Feature Extraction Using Information-Theoretic Learning

Kenneth E. Hild II, *Senior Member*, *IEEE*, Deniz Erdogmus, *Member*, *IEEE*, Kari Torkkola, and Jose C. Principe, *Fellow*, *IEEE*

Abstract—A classification system typically consists of both a feature extractor (preprocessor) and a classifier. These two components can be trained either independently or simultaneously. The former option has an implementation advantage since the extractor need only be trained once for use with any classifier, whereas the latter has an advantage since it can be used to minimize classification error directly. Certain criteria, such as Minimum Classification Error, are better suited for simultaneous training, whereas other criteria, such as Mutual Information, are amenable for training the feature extractor either independently or simultaneously. Herein, an information-theoretic criterion is introduced and is evaluated for training the extractor independently of the classifier. The proposed method uses nonparametric estimation of Renyi's entropy to train the extractor by maximizing an approximation of the mutual information between the class labels and the output of the feature extractor. The evaluations show that the proposed method, even though it uses independent training, performs at least as well as three feature extraction methods that train the extractor and classifier simultaneously.

Index Terms—Feature extraction, information theory, classification, nonparametric statistics.

1 Introduction

Feature extraction can be used as a preprocessor for applications including visualization, classification, detection, and verification. Herein, feature extraction is investigated as it applies to classification. Classification consists of associating each incoming exemplar, having N_I features, with one of N_C class labels. It is a supervised process, which implies that a set of N_T exemplars are available for which the true class labels are known. The designer of a classification system does not usually know a priori which features will yield acceptable classification performance and, in theory, the classification performance is a nondecreasing function of the number of features. Hence, the designer might choose to use all available features. However, using a large number of features can be wasteful of both computational and memory resources. In addition, due to practical problems associated with training a classifier with a finite amount of data, using a large number of features can actually cause degradation of classification performance [1]. Reduction of the number of input features can be done by linear or nonlinear transformations. The use of a linear transformation to reduce the number of features is known as (linear) feature extraction or subspace projection. A constrained linear transformation can also be used. The selection of a subset of the input features is a common method of constraining the linear transformation.

Fig. 1 shows a block diagram of a generic classification system. In this figure, $\boldsymbol{s}_j(n)$, $\boldsymbol{x}_j(n)$, and $\boldsymbol{y}_j(n)$ are the size $(N_I \times 1)$ input features, $(N_O \times 1)$ output features, and the $(N_C \times 1)$ outputs of the classifier for the nth exemplar and having class j $(j = 1, 2, \ldots, N_C)$, respectively. Each of these variables represents a vector the elements of which are denoted using two subscripts, e.g., $s_{j,i}(n)$ is the ith element of the input feature vector for the nth exemplar and having class j. A single subscript is used when referring to the vector having a particular class label, whereas no subscript is used to denote the corresponding vector when the class label is unknown. The maximum (MAX) operator selects the single output of the N_C classifier outputs that has the largest value, the index of which provides the estimate of the class label. Likewise, $c(n)$ and $e(n)$ denote, respectively, the (true) class label and the error for the nth exemplar (other aspects of Fig. 1 are described below). Only linear transformations are considered herein. In this case, the feature extraction is performed using a $(N_O \times N_I)$ matrix $\boldsymbol{R}$, where $\boldsymbol{x}_j(n) = \boldsymbol{R}\boldsymbol{s}_j(n)$.

The feature extractor and the classifier shown in Fig. 1 can be trained simultaneously or independently. Simultaneous training is used in several recent approaches (listed below), which involve the minimization of criteria that resemble the misclassification rate and are expected to outperform methods that train the extractor and classifier independently. Herein, we introduce an information-theoretic method that trains the extractor in an independent fashion, we show that it performs better than several simultaneously-trained systems on five randomly chosen data sets, and we explain

- *K.E. Hild II is with the Biomagnetic Imaging Laboratory, University of California at San Francisco, Room C-324B, San Francisco, CA 94122. E-mail: k.hild@ieee.org.*
- *D. Erdogmus is with the Departments of Computer Science and Engineering, and Biomedical Engineering, OGI School of Science and Engineering, Oregon Health and Science University, Beaverton, OR 97006. E-mail: derdogmus@ieee.org.*
- *K. Torkkola is with the Center for Human Interaction Research, Motorola Labs, 2900 South Diablo Way, MD DW286, Tempe, AZ 85282. E-mail: kari.torkkola@motorola.com.*
- *J.C. Principe is with the Computational NeuroEngineering Laboratory, University of Florida, Room NEB 451, Gainesville, FL 32611. E-mail: principe@cnel.ufl.edu.*

Manuscript received 24 Mar. 2005; revised 7 Jan. 2006; accepted 30 Jan. 2006; published online 13 July 2006.
Recommended for acceptance by A. Srivastava.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0160-0305.

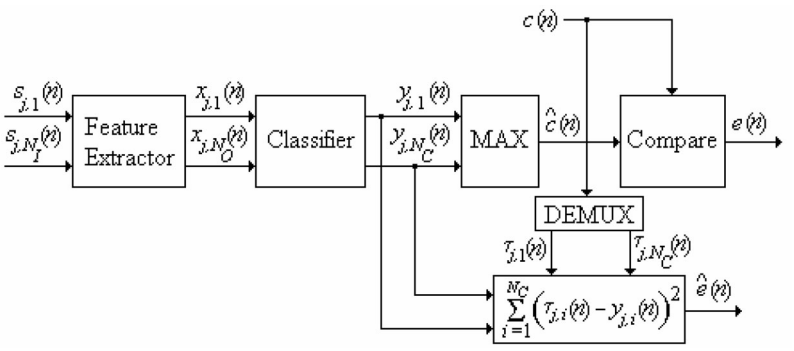


Fig. 1. Block diagram of a generic classification system.

why simultaneously-trained systems do not necessarily outperform independently-trained systems.

