Estimating Mutual Information Using Gaussian Mixture Model for Feature Ranking and Selection

Tian Lan, Deniz Erdogmus, Umut Ozertem, Yonghong Huang

Abstract—**Feature selection is a critical step for pattern recognition and many other applications. Typically, feature selection strategies can be categorized into wrapper and filter approaches. Filter approach has attracted much attention because of its flexibility and computational efficiency. Previously, we have developed an ICA-MI framework for feature selection, in which the Mutual Information (MI) between features and class labels was used as the criterion. However, since this method depends on the linearity assumption, it is not applicable for an arbitrary distribution. In this paper, exploiting the fact that Gaussian Mixture Model (GMM) is generally a suitable tool for estimating probability densities, we propose GMM-MI method for feature ranking and selection. We will discuss the details of GMM-MI algorithm and demonstrate the experimental results. We will also compare the GMM-MI method with the ICA-MI method in terms of performance and computational efficiency.**

I. Introduction

FEATURE selection and dimensionality reduction is an important problem for pattern recognition and many other applications. For example, in communication, transmitting low dimensional data that contains most of the information is more desirable than directly sending the original high dimensional counterpart due to bandwidth limitations. In pattern recognition context, feature selection and dimensionality reduction can address the salient features, and eliminate the irrelevant features; hence, increase the robustness and improve the generalization performance of the classification system. Specifically, in geographical and biomedical signal processing, the dimension of feature space can be hundreds or thousands, and it is impractical to analyze these data directly without a dimensionality reduction that improves generalization.

Dimensionality reduction can be achieved by subspace projection or feature selection. In subspace projection, the original features are projected linearly or non-linearly to a low dimensional space, which represents major statistics of the data. There are many existing subspace projection methods, such as PCA, ICA and LDA [1-5]. However, the projections that PCA and ICA seek are not necessarily related to the classification performance, hence are not necessarily useful in pattern recognition. LDA overcomes this shortcoming by finding the projections that maximize class separability. On the other hand, this method relies on a Gaussian distribution assumption; so it is not applicable for an arbitrary distribution. Lan et al. developed a subspace projection framework, which applies linear ICA transformation and mutual information maximization for dimensionality reduction in EEG signal classification [6]. This method exhibits several advantages, such as it is computationally efficient and flexible, and it is also suitable for high dimensional data; however, the linearity assumption essentially resulting from ICA limits its applications.

Although subspace projection can effectively remove the redundant features, the relationship between the projected features and the original features becomes vague. In some applications, such as multi-sensor array target detection, and EEG signal processing, the system can only transmit and process signals from a certain number of sensors in real-time, due to the limitation of bandwidth and computation capacity. In these particular cases, feature selection is more suitable, which selects a subset from the original feature space. It is widely accepted that some classification algorithms, such as decision tree, multi-layer perceptron neural networks have inherent ability to focus on relevant features and ignore irrelevant ones [7]. In general, feature selection is achieved by a feature ranking procedure. Feature selection methods can be divided into wrapper and filter approaches. Wrapper approach uses classification accuracy as criterion coupled with a specific classifier, which requires re-training the classifier for different combinations of feature sets; hence, it is slow and inflexible. Filter approach, on the other hand, ranks and selects features by optimizing some criteria independent of the classifier, and is more flexible and suitable for adaptive learning.

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 the MI between the selected features and the class labels, motivated by lower and upper bounds in information theory that relate this quantity to probability of error [8,9]. As opposed to linear and second-order statistics such as correlation and covariance, MI

Tian Lan is with the Biomedical Engineering Department, OGI School of Science and Engineering, Oregon Health and Science University, Beaverton, OR 97006 USA, (phone: 503-748-1564; e-mail: lantian@bme.ogi.edu).

Deniz Erdogmus is with the Computer Science and Electrical Engineering Department and Biomedical Engineering Department, OGI School of Science and Engineering, Oregon Health and Science University, Beaverton, OR 97006 USA, (e-mail: derdogmus@ieee.org).

Umut Ozertem is with the Computer Science and Electrical Engineering Department, OGI School of Science and Engineering. Oregon Health and Science University, Beaverton, OR 97006 USA. (e-mail: ozertemu@csee.ogi.edu).

