Evaluation of Bayes, ICA, PCA and SVM Methods for Classification

V. C. Chen
Radar Division, US Naval Research Laboratory
4555 Overlook Avenue, S.W.
Washington DC 20375
USA

vchen@radar.nrl.navy.mil

ABSTRACT

In this paper, we introduce the basic concepts of some state-of-the-art classification methods, including independent component analysis (ICA), principal component analysis (PCA), Bayes method, and support vector machine (SVM) or kernel machine. We discuss their function in the classification and evaluate their performance for different applications.

1 STATISTICAL CLASSIFICATION

Classification means to resolve the class of an object, e.g., a ground vehicle vs. an aircraft. *Recognition* means to determine whether the ground vehicle is a truck, a school bus, or a tank. *Identification* means to identify the type or model of the target (T72 tank or M60 tank). *Statistical classification* utilizes the statistical *pattern recognition* method for classification, recognition and identification [1]. A *pattern* is a characteristic of an observation, such as a speech signal or a human face image. A structural characteristic extracted from a pattern is called a *feature*. It can be a distinctive measurement, a transformation, or a structural component. The process of converting a pattern to features is called *feature extraction*. Each pattern can be viewed as a point (or a vector) in the *feature space*. The best features are selected using a *feature selection* algorithm. The selected features should best represent the classes or best represent the distinction between classes. The dimensionality of the selected feature space can also be greatly reduced compared to the full feature space.

The statistical classification process based on the probability distributions of the feature vectors can be described as follows:

(1) First, define the classes of patterns:

$$(C_1, C_2, ...C_M)$$

(2) Then, extract and select the best features from a pattern:

$$x = (x_1, x_2, ..x_N)$$

(3) Then, specify or learn the conditional probability function of a feature vector x belonging to class C_i:

$$p(x|\ C_i)$$

(4) Then, chose a decision rule (Bayes rule, maximum likelihood rule, Neyman-Pearson rule, or other rules).

(5) Finally, find the decision boundaries.

Paper presented at the RTO SET Symposium on "Target Identification and Recognition Using RF Systems", held in Oslo, Norway, 11-13 October 2004, and published in RTO-MP-SET-080.

The complete statistical classification process, as shown in Figure 1, includes pre-processing of observed or sensed data (such as segmentation, noise removal, filtering, spatial or temporal localization, and normalization of patterns), feature extraction, feature selection, learning, and classification. Feature extraction is accomplished with the principal component analysis (PCA) or independent component analysis (ICA). Then, in feature selection, the methods used include branch and bound search (B & B), sequential forward selection (SFS), sequential backward selection (SBS), sequential forward floating search (SFFS) and sequential backward floating search (SBFS). Finally, learning and classification are accomplished with Bayes classifier, k-nearest neighbor (k-NN) classifier, linear discrimination classifier (LDC) and support vector machine (SVM) as indicated in Figure 1.

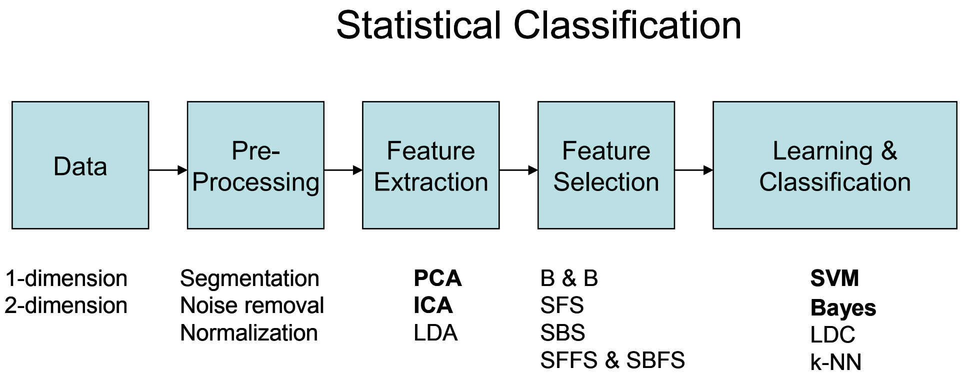

Figure 1. Basic stages of the statistical classification process.