2 Information-Theoretic Feature Extraction

Methods that use second-order statistics compare the linear relationship between random variables, whereas information-theoretic methods compare the nonlinear relationships between random variables, i.e., between a vector of features and the class label. A possible criterion for the latter approach is mutual information (MI), which may be described as the amount of information the random output feature vector, $\boldsymbol{X}$, carries about the class, C (where realizations of $\boldsymbol{X}$ and C are given by $\boldsymbol{x}(n)$ and $c(n)$, respectively). Mutual information is defined by [2],

$$I(\boldsymbol{X};C) = H(\boldsymbol{X}) - H(\boldsymbol{X}|C), \tag{1}$$

where $H(\boldsymbol{X})$ is Shannon's (differential) entropy [2]. Upper and lower bounds on the Bayes error rate exist that are minimized by maximizing MI [3], [4], [5]. However, MI is not in common use since the estimation of Shannon's entropy is computationally intensive [6] (estimation of Shannon's entropy can be accomplished by discretizing the variables [6], [7], [8], [9], [10] or by using an $\mathrm{O}(N_T^2)$ nonparametric estimator [11]).

2.1 Proposed Method, MRMI-SIG

The proposed method replaces the two (Shannon) entropy terms in (1) with entropy terms introduced by Renyi [12]. This substitution produces the following estimate of MI,

$$I(\boldsymbol{X};C) \cong H_2(\boldsymbol{X}) - H_2(\boldsymbol{X}|C), \tag{2}$$

where $H_2(\boldsymbol{X})$ is Renyi's quadratic entropy [13]. This substitution is chosen since there exists a nonparametric estimate of Renyi's quadratic entropy that reduces the computational complexity from $\mathrm{O}(N_T^2)$ to $\mathrm{O}(N_T)$. This entropy estimator is given by [14], [15],

$$H_2(\boldsymbol{X}) \cong -\log \frac{1}{N}\sum_{n=1}^{N} G(\boldsymbol{x}(n) - \boldsymbol{x}(n-1), 2\sigma^2\boldsymbol{I}), \tag{3}$$

which is based on Parzen windows [16] and where N is the number of exemplars and $G(\boldsymbol{x}, \sigma^2\boldsymbol{I})$ is a Gaussian kernel evaluated at $\boldsymbol{x}$ and having a diagonal, isotropic covariance matrix.

Both classifiers considered herein are invariant under an invertible, linear transformation. This invariance allows a reduction in the number of free parameters that must be adapted without unnecessarily restricting the possible set of decision surfaces that can be produced by a (linear) projection. The reduction is performed by constraining the feature extraction matrix, $\boldsymbol{R}$, to be a pure rotation matrix. Hence, $\boldsymbol{R}$ can be expressed as a function of a vector of rotation angles, $\boldsymbol{\theta}$, as follows:

$$\boldsymbol{x}_j(n) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{s}_j(n), \tag{4}$$

$$\boldsymbol{R}(\boldsymbol{\theta}) = \left[\prod_{i=1}^{N_O}\prod_{m=N_O+1}^{N_I} \boldsymbol{R}_{i,m}(\theta_{i,m})\right]_{N_O}, \tag{5}$$

where the notation $[\boldsymbol{A}]_{N_O}$ corresponds to keeping only the first N_O rows of matrix $\boldsymbol{A}$, $\theta_{i,m}$ is a single element of the rotation angle vector, and $\boldsymbol{R}_{i,m}(\theta_{i,m})$ is an individual Given's rotation matrix [17]. Constraining the transformation in this manner reduces the number of parameters from $N_O N_I$ to $N_O(N_I - N_O)$. For the classifiers used here, rotations between retained output features have no effect on classification nor do rotations between rejected outputs. Only rotations between a feature that is retained and a feature that is rejected have an effect on classification. Therefore, only these rotation angles are trained, as indicated by the limits in (5). The proposed method can also be used with unconstrained linear or nonlinear transformations, the choice of which is dictated by the classifier (e.g., it is, in general, overly restrictive to use a linear feature extractor with a classifier that can only generate linear decision surfaces).

The combination of (2) and (3) results in the proposed information-theoretic criterion for feature extraction. This criterion, known as MRMI-SIG, is given by,

$$\begin{aligned} J = &-\log \frac{1}{N_T}\sum_{n=1}^{N_T} G(\boldsymbol{x}(n) - \boldsymbol{x}(n-1), 2\sigma^2) \\ &+ \sum_{j=1}^{N_C}\left(\frac{N_j}{N_T}\log\frac{1}{N_j}\sum_{n=1}^{N_j} G(\boldsymbol{x}_j(n) - \boldsymbol{x}_j(n-1), 2\sigma^2)\right), \end{aligned} \tag{6}$$

where $\boldsymbol{x}_j(n) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{s}_j(n)$, $\boldsymbol{\theta}$ is the $(N_O(N_I - N_O)x1)$ vector of rotation angles adapted to maximize J, N_j is the number of class labels in the training set having class j, and N_T is the

length of the training set. The parameters are updated using gradient ascent optimization,

$$\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) + \eta \nabla_{\theta} J_n, \tag{7}$$

where η is the step size and the subscript on J is used to denote that the order in which the data are presented is shuffled every iteration, the need for which is explained next.

