Mixture of Experts for Automated Detection of Phasic Arousals in Sleep Signals

G. Viardot[†], R. Lengellé[‡] and C. Richard[‡]

[†] FORENAP Association
Centre hospitalier de Rouffach, 27 rue du 4ème RSM
68250 Rouffach, France
tel: +33.3.89.78.70.18 fax:+33.3.89.78.51.24
geoffrey.viardot@forenap.asso.fr

[‡] Laboratoire LM2S
Université de Technologie de Troyes, 12 rue Marie Curie, BP 2060
10010 Troyes cedex, France
tel: +33.3.25.71.58.47 fax: +33.3.25.71.56.99
regis.lengelle@utt.fr cedric.richard@utt.fr

Abstract— **Detector design from training data generally relies on the fact that data labels are provided unambiguously by experts. However this optimistic situation is not always realistic, particularly when the phenomena to be detected are complex and poorly understood. The aim of this work is to propose a method for the mixture of experts which authorizes a posteriori analysis of each expert relevance. Our method consists in optimizing the normalized mutual information between the observations and a function of the labels provided by the pool of experts. The method is successfully applied to transient detection in physiological sleep signals.**

Keywords— **Mixture of experts, normalized mutual information, phasic events detection.**

I. Introduction

In the past 50 years, polygraphic sleep technics have allowed human sleep understanding to be greatly improved [9]. Unfortunately, visual scoring of whole night EEG, EOG, and EMG records is still time-consuming and expensive, playing the role of a limiting factor in sleep studies development. This justifies the growing interest for automated sleep data analyzers. In the past two decades, significant advances have been made in this domain, starting with hybrid systems has in [4]. Recently, a large number of algorithms based on statistical pattern recognition techniques [11] and some expert systems have been proposed [8]. Automated detection of waveforms such as alpha, delta, and K-complex waves is an important component of sleep stage monitoring. In particular, phasic sleep arousals are key features that contribute to sleep stages assessment. Unfortunately, their objective description often involves disagreement between experts. [5]

Methods of designing detectors from training data are potentially of great benefit when few *a priori* information is available on the signal to be detected. Amongst the myriad of solutions that have been proposed (see, e.g., [3], [13] and ref. therein), there exist methods for deriving detection structures from training samples with ambiguous labels, e.g., due to disagreements between experts. Some of them consist in combining labels provided by experts and designing the decision device, simultaneously. Others involve the following successive steps: 1) mixture of experts, and 2) derivation of the decision rule. Note that such a strategy allows us to characterize *a posteriori* the behavior of experts during the labelling stage since it suffises to analyze the results of mixture modelling in relation to the data. In this paper a new approach to mixture of experts is presented. This method is based on the optimization of the mutual information between the observations and a combination of the individual expert decisions. The paper is organized as follows. First, the different kinds of uncertainty are described. Second, we present a short taxonomy of fusion operators, regarding their main properties. Then our method is presented and successfully applied to mixture of experts for transient detection in human sleep electrophysiological signals.

II. Fuzzy information

Uncertain information can be partitioned into three main categories. The next paragraph give some description of these categories, according to [7].

A. Kinds of uncertainty

Probabilistic uncertainty is related to randomness of observations. It is often associated to measurement noise or random fluctuation of the observed system. This kind of uncertainty is usually assessed using statistical decision methods.

Resolutional uncertainty corresponds to the limitations which make the observation not exactly perceptible. This kind of uncertainty is, up to some extent, responsible for possible conflict between experts.

Fuzziness is the part of uncertainty which comes from the information coding scheme. As an example, fuzziness is associated with the language used by experts to characterize observations. Examples of fuzziness in an expert judgement are: "possible apparition of ...", "high level of ...", "weak presence of ...". These three kinds of uncertainty are responsible, each of which up to some extent, to the need for mixture of experts.

B. Taxonomy of fusion operators for mixture of experts

In this section, we shall assume that the values $(d_1, \ldots, d_n)$ to combine belong to the following range: $I^N = [0,1]^N$. Different approaches to mixture of experts can be found in the literature. The simplest solution relies on the elaboration of either a "reference" expertise or a consensual prediction [2], [10].

