Proceedings of the 4th International
IEEE EMBS Conference on Neural Engineering
Antalya, Turkey, April 29 - May 2, 2009

Feature Down-Selection in Brain-Computer Interfaces

Dimensionality Reduction and Discrimination Power

N.S. Dias, L.R. Jacinto, P.M. Mendes, J.H. Correia
Dept. of Industrial Electronics
University of Minho
Guimaraes, Portugal
ndias@dei.uminho.pt

Abstract—**Current non-invasive Brain-Computer Interface (BCI) designs use as much electroencephalogram (EEG) features as possible rather than few well known motor-reactive features (e.g. rolandic μ-rhythm picked from C3 and C4 channels). Additionally, motor-reactive rhythms do not provide BCI control for every subject. Thus, a subject-specific feature set needs to be determined from a large feature space. Classifier over-fitting is likely for high-dimensional datasets. Therefore, this study introduces an algorithm for feature down-selection on a subject basis based on the across-group variance (AGV). AGV is evaluated in comparison with three other algorithms: recursive feature elimination (RFE); simple genetic algorithm (GA); and RELIEF algorithm. High-dimensional data from 5 healthy subjects were first reduced by the algorithms under experiment and then classified on the alternative right hand or foot movement imagery tasks. AGV outperformed the other tested methods simultaneously selecting the smallest feature subsets. Effective dimensionality reduction (as low as 8 features out of 118) with high discrimination power (as high as 90.4) was best observed on AGV's performance.**

Keywords-feature selection; neural signal processing; brain-computer interface

I. INTRODUCTION

Brain-Computer Interfaces (BCI) enable movement independence for the physically disabled by translating their thoughts into device commands. Electroencephalogram (EEG), as control signal, is usually preferred to invasive recordings due to its ease of acquisition. However, EEG patterns produced in response to movement imagery performance are subject-dependent and the translation algorithm needs to be trained on a subject basis. An effective implementation of BCI requires a previous calibration session (no feedback is provided to the subject) whose data is employed to train the translation algorithm. The set of features (e.g. event-related desynchronizations, spectral band power, movement-related potentials) extracted from the EEG channels may be larger than the subset that optimally translates movement imagery performance for each subject. Therefore, the feature set dimensionality should be reduced by determining the subject-specific feature subset to include in the classification model. Two main methodologies have been adopted in BCI research. The transformation of original feature spaces into lower dimensional spaces has been often tested [1]. An alternative methodology is feature down-selection which produces a subset of original features that is most relevant to discriminate subject performance. The greatest advantage of the latter methodology is the effective reduction of BCI computational complexity. The methods proposed in previous studies to down-select feature sets are commonly categorized as wrapper or filter methods based on dependence on a learning technique. Wrapper methods use the predictive accuracy of a pre-selected classifier to evaluate a feature subset. Among the state-of-the-art exemplars, the recursive feature elimination (RFE) [2] and genetic algorithms (GA) [3] are popular in BCI research. Filter methods separate feature selection from classifier training and produce feature subsets independent of the selected classifier. The RELIEF algorithm is often used as a filter method [4].

The current work introduces a filter algorithm based on a formulation of principal component analysis that accommodates the group structure of the dataset. This algorithm uses the concept of across-group variance (AGV) to reduce dataset dimensionality. The proposed algorithm, as well as RFE, GA and RELIEF, was tested on EEG data collected during the imagery of right hand and foot movements performed by five subjects. Both dimensionality reduction ability and discrimination power were assessed for comparison.

II. DATA

The data set IVa from the BCI competition III [5] was recorded from 5 healthy subjects and used for algorithm performance comparison. These data were recorded during 4 calibration sessions. The subject was instructed to perform right hand and foot movement imagery for 3.5 s periods. Data were recorded from 118 EEG channels at positions of the extended international 10/20-system. Although signals were digitized at 1000 Hz with 16 bit (0.1 μV) accuracy, a 100 Hz version of the data (by picking each 10th sample) was used for further analysis.

