.

3. Evaluation

3.1. EEG signals: BCI competition 2003, data set IV

We used data set IV from BCI competition 2003 for evaluation [14]. It contains EEG signals recorded while a subject performed self-paced left and right hand finger tapping.

EEG signals were recorded using 28 electrodes and comprised the 500 ms before each actual movement. The classification task consists in predicting the hand that the subject will use for tapping. The training and testing sets provided contained 314 and 100 EEG trials respectively.

3.2. Results

We trained FuRIA and FIS on the available training set and tested the resulting BCI on the testing set. The rules automatically extracted are shown in Figure 3. In this figure, each row represents a fuzzy "if-then" rule and each column represents a feature. As such, the fuzzy sets or words displayed in the tables describe the value of the power in the brain region (in red color) and frequency band displayed on top, for the mental state inferred by the corresponding rule.

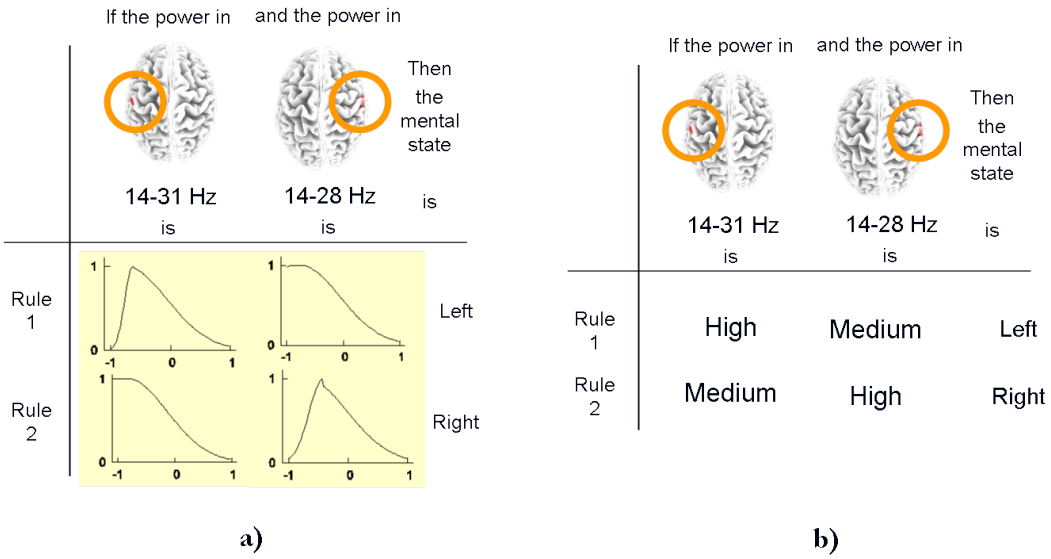

Figure 3. Rules extracted automatically by the BCI: a) raw rules without any linguistic approximation, b) linguistic rules using a vocabulary with $N_t = 3$ terms: Low, Medium and High.

First, it appears that the linguistic approximations are clearly easier to read than the raw rules. More importantly, these linguistic approximations are accessible to people who do not know what a fuzzy set is, e.g., neurophysiologists. Then, Rule 1 suggests that during an intention of left hand movement (mental state "Left"), the Beta band (here 14-28 Hz or 14-31 Hz) power is lower (label "Medium") in the right motor cortex than in the left motor cortex (label "High"). Rule 2 suggests a symmetric behavior for right hand movement intention. This is consistent with the literature, as hand movement intention is known to trigger a power decrease (event related desynchronization), in the motor cortex contralateral to the hand concerned, in the Mu and Beta bands [1]. Finally, our BCI (using the raw rules) reached an accuracy of 85 % on the test set, i.e., a slightly better score than that of the competition winners, who reached a score of 84

% [14]. This suggests our BCI is both efficient and interpretable.

We also evaluated this approach on EEG signals recorded during a visual spatial attention experiment, in collaboration with Dr. Tzelepi, a neurophysiologist from the National Technical University of Athens in Greece, and Dr. Ron-Angevin from Malaga university in Spain (details not reported here due to space limitations). Inspection of the obtained linguistic rules by Dr. Tzelepi revealed that out of 5 rules and 5 features automatically extracted, 4 rules and 4 features were consistent with the literature whereas 1 rule and 1 feature were not. Interestingly enough, removing this rule and feature slightly increased the BCI classification performances. This reveals yet another advantage of interpretable BCI: they can be checked to be potentially improved. Interested readers may refer to the first author's PhD thesis [15] for more details.