2 FEATURE EXTRACTION

2.1 Feature Extraction and Dimensionality Reduction

Feature extraction converts data patterns to features, which are condensed representations of patterns and contain only salient information (as shown in Figure 2). The converted features should represent patterns with minimal loss of the information required for best classification. Features include non-transformed structural characteristics, transformed structural characteristics, and structures (such as lines, slopes, corners, or peaks). Non-transformed structural characteristics are obtained directly from sensor observations such as amplitudes, phases, time durations, or moments. Transformed structural characteristics are obtained from transformations such as the Fourier transform, wavelet transform, time-frequency transform, singular value decomposition, or Karhunan-Loeve transform.

Linear transforms, such as PCA and linear discrimination analysis (LDA), are widely used for feature extraction and dimensionality reduction. PCA is the best-known unsupervised linear feature extraction algorithm; it is a linear mapping which uses the eigenvectors with the largest eigenvalues. LDA is a supervised linear mapping based on eigenvectors, and it usually performs better than PCA for classification. ICA [2-4] is also a linear mapping but with iterative capability, which is suitable for non-Gaussian distributions. ICA decomposes a set of features into a basis whose components are statistically independent. It searches for a linear transformation W_{ICA} (or weight matrix) to express a set of feature vectors $X = (x_1, x_2, \dots x_N)$ as a linear combination of statistically independent vectors $Y = (y_1, y_2, \dots y_N)$, so that the transformed components $Y = W_{ICA}{}^T X$ are independent, that is, knowledge of the value of y_i provides no information on the value of y_j for $i \neq j$. There is no closed form solution for finding the weight matrix W_{ICA}. Therefore, iterative algorithms have been proposed to search for a weight matrix. PCA only requires that the coefficients y_i and y_j be uncorrelated, i.e.

$$\text{cov}(y_i, y_j) = E\{y_i, y_j\} - E\{y_i\}E\{y_j\} = 0$$

However, independence is a stronger requirement, because independent components are uncorrelated, but uncorrelated components may not be independent. Thus, the ICA accounts for higher order statistics and provides a more powerful data representation than PCA.

Kernel PCA is a nonlinear feature extraction method based on eigenvectors, which maps input patterns into a new feature space through a nonlinear function, and then performs a linear PCA in the mapped space.

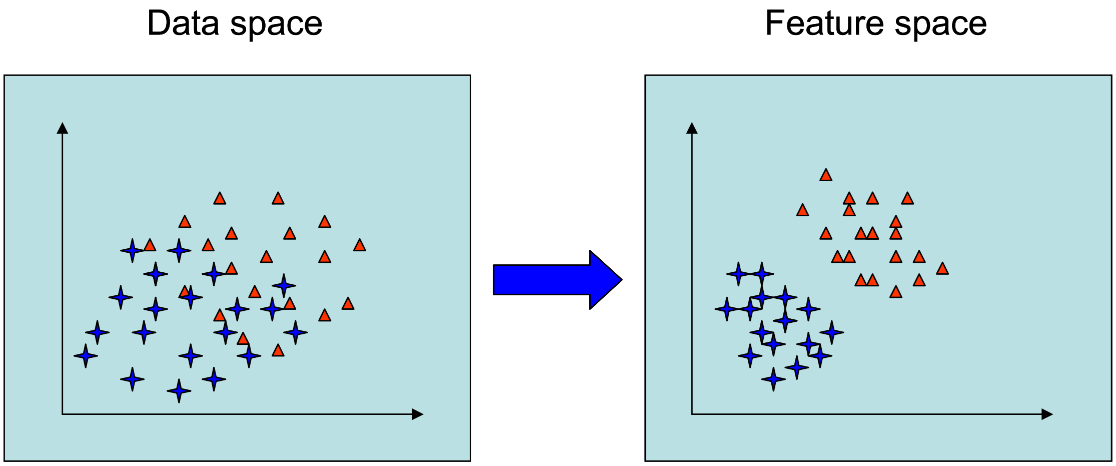

Figure 2. Feature extraction converts data pattern space to feature space.

