A NEW ALGORITHM FOR EEG FEATURE SELECTION USING MUTUAL INFORMATION

Mohamed Deriche and Ahmed Al-Ani

Signal Processing Research Centre
Queensland University of Technology
GPO Box 2434, Brisbane, Q 4001, Australia
email: m.deriche@qut.edu.au and a.alani@qut.edu.au

ABSTRACT

An EEG feature selection technique for the purpose of classification is developed. The technique selects those features that have maximum mutual information with the specified classes of interest (two classes in this case). Obviously, the simplest way is to consider all possible feature subsets (M out of N). However, even with a small number of features, this procedure is *computationally impossible* and can not be used in practice. Given the fact that most features used to represent EEG signal are *sets* of features (such as AR parameters), our technique considers a trade off between computational cost and chosen feature combination. This contrasts other techniques which select features individually. The classification accuracy of features obtained by applying our technique outperforms those obtained by applying individual feature selection methods when applied to EEG signals.

1. INTRODUCTION

Automated EEG analysis has been extensively used in psychotropic drug research, sleep studies, and seizure detection [2]. Most of the automated procedures include a feature extraction stage followed by a classification stage. Several feature extraction methods have been used to represent EEG signals, among others are: spectral analysis based on parametric (AR modelling) [4], non-parametric (FFT), and multi-scale (WVT) [3], where the relative powers are calculated from the various EEG frequency bands (delta, theta, alpha, sigma and beta). In addition to the above, spectral moments, and morphological features [5] such as mean, number of zero-crossing, extrema, duration of monotonies, duration of phases, and amplitude of monotonies, have also been used.

To improve classification performance, a large size of feature vector is desirable, however, this size cannot be increased indefinitely, nor is it practical to run experiments on *combinatorially chosen* subsets from the feature set to pick the best performing subspace with the desired dimensionality. Therefore, to reduce the number of model parameters while keeping a low computational load, it is clearly desirable to understand which of the features provide the greatest contribution to classification performance, and to discard the others.

Recently, Battiti [1] proposed a promising method for feature selection, based on the concept of mutual information (MI), which he named MIFS. MI measures strength in dependencies between random variables. It is suitable for assessing the "information content", where methods based on linear relations are prone to mistakes.

In this paper we expand the MIFS by considering combination sets of features, by doing so, we can select those features that work well *together* instead of just looking at these *individually* as the MIFS algorithm does. To reduce the number of subsets searched, our algorithm only considers those subsets that are more likely to have maximum MI. This modification is shown to improve classification accuracy.

The following section gives some background about the concept of MI. The MIFS algorithm is described in section three. Section four explains our proposed algorithm. Experimental results on EEG data are presented in section five, Section six gives a conclusion for the paper.

2. BACKGROUND

The MI between random variables X and Y, $I(X;Y)$, is defined as:

$$I(X;Y) = \int P_{XY}(x,y) log[P_{XY}(x,y)/P_X(x)P_Y(y)]dx\,dy \quad (1)$$

The main problem with experimental data is estimating P_{XY} from the histogram. One approach, which we use here, is to divide the XY plane into boxes of size $\Delta x \Delta y$. By doing so, we are able to estimate the $discrete$ value of P_{XY}.

Under the above assumptions, the MI can be re-written

as:

$$I(x;y) = \sum_{r_x}\sum_{r_y} P_{XY}(r_x, r_y) \; log[P_{XY}(r_x, r_y)/P_X(r_x)P_Y(r_y)] \quad (2)$$

where, r_x and r_y are the discrete levels of X and Y respectively. If $r_x = r_y = R$, then we will need R^2 boxes to estimate P_{xy}. In the case of three variables, we will need R^3 to estimate P_{xyz}, and so on. It is clear that this number becomes very large as the number of variables increases.

3. THE MIFS ALGORITHM

Instead of calculating MI between a feature vector $\mathbf{f}$ and output classes $\mathbf{C}$, the MIFS algorithm only computes $I(\mathbf{C}; f)$ and $I(f; \dot{f})$, where f and $\dot{f}$ are individual features. The algorithm chooses one feature at a time; the one maximising the information with output classes. This MI expression is corrected by subtracting a quantity proportional to the average MI with the selected features. The MIFS algorithm is formalised as follows:

1. Set $F \leftarrow$ "initial set of N features"; $S \leftarrow \{\emptyset\}$.
2. For each feature $f \in F$, compute $I(\mathbf{C}; f)$.
3. Find feature f that maximises $I(\mathbf{C}; f)$; set $F \leftarrow F \setminus \{f\}$; set $S \leftarrow \{f\}$.
4. repeat until $|S| = M$ (M chosen a priori),
   (a) For all couples of variables $(f, \dot{f})$ with $f \in F$, $\dot{f} \in S$, compute $I(f; \dot{f})$.
   (b) Choose feature f that maximises $I(\mathbf{C}; f) - \beta / |S| \sum_{\dot{f} \in S} I(f; \dot{f})$; set $F \leftarrow F \setminus \{f\}$; set $S \leftarrow S \cup \{f\}$.