The EEG signals were filtered differently for each subject. The band-pass filter ranges 8-30 Hz, 8-14 Hz or 15-30Hz were used depending on the best group membership prediction achieved for each subject. The signal epoch was defined from the cue presentation instant (i.e. 0 s) to the end of the imagery period (i.e. 3.5 s after cue presentation). The epoch data was assessed in 1 s long windows with 0.5 s overlap. In each time window, the sum of the squared filtered signals was

calculated. The feature matrices had 280 samples available with 118 features.

III. METHODS

The original feature matrix $X_{n\times p}$ has samples (n) in rows and features (p) in columns. The risk of classifier over-fitting to the training data is larger for high-dimensional datasets. Additionally, just a few features (p_{opt}) are generally relevant for discrimination and the optimal feature subset is subject-dependent. Thus, a feature down-selection algorithm is required in order to promote robust and effective discrimination by reducing data dimensionality.

A recently developed algorithm, as in [6], is compared with three other algorithms in common use: RELIEF [4]; recursive feature elimination (RFE) [2]; and genetic algorithm (GA) [3]. A linear discriminant classifier was employed to predict group membership for all algorithms but RFE. Instead, a standard support vector machine (SVM) was used. The feature down-selection algorithms were tested in a 10-fold cross-validation scheme since the average of the folds' prediction accuracy is indicative of the classifier's online performance. The 10-fold cross-validation scheme was run 10 times in order to compensate for performance variability (100 classification error values were calculated). Besides the 10-fold validation loop, the cross-validation also comprises an inner 10-fold loop that partitions the training dataset into new training and validation subsets. The inner loop optimized algorithm parameters such as the number of features to select (p_{opt}).

A. Across-Group Variance (AGV) Algorithm

The principal components (PCs) are linear projections of the features onto the orthogonal directions that best describe the dataset variance. The component orthogonal matrix $U_{n\times c}$ (c is the number of PCs) is calculated through singular value decomposition of X. Although the PCs are already organized by decreasing order of the total variance accounted for, this order is optimized for orthogonality rather than discrimination between groups. Additionally, in the presence of group structure, the variance information provided by a component comprises two parcels as in (1): a function of the sample distances to their respective group mean; and a function of the distances between the respective group means.

$$\lambda_i = v_i^T \Psi_{within} v_i + v_i^T \Psi_{between} v_i \tag{1}$$

Ψ_{within} is the pooled covariance matrix, $\Psi_{between}$ represents the between-group covariance matrix and is calculated through the total covariance (Ψ) decomposition in (2). λ_i is the eigenvalue correspondent to the i^{th} eigenvector v_i.

$$\Psi = \Psi_{within} + \Psi_{between} \tag{2}$$

In a discrimination context, only the second parcel in (1) comprises useful variance information. Therefore, the distance between groups given by the i^{th} component, normalized by its total variance, provides a relative measure to calculate the across-group variance (AGV) as in (3).

$$AGV_i = v_i^T \Psi_{between} v_i \big/ \lambda_i \tag{3}$$

In order to take the data group structure into account, the principal components were ordered according to the AGV score in (3), instead of the eigenvalues λ that account for the total variance.

The dimensionality reduction results from the truncation of the c principal components previously ranked as in (3). The truncation criterion is a cumulative sum percentage of the descending ordered AGV scores and was defined to take one of the following values: 60%, 70%, 80% or 90%. These threshold values are commonly used for component truncation. The principal components k that met the truncation criterion compose a truncated version of the component matrix ($U_{n\times k}$ with $k < c$) which is a lower dimensional representation of the original feature space, more suitable for group discrimination.

In order to determine the features which resemble the retained components with minimal information loss, a modified version of the spectral decomposition property is used to calculate an across-group covariance matrix (Ψ_{AGV}) as in (4).

$$\Psi_{AGV} = \sum_{i=1}^{k} AGV_i v_i v_i^T \tag{4}$$

Note that AGV_i is used instead of λ_i on the spectral decomposition equation. Each diagonal value of Ψ_{AGV} represents the variance of a particular feature accounted for the k retained principal components and measures feature discrimination ability. A list with the p features in descending order of discrimination ability is determined. Finally, the top listed p_{opt} features comprise the optimal subset.