According to [1], the corresponding operator is usually defined as "context independent constant behavior" if it only uses the experts assessments and verifies one of the three (mutually exclusive) following properties, for all $(d_1, \ldots, d_n)$:

1. conjunction:

$$F(d_1, \ldots, d_n) \leq \min(d_1, \ldots, d_n)$$

2. disjunction:

$$F(d_1, \ldots, d_n) \geq \max(d_1, \ldots, d_n)$$

3. compromise:

$$\min(d_1, \ldots, d_n) \geq F(d_1, \ldots, d_n) \geq \max(d_1, \ldots, d_n)$$

If the fusion operator is context independent and its behavior depends on the values $(d_1, \ldots, d_n)$, it will be qualified as "context independent variable behavior". Finally, when not only $(d_1, \ldots, d_n)$ but additional information sources are taken into account in the fusion process, the corresponding operator is called "context dependent". Such an operator, for example, is able to perform gradual fusion. However, when the fusion operator deals with the experts reliability, one must be a priori able to provide an indicator of each expert relevance, or an order relationship for the individual reliability. This kind of information is not easily accessible, while it is contained in the labelled training data set and can be a posteriori extracted. The originality of our paper results is twofold :

1. We propose the elaboration of a context independent constant behavior operator via the maximization of the mutual information between a combination of the experts assessments and observed data;
2. We present a method for extracting experts individual performance after the optimization process.

In the last section, we shall briefly explain how this method can also be used to perform feature space optimization.

III. Principe of the proposed method

When an expert is asked to label a set of events, it is generally assumed that there exists a stochastic relationship between the expert's decision and the event considered. The characteristics used by the expert in order to perform labelling are generally not accessible. However, we shall assume that some stochastic relationship exists between the label given by the expert and a set of features extracted from the events considered. This affirmation remains true when the labels are provided by a pool of experts. In this case, the mixture of experts output is considered as some *unified* label. For a very efficient expert (or expert pool), and assuming that the features extracted are pertinent to the decision making process, the stochastic dependency between the label and the features will be very strong. Picking up the best fusion rule will be achieved by the optimization of some measure of this stochastic dependency. To go further, we must define :

- a fusion rule;
- an efficient measure of the stochastic dependency between 2 random variables.

Let $\Phi = \{\phi(d_1(X), \ldots, d_M(X))\}$ be a family of fusion rules, where $d_i(X)$ denotes the label provided by expert $\#i$ for the observation X. Let $U(X)$ be a set of features extracted from X and describing the event of interest. As indicated previously, selection of the best element $\phi^\star \in \Phi$ must result from the optimization of a measure of the stochastic dependency of the mixture of experts output and the features selected. The measure considered in this paper is the normalized mutual information between ϕ and the feature vector U.

The normalized mutual information is defined as :

$$\widehat{\mathcal{I}}\{\phi, U\} = \frac{\mathcal{I}\{\phi, U\}}{\sqrt{\mathcal{H}\{\phi\}\,\mathcal{H}\{U\}}}, \tag{1}$$

where

$$\mathcal{I}\{A, B\} = \sum_{a \in \Omega_A} \sum_{b \in \Omega_B} p(a,b) \log \frac{p(a,b)}{p(a)\,p(b)} \tag{2}$$

$$\mathcal{H}\{A\} = -\sum_{a \in \Omega_A} p(a) \log p(a). \tag{3}$$

Then we have

$$\phi^\star = \arg\max_{\phi \in \Phi} \widehat{\mathcal{I}}\{\phi(d_1(X), \ldots, d_M(X)), U(X)\}. \tag{4}$$

Mutual information is widely used in information theory. It has the great advantage, compared to correlation measures, of exhibiting nonlinear dependencies. In order to decrease the influence of the features entropies on the dependency measure, some kind of normalization has been used (appearing via the denominator in the definition (1)). For practical reasons, we have focused our attention on the class Φ of linear fusion rules defined as :

$$\phi(d_1(X), \ldots, d_K(X)) = \sum_{i=1}^{K} \theta_i\, d_i(X), \tag{5}$$

where K is the number of experts involved in the study. This fusion rule is a context-independent constant-behavior operator (see, e.g., [Blo96] and ref. therein).

It is "consensual" in the sense that it verifies the third property proposed in section II-B (compromise). However, other fusion rules could be considered as well.