2.2 Discussion of MRMI-SIG

The second term on the right-hand side of (6) is maximized by minimizing $(\boldsymbol{x}_j(n) - \boldsymbol{x}_j(n-1))^2$, which is accomplished by choosing $\boldsymbol{R}$ such that all the consecutive exemplars from a given class are as near as possible to each other in the space of the output features. This equates to minimizing the within-class spread in the limit as long as the data order is shuffled during adaptation. A trivial solution for minimizing the total within-class spread is to set $\boldsymbol{R}$ equal to an all-zeros matrix. This, however, causes the features from all classes to overlap perfectly. The first term on the right-hand side of (6) prevents this undesirable solution since it is maximized by maximizing $(\boldsymbol{x}(n) - \boldsymbol{x}(k))^2$, which is a measure of the spread of the data (in the space of the output features) irrespective of the class. There are other ways to construct a criterion that attempts to minimize within-class spread and maximize overall spread, e.g., LDA [18] (which is based on second-order statistics). MRMI-SIG, which has an information-theoretic interpretation, represents another possibility.

The proposed criterion is similar to one that was previously used to minimize the mutual information between a set of outputs for the application of blind source separation (BSS) [19]. There are, however, several notable differences between the formulation above and that used for BSS:

1. the criterion for feature extraction uses supervised training as opposed to unsupervised training,
2. only N_O of the outputs of the rotation matrix are kept,
3. mutual information is measured between the output feature set and the class label (instead of between the outputs), and
4. the criterion is based on the entropies of multi-dimensional random vectors.

Each of these items, as discussed next, has an important implication on the performance of MRMI-SIG for feature extraction.

For the BSS application, which involves unsupervised training, the sign of each entropy term in the MRMI-SIG criterion is determined by the shape of the probability density function (pdf) of the associated output [11]. No such sign change is necessary for feature extraction since it involves supervised training. The knowledge of the true class labels (from the training set) is sufficient to prevent/resolve any sign ambiguities.

The requirement that only N_O rows of the rotation matrix be kept for feature extraction impacts asymptotic analyses. For BSS, it is possible to prove that all the elements of the gradient expression for MRMI-SIG are zero at the separating solution without requiring σ to approach 0 (which is the first step in proving that MRMI-SIG provides an unbiased and consistent estimate of the rotation angles required for separation). This proof requires that all rows of the separation matrix be kept and is, therefore, not applicable for feature extraction. The only known proof that dictates when the right-hand side of (2) produces the same solution as (1) is very restrictive (it is limited to the two-class case and it requires that the class covariance matrices are identical). We believe that the proposed method is useful even when these conditions are not met. Nevertheless, this highlights the first of two possible drawbacks to using MRMI-SIG for the present application. Namely, there is no general guarantee that maximizing (2) using Renyi's definition of entropy is equivalent to maximizing (1) using Shannon's definition.

The MI given by (1), on which MRMI-SIG is based, may be written in one of three equivalent expressions,

$$\begin{aligned} &H(\boldsymbol{X}) - H(\boldsymbol{X}|C), \\ &H(C) - H(C|\boldsymbol{X}), \\ &H(\boldsymbol{X}) + H(C) - H(\boldsymbol{X}, C). \end{aligned} \tag{8}$$

The second formulation is not convenient for the entropy estimator of (3) since the given information, $\boldsymbol{X}$, has a continuous distribution, thus requiring integration. The third form is used for BSS since the joint entropy can easily be made to be invariant to the adaptation [19]; however, this simplification does not apply when extracting features since the MI is measured between the output feature set and the class label. The entropy estimator of (3) has an associated gain that is a function of the dimensionality of the underlying random vector [20]. This gain is irrelevant for the maximization or minimization of entropy in some cases (e.g., when there is a single entropy term), but it plays an important role when the criterion consists of a summation of two or more entropies if the random vectors on which they are based have different dimensionalities [20]. To avoid this problem (the dimensionality-dependent gain), for feature extraction, the first formulation of (8) is used since the two entropy terms have an identical dimensionality of N_O (the problem of dimensionality-dependent gain is avoided in BSS by using the third formulation above with the joint entropy term removed). This, however, brings up the second of two possible drawbacks of using MRMI-SIG for feature extraction. The fact that the entropy terms are based on N_O-dimensional random vectors implies that the pdf estimation, which is required to prove the validity of the entropy estimator, is subject to the curse of dimensionality. Notice, however, the dimensionality of the implicit pdf estimation is not determined by the number of input features, but by the (smaller) number of output features.

3 Comparisons

The performances of several different methods are compared using the rate of correct classification of five different data sets. The two Bayes classifiers that are used are the Bayes-G and the Bayes-NP classifiers, both of which generate nonlinear decision surfaces. The Bayes-G classifier is a parametric classifier that assumes that the set of output features, for each class j, has a multivariate Gaussian distribution [21]. The Bayes-NP is a nonparametric classifier that uses Parzen windows [16] to estimate each of the

TABLE 1
Description of the Data Sets Used in the Comparison

Dataset	N_I	N_C	N_T	Test Size	Outliers
Pima Indians	8	2	500	268	8.0%
Landsat Satellite Image (Statlog)	36	6	4435	2000	0%
Letter Recognition	16	26	16000	4000	0%
Musk	166	2	300	176	0%
Arrhythmia	279	16	300	152	0.3%

a posteriori distributions [22]. Unlike the Bayes-G classifier, the Bayes-NP classifier makes no assumptions on the distribution of the output features so that it is able to take into account higher-order statistics of the output features, including multiple-modality. All reported results for the Bayes-NP classifier are based on using a kernel size of 0.25.

