Feature Selection by Independent Component Analysis and Mutual Information Maximization in EEG Signal Classification

Tian Lan, Deniz Erdogmus, Andre Adami, Michael Pavel
Department of Biomedical Engineering
OGI School of Science and Engineering
Oregon Health & Science University
Beaverton, Oregon 97006, USA
E-mail: {lantian,deniz,adami,pavel}@bme.ogi.edu

***Abstract* — Feature selection and dimensionality reduction are important steps in pattern recognition. In this paper, we propose a scheme for feature selection using linear independent component analysis and mutual information maximization method. The method is theoretically motivated by the fact that the classification error rate is related to the mutual information between the feature vectors and the class labels. The feasibility of the principle is illustrated on a synthetic dataset and its performance is demonstrated using EEG signal classification. Experimental results show that this method works well for feature selection.**

***Keywords* — Feature Selection, Independent Component Analysis, Mutual Information, Entropy Estimation, EEG, Brain-Computer Interface**

I. INTRODUCTION

Feature selection and dimensionality reduction are important steps in pattern recognition tasks and many other applications. In practice, the relevant information about the data structure can often be represented by a lower dimensional manifold embedded in the original Euclidian data space. Specifically, in pattern recognition, a high dimensional feature vector is available, but usually the classification task can be achieved equally well by a feature vector of reduced dimensionality. Furthermore, reducing the number of features will also help the classifier learn a more robust solution and achieve a better generalization performance. This is due to the fact that irrelevant feature components are eliminated by the optimal subspace projection.

Dimensionality reduction by subspace projection is typically achieved by feature transformation methods. This transformation generates either a new feature space, or a sub-set of the original feature space, which can be treated as a special case of the former situation. The transformation can be linear or non-linear. Linear transformations have been widely used due to their simplicity. While nonlinear transformations attract increasingly more attention due to their ability to capture the nonlinear relationships within the data, the complexity of finding robust regularized nonlinear transformations makes them a second choice in most of applications. In this paper, we will focus on linear transformations leaving the nonlinear transformations for future study.

There are many existing linear transformation methods for dimensionality reduction. Principle component analysis (PCA) is a widely used dimensionality reduction technique [1,2]. However, since the projections it finds are not necessarily related to the class labels, it is not particularly useful in pattern recognition. Linear discriminant analysis (LDA) attempts to eliminate this shortcoming of PCA by finding linear projections that maximize class separability under the Gaussian distribution assumption [3]. The LDA projections are optimized based on the means and the covariance matrices of classes, which are not descriptive of an arbitrary probability density function (pdf). Independent component analysis (ICA) has also been used as a tool to find linear transformations that maximize the statistical independence of random variables [4,5]. However, like PCA, the projection that ICA finds also has no necessary relationship with class labels, and it is not able to enhance class separability [6].

Optimal feature selection coupled with a specific classifier topology, namely the wrapper approach, results in a combinatorial computational requirement; thus, is unsuitable for adaptive learning of feature projections. On the contrary, the filter approach, which selects features by optimizing some criterion is independent of the classifier, hence is more flexible.

In the filter approach, it is important to optimize a criterion that is relevant to Bayes risk, which is typically measured by the probability of error. A suitable criterion is mutual information (MI) between the projected features and the class labels, which is motivated by lower and upper

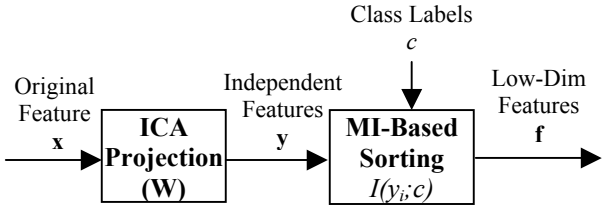

Fig. 1. Feature selection using ICA preprocessing and mutual information sorting

bounds in information theory that relate this quantity to probability of error [7,8]. In principle, MI measures non-linear dependencies between a set of random variables taking into account higher order statistical structures existing in the data, as opposed to linear and second-order statistical measures such as correlation and covariance.