Yonghong Huang is with the Computer Science and Electrical Engineering Department, OGI School of Science and Engineering. Oregon Health and Science University, Beaverton, OR 97006 USA. (e-mail: huang@csee.ogi.edu).

measures non-linear dependencies between a set of random variables taking into account higher order statistical structures existing in the data.

Many feature selection methods have been developed in the past years [10-12]. Guyon & Elisseeff also reviewed several approaches used in machine learning context [13]. In our previous work, we have proposed an ICA-MI method for feature selection, and applied it on EEG channel selection [14]. This method exploits the fact that an invertible linear transformation does not change the MI, and assumes that linear ICA transformation yields independent features, so that the MI between feature vectors and class labels can be conveniently estimated by the direct summation of MI between each independent projected feature vector and class labels. However, since the accuracy of this method highly depends on the performance of linear transformation, it is not applicable for arbitrary distribution. Actually, if we know the distribution of the feature vectors, we can directly estimate the MI between feature vectors and class labels by definition. There are several density estimation methods, such as histogram, GMM, and Kernel density estimation (KDE). GMM is widely used because (1) it is a more powerful tool as compared to parametric estimators that only can estimate a family of density functions; (2) it results in a continuously differentiable estimation, which is appropriate for gradient based adaptive learning approaches; (3) it is less computationally intensive as compared with KDE. This motives us to use GMM estimating MI for feature selection.

In this paper, we propose a GMM-MI method for feature ranking and selection. In the next section, we will discuss the algorithm in detail. In the experimental result part, we apply this method on several datasets of UCI machine learning repository, and EEG dataset collected by Honeywell for the AugCog project. We also compare this method with the previous ICA-MI method in terms of accuracy and speed.

II. GMM-MI ALGORITHM

The goal of feature selection is to improve the generalization performance of the classification system by selecting the informative features, without compromising classification accuracy by throwing away components. Therefore the feature selection criterion must minimize the Bayes risk, which typically is classification error in pattern recognition problem. 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 [8, 9]. 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 [15,16]. Hellman & Raviv [7] showed that the upper bound on Bayes error is given by ($H_S(c)$-$I_S(\mathbf{x},c)$)/2, where $H_S(c)$ is the Shannon entropy of the *a priori* probabilities of the classes and $I_S(\mathbf{x},c)$ is the Shannon MI between the continuous-valued feature vectors and the discrete-valued class labels. Consequently, maximizing the MI between the selected features and the class labels potentially improves classification performance, and has drawn much attention [17, 18].

Shannon MI between feature vectors **x** and c is defined in terms of the entropies of the overall data and individual classes as

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

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

$$\begin{aligned} H_S(x) &= -\int p(x)\log\, p(x)dx \\ H_S(x \mid c) &= -\int p(x \mid c)\log\, p(x \mid c)dx \end{aligned} \tag{2}$$

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

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

Above equations show that the critical step for this feature selection method is the entropy estimation. Previously in ICA-MI method, we estimate entropy by an indirect method. While in GMM-MI method, since one can approximate an arbitrary distribution by limited number of Gaussian components with sufficient amount of data, one can estimate entropy directly by definition (2). The GMM density estimation can be written as:

$$p(x) = \sum_{m=1}^{M} \alpha_m G_m(x) \tag{4}$$

where $G_m(x)$ is the distribution of each Gaussian component, and α_m is the corresponding component prior. So the estimation of overall entropy can be written as:

$$\hat{H}_S(x) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\sum_{m=1}^{M} \alpha_m G_m(x_i)\right) \tag{5}$$

where the class conditional entropy is given by:

$$\hat{H}_S(x \mid c) = -\frac{1}{N_c}\sum_{i=1}^{N_c} \log\left(\sum_{m=1}^{Mc} \alpha_m G_m(x_i^c)\right) \tag{6}$$

where N is the overall data samples and N_C is the data samples for class *C*, x^c is the data samples from class *C*. Combining (1), (5) and (6), the MI estimation can be written as:

$$\begin{aligned} I(x;c) = &-\frac{1}{N}\sum_{i=1}^{N} \log\left(\sum_{m=1}^{M} \alpha_m G_m(x_i)\right) \\ &+ \sum_c p_c\left(\frac{1}{N_c}\sum_{i=1}^{N_c} \log\left(\sum_{m=1}^{M_c} \alpha_m G_m(x_i^c)\right)\right) \end{aligned} \tag{7}$$