Results are shown for a total of six methods. Three of these train the extractor and classifier independently, namely, the proposed method (MRMI-SIG), Principal Components Analysis (PCA) [18], and a method that is based on maximizing quadratic mutual information (QMIE) [23], [24], [25], [26], [27] (this last method bears some similarity to MRMI-SIG in that both use Parzen windows and both are a function of a squared pdf). The remaining three methods train the feature extractor and the classifier simultaneously. These methods include Minimum Classification Error (MCE) [28], [29], [30], [31], [32], [33] (which is related to a method by Nedeljkovic [34]), Mean Square Error (MSE) [35], and a method that ranks features based on classification performance of a validation set (FR-V) [35]. For the sake of perspective, the classification results of random projections are also included for the lower-dimensional data sets, the coefficients of which are chosen uniformly in $[-1, 1]$. The results of the random projection are represented in the plots using a dashed line.

The MRMI-SIG, MCE, and MSE methods all have computational complexity $\mathrm{O}(N_T)$ and QMIE has computational complexity $\mathrm{O}(N_T^2)$, whereas the computational complexity of FR-V depends only on the classifier and PCA has an analytical solution. For the two high-dimensional data sets, MRMI-SIG is used only to rank the input features so that the comparison between MRMI-SIG and FR-V is between two methods having similar computational complexity. Feature ranking, which is suitable for data sets having extremely high dimensionality, is used for demonstrative purposes only. We expect that using PCA to reduce the dimensionality to a manageable intermediate value or using a multistage (semigreedy) approach will provide better classification results than ranking.

MRMI-SIG uses a kernel size of $\sigma = 0.5$, the value of which is based on a theoretical analysis of nonparametric entropy estimators (for unit-variance random vectors, kernel sizes between 0.2 and 0.5 minimize the sensitivity of the entropy estimator due to the kernel size [36]). The MSE criterion uses the 1 of the N_O scheme to define the targets, $\boldsymbol{\tau}_j(n)$, which is defined as setting the target for the nth exemplar associated with class j to 1 and the other $N_C - 1$ targets to 0 (this is represented in Fig. 1 using the demultiplexer). MCE has two user-defined parameters, α and ν, which are set to 10 and 2, respectively. Local minima of the MCE algorithm can, if performed properly, be avoided by scheduling the value of the smoothing parameter, α. This can also be accomplished with the proposed information-theoretic criterion by annealing the kernel size, σ [37]. In order to simplify the experimental procedure this is not done for either MCE or MRMI-SIG. The FR-V method uses a validation set that is found by randomly selecting a (disjoint) subset of the original training set.

Table 1 shows the important characteristics of the five data sets used herein. The first three were randomly selected from the list of all data sets at the UCI Machine Learning Repository, whereas the Musk and Arrhythmia data sets were selected for their large input dimensionality (all data sets may be found at http://www.ics.uci.edu/~mlearn/MLRepository.html). For all methods except PCA it is assumed (without loss of generality due to the properties of the two classifiers used here) that the original features of each data set have been shifted, rotated, and scaled so that the resulting $(N_I \times 1)$ input features, $\boldsymbol{s}(n)$, are zero-mean, (spatially) uncorrelated, and have unit variance. Since the transform for PCA depends on the eigenvalues of the autocorrelation matrix, sphering should not be used with PCA. The Pima data set has numerous invalid data points, e.g., features that have a value of 0 even though a value of 0 is not meaningful or physically possible. These correspond to points in feature space that are statistically distant from the mean calculated using the remaining data (with the points in question removed). No attempt was made to remove or correct for these outliers. Likewise, the Arrhythmia data set has missing values, all of which are set to 0 for the comparison.

The Bayes-G classifier produces the best classification performance for the Pima, Landsat, and Musk data sets, whereas the Bayes-NP classifier performs the best for the Letter Recognition and Arrhythmia data sets. Therefore, the results shown are restricted to these combinations of data sets and classifiers. This choice has no effect on the relative performance of the different feature extraction methods. The training is performed using N_T randomly-selected samples of the data set and is tested on the remaining (disjoint) data. All results are reported using 10-fold cross validation. The results for all algorithms are the same whenever $N_O = N_I$. This is because both classifiers are invariant under full-rank linear transformations.

Figs. 2, 3, and 4 show the correct classification rates for the Pima, Landsat, and Letter Recognition data. Each figure includes error bars that represent one standard error. Results are not shown for the Landsat data for $N_O > 9$ since the performance of the methods becomes indistinguishable

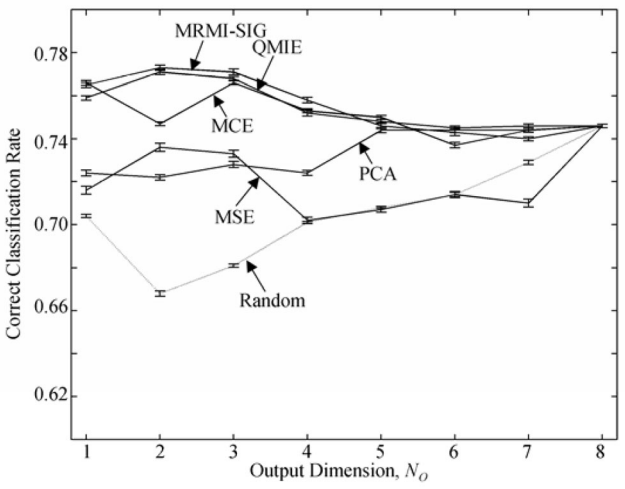


Fig. 2. Classification performance versus output dimension, N_O, for the Pima data set.