2.2 PCA vs. ICA

PCA is a classical projection method used in signal analysis. ICA was originally used for separating mixed signals into independent components; this process is called *blind source separation* (BSS). The goal of PCA is to minimize the projection error, but the goal of ICA is to minimize the statistical dependence between basis feature vectors. Recently, ICA has been applied to image analysis. Some results show that ICA outperforms PCA, and others show that there is not much performance difference between ICA and PCA. We should realize that the nature of our classification task affects the evaluation. For some classification tasks, if the global properties such as width and length are more important, then they are more easily extracted by PCA than ICA. If features such as time-frequency signatures are more spatially localized, ICA is better than PCA. For small ship classification, the global features are more important than localized spatial features, as illustrated in Figure 3. Thus, PCA is good enough for feature extraction. However, for micro-Doppler time-frequency signatures, the localized spatial features are more important, and the ICA should be used in feature extraction as illustrated in Figure 4.

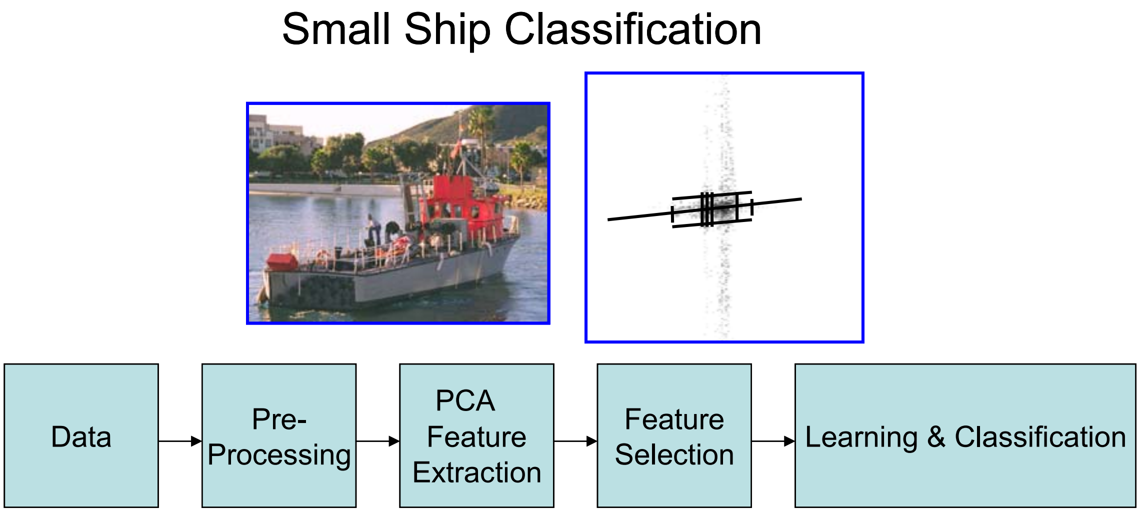

Figure 3. Using PCA for small ship feature extraction.

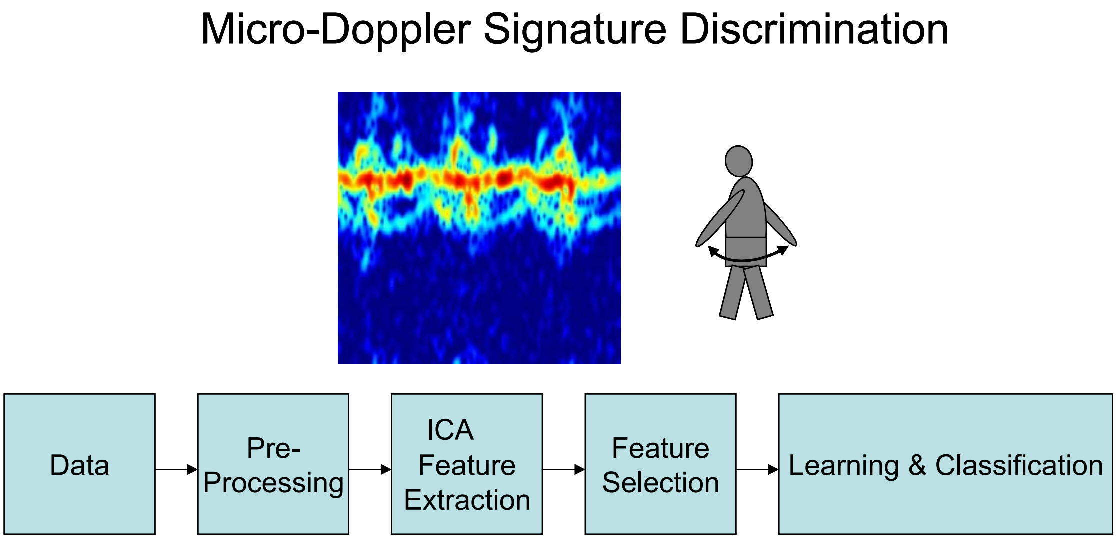
Micro-Doppler Signature Discrimination
Data
Pre-Processing
ICA Feature Extraction
Feature Selection
Learning & Classification