It can be shown that the normalized mutual information is invariant to any nonsingular linear transformation, so the following constraint should be imposed:

$$\sum_{i=1}^{K} \theta_i = 1 \tag{6}$$

Due to the high non-convexity of the mutual information with respect to $\{\theta_i\}$, it was optimized using a genetic algorithm. Obviously, analysis of the optimal fusion parameters θ_i provides an easy way to identify experts which present a particular behavior.

IV. Application to phasic arousals detection in sleep signals

In this paper, we address the problem of automated detection of phasic arousals in sleep signals [6]. The main issue deals with the necessary mixture of experts resulting from the high disagreement observed during the preliminary experiments.

A. Description of the electrophysiological signals

Several electrophysiological signals, i.e., EEG (electroencephalogram: 2 channels), EOG (electrooculogram: 2 channels), EMG (electromyogram) and ECG (electrocardiogram) were used by 4 physiologists to label phasic arousals. An example of these signals are shown in figure 1. As can be seen, there exists disagreement between experts (location of the phasic arousal, confidence degree in the decision). Features $U(X)$ were extracted from time-frequency representations of EEG, EMG and ECG signals, resulting from different bidimensional filtering of these time frequency representations. In the first experiments, we used relative power in frequency bands [8 Hz, 13 Hz[of EEG channel C3-A2 and in the range [10 Hz, 60 Hz] for the EMG signal. Those features have been recommended in [12], [2]. They are associated with a simple bidimensional rectangular window in the time frequency plane. The bidimensional impulse response of these filters can also be optimized as it will be shown in section IV-D.

B. Fuzzy labels coding scheme

Considering the high level of uncertainty in the labelling process, experts were asked to give a confidence degree for each detected phasic sleep arousal according to: *dubious* (dub.), *possible* (pos.) and *certain* (cer.) (see Table I). Using this coding scheme implies a new definition of Φ :

$$\phi\left(d_{1,1}(X), \ldots, d_{K,M}(X)\right) = \sum_{i=1}^{K} \sum_{j=1}^{M} \theta_{i,j}\, d_{i,j}(X), \tag{7}$$

where K denotes the number of experts and M the number of confidence degrees considered. The constraint defined by 6) now becomes:

$$\sum_{i=1}^{K} \sum_{j=1}^{M} \theta_{i,j} = 1. \tag{8}$$

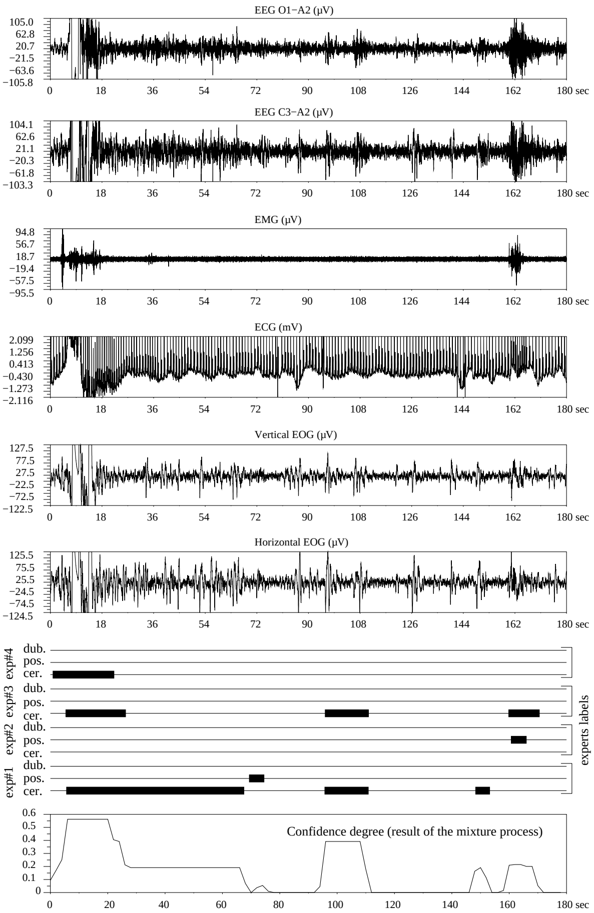

Fig. 1. Phasic sleep arousal: main electrophysiological signals, experts labels and result of the mixture process

TABLE I
Coding scheme for the expert $\#i$

	$\mathbf{d}_{i,1}$	$\mathbf{d}_{i,2}$	$\mathbf{d}_{i,3}$
certain	1	0	0
possible	0	1	0
dubious	0	0	1

Optimization of $\widehat{\mathcal{I}}\{\phi, U\}$ was performed using a genetic algorithm, as explained in section III.