Several MI based methods have been developed for feature selection [9-13]. Estimating MI requires the knowledge of joint pdf of the data in feature space. Evaluating the MI between two scalar random variables (one being the discrete class labels) using histograms is studied in literature [9,13]. However, this approach fails when dealing with high dimensional variables. Torkkola [6] proposed an approach using a quadratic divergence measure to find an optimal transformation that maximizes the MI between features and class labels. This approach, being dependent on Parzen density estimation, is inefficient for subspace projections to high dimensionalities due to the joint density estimation requirement.

A shortcoming of existing MI-based feature selection methods is that, since features are generally mutually dependent, feature selection in this manner is typically suboptimal in the sense of maximum joint mutual information principle. In practice, the mutual information must be estimated nonparametrically from the training samples [14]. Although this is a challenging problem for multiple continuous-valued random variables, the class labels are discrete-valued in the feature transformation setting. This reduces the problem to just estimating entropies of continuous random vectors. Furthermore, if the components of the random vector are independent, the joint entropy becomes the sum of marginal entropies. Thus, the joint mutual information of a feature vector with the class labels is equal to the sum of marginal mutual information of each individual feature with the class labels, provided that the features are independent. In this paper, we exploit this fact by combining independent component analysis (ICA) preprocessing with a sample-spacing based entropy estimator [15] for feature selection (see Fig. 1).

The contributions of this paper are: (i) theoretical motivation of feature selection using ICA preprocessing and marginal mutual information sorting, (ii) a computationally efficient training algorithm for this paper that employs a fast analytical solution for ICA and simple and consistent sample-spacing estimators for mutual information, (iii) the application of this technique to the classification of EEG signals for cognitive load assessment.

II. Theoretical Background

The goal of feature subspace projections is to improve classifier robustness by reducing data dimensionality in order to facilitate better generalization, as well as reducing the learning and operating complexity of the classifiers. While doing so, classification performance must not be compromised by throwing away components that provide useful information regarding the class labels. Theoretically, optimal feature projections should minimize the Bayes risk function for the given problem; the average probability of error is a widely used and accepted risk function and merits special attention. For different risk functions, the following theoretical and practical results can be easily modified.

The average probability of error has been shown to be related to MI between the feature vectors and the class labels. Specifically, Fano's and Hellman & Raviv's bounds demonstrate that probability of error is bounded from below and above by quantities that depend on the Shannon MI between these variables [7,8]. Maximizing this MI reduces both bounds, therefore, forces the probability of error to decrease. A similar result was also obtained by Erdogmus & Principe using Renyi's MI; a parametric family of lower and upper bounds for the probability of error was provided [16,17]. Hellman & Raviv [7] showed that the upper bound on Bayes error is given by $(H_S(C)-I_S(\mathbf{Y},C))/2$, where $H_S(C)$ is the Shannon entropy of the a priori probabilities of the classes and $I_S(\mathbf{Y},C)$ is the Shannon MI between the continuous-valued feature vector and the discrete-valued class label. Consequently, maximizing the MI between the projected features and the class labels potentially improves classification performance, and has drawn much attention [6,9-12].

Mutual information was first introduced by Shannon in the context of digital communications between discrete random variables and was generalized to continuous random variables. In feature extraction, we are interested in the MI between the continuous-valued feature vector $\mathbf{y}$ and the discrete-valued class labels c. Shannon MI between $\mathbf{y}$ and c is defined in terms of the entropies of the overall data and the individual classes as

$$I_S(\mathbf{y};c) = H_S(\mathbf{y}) - \sum_c p_c H_S(\mathbf{y}\mid c) \tag{1}$$

where p_c are the prior class probabilities. The entropy is given by

$$\begin{aligned} H_S(\mathbf{y}) &= -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y} \\ H_S(\mathbf{y}\mid c) &= -\int p(\mathbf{y}\mid c) \log p(\mathbf{y}\mid c) d\mathbf{y} \end{aligned} \tag{2}$$

where $p(\mathbf{y}|c)$ are the class conditional distributions and the overall data distribution is

$$p(\mathbf{y}) = \sum_c p_c\, p(\mathbf{y}\mid c) \tag{3}$$

Assume that features y are mutually independent, we have:

$$I_S(\mathbf{y};c) = \sum_{i=1}^{n} I_S(y_i;c) \tag{4}$$

where $I_S(y_i;c) = H_S(y_i) - \sum_c p_c H_S(y_i \mid c)$, and y_i is the i^{th} component of feature space.