B. RELIEF

The RELIEF algorithm is a filter method that assigns a relevance value to each feature producing a ranking that permits the selection of the top ranked features according to a previously chosen threshold or criterion [4]. The relevance value, or feature weight (W), is iteratively estimated according to how well a feature distinguishes among instances that are near each other. In each iteration, a sample x is randomly selected and the weight of each feature is updated from the difference between the selected sample and two neighbouring samples: one from the same group $H(x)$ (named nearest hit) and another from a different group $M(x)$ (named nearest miss). The weight of each feature p is updated as in (5).

$$W_p = W_p - \left| x_p - H(x)_p \right| + \left| x_p - M(x)_p \right| \tag{5}$$

The weights are calculated along n (number of available training samples) sequential iterations. Iteratively, the feature with the lowest weight was removed and the classification accuracy of the resulting subset evaluated by a linear discriminant classifier. The selection stops when p_{opt} features are left.

C. *Recursive Feature Elimination (RFE)*

The recursive feature elimination (RFE) algorithm based on a support vector machine classifier is a wrapper method that uses the feature weights of the SVM training process to perform backward feature elimination [2]. A linear kernel machine was used with parameters set to Matlab® Bioinformatic Toolbox defaults. RFE ranking criterion $\| W_p^2 \|$ for feature p is calculated from (6) which depends on the weighted sum of support vectors that define the separation between groups as optimized by the SVM for every sample n.

$$W = \sum_n \alpha_n y_n x_n \tag{6}$$

α_n is the sample weight, x_n is the p-dimensional training sample and y_n is the group label. The samples with non-zero weights are the support vectors. The features with the lowest ranking, thus contributing less to group separation, are removed iteratively. This procedure stops when the optimum subset size (p_{opt}) is reached.

D. *Genetic Algorithm (GA)*

This is a wrapper method that uses a simple genetic algorithm to search the space of possible feature subsets. The genetic algorithm is a global search method based on the mechanics of natural selection and population genetics and has been successfully applied to BCI problems [3]. It starts with the generation of an initial random population, where each individual (or chromosome) encodes a candidate solution to the feature subset selection problem. The individual is constituted by various genes represented by a binary vector of dimension equal to the total number of features. A fitness measure is evaluated for each individual after which, selection and genetic operators (recombination and mutation) are applied. In this study, the classification accuracy of a linear discriminant classifier was the fitness measure. Starting with conventional values, the parameter calibration was based on empirical tests executed beforehand and were set to the following: the population size was 30, the number of generations was 50; the selection rate was 0.5; elite children (chromosomes that pass unchanged, without mutation, to the next generation) was 2; the mutation rate was set to 0.05 and the crossover probability to 0.5. The selection of chromosomes to be recombined was done by tournament selection (with tournament size equal to 2). Crossover and mutation were uniform. The most frequently selected features within the inner loop up to the number of features p_{opt} were tested on validation data.

IV. Results

A newly developed algorithm and three other popular algorithms in BCI were tested for feature down-selection in high dimensional datasets that are publicly available [5]. The performance measures used for comparison were the cross-validation error for 100 folds, its standard deviation and the average number of features selected. According to table I, the proposed algorithm (AGV) achieved the best average performance among the tested algorithms. AGV achieved the lowest average error and standard deviation for the smallest subsets. RFE, GA and RELIEF algorithms ranked next in error increasing order. For subject AL, RFE achieved lower classification error than AGV. However, the former selected more features than the latter. Since the AGV ranking algorithm is a filter method, it was alternatively tested with a SVM classifier (i.e. the same employed in RFE) for validation. The error average was 6.96 % thus, lower than RFE's and still maintaining a small number of features selected. The statistical confidence of the classification results was assessed by a paired t-test with a confidence level of 95%. A significant difference between the methods was found (p-values << 0.05).