4. Conclusion

We have presented an algorithm to design a fully interpretable BCI system. This algorithm relies on the combination of inverse-solutions, fuzzy inference systems and linguistic approximation. This system can explain what power in which brain regions and frequency bands corresponds to which mental state, using "if-then" rules expressed with simple words.

Evaluations of our algorithm suggested that knowledge from the literature was actually reflected by the rules automatically extracted. They also suggested that being able to interpret the BCI may help improve it. Incidentally, this BCI also appeared to have high classification performances. Therefore, the proposed method appears as a useful tool to 1) verify what has been learnt by the BCI and 2) to display and discuss the knowledge automatically extracted by the BCI with non technical people, e.g., medical doctors. It might also prove useful to gain knowledge about the brain dynamics when used to analyze new neurophysiological signals.

Future work could deal with additional evaluations on more subjects and other types of neurophysiological signals. It could also aim at improving the method by integrating additional information such as the time course of brain activity. Finally, it could also be interesting to study new ways of representing the rules, to make the BCI even more intuitively interpretable.

Acknowledgments

Authors would like to thank Dr. Areti Tzelepi, Dr. Brahim Hamadicharef and Ms. Morgane Rosendale for their help related to this work.

References

[1] G. Pfurtscheller and C. Neuper. Motor imagery and direct brain-computer communication. *proceedings of the IEEE*, 89(7):1123–1134, 2001.

[2] M. Kubat, D. Flotzinger, and G. Pfurtscheller. Discovering patterns in EEG-signals: Comparative study of a few methods. In *European Conference on Machine Learning*, pages 366–371, 1993.

[3] D. J. McFarland, C. W. Anderson, K.-R. Müller, A. Schlögl, and D. J. Krusienski. BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):135 – 138, 2006.

[4] C. W. Anderson and Z. Sijercic. Classication of EEG signals from four subjects during five mental tasks. In *International Conference on Engineering Applications of Neural Networks*, 1996.

[5] M. Besserve, K. Jerbi, F. Laurent, S. Baillet, J. Martinerie, and L. Garnero. Classification methods for ongoing EEG and MEG signals. *Biol. Res.*, 40(4):415–437, 2008.

[6] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Proc Magazine*, 25(1):41–56, 2008.

[7] B. Blankertz, G. Curio, and K. R. Müller. Classifying single trial EEG: Towards brain computer interfacing. *Advances in Neural Information Processing Systems*, 14:157–164, 2002.

[8] J. del R. Millán, M. Franzé, J. Mouriño, F. Cincotti, and F. Babiloni. Relevant EEG features for the classification of spontaneous motor-related tasks. *Biological Cybernetics*, 86(2):89–95, 2002.

[9] F. Lotte, A. Lécuyer, and B. Arnaldi. FuRIA: An inverse solution based feature extraction algorithm using fuzzy set theory for brain-computer interfaces. *IEEE transactions on signal processing*, 57(8):3253–3263, 2009.

[10] R.D. Pascual-Marqui. Standardized low resolution brain electromagnetic tomography (sLORETA): technical details. *Methods and Findings in Experimental and Clinical Pharmacology*, 24D:5–12, 2002.

[11] F. Lotte, A. Lécuyer, F. Lamarche, and B. Arnaldi. Studying the use of fuzzy inference systems for motor imagery classification. *IEEE Transactions on Neural System and Rehabilitation Engineering*, 15(2):322–324, 2007.

[12] L.A. Zadeh. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2):103–111, 1996.

[13] R.R. Yager. On the retranslation process in Zadeh's paradigm of computing with words. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(2):1184–1195, 2004.

[14] B. Blankertz et al. The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials. *IEEE Transactions on Biomedical Engineering*, 51(6):1044–1051, 2004.

[15] F. Lotte. *Study of Electroencephalographic Signal Processing and Classification Techniques towards the use of Brain-Computer Interfaces in Virtual Reality Applications*. PhD thesis, INSA de Rennes, France, 2008.