Using the MI estimation given by (5)-(7), the GMM-MI feature ranking algorithm can be described as Procedure 1:

A. Estimate both class densities and overall density for each feature vector using GMM separately.
B. Estimate the MI between each feature vector and class labels. Find the feature with maximum MI, and mark it as opt-sub1 (optimal subset of 1 feature).
C. Select one of the remaining feature vectors, combine it with opt-sub1 to form sub2 (subset of 2 features). Estimate both class density and overall density of sub2, and then estimate MI between sub2 and class

labels. Repeat this process for all remaining features, find the features with maximum MI, and mark them as opt-sub2.

D. Repeat Step C by increasing one feature at a time, until all features are ranked in the sense of MI maximization.

This procedure results in an ordering of features such that the first *d* features have maximal MI with class labels. The choice of *d* to be used in the application is dependent on the requirement for classification performance and computational cost. Using this search strategy, the computational complexity is $(n+1)n/2$ (n is the total number of features) instead of the 2^n of exhaustive evaluation.

In Procedure 1, GMM with certain number of components is fitted to the data from using the Expectation-Maximization algorithm [19]. This GMM fitting is required for all combinations of feature vectors. What's more, to determine the optimal number of components, we apply cross-validation and use several restarts to achieve maximum likelihood. Therefore, Procedure 1 is time-consuming. We implemented Procedure 1 with Matlab 7.0.1 on Dell Precision 370 with single P4 2.8G CPU, 1GB memory. The training datasets contain 30 dimensional EEG signals, with about 300 samples. The whole feature ranking procedure took about 125 hours, which makes GMM-MI algorithm almost impracticable in real world applications. To overcome this difficulty, we first use spherical covariance matrix, and assume that the number of optimal components for all features is identical to that for different combinations of feature subsets. In this way, one only needs to do the cross-validation once for all features at the beginning, and pick rows and columns from the mean vectors and covariance matrices for the corresponding features. Under this assumption, the GMM-MI algorithm is revised as procedure 2:

A. Use cross-validation to determine the optimal number of components for each class and overall data for all feature vectors. Estimate both class densities and overall density for all feature vectors using GMM. Get the mean vectors and covariance matrices for each component.

B. Pick the corresponding rows and columns from mean vectors and covariance matrices (generated in A), and estimate density for each feature vector. Estimate the MI between each feature vector and class labels. Find the feature with maximum MI, and mark it as opt-sub1 (optimal subset of 1 feature).

C. Pick one in the remaining feature vectors, combine it with opt-sub1 to form sub2 (subset of 2 features). Find the corresponding rows and columns from mean vectors and covariance matrices (generated in A), form the new mean vectors and covariance matrices, and estimate both class densities and overall density of sub2, and then estimate MI between sub2 and class labels. Repeat this process for all remaining features, find the features with maximum MI, and mark it as opt-sub2.

Repeat Step C by increasing one feature at a time, until all features are ranked in the sense of MI maximization.

We implemented Procedure 2 on the same software and hardwire platform. The training datasets are identical to above experiment. Resulting in a better computational efficiency, the whole feature ranking procedure took about 10 minutes for this simulation. In the next section, we will show the different feature ranking results from both procedures on the same training set.

III. Experiments And Results

In this section, we will show experimental results by applying the proposed GMM-MI algorithm on Iris data and Wisconsin breast cancer data from UCI machine learning repository [20]. We also implemented the algorithm on EEG data collected by Honeywell in AugCog project. As a comparison, we also implemented the previous ICA-MI algorithm on the same datasets.

A. *UCI machine learning repository*

In this experiment, we applied both GMM-MI and ICA-MI on Iris and Wisconsin breast cancer datasets. For experiments using GMM-MI, the applied procedure that can be described as:

1. Randomly split data into training and testing sets.
2. Use training data to fit GMM by 5-fold cross-validation. Use the trained GMM model, together with training data to do feature ranking (If not mentioned, procedure 2 is used throughout the simulations).
3. Use the trained GMM model to form a parametric Bayes classifier, and use this classifier to do the classification on the ranked features.
4. Repeat 1-3 for 10 Monte Carlo runs, record the feature ranking indices for each time, and average the classification accuracy over 10 times.