There exist a number of entropy estimators for one-dimensional variables. Here, we will use the sample-spacing estimator for its simplicity.

The independence assumption can be acquired by ICA transformation. After that, the MI between each feature and the class labels, $I_S(y_1,c),\ldots,I_S(y_n,c)$ can be estimated by the projected data samples. We rank $I_S(y_i,c)$ according to the value, and choose the m features with largest MI that account for the majority of the total MI between the feature vector and the class label.

In principle, any ICA algorithm followed by any MI estimator could be employed during the feature selection procedure described above. In the next section, we discuss the specific ICA transformation and MI estimator that are employed in our experiments.

III. ICA Transformation and MI Estimation

ICA Using Generalized Eigenvalue Decomposition: The square linear ICA problem is expressed in (5), where **X** is the $n \times N$ observation matrix, **A** is the $n \times n$ mixing matrix, and **S** is the $n \times N$ independent source matrix.

$$\mathbf{X} = \mathbf{AS} \tag{5}$$

Each column of **X** and **S** represents one sample of data. If we consider each column as a sample in time, (5) becomes:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{6}$$

Many effective and efficient algorithms based on a variety of assumptions including maximization of nonGaussianity, minimization of mutual information, nonstationarity of the sources, etc. exist to solve this ICA problem [14,15,18]. All these could be compactly formulated in the form of a generalized eigendecomposition problem that gives the ICA solution in an analytical form [19]. Therefore, this formulation reviewed by Parra & Sajda in [19] will be employed in this paper.

According to this formulation, one possible assumption set that leads to an ICA solution utilizes the higher order statistics (specifically fourth-order cumulants). Under this set of assumptions, the separation matrix **W** is the solution to the following generalized eigendecomposition problem:

$$\mathbf{R_x W} = \mathbf{Q_x W \Lambda} \tag{7}$$

where $\mathbf{R_x}$ is the covariance matrix and $\mathbf{Q_x}$ is the cumulant matrix estimated using sample averages: $\mathbf{Q_x}$=E[$\mathbf{x^T x x x^T}$]-$\mathbf{R_x}$tr($\mathbf{R_x}$)-E[$\mathbf{x x^T}$]E[$\mathbf{x x^T}$]-$\mathbf{R_x R_x}$. Given the estimates for these matrices, the ICA solution can be easily determined using efficient generalized eigendecomposition algorithms (or using the *eig* command in Matlab).

Estimating MI Using Sample-Spacings: Recall that in the case of feature selection for classification, the mutual information estimation reduces to the sum of marginal and conditional entropies as shown in (1) and (4). Therefore, we only need to estimate marginal entropies. There exist many entropy estimators in the literature for single-dimensional variables. Here, we use an estimator based on sample-spacings, which stems from order statistics. This estimator is selected because of its consistency, rapid asymptotic convergence, and simplicity.

Consider a one dimensional random variable Y. Given a set of iid samples of Y $\{y_1,\ldots,y_N\}$, first these samples are sorted in increasing order such that $y_{(1)} \le \ldots \le y_{(N)}$. The m-spacing entropy estimator is given by:

$$\hat{H}(Y) = \frac{1}{N-m} \sum_{i=1}^{N-m} \log \frac{(N+1)(y_{(i+m)} - y_{(i)})}{m} \tag{8}$$

This estimator uses two assumptions: the true density $p(y)$ is approximated by a piecewise uniform density determined by m-neighbor distances and outside of the sample range, the true density has the same mean log probability density as the rest of the distribution.