Figure 4. Using ICA for micro-Doppler time-frequency signatures discrimination.

3 FEATURE SELECTION

The purpose of feature selection is to determine a subset within the set of features in order to minimize the classification error based on various criteria [1]. A straightforward method of feature selection is the exhaustive search that selects the best features and minimizes the classification error. Another efficient feature selection method is the *sequential forward and backward selection* (SFBS). Forward selection means a bottom up process that begins with an empty set and selects the first feature that is the best feature. Then, at each step, it selects the best feature from the remaining set, which, combined with the features already selected, gives the best value under the selection criterion. Backward selection is a top down process which removes features from the feature set. However, it cannot re-select those removed features even if they would be useful for further processing.

Suppose there is a set of N features represented by Y and M features of a subset represented by X. Let $J(X)$ to be a *criterion function* for selecting X from Y. Then, the selection procedure can be summarized as (1) searching to find all possible subsets of size M from N features and (2) selecting the subset X with the largest value of $J(X)$ as the optimal subset. Most selection methods use the classification error of a selected feature subset to evaluate the effectiveness of the selection method.

The searching methods include:

(1) Exhaustive search.
(2) Branch and bound search (B & B): The criterion function is monotonic and the performance of a subset can be improved when adding a feature to it.
(3) Sequential forward selection (SFS): Evaluate a feature set by adding one feature at a time. Once a feature is added, it cannot be discarded.
(4) Sequential backward selection (SBS): Evaluate a feature set by deleting one feature at a time. Once a feature is deleted, it cannot be re-entered into the feature subset.
(5) Sequential forward floating search (SFFS) and sequential backward floating search (SBFS): Backtrack as long as there are improvements of the current feature set compared to the previous feature set. Performance is comparable to the B & B method with a lower computational cost.

4 LEARNING AND CLASSIFIER

The effectiveness of the feature space depends on how well different classes can be separated in the space. The objective of classification is to find decision boundaries between classes in the feature space that can best separate different classes. These decision boundaries are determined by the probability distributions of the patterns associated with each class. The probability distributions can be either specified or learned, i.e., boundaries can be found by either specifying the parametric format of the boundaries (such as linear or quadratic) or by finding them by learning through a training process.

The performance of a classifier depends on the number of available training samples. Learning includes supervised learning and unsupervised learning. Supervised learning requires that the training samples be labelled by their classes. Unsupervised learning does not require labelled training samples and the number of classes must be learned.

Classical classification methods include the Bayes, k-NN, LDC, and others. *Support vector machine* (SVM) is a modern classification method with a nonlinear classification function using an iterative method [5-7]. It can maximize the margin between the classes by selecting a minimum number of support vectors.

4.1 Bayes Classifier

The Bayes classifier assigns a pattern to the class that has the maximum estimated posterior probability. Given a pattern x, the posteriori conditional probability that the pattern belongs to the class C is determined by

$$P(C \mid x) = \frac{P(x \mid C) P(C)}{P(x)},$$

where $P(x)$ is the a priori probability that a pattern is x, $P(C)$ is the a priori probability that a pattern belongs to class C, and $P(x|C)$ is the conditional probability that a pattern is x if the pattern belongs to class C.