C. Results obtained

Figure 1 shows an example of electrophysiological signals where phasic arousals have been detected by the experts. It also shows the label (confidence degree) given by each expert of the pool as a function of time, and the optimized mixture output. First, it can be observed that the result of the mixture process (lower curve) fits well the physiological data. Secondly, comparing the mixture output with the individual expert labels, we can conclude that the fusion rule presents a *consensual* behavior, as expected.

However, one can notice that the mixture output does not reach the hypothetical maximum value 1. This not only results from the fact that expert #2 does not identify a phasic arousal around time 10 sec. It can be explained by the normalization process and because the label values are mutually exclusive. Thus, we have:

$$\phi\left(d(X)\right) \leq \sum_{i=1}^{K} \max_{j=1,\ldots,M} \theta_{i,j} \tag{9}$$

Figure 2 shows the repartition of the mixture parameters. Results have been averaged after different optimization processes (the vertical intervals represent the optimized parameters standard deviation). As can be seen, the fusion rule parameters θ_i allows us to characterize the experts individual behavior (see the particular parameters values for expert #2). It should be noted that these behaviors have been automatically extracted from the labelled training data.

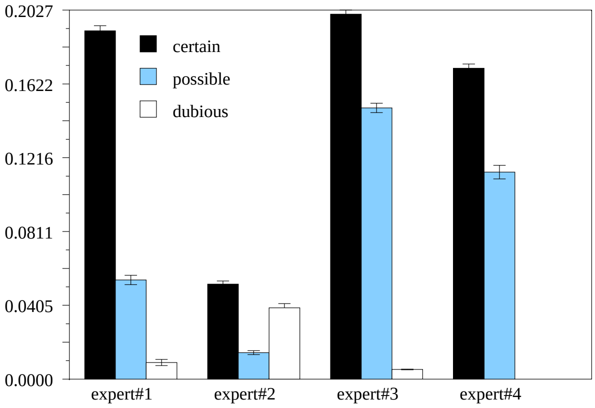

Fig. 2. Average parameters of mixture

D. Feature space optimization

As explained in Section IV-A, the features were extracted from the electrophysiological signals by filtering the time frequency representations of these signals. Feature $\#m$ evaluated at time n can be expressed as :

$$U_m(n) = \sum_{k,l} S_{TF}(n-k,l)H_m(k,l), \tag{10}$$

where S_{TF} denotes the time frequency representation of any signal used, and H_m is the impulse response of the filter used to evaluate the feature m. $H_m(k,l)$ can be considered as new parameters that should be optimized as well as the fusion parameters $\theta_{i,j}$ during optimization. This joint optimization process obviously results in an increase of the dependency measure (see figure 3). The new fusion rule parameters are shown in figure 4 and must be compared to 2. The solution obtained shows that there exist optimized features that are strongly related to the assessment of the first expert, who is surprisingly the reference expert in the pool.

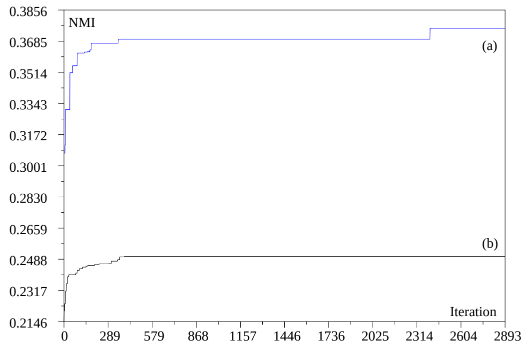

Fig. 3. Evolution of normalized mutual information (NMI) during the optimization process with optimized (a) and fixed (b) filters