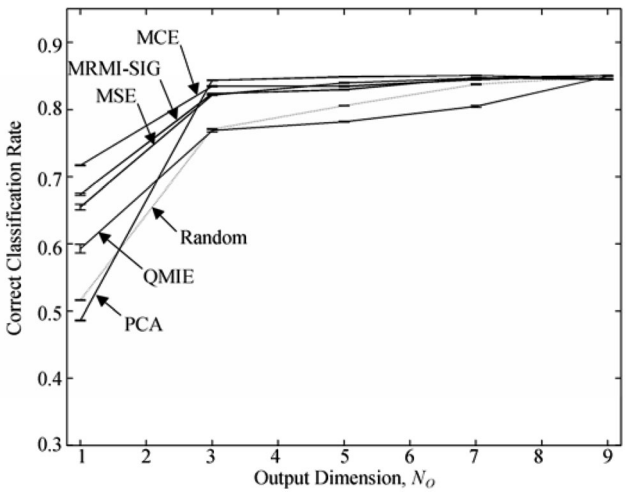


Fig. 3. Classification performance versus output dimension, N_O, for the Landsat data.

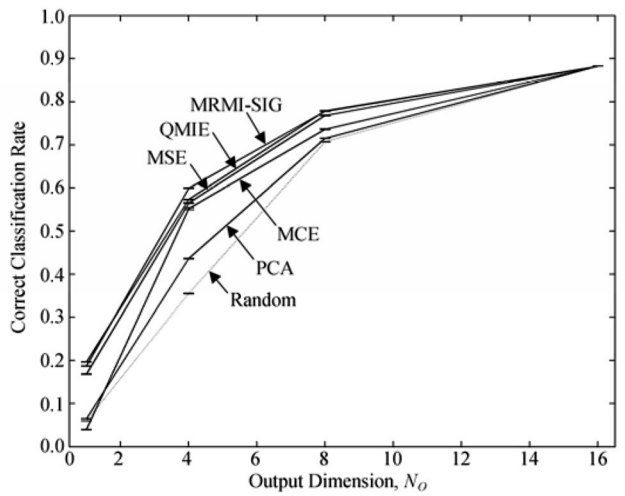


Fig. 4. Classification performance versus output dimension, N_O, for the Letter Recognition data.

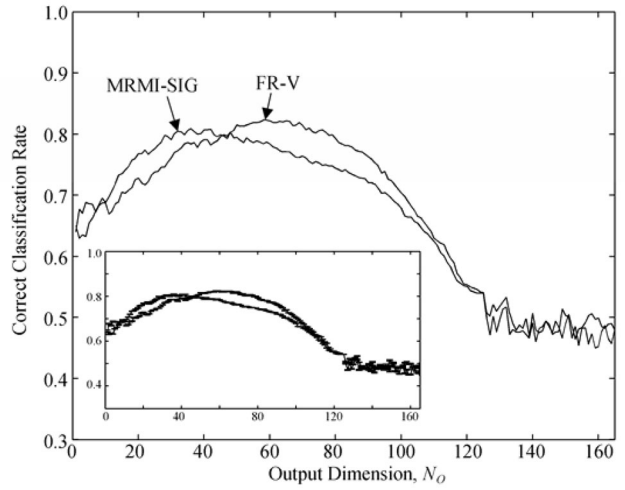


Fig. 5. Classification performance versus output dimension, N_O, for the Musk data.

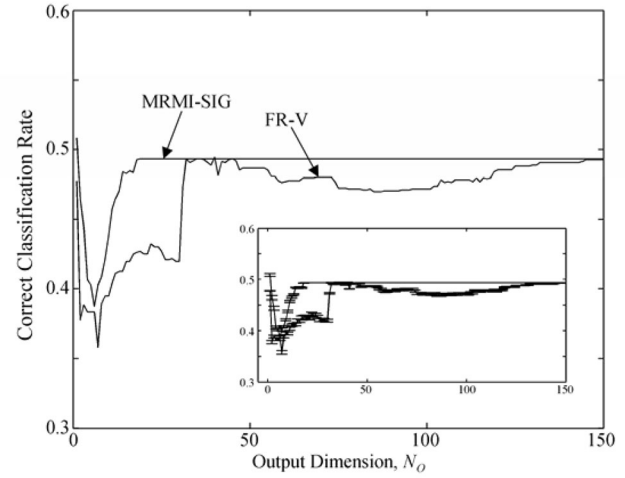


Fig. 6. Classification performance versus output dimension, N_O, for the Arrhythmia data.

beyond this point. All five methods perform well on these three data sets, with the exception that MSE has trouble with the Pima data set and PCA performs poorly for the Landsat data for small N_O. The proposed method has the best performance for the Pima and Letter Recognition data sets and the second best for the Landsat data.