The classification error vs. number of features average curves for subject AL are presented in Figure 1 and further illustrates the comparison results (see Table I). Although AGV and RFE algorithms achieve comparable minimum classification errors, the former selects subsets considerably smaller than the latter. RELIEF obtained the highest classification errors. GA did not produce a sequential selection curve due to its intrinsic search design. GA produces generation dependent feature subsets rather than nested ones.

AGV was also tested on the training/validation data splits provided for the BCI competition (see [5]) in order to evaluate algorithm ability to deal with small training sets. The amount of data for classifier training is in descending order on table II. As for the results on Table I, the best classification results were achieved for subjects AL and AY. Note that values on Table I result from datasets with 252 training samples. The number of features selected decrease with the available training data.

V. Discussion and Conclusions

The feature subsets calculated by AGV achieved highest prediction accuracy (either with linear discriminant or support vector machine predictors), at the 95% level of confidence, with the smallest number of features (see Table I). These results were further illustrated for subject AL (Figure 1) by the cross-validation error curves. The minimum classification error achieved by backward elimination of task irrelevant features (or forward addition of task relevant features) establishes the optimum number of features to be selected. All the remaining features are deemed task relevant. Thus, among the tested algorithms, AGV's error curve seems most suitable for feature down-selection since it achieved the best accuracy with the fewest features. RFE ranked second best in our algorithm comparison.

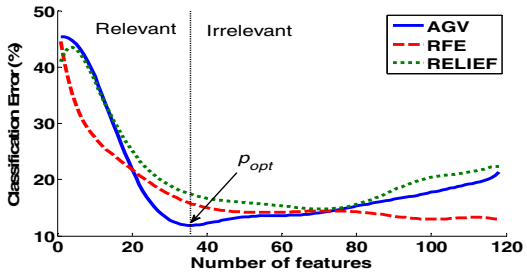

Figure 1. Mean error vs. number of features average curve, for subject AL, calculated by the across-group variance with a linear discriminant classifier (AGV), recursive feature elimination (RFE) and RELIEF algorithms. The marker p_{opt} defines the optimum subset size calculated by AGV and separates relevant from irrelevant features.

TABLE I. RESULTS COMPARISON OF THE TESTED FEATURE DOWN-SELECTION ALGORITHMS

Subject	AGV			RFE			GA			RELIEF		
	Error[a] (%)	SD[b]	p_{opt}[c]	Error (%)	SD	p_{opt}	Error (%)	SD	p_{opt}	Error (%)	SD	p_{opt}
AA	**25.1**	7.2	21	27.4	9.5	103	31.6	9.7	58	34.3	9.9	55
AL	9.6	6.0	37	**7.6**	5.3	76	13.6	5.8	58	16.0	6.6	53
AV	**28.5**	7.4	9	33.2	7.8	78	37.6	10.0	58	36.4	9.2	62
AW	**19.9**	7.1	20	29.8	7.4	93	28.7	8.6	59	30.2	8.3	70
AY	**9.2**	4.8	8	14.3	6.4	101	17.2	6.9	58	21.8	7.8	89
Average	**18.5**	6.5	19	22.5	7.3	90	25.7	8.2	58	27.7	8.3	66

[a]Error is the mean classification error; [b]SD is the error standard deviation; p_{opt} is the mean number of features selected

The lowest classification error for each subject was printed in **bold**

TABLE II. CLASSIFICATION RESULTS FOR AGV FEATURE SUBSETS CALCULATED FROM DATASETS WITH DIFFERENT PROPORTIONS OF TRAINING DATA

Subject	Training[a]	Validation[a]	Error (%)	p_{opt}
AL	224	56	10.7	24
AA	168	112	30.4	14
AV	84	196	29.1	7
AW	56	224	31.7	9
AY	28	252	16.3	2
Average	-	-	23.6	11

[a]column values represent the number of samples employed either for Training or Validation procedures