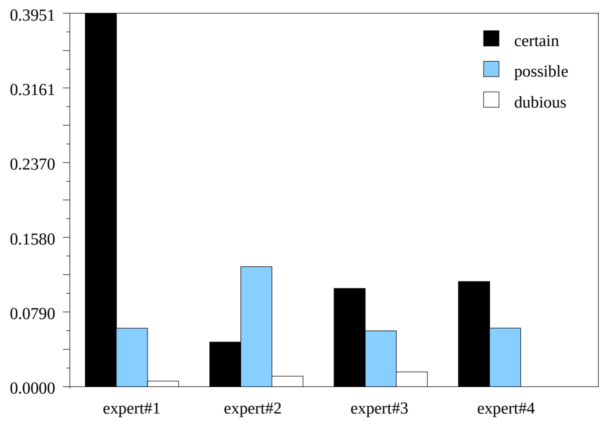

Fig. 4. Mixture parameters obtained after joint optimisation

V. Conclusion

During the constitution of a labelled training set used for designing a decision device, we generally suppose that the expert assessment is unambiguously known. In complex cases, when the training set is labelled by a pool of experts, we can generally observe some disagreement that should be taken into account during the design process. In this paper, we proposed a new approach for mixture of experts, which is based on the optimization of a dependency measure between the mixture output and the feature space. Our method enables a posteriori analysis of each expert relevance, which is automatically extracted from the concept to be learnt. Furthermore, our method takes into account every kind of uncertainty (probabilistic, resolutional, fuzziness), whose individual contribution, however, cannot be isolated. We successfully applied this method to Phasic Arousal detection in human sleep signals. The device described in this paper has great potential since it provides satisfactory performance in experimental conditions. However, we used here the simplest kind of fusion rules, i.e., a linear combination of the $d_{i,j}(X)$'s. Nonlinear and context dependent cases $\theta_{i,j} = \theta_{i,j}(X)$ should be considered in a further work.

Acknowledgments

This work was funded by the French ministry of Research through the DAMES project. (ACI n°99B0450).

References

[1] I. Bloch. Information combination operators for data fusion: a comparative review with classification. *IEEE Trans. on Man. Cyber. - Part I*, vol. 26, n°1, pp. 52–67, 1996.

[2] F. De Carli, L. Nobili, P. Gelcich and F. Ferrillo. A method for the automatic detection of arousals during sleep. *Sleep*, vol. 22, n°5, pp. 561–572, 1999.

[3] L. Devroye, L. Gyorfi and G. Lugosi, *A probabilistic theory of pattern recognition*, New-York: Springer-Verlag, 1996.

[4] J. M. Gaillard and R. Tissot. Principles of automatic analysis of sleep records with hybrid system. *Comput. Biomed. Res.*, vol. 6, pp. 1–13, 1973.

[5] J. Loredo, J. Clausen, S. Ancoli-Israel and J. Dimsdale. Night to night arousal variability and interscorer reliability of arousal measurements. *Sleep*, vol. 22, n°7, pp. 916–920, 1999.

[6] A. Nicolas. Etude des caractéristiques électrophysiologiques des phases d'activation transitoires spontanées au cours du sommeil humain. Ph.D. Thesis, ULP Strasbourg, 1995.

[7] N. R. Pal and J. C. Bezdek. Measuring fuzzy uncertainty. *IEEE Trans. on Fuzzy Systems*, vol. 2, n°2, pp. 107–118, 1994.

[8] S. Ray, W. D. Lee, C. D. Morgan and W. Airth-Kindree. Computer sleep stage scoring. An expert system approach. *Int. J. Bio-Med. Comput.*, vol. 19, pp. 43, 1986.

[9] C. Reynolds, C. F. Kupfer, D. J. Houck, C. C. Stack, S. R. Berman and B. Zimer. Sleep research in affective illness: state of the art. *Sleep*, vol. 10, pp. 189, 1987.

[10] M. Shaw and B. Gaines. Comparing conceptual structures: consensus, conflict, correspondance and contrast. *Knowledge acquisition*, vol. 1, pp. 341–363, 1989.

[11] N. Schaltenbrand, R. Lengellé and J. P. Macher. Neural network model: application to automatic analysis of human sleep. *Comput. Biomed. Res.*, vol. 26, pp. 157–171, 1993.

[12] J. Schieber, A. Muzet and J. Ferriere. Les phases d'activation transitoires spontanées au cours du sommeil normal chez l'homme. *Archive des sciences physiologiques*, vol. 25, n°4, pp. 443–465, 1971.

[13] V. N. Vapnik, *The nature of statistical learning theory*, New-York: Springer, 1999.