Figs. 5 and 6 show the results for the two high-dimensional data sets, Musk and Arrhythmia. For these two plots, MRMI-SIG is only used to rank the features. Results are shown for both MRMI-SIG and FR-V and the inset in each figure shows the corresponding error bars. In Fig. 5, the best performances for MRMI-SIG and FR-V are similar. Notice, however, that MRMI-SIG concentrates the group of features most useful for classification into the top 20 percent of the highest-ranked features. Hence, it requires roughly 25 fewer features to reach the peak performance. The jaggedness of the curves for $N_O > 120$ corresponds to the point at which at least one of the class covariance matrices used in the Bayes-G classifier becomes ill-conditioned. The curves in Fig. 6 have a very different characteristic due to the use of the Bayes-NP classifier on data where roughly 1/4 of the 16 class labels are poorly

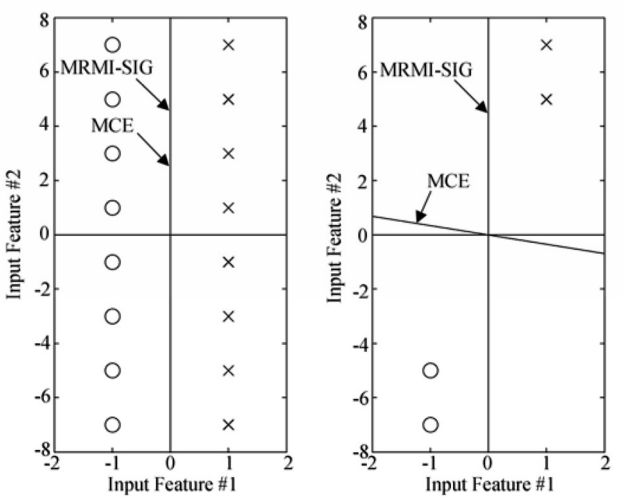


Fig. 7. The left subplot shows all possible locations of input features for an example two-class problem and the right subplot shows an example realization of $N_T = 4$ data drawn from this distribution. Also shown are the resulting decision boundaries of a linear classifier based on using MRMI-SIG and MCE to extract a single feature.

represented (fewer than five instances each). Classification results for both methods are flat and very nearly equal for the portion of N_O not shown in the figure. As can be seen, the performance of MRMI-SIG is better than or equal to the performance of FR-V for all values of N_O.

4 Discussion

Methods that simultaneously train the extractor and the classifier place a stronger emphasis on the locations in feature space that are near the decision boundary. This does not occur for methods that train the extractor independently of the classifier since the boundary is necessarily defined by the classifier. Instead, they are based on the structure of the overall data distribution and the within-class distributions. Consequently, if a particular data set has few data near the boundary (relative to the complexity of the optimum discriminant function), then the generalization of the methods that simultaneously train the extractor and classifier may suffer relative to methods that train in an independent fashion.

A trivial example of this is shown in Fig. 7. The left subplot shows all 16 possible locations of the two-dimensional features for a fictitious two-class data set. Both MRMI-SIG and MCE are used to train a rotation matrix to reduce the dimensionality to one. For the left subplot (which pertains to infinite training), feature reduction using either method amounts to selecting the first feature. The resulting decision boundary produced by a linear classifier, shown in the space of the input features, is included in the figure. Notice that both produce perfect classification. On the other hand, if only $N_T = 4$ data are drawn from this distribution, as shown in the right subplot, the resulting decision boundaries for MRMI-SIG and MCE are noticeably different. In this case, maximizing the margin for MCE produces a decision boundary that yields a 33 percent correct classification rate on the test data. MRMI-SIG still selects the first feature since doing so places all features of each class in the exact same location in the space of the output features (minimizing within-class spread) while maintaining separation of the two classes. In fact, this same solution is always produced by MRMI-SIG if the training set includes at least two distinct samples from each class. If MSE is used in a simultaneously-trained fashion, the results are similar to that obtained using MCE. If MSE is used in an independently-trained fashion and supposing the targets were chosen correctly ($\boldsymbol{\tau}_o(\mathrm{n}) = [-1, 0]$ and $\boldsymbol{\tau}_{\mathrm{x}} = [1, 0]$, where $\boldsymbol{\tau}_o(n)$ and $\boldsymbol{\tau}_{\mathrm{x}}(n)$ are the targets for class "o" and class "x" for the $n\mathrm{th}$ exemplar, respectively), then the results are identical to that obtained using MRMI-SIG (however, in general, it is not possible to know good locations for the targets in the output feature space). This example favors methods that take into account the structure of the data, but it is just as simple to construct an example that favors methods that emphasize features near the boundary.

To the extent that the preceding argument holds for the five data sets considered here, it could be argued that MRMI-SIG performs well compared to the methods that train the extractor and classifier simultaneously because the Pima, Letter Recognition, Musk, and Arrhythmia data sets have optimum decision boundaries that are not represented sufficiently well by the (finite) training data, whereas the optimum decision boundary for the Landsat data is represented marginally well by the training data.

5 Conclusion

Interest in the distinction between training the extractor simultaneously or independently from the classifier (also known as the wrapper and filter approaches, respectively) has heightened in recent years due to the paper by Biem et al. [28] that introduced the MCE method. A second paper that deals with this topic and is important for this discussion is a paper by LeCun et al. [38], where it is argued that discriminative methods, i.e., simultaneously-trained methods, are preferred. Nevertheless, the proposed feature extraction method performs, on average, as well as or better than MCE, MSE, and FR-V even though all three train the extractor and classifier simultaneously and two of them (MCE and FR-V) are optimized by minimizing classification error directly. This is possible since the optimization is necessarily performed on the finite training set, not on the disjoint finite test set. In order to prove that any method minimizes the probability of error, on the disjoint test set, "infinite training" is required, as acknowledged by Watanabe et al. [29] and Katagiri et al. [30]. This is essentially equivalent to knowing the underlying distributions, which would allow any of a number of methods to produce the optimum (Bayes) solution.