Parameter β regulates the relative importance of MI between a candidate feature and the already-selected ones with respect to MI with output classes. According to the work in [1], β is chosen between 0.5 and 1. The number of calculated I's = $\binom{N}{2} + N$.

It can be seen that the MIFS algorithm only considers those features that have maximum MI with the output classes, and are loosely correlated. Of course, this does not guarantee that these features *work well together*, where the ultimate objective of performing feature selection is to choose the best possible subset of features and *not to rank features individually* according to their importance. To solve this problem, we propose the modified MI feature selection (MMIFS) algorithm explained below.

4. THE MODIFIED MUTUAL INFORMATION FEATURE SELECTION (MMIFS) ALGORITHM

As explained above, the exact MI between all possible subsets and output classes is *computationally impossible*. On the other hand, considering features individually (as the MIFS algorithm does) is not an appropriate solution, as it considers only individual features rather than sets of features.

The MMIFS algorithm, on the other hand, finds first the best four features, by considering MI between output classes and subsets of four features (not all subsets are considered). To select the fifth feature, four values of MI are calculated between outputs and subsets of four features (three selected features and new one). The feature that gives the maximum sum of these four MI's is selected. From the original four selected features, the three features that work best with the new selected one will be used with the new one in selecting the sixth feature. This procedure is repeated untill we reach the desired number of features, M. The MMIFS algorithm reduces the number of possible subsets drastically. Below is the selection procedure:

1. Set $F \leftarrow$ "initial set of N features"; $S \leftarrow \{\emptyset\}$.
2. Compute $I(\mathbf{C}; f), f \in F$. Set $F_1 \leftarrow$ "the $M/2$ features that maximise I". (No. of calculated I's $= N$).
3. For each $\dot{f} = F_1(j), j = 1 : M/2$, Compute $I(\mathbf{C}; f, \dot{f})$, $f \in F, f \neq \dot{f}$. Set $F_2 \leftarrow$ "the $M/2$ subsets (of 2 features) that maximise I". (No. of calculated I's $= M/2(N-1)$).
4. For each subset $\{\dot{f}, \ddot{f}\} = F_2(j), j = 1 : M/2$ compute $I(\mathbf{C}; f, \dot{f}, \ddot{f})$, $f \in F, f \notin F_2(j)$. Set $F_3 \leftarrow$ "the $M/2$ subsets (of 3 features) that maximise I". (No. of calculated I's $= M/2(N-2)$).
5. For each subset $\{\dot{f}, \ddot{f}, \dddot{f}\} = F_3(j), j = 1 : M/2$ compute $I(\mathbf{C}; f, \dot{f}, \ddot{f}, \dddot{f})$, $f \in F, f \notin F_3(j)$. Using the maximum value of I, set $F_4 = S \leftarrow \{\dot{f}, \ddot{f}, \dddot{f}, \ddddot{f}\}$. (No. of calculated I's $= M/2(N-3)$).
6. Repeat until $|S| = M$

   For each $f \in F, f \notin S$, compute $I(\mathbf{C}; f, \dot{f}, \ddot{f}, \dddot{f}) + I(\mathbf{C}; f, \dot{f}, \ddot{f}, \ddddot{f}) + I(\mathbf{C}; f, \dot{f}, \dddot{f}, \ddddot{f}) + I(\mathbf{C}; f, \ddot{f}, \dddot{f}, \ddddot{f})$. (No. of calculated I's $= 4(N - |S|)$). Substitute f that gives the maximum value with one of the F_4 elements that has less I compared to the other three; set $S \leftarrow S \cup \{f\}$.