According to the Bayes rule, assign a pattern x to class C_i if the risk function, given by

$$risk(C_i \mid x) = \sum_{j=1}^{M} l(C_i, C_j) P(C_j \mid x)$$

is minimum, where $l(C_i,C_j)$ is the loss function when C_i is chosen if the true class is C_j, and $P(C_j |x)$ is the posterior probability of C_j. The Bayes classifier has the minimum classification error when the probability density functions are known.

4.2 k-Nearest Neighbor (k-NN) Classifier

The 1-Nearest Neighbor rule assigns a pattern to the class of the nearest training pattern without a training process. The classifier using the k-NN rule assigns a pattern to the majority class among k nearest neighbor.

4.3 Linear Discrimination Classifier (LDC)

Assume x_i is a feature vector with d dimensions, and $X = (x_1, x_2, \dots x_N)$ is the training set with N classes. Given a transformation matrix W, the original feature vector is transformed to a projection feature vector $Y = (y_1, y_2, \dots y_N)$ with a reduced dimension of d_l ($d_l < d$):

$$Y = W^T X .$$

Define a scatter matrix S:

$$S = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T .$$

where μ is the mean target feature vector. The LDC uses the transformation matrix W_{LDC} that satisfies

$$W_{LDC} = \arg\max_{W} \frac{W^T S_{between} W}{W^T S_{within} W}$$

where the between-class scatter matrix is defined by

$$S_{between} = \sum_{i=1}^{K} N_i (x_i - \mu_i)(x_i - \mu_i)^T ,$$

and the within-class scatter matrix is defined by

$$S_{within} = \sum_{i=1}^{K} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T ,$$

where N_i is the number of training samples in class i, K is the number of distinct classes, μ_i is the mean vector of samples that belongs to class i, and X_i is the set of samples that belongs to class i. To reduce the dimensionality, the LDC should apply the PCA first.

4.4 Support Vector Machine

SVM is an unsupervised approach based on statistical learning theory. It estimates the optimal boundary in the feature space by combining a maximal margin strategy with a kernel method; this process is called a *kernel machine*. The machine is trained according to the *structural risk minimization* (SRM) criterion [5,6]. The decision boundaries are directly derived from the training data set by learning.

The SVM maps the inputs into a high-dimensional feature space through a selected kernel function. Then, it constructs an optimal separating hyper-plane in the feature space. The dimensionality of the feature space is determined by the number of support vectors extracted from the training data (see Figure 3). The SVM can locate all the support vectors, which exclusively determine the decision boundaries. To estimate the misclassification rate (risk), the so called *leave-one-out* procedure is used. It removes one of N_i training samples, performs training using the remaining training samples, and tests the removed sample with the newly derived hyperplane. It repeats this process for all of the samples, and the total number of errors becomes the estimation of the risk.

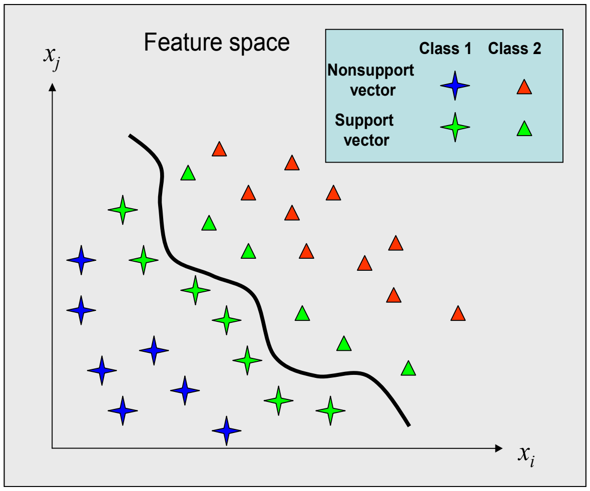

Figure 3. Optimal boundary serached by the SVM.

4.5 Classifier Evaluation

Pattern recognition software packages and toolboxes are widely available [8]. To evaluate different classifiers, we create a 2 dimensional dataset generated with 100 samples for each class. Among the 100 samples, 20 samples are used for training and 80 samples are for testing. Four classifiers (LDA, k-NN, Bayes, and SVM) are compared. Note that here the data is generated by a random generator.