The selection of the parameter m is determined by a bias-variance trade-off and typically $m = \sqrt{N}$. In general, for asymptotic consistency the sequence $m(N)$ should satisfy

$$\lim_{N\to\infty} m(N) = \infty \qquad \lim_{N\to\infty} m(N)/N = 0 \tag{9}$$

IV. Experiments and Results

Synthetic Dataset: In order to illustrate the feasibility and the performance of the proposed feature selection method, we apply it to a simple synthetic dataset. This problem consists of classifying two one-dimensional Lapalacian classes where a second confusing irrelevant Gaussian feature is introduced. Specifically, The 2-dimensional feature vector **x** is a random linear combination (determined by a matrix **A**) of the 2 independent features s_1 and s_2, where s_1 obeys the distribution given in (10) determining the class labels completely and s_2 is redundant zero-mean unit-variance Gaussian noise independent from the class label.

$$s_1 \sim p f_1(s_1) + (1-p) f_2(s_1) \tag{10}$$

The class distributions are

$$f_c(s_1) = \frac{1}{\sqrt{2}\sigma} \exp(-\sqrt{2}|s_1 \pm 1|/\sigma) \tag{11}$$

A Monte Carlo experiment is performed with p=0.5, σ varying from 0.1 to 2, and the number of training samples selected as 10^2, 10^3, and 10^4. For each combination the following process is repeated 100 times: a random mixing matrix **A** is selected (each entry uniform in [0,1]), a new training set is generated with the specified number of samples, and a new testing set of 10^6 samples is generated. The ICA solution and MI-based feature selection are performed using the training data, and a simple threshold classifier (also determined from the training data) is employed on the test data. For reference, the true optimal Bayes classifier (simple threshold of zero on s_1) is also applied to the test data in every case.

The results averaged over the 100 Monte Carlo runs are shown in Fig. 2. As expected, the performance approaches the theoretical optimal as the training set size increases. In order to evaluate the performance of selected ICA transformation and the feature selection method, we introduce a parameter: cosα, which is defined as:

$$\cos\alpha = \boldsymbol{\alpha}^T \mathbf{e} / \|\alpha\| \tag{12}$$

where $\boldsymbol{\alpha}^T = \mathbf{e}^T\mathbf{W}\mathbf{A}$ is the actual selection matrix, and $\mathbf{e}^T$ is the ideal selection matrix with value $[1,0]^T$ or $[0,1]^T$. Ideally, we expect the value of cosα to be as close as 1 when number of training samples increases. Fig. 3 shows the cosα value averaged over 100 Monte Carlo runs. As we expected, the value of cosα keeps unchanged for different σ. However, unexpectedly, it does not increase as the size of the training set increases.

Cognitive State Classification Using EEG Signals: In this example, the proposed method for feature selection is applied to the classification of cognitive state using features extracted from EEG signals collected while the subject performs a mental task. The data is collected as part of an augmented cognition project, in which the estimated cognitive state is used to assess the mental load of the subject in order to modify the interaction of the subject with a computer system with the goal of increasing user performance. In this experimental setup, the EEG signals measured at 256Hz by seven electrodes located at salient sites (CZ, P3, P4, PZ O2, P04, F7) are used to generate power-spectral features (1-second sliding window integrated over 5 frequency bands: 4–8Hz, 8–12Hz, 12–16Hz, 16–30Hz, 30–44Hz). The novelty in this application is that the subjects are freely moving around in contrast to the typical brain-computer interface (BCI) experimental setups where the subjects are in a strictly controlled setting. The assessment of cognitive state in ambulatory subjects is particularly difficult, since the movements introduce strong artifacts irrelevant to the mental task/load. The mobility of the operator increases the complexity of the design, because the measurement of the physiological states is extremely difficult in situations where the body of the subject is in motion. Feature selection becomes important in this task due to its abilities to keep the useful information and eliminate the irrelevant information for classification, in order to increase the robustness of the classification performance of the system.