For experiments using ICA-MI, the only difference is the second step, in which the ICA-MI algorithm was applied to rank the features. The feature ranking results for both datasets are shown in Table I and II, and the classification accuracies in Fig. 1 and 2. As a comparison, we also apply the Bayes error based wrapper approach for feature ranking. The results are shown in Fig. 1 and 2.

Experimental results show that in Iris and Wisconsin breast cancer data, GMM-MI exhibits more accuracy than ICA-MI. Visualizing each feature in iris data, feature 3 and 4 exhibit much higher separability than feature 1 and 2. Table I shows that GMM-MI reflects this fact without any problem; while ICA-MI yields inaccurate results (rank feature 4 as the least important feature). The ranking results of Wisconsin breast cancer (Table II) are consistent with the separability of the each feature. The classification accuracy curve in Fig. 1 and 2 also demonstrates that GMM-MI works better than ICA-MI for these two datasets. Despite the better performance, GMM-MI is very slow compared with ICA-MI (about 100 times slower according to Table I and II). Fig. 1 and 2 also show the feature ranking results using Bayes error as criterion. Obviously, the results are better than GMM-MI and ICA-MI methods, because this method is optimal to selected GMM

TABLE I

FEATURE RANKING RESULTS FOR IRIS DATA

	Average CPU time (second)	Ranking indices
GMM-MI	50.69	3 4 2 1 (9 times) 3 4 1 2 (1 time)
ICA-MI	0.07	3 1 2 4 (10 times)

The second row shows the feature ranking results by GMM-MI method; the third row shows feature ranking results by ICA-MI method.

The second column shows the run-time for two methods (based on P4 2.8G CPU). These numbers only give the reader a concept about the efficiency of two methods.

The third column shows the feature indices for 10 Monte Carlo runs.

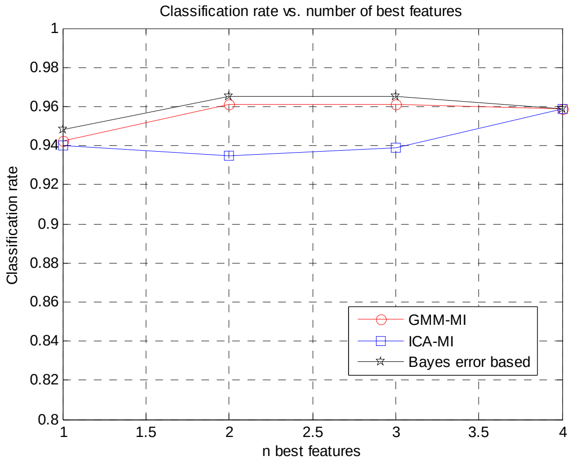

Fig. 1. Classification accuracy for Iris data by GMM-MI, ICA-MI and Bayes error based algorithms. The classification accuracy is the average over 10 Monte Carlo simulations.

TABLE II

FEATURE RANKING RESULTS FOR WISCONSIN BREAST CANCER DATA

	Average CPU time (second)	Ranking indices

GMM-MI	314.99	2 6 5 4 9 3 7 1 8
		2 6 1 8 4 9 5 7 3
		2 6 1 8 5 9 7 4 3
		3 6 1 8 4 9 5 7 2
		2 6 1 4 5 9 7 3 8
		2 6 1 8 4 5 7 9 3
		2 6 8 5 4 9 1 3 7
		2 6 1 4 5 7 9 8 3
		2 6 1 8 4 9 5 7 3
		2 6 1 4 5 9 7 3 8

ICA-MI	2.35	3 6 9 5 8 4 7 2 1
		3 6 9 5 8 4 7 2 1
		3 6 9 5 8 4 1 2 7
		3 6 9 5 8 4 7 2 1
		3 6 9 5 8 4 7 2 1
		3 6 9 5 8 4 7 2 1
		3 6 9 5 8 7 2 4 1
		3 6 9 5 8 4 1 2 7
		3 6 9 5 8 4 7 2 1
		3 6 9 5 8 4 7 2 1

The second row shows the feature ranking results by GMM-MI method; the third row shows feature ranking results by ICA-MI method.