In the first example, two classes of samples are overlapped in the 2-D feature space. Figure 4 shows the decision boundaries found by (a) LDC, (b) Bayes, (c) k-NN and (d) SVM. Figure 5 shows classification error rates or the receiver operation curve (ROC) of the corresponding four classifiers.

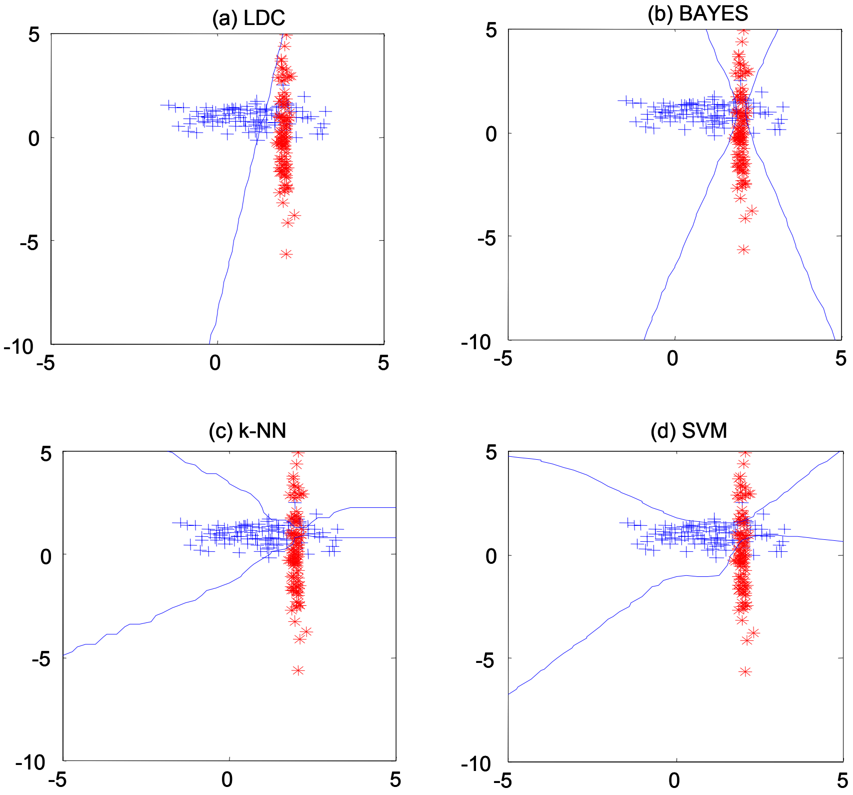

Figure 4. Classifier boundary found by (a) LDC, (b) Bayes, (c) k-NN, and (d) SVM for two overlapped classes in 2-D feature space.

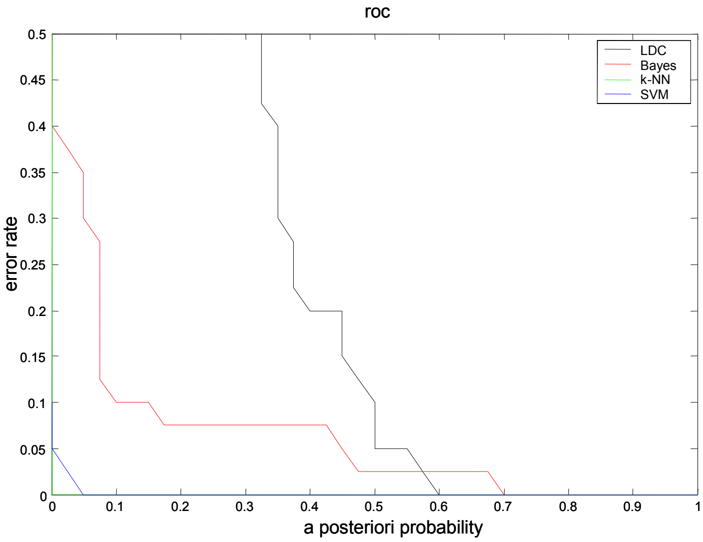

Figure 5. Two classes' classification error rates for the four classifiers.