During data collection, the subject is outfitted with the suite of sensors and performs a predetermined set of tasks: *slow walking, navigating and counting, communicating with radio,* and *studying mission map*. The EEG data is collected for training and testing. The whole classification system contains four parts: preprocessing, feature extraction and selection, classification, and postprocessing. Preprocessing is used to filter out noise and remove the artifacts. Feature extraction and selection generates features from the clean EEG signal, and selects useful features using the proposed method. For classification, the K-Nearest-Neighbor (KNN) classifier is utilized. The postprocessing uses the assumption

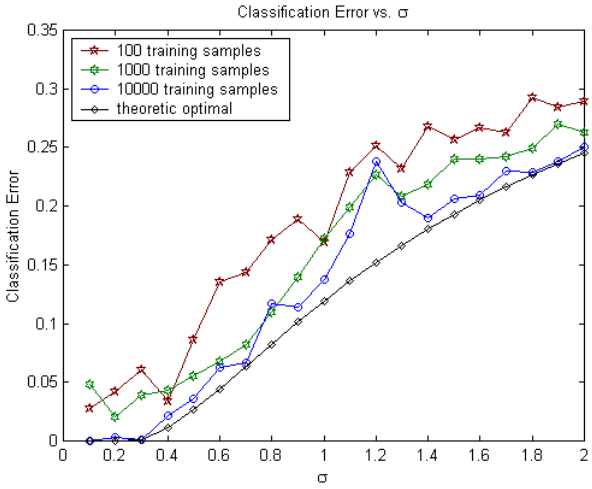

Fig. 2. Classification errors vs. σ for different sizes of training sets compared with the theoretically optimal classifier.

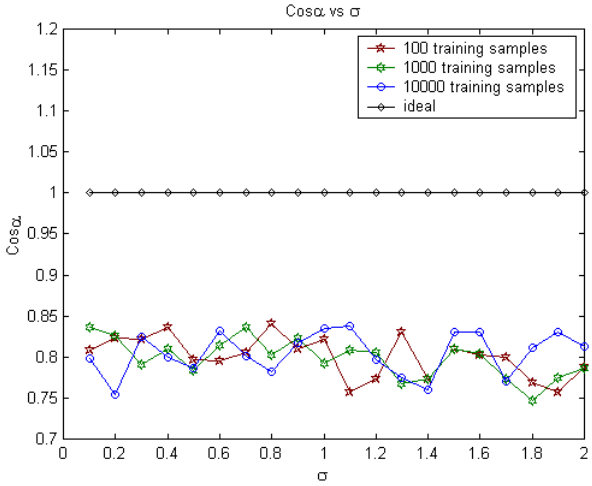

Fig. 3. Cosα vs. σ for different sizes of training sets compared with the ideal value of 1.

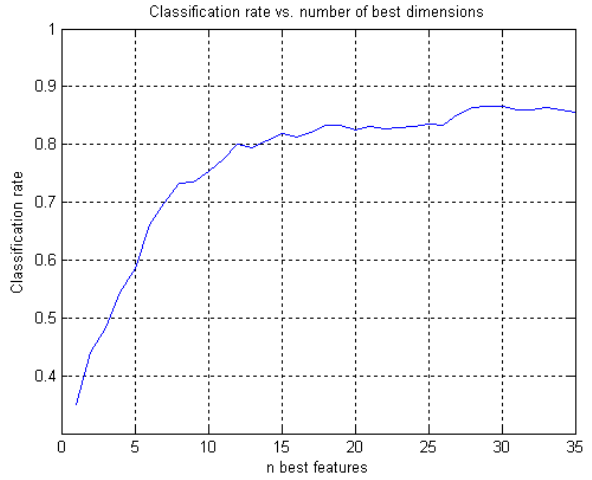

Fig. 4. Correct classification rate (vertical axis) vs. dimensionality of optimally selected features (horizontal axis).

that the variations in cognitive state for a given continuous task will be slowly varying in time. A median filter operating on a window of 2-second decisions recently generated by the classifier is used to eliminate a portion of erroneous decisions made by the classification system.

Using the first 1/3 of the collected data for training and the remaining 2/3 for testing, the correct classification rate of the system on the test data over four classes is shown in Fig. 4 for different feature subspace projection dimensions. An accuracy of 80% is achieved with 12 dimensions, while the remaining 23 dimensions do not significantly contribute to the classification accuracy. These results demonstrate that the proposed method of feature dimensionality reduction is able to capture the low-dimensional relevant components of the feature vector.