The second column shows the run-time for two methods (based on P4 2.8G CPU). These numbers only give the reader a concept about the efficiency of two methods.

The third column shows the feature indices for 10 Monte Carlo runs.

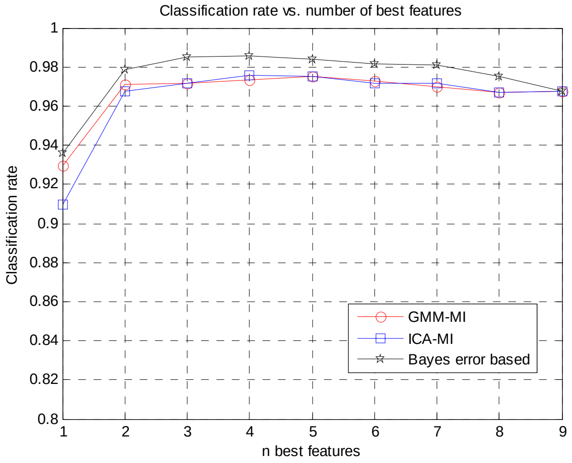

Fig. 2. Classification accuracy for Wisconsin breast cancer data by GMM-MI, ICA-MI and Bayes error based algorithms. The classification accuracy is the average over 10 Monte Carlo simulations.

classifier. However, since it requires combinational learning, it is much slower than both GMM-MI and ICA-MI. And the classification results obtained by ranked feature have only slight difference compared with error based approach.

B. *EEG dataset*

We also applied the proposed GMM-MI method on EEG dataset collected by Honeywell in Augmented Cognition (AugCog) project. The aim of AugCog project is to enhance the task-related performance of a human user through computer mediated assistance based on assessments of cognitive state from EEG signals. In this experiment, the subject executed the predefined tasks, which correspond to the different level of brain activities (high and low). The EEG data was pre-processed by removing the muscular artifacts, filtering out irrelevant frequency bands. Power Spectrum Density features are extracted in 30 dimensions (for more information about AugCog project, please refer to [6, 14, 21]). The experimental procedure is similar to that for Iris and Wisconsin breast cancer datasets, except for two differences: 1) in AugCog projection, we have two data files, one is used as training, the other is used as testing; 2) we did not use Monte Carlo procedure. The ranking results and classification accuracy are shown in Table III and Fig. 3. As reference, we also list the ranking results by GMM-MI Procedure 1 mentioned in section II.

1) The experimental results on EEG dataset also validate that ICA-MI is much faster than GMM-MI. From performance point of view, both GMM-MI and ICA-MI exhibit certain degree of consistency. However, none of them is superior to the other: if we only want to select 3-5 features, ICA-MI yields better performance. If we want to select 15-20 features, GMM-MI yields better performance. For the ICA-MI algorithm, as we mentioned before, linear assumption might degrade the accuracy of MI estimation. For GMM-MI algorithm, there could be two reasons: 1) by using Procedure 2 to replace Procedure 1, we assume that the optimal numbers of components are identical for different combination of features, this might not hold in some cases; 2) we do not have enough data. Consider we are working on 30 dimensions, but we only use about 300 data samples to fit GMM model

TABLE III

FEATURE RANKING RESULTS FOR AUGCOG EEG DATA

	Average CPU time (second)	Ranking indices
GMM-MI (Procedure 2)	558.86	24 3 23 7 11 8 30 2 4 20 26 9 6 1 29 17 16 10 5 18 14 12 19 21 28 25 27 15 13 22
ICA-MI	2.86	24 18 3 22 13 23 29 7 16 28 2 5 14 27 25 6 1 10 11 19 8 20 4 30 9 15 21 17 12 26
GMM-MI (Procedure 1)	About 125 hours	24 4 22 18 14 9 8 23 16 29 27 20 3 12 19 17 26 7 6 2 15 25 5 28 21 30 11 1 10 13

The second row shows the feature ranking results by GMM-MI method; the third row shows feature ranking results by ICA-MI method. the fourth row also show the feature ranking results by using GMM-MI method procedure 1.

The second column shows the run-time for two methods (based on P4 2.8G CPU). These numbers only give the reader a concept about the efficiency of two methods.