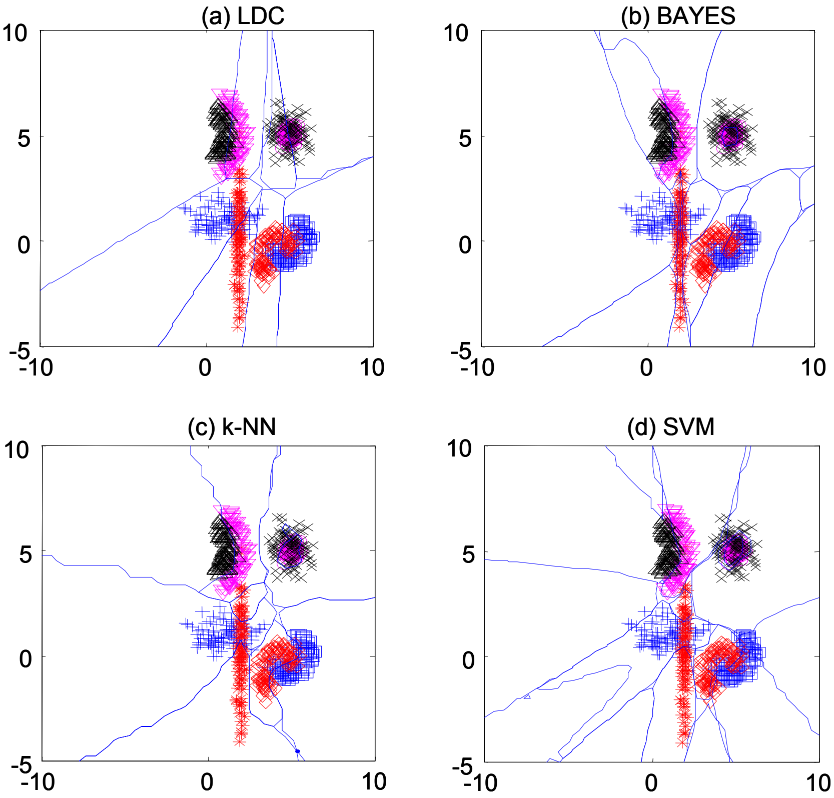

Figure 6. Classifier boundary found by (a) LDC, (b) Bayes, (c) k-NN, and (d) SVM for eight overlapped classes in 2-D feature space.

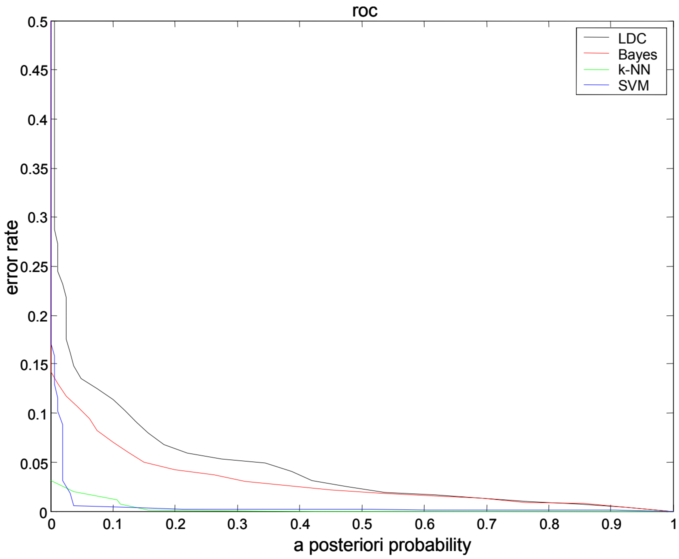

Figure 7. Eight classes' classification error rates for the four classifiers.

The second example is eight mixed classes of samples overlapped in the 2-D feature space. Figure 6 shows the decision boundaries found by (a) LDC, (b) Bayes, (c) k-NN and (d) SVM. Figure 7 shows the classification error rates of the corresponding four classifiers.

From the above two examples, we see that SVM can find more complicated decision boundaries, and