The purpose of the first five steps is to consider only the most likely subsets of four features that have maximum MI with the output, where we only consider $N + M/2 \sum_{k=1}^{3}(N-k)$ instead of $\binom{N}{4}$ combinations. In the last

step, adding the four values of MI for each three of the four features in F_4 with each of the non-selecting features and choosing the one that gives the maximum value, guarantees that it has, on average, the maximum MI with the previously selected four features. As the number of selected features increase, it becomes computationaly very costly repeating the same procedure, where $\binom{|S|}{3}$ MI values need to be calculated each time we select a new feature. Therefore, we only consider here the newly selected feature and the three features in F_4 that work best with it in measuring the MI values (the fourth feature will be removed from F_4). By doing so, the number of calculated I's in this step will be $4 \times (N - |S|)$. This makes the total number of calculated I's $= N + M/2 \sum_{k=1}^{3}(N-k) + 4\sum_{k=4}^{M-1}(N-k)$.

For $N = 20$ and $M = 15$, the MMIFS algorithm will compute 909 MI values, compared to 210 for the MIFS algorithm and 15504 considering all possible subsets! Also, we have to mention that the MIFS algorithm uses R^2 boxes to calculate P, compared to R^4 for the MMIFS and R^M if all subsets are considered (refer to section 2). If $R = 10$, these numbers are 100, 10000 and 10^{15} respectively. It is clear that the MIFS algorithm requires the least number of computations, followed by MMIFS, and considering all subsets is obviously impossible.

The MMIFS calculates MI using 4-combined features as this is seen a good compromise between high computational cost and finding subsets of features that *work well together*. As explained above, the computational cost is reasonable but will increase rapidly if we use more than 4 features. On the other hand, we could not find any noticeable difference in performance when using 5 features instead of 4.

5. FEATURE EXTRACTION AND SELECTION FROM EEG DATA

Some of the paralysed patients are aware of their environment, but are not able to communicate, for example by speaking words or moving their eyes to signal "yes" or "no". The only way to answer questions is to use signals from the brain in order to develop a kind of response code. A simple binary response could be used, for example to select a letter or a word on a computer monitor. This means that the brain signals have to be modified by the patients through special "thoughts". Further, the brain signals have to be analysed and classified in real-time and the classification results used to control cursor movement on a monitor. Such a system, which transforms signals from the brain into control signals, is known as a "brain-computer interface" (BCI) [4]. Different brain signals can be used as input to a BCI: evoked potential, slow cortical potential shifts, or electroencephalogram (EEG).

The EEG data used here was obtained from Graz University of Technology, Austria. Three subjects were instructed not to move and to keep their arms and hands relaxed. Based on a visual stimulus presented on a computer monitor, each subject was asked to imagine a movement of the right or left hand [4].

The features used to represent the EEG data were:

- dominant frequency and its amplitude (which are critical in the characterisation of rhythmic discharge).
- average power in main lobe (used to reflect the concentration of energy in a spectrum, average half-waves amplitude and duration.
- energy, zero crossing and number of extrema of each segment.
- 14 AR parameters and 5 poles.
- Energy of 8 wavelet subbands (characterising the different EEG bands).
- Fractal dimension.

In this experiment, we applied both the MIFS and MMIFS algorithms to select M features ($M = 4 : 20$). Tables 1 and 2 contain the selected features using both methods.

Table 1: Selected features using the MIFS algorithm

No.	Selected features
5	$AR_{par}(5), 1AR_{pol}, F_{Dim}, F_{dom}, N_{Ext}$
10	$AR_{par}(5,6,8), Wvt_{en}(4), 2AR_{pol}, F_{Dim}$, $F_{dom}, AF_{dom}, N_{Ext}$
15	$AR_{par}(4,5,6,8,11), Wvt_{en}(2,4,7), 3AR_{pol}$, $F_{Dim}, F_{dom}, AF_{dom}, N_{Ext}$
20	$AR_{par}(4,5,6,7,8,11,12,14), Wvt_{en}(1,2,3,4,7)$, $3AR_{pol}, F_{Dim}, F_{dom}, AF_{dom}, N_{Ext}$

Table 2: Selected features using the MMIFS algorithm

No.	Selected features
5	$AR_{par}(5,13), Wvt_{en}(2,3,6)$
10	$AR_{par}(5,6,8,9,12,13), Wvt_{en}(2,3,6), 1AR_{pol}$
15	$AR_{par}(3,5,6,8,9,10,12,13,14)$, $Wvt_{en}(2,3,4,5,6), 1AR_{pol}$
20	$AR_{par}(3,4,5,6,8,9,10,11,12,13,14)$, $Wvt_{en}(1,2,3,4,5,6), 2AR_{pol}, F_{Dim}$

It is clear from these two tables that the MMIFS tends to select combined features, which is reflected by the selection of most AR parameters and wavelet bands energies, while the MIFS algorithm looks at features individually! This explains the reason behind selecting the dominant frequency, amplitude at the dominant frequency and number of extrema.