The third column shows the feature indices for 10 Monte Carlo runs.

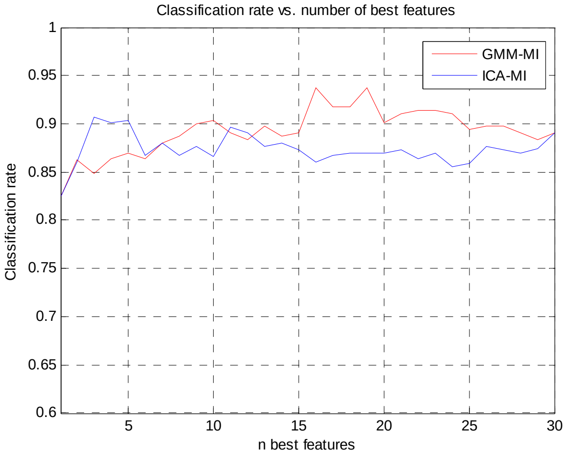

Fig. 3. Classification accuracy for EEG data by GMM-MI and ICA-MI algorithms.

IV. CONCLUSION

In this paper, we proposed a feature ranking and selection method: GMM-MI algorithm. This method exploits the fact that GMM is a widely accepted density estimator for an arbitrary distribution, and can be used for entropy estimator. In section II, this method is described in detail, and an approximation is used for reducing the computational requirement. Experiments on Iris, Wisconsin breast cancer, and EEG data show that GMM-MI can partly overcome the non-linear problem that ICA-MI has. However, because it is much slower than ICA-MI, and its accuracy will be impaired in higher dimension when there is no enough training data, this method can be used as a supplement to our existing method. We also show the feature ranking results by using Bayse error criterion for Iris and Wisconsin breast cancer data, experimental results show that this method is optimal to selected GMM classifier; however, it is too slow to use in real time AugCog application.

Knowing the shortcomings of both GMM-MI and ICA-MI algorithms, the future work will focus on two aspects: 1) improving the performance of ICA transformation, in order to increase the accuracy of MI estimation; 2) accelerating the GMM algorithm to reduce the computational requirement for GMM fitting.

ACKNOWLEDGMENT

This work was supported by DARPA under contract DAAD- 16-03-C-0054 and by NSF under grant ECS-0524835. 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 Independent Component Analysis",Proceedings of ICPR'98, pp. 1268-1273, 1998.

[6] T. Lan, D. Erdogmus, A. Adami, M. Pavel, "Feature Selection by Independent Component Analysis and Mutual Information Maximization in EEG Signal Classification," Proceedings of IJCNN'05, Montreal, Canada, pp. 3011-3016, Aug. 2005.

[7] W. Duch, T. Wieczorek, J. Biesiada, M. Blachnik, "Comparison of feature ranking methods based on information entropy", Proc. of International Joint Conference on Neural Networks (IJCNN), Budapest 2004, IEEE Press, pp. 1415-1420.

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

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

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

[11] Kira, K. and Rendell,L., "The feature selection problem: Traditional methods and a new algorithm," *In Proceedings of the Tenth National Conference on Artificial Intelligence* (AAAI-92), pages 129–134, Menlo Park, CA, USA, 1992. AAAI Press.

[12] John,G.H., Kohavi,R., & Pfleger,K., "Irrelevant features and the subset selection problem". *In Proceedings of the 11th International Conference on Machine Learning*, pp 121-129, San Mateo, CA, Morgan Kaufmann, 1994.

[13] Guyon,I., and Elisseeff, A., "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research (Special Issue on Variable and Feature Selection)*, 2003.

[14] T. Lan, D. Erdogmus, A. Adami, M. Pavel, S. Mathan, "Salient EEG Channel Selection in Brain Computer Interfaces by Mutual Information Maximization," Proceedings of EMBC'05, Shanghai, China, Sept. 2005.

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

[16] 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.

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

[18] 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.

[19] A.P. Dempster, N.M. Laird, D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977.

[20] http://www.ics.uci.edu/~mlearn/MLRepository.html

[21] T. Lan, A. Adami, D. Erdogmus, M. Pavel, “Estimating Cognitive State Using EEG Signals,” Proceedings of EUSIPCO’05, Sep 2005.