Much more theoretical work is required to validate the mutual information approach for feature extraction in classification. The fundamental difficulty is related to the implicit link between mutual information and classification error, where the only known results are expressed in the form of upper and lower bounds on the Bayes classification error. Perhaps a more productive approach with

information-theoretic learning is to reverse the question of tuning the classifier topology to the feature extractor and seek classifiers that will meet the minimization of the Bayes error with MI-derived features.

ACKNOWLEDGMENTS

This work was partially supported by US National Science Foundation ECS #9900394.

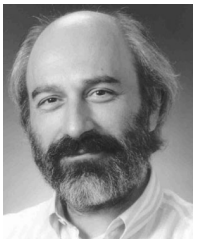
REFERENCES

[1] B.D. Ripley, *Pattern Recognition and Neural Networks.* Cambridge Univ. Press, 1995.
[2] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* John Wiley & Sons, 1991.
[3] J.C. Principe, D. Xu, Q. Zhao, and J.W. Fisher III, "Learning from Examples with Information Theoretic Criteria," *J. VLSI Signal Proc. Systems,* vol. 26, nos. 1/2, pp. 61-77, Aug. 2000.
[4] D. Erdogmus and J.C. Principe, "Lower and Upper Bounds for Misclassification Probability Based on Renyi's Information," *J. VLSI Signal Processing,* vol. 37, nos. 2-3, pp. 305-317, June 2004.
[5] M.E. Hellman and J. Raviv, "Probability of Error, Equivocation, and the Chernoff Bound," *IEEE Trans. Information Theory,* vol. 16, no. 4, pp. 368-372, July 1970.
[6] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks,* vol. 5, no. 4, pp. 537-550, July 1994.
[7] H.H. Yang and J. Moody, "Feature Selection Based on Joint Mutual Information," *Proc. Conf. Advances in Intelligent Data Analysis, Computational Intelligence Methods, and Applications,* June 1999.
[8] K.D. Bollacker and J. Ghosh, "Mutual Information Feature Extractors for Neural Classifiers," *Proc. Int'l Conf. Neural Networks (ICNN '96),* pp. 1528-1533, June 1996.
[9] N. Kwak and C.-H. Choi, "Improved Mutual Information Feature Selector for Neural Networks in Supervised Learning," *Proc. Int'l Joint Conf. Neural Networks,* vol. 2, pp. 1313-1318, July 1999.
[10] R. Rajagopal, K.A. Kumar, and P.R. Rao, "An Integrated Approach to Passive Target Classification," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 2, pp. 313-316, Apr. 1994.
[11] K.E. Hild II, D. Erdogmus, and J.C. Principe, "An Analysis of Entropy Estimators for Blind Source Separation," *Signal Processing,* vol. 86, no. 1, pp. 182-194, Jan. 2006.
[12] A. Renyi, *Probability Theory.* Amsterdam: North-Holland Publishing Company, 1970.
[13] K.E. Hild II, D. Erdogmus, and J.C. Principe, "On-Line Minimum Mutual Information Method for Time-Varying Blind Source Separation," *Proc. Int'l Workshop Independent Component Analysis and Signal Separation,* pp. 126-131, Dec. 2001.
[14] D. Erdogmus, K.E. Hild II, and J.C. Principe, "On-Line Entropy Manipulation: Stochastic Information Gradient," *IEEE Signal Processing Letters,* vol. 10, no. 8, pp. 242-245, Aug. 2003.
[15] J. Beirlant, E.J. Dudewica, L. Gyofi, and E. van der Meulen, "Nonparametric Entropy Estimation: An Overview," *Int'l J. Math. Statistics Sciences,* vol. 6, no. 1, pp. 17-39, 1997.
[16] E. Parzen, "On Estimation of a Probability Density Function and Mode," *Annals of Math. Statistics,* vol. 33, no. 3, pp. 1065-1076, Sept. 1962.
[17] G.H. Golub and C.F. Van Loan, *Matrix Computations,* third ed. Baltimore: John Hopkins Univ. Press, 1996.
[18] S. Theodoridis and K. Koutroumbas, *Pattern Recognition.* San Diego, Calif.: Academic Press, 1999.
[19] K.E. Hild II, D. Erdogmus, and J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information," *IEEE Signal Processing Letters,* vol. 8, no. 6, pp. 174-176, June 2001.
[20] R.A. Morejon, "An Information-Theoretic Approach to Sonar Automatic Target Recognition," PhD dissertation, Univ. of Florida, 2003.
[21] C. Bishop, *Neural Networks for Pattern Recognition.* Oxford, U.K.: Oxford Univ. Press, 1995.
[22] S.C. Fralick and R.W. Scott, "Nonparametric Bayes-Risk Estimation," *IEEE Trans. Information Theory,* vol. 17, no. 4 pp. 440-444, July 1971.
[23] K. Torkkola, "On Feature Extraction by Mutual Information Maximization," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 821-825, May 2002.
[24] K. Torkkola, "Learning Discriminative Feature Transforms to Low Dimensions in Low Dimensions," *Proc. Conf. Advances in Neural Information Processing Systems,* Dec. 2001.
[25] K. Torkkola and W.M. Campbell, "Mutual Information in Learning Feature Transformations," *Proc. Int'l Conf. Machine Learning,* pp. 1015-1022, June 2000.
[26] K. Torkkola, "Visualizing Class Structure in Data Using Mutual Information," *Proc. Conf. Neural Networks for Signal Proc. (NNSP '00),* pp. 376-385, Dec. 2000.
[27] D. Xu and J.C. Principe, "Feature Evaluation Using Quadratic Mutual Information," *Proc. Int'l Joint Conf. Neural Networks,* vol. 1, pp. 459-463, July 2001.
[28] A. Biem, S. Katagiri, and B.-H. Juang, "Pattern Recognition Using Discriminative Feature Extraction," *IEEE Trans. Signal Processing,* vol. 45, no. 2, pp. 500-504, Feb. 1997.
[29] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative Metric Design for Robust Pattern Recognition," *IEEE Trans. Signal Processing,* vol. 45, no. 11, pp. 2655-2662, Nov. 1997.
[30] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern Recognition Using a Family of Design Algorithms Based upon the Generalized Probabilistic Descent Method," *Proc. IEEE,* vol. 86, no. 11, pp. 2345-2373, Nov. 1998.
[31] B.-H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Processing,* vol. 40, no. 12, pp. 3043-3054, Dec. 1992.
[32] A. Biem, S. Katagiri, and B.-H. Juang, "Discriminative Feature Extraction for Speech Recognition," *Proc. Conf. Neural Networks for Signal Processing (NNSP '93),* pp. 392-401, Sept. 1993.
[33] Q. Li and B.-H. Juang, "A New Algorithm for Fast Discriminative Training," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '02),* vol. 1, pp. 97-100, May 2002.
[34] V. Nedeljkovic, "A Novel Multilayer Neural Networks Training Algorithm that Minimizes the Probability of Classification Error," *IEEE Trans. Neural Networks,* vol. 4, no. 4, pp. 650-659, July 1993.
[35] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Boston: Academic Press, 1990.
[36] D. Erdogmus, K.E. Hild II, and J.C. Principe, "Kernel Size Selection in Parzen Density Estimation," *J. VLSI Signal Processing Systems,* submitted.
[37] D. Erdogmus and J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," *IEEE Trans. Neural Networks,* Sept. 2002.
[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE,* vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