The selected features were then fed to an artificial neural network (ANN) to perform classification, where the training set contained 45 channels of three subjects while the

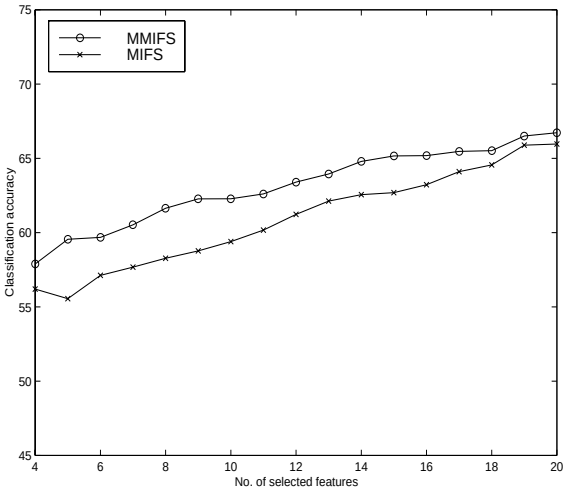

Fig. 1. Classification accuracy from training set using MMIFS and MIFS

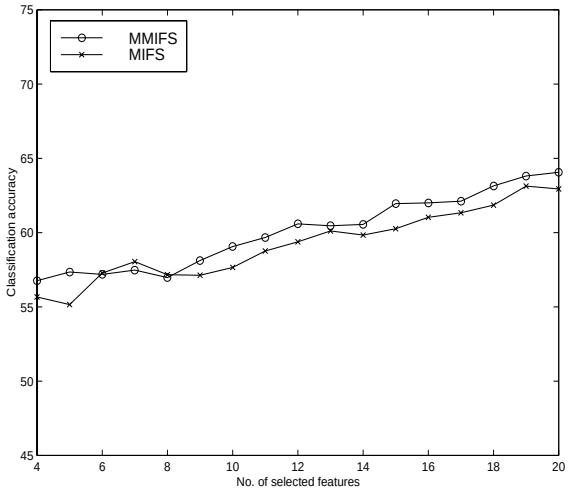

Fig. 2. Classification accuracy from test set using MMIFS and MIFS

test set contained 11 channels. Figures 1 and 2 show the classification accuracy of the training and testing sets for both MMIFS and MIFS algorithms. The superiority of the MMIFS is very clear in the training set. Due to the limited amount of the test set, we got some fluctuations in the results, but the better performance of the MMIFS algorithm is still clear. Note that the improvement in performance obtained here comes with minor additional computational load. It is worth noting that our focus, here, is on selecting appropriate features rather than finding the best performing classification algorithm, and as such, good classification accuracy has not been fully investigated yet. In this paper, we wanted to show the potential of Mutual Information Concepts in the optimal selection of "sets" of features. This concept as presented here is new and has a tremendous potential in enhancing the power of more advanced classification techniques. Based on the very promising results we obtained here, we are planning to investigate the possibility of developing a hybrid scheme, which optimises both feature selection and classification stages. Applications other than EEG signals have been considered as well, such as speech signals and texture images.

6. CONCLUSION

An EEG feature selection algorithm based on maximising MI has been developed. The algorithm takes into consideration how features work together by calculating MI between output classes and subsets of four features. It has been found that choosing four features to measure MI is a reasonable choice representing the best compromise between computational cost and how combined features work together. When tested on real EEG data, the algorithm outperforms the original MIFS algorithm. With the promising results we obtained here, we expect the proposed algorithm to form a cornerstone in the area of *feature set selection* when used in conjunction with advanced classification techniques for on-line monitoring of EEG.

7. REFERENCES

[1] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994.

[2] I. Gath and L. Schwartz. Syntactic pattern recognition applied to sleep EEG staging. *Pattern Recognition Letters*, 10:265–272, 1989.

[3] A.B. Geva and D.H. Kerem. Forecasting generalized epilileptic seizures from the EEG signal by wavelet analysis and dynamic unsupervised fuzzy clustering. *IEEE Transaction on Biomedical Engineering*, 45:1205–1216, 1998.

[4] G. Pfurtscheller, C. Neuper, A.Schlogl, and K. Lugger. Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. *IEEE Transaction on Rehabilitation Engineering*, 6:316–325, 1998.

[5] F. Wendling, J.J. Bellanger, J.M. Badier, and J.L. Coatrieux. Extraction of spatio-temporal signatures from depth EEG seizure signals based on objective matching in warped vectorial observations. *IEEE Transaction on Biomedical Engineering*, 43:990–1000, 1996.