Although RFE's classification errors were comparable to AGV's for two subjects (AA and AL) the subset sizes were largely different. On average, RFE selected 90 features out of 118, while AGV only kept 19. As suggested in [7], SVM tends to calculate similar weights (W) for highly correlated (e.g. redundant) features. Thus, these features are eliminated or kept simultaneously during the RFE selection. As depicted on Figure 1, RFE finds the minimum error early on the backward elimination. The following error raise is slow which indicates that the features being eliminated are not highly relevant. On the contrary, the slope of the AGV error increase is much higher. SVM achieves low classification errors even for large feature sets (see Figure 1) and seems to perform well with many redundant features. However, large datasets increase the risk of classifier over-fitting and decreases its generalization ability. Additionally, as claimed in [8], the latter elimination of redundant features might promote premature elimination of more relevant ones and mislead the subset optimization. On the other hand, AGV ranks each feature based on its covariance with the truncated component space rather than its covariance with other features. Therefore, linear correlations between features are implicitly considered but not determinant for feature selection. Although, both RFE and AGV were able to order features by relevance, the latter seems more capable of dealing with redundant features. Moreover, as expected for filter methods, AGV ran 8 times faster than RFE on average.

The genetic algorithm ranked third best on comparison. Surprisingly, the GA tested never achieved classification results comparable with AGV's. Moreover, on all tested subjects it doesn't seem to accomplish an effective feature reduction and the generalization error is high, thus suggesting that a premature convergence phenomenon is occurring.

As expected, RELIEF achieved the poorest classification accuracy. Unlike AGV, RELIEF evaluates feature relevance independently of other features and thus is incapable of dealing with redundant features. Another drawback is that RELIEF is highly susceptible to select irrelevant channels in the presence of outliers on noisy channels.

The competition datasets (Table II) led to an error increase as a consequence of less training data available. However, AGV appears able to reduce the subset size when less task information is available thus avoiding over-fitting.

In this study, the across-group variance algorithm outperformed other popular methods in feature down-selection for BCI. AGV seems a valuable solution to decrease prosthesis computational complexity for the physically disabled.

ACKNOWLEDGMENT

N. S. Dias is supported by the Portuguese Foundation for Science and Technology under Grant SFRH/BD/21529/2005 and Center Algoritmi. L. R. Jacinto is supported by the Portuguese Foundation for Science and Technology under Grant SFRH/BD/40459/2007 and Center Algoritmi.

REFERENCES

[1] G. Blanchard and B. Blankertz, "BCI competition 2003–data set IIa: spatial patterns of self-controlled brain rhythm modulations" IEEE Trans. Biomed. Eng. vol. 51 pp. 1062–6, 2004.

[2] L. T. Schröder, T. Weston, J. Bogdan, M. Birbaumer and N. B. Schölkopf, "Support vector channel selection in BCI", IEEE Trans. Biomed. Eng., Vol. 51(6), pp. 1003-1010, 2004.

[3] M. Schröder, M. Bogdan, W. Rosenstiel, T. Hinterberger and N. Birbaumer, "Automated EEG feature selection for brain computer interfaces", Proc. 1st IEEE EMBS Neural Eng., 2003, pp. 626 – 629.

[4] J. Millán, M. Franzé, J. Mouriño, F. Cincotti and F. Babiloni, "Relevant EEG features for the classification of spontaneous motor-related tasks", Biol. Cybern., vol. 86, pp. 89-95, 2002.

[5] B. Blankertz *et al*, "The BCI competition III: validating approaches to actual BCI problems", IEEE Trans. Neural Sys. Rehab. Eng., vol. 14(2), pp. 153-159, 2006.

[6] N.S. Dias, P.M. Mendes and J.H. Correia, Feature Selection for Brain-Computer Interface, IFMBE Proceedings vol.22, 23-27 November 2008, Antwerp, Belgium, pp. 318–321

[7] Z. Xie, Q. Hu, D. Yu, "Improved feature selection algorithm based on SVM and correlation", Lecture Notes Comput. Sci., vol. 3971, 2006, pp. 1373-1380.

[8] M. Yousef, S. Jung, L. Show and M. Showe, "Recursive cluster elimination (RCE) for classification and feature selection from gene expression data", BMC Bioinformatics, vol. 8c, 2007, 144.