Kenneth E. Hild II received the BS and MSc degrees in electrical engineering, with emphasis in signal processing, communications, and controls, from the University of Oklahoma, Norman, in 1992 and 1996, respectively. He received the PhD degree in electrical engineering from the University of Florida, Gainesville, in 2003, where he studied information theoretic learning and blind source separation in the Computational NeuroEngineering Laboratory. Dr. Hild has also studied biomedical informatics at Stanford University, Palo Alto, California. From 1995 to 1999, he was employed at Seagate Technologies, Inc., where he served as an advisory development engineer in the Advanced Concepts Group. From 2000 to 2003, he taught several graduate-level classes on adaptive filter theory and stochastic processes at the University of Florida. He is currently employed at the Biomagnetic Imaging Laboratory in the Department of Radiology at the University of California at San Francisco, where he is applying variational Bayesian techniques for processing encephalographic and cardiographic data. He is a senior member of the IEEE.

Deniz Erdogmus received the BS degree in electrical and electronics engineering and mathematics in 1997 and the MS degree in electrical and electronics engineering, with emphasis on systems and control, in 1999, both from the Middle East Technical University, Ankara, Turkey. He received the PhD degree in electrical and computer engineering from the University of Florida, Gainesville, in 2002. He was a research engineer at the Defense Industries Research and Development Institute (SAGE), Ankara, from 1997 to 1999. From 1999 until 2004, he worked at the Computational NeuroEngineering Laboratory, University of Florida, under the supervision of Dr. J.C. Principe, the last two years of which he was a postdoctoral fellow. He is currently an assistant professor with a joint appointment in the Department of Computer Science and Engineering and the Department of Biomedical Engineering at the Oregon Health and Science University. His current research interests include information theory and adaptive systems for signal processing, communications, and control. Dr. Erdogmus is a member of the IEEE, Tau Beta Pi, and Eta Kappa Nu.

Kari Torkkola received the Dipl.Eng., Lic.Tech., and Dr.Tech. degrees in computer science from Helsinki University of Technology, Finland, in 1985, 1988, and 1991, respectively. From 1985 and 1992, he held various teaching and research positions at Helsinki University of Technology. In 1992, he joined IDIAP in Switzerland as a speech recognition research director. In 1994, he joined Motorola Labs, Tempe, Arizona, where he is currently a distinguished member of the technical staff. He also teaches at Arizona State University as an adjunct faculty member. His research interests include machine learning and intelligent systems.



Jose C. Principe is a Distinguished Professor of Electrical and Biomedical Engineering at the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANN's) modeling. He is the BellSouth Professor and founder and director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). He has been involved in biomedical signal processing, in particular the electroencephalogram (EEG), and the modeling and applications of adaptive systems. Dr. Principe is editor-in-chief of the *IEEE Transactions on Biomedical Engineering*, president-elect of the International Neural Network Society, and formal secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. He is also a member of the scientific board of the US Food and Drug Administration, and a member of the advisory board of the University of Florida Brain Institute. He has more than 90 publications in refereed journals, 10 book chapters, and over 190 conference papers. He has directed 39 PhD degree dissertations and 57 master's degree theses. He is a fellow of the IEEE.