V. Conclusions

In this paper, we presented a feature selection method based on the maximum mutual information principle. The technique combines the analytical solution for linear ICA transformations and a computationally efficient mutual information estimator, by taking into account the fact that minimization of Bayes classification error can be approximately achieved by maximizing the mutual information between the features and the class labels. The linear ICA transformation is used to separate the mixed features into approximately independent features so that single-dimensional mutual information estimation can be conveniently employed. The ICA transformation is determined by solving a generalized eigendecomposition problem, which is also computationally efficient and effective. The current method relies on linear ICA, which does not necessarily yield independent features, which violates the assumption of additive decomposition of mutual information that we have employed. Future work will expand this technique using nonlinear ICA in order to improve performance. Another alternative research direction is to use linear independent subspace analysis combined with efficient joint entropy estimators.

Experiments using synthetic and real (EEG) data demonstrate the validity and the effectiveness of the proposed technique. The results on the EEG data set reveal the fact that this method is able to determine relevant low-dimensional structures in data in the broad context of brain computer interfaces. The method exhibits the following appealing properties:

- Intuitively motivated by information theory.
- Easy to implement with low computational requirements.
- Robust and accurate in classifier design due to the selection of salient features.

Acknowledgement

This work was supported by DARPA under contract DAAD-16-03-C-0054. The EEG data was collected at the Human-Centered Systems Lab., Honeywell, Minneapolis, Minnesota.

References

[1] E. Oja, *Subspace Methods of Pattern Recognition*, Wiley, New York, 1983.

[2] P.A. Devijver, J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice Hall, London, 1982.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, 1990.

[4] R. Everson, S. Roberts, "Independent Component Analysis: A Flexible Nonlinearity and Decorrelating Manifold Approach," Neural Computation, vol. 11, no. 8, pp. 1957-1983, 2003.

[5] A. Hyvärinen, E. Oja, P. Hoyer, J. Hurri, "Image Feature Extraction by Sparse Coding and Icomponent Analysis," Proceedings of ICPR'98, pp. 1268-1273, 1998.

[6] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization," Journal of Machine Learning Research, vol. 3, pp. 1415-1438, 2003.

[7] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*. Wiley, New York, 1961.

[8] M.E. Hellman, J. Raviv, "Probability of Error, Equivocation and the Chernoff Bound," IEEE Transactions on Information Theory, vol. 16, pp. 368-372, 1970.

[9] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Networks learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, 1994.

[10] A. Ai-ani, M. Deriche, "An Optimal Feature Selection Technique Using the Concept of Mutual Information," Proceedings of ISSPA, pp. 477-480, 2001.

[11] N. Kwak, C-H. Choi, "Input Feature Selection for Classification Problems," IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 143-159, 2002.

[12] H.H. Yang, J. Moody, "Feature Selection Based on Joint Mutual Information," in Advances in Intelligent Data Analysis and Computational Intelligent Methods and Application, 1999.

[13] H.H. Yang, J. Moody, "Data Visualization and Feature Selection: New Algorithms for Nongaussian Data," Advances in NIPS, pp. 687-693, 2000.

[14] K.E. Hild II, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information," IEEE Signal Processing Letters, vol. 8, no. 6, pp. 174-176, 2001.

[15] E.G. Learned-Miller, J.W. Fisher III, "ICA Using Spacings Estimates of Entropy,".Journal of Machine Learning Research, vol. 4, pp. 1271-1295, 2003.

[16] D. Erdogmus, *Information Theoretic Learning: Renyi's Entropy and its applications to Adaptive System Training*, PhD Dissertation, University of Florida, 2002.

[17] D. Erdogmus, J.C. Principe, "Lower and Upper Bounds for Misclassification Probability Based on Renyi's

Information," Journal of VLSI Signal Processing Systems, vol. 37, no. 2/3, pp. 305-317, 2004.
[18] A. Hyvärinen, E. Oja, "A Fast Fixed Point Algorithm for Independent Component Analysis," Neural Computation, vol. 9, no. 7, pp. 1483-1492, 1997.
[19] L. Parra, P. Sajda, "Blind Source Separation via Generalized Eigenvalue Decomposition," Journal of Machine Learning Research, vol. 4, pp. 1261-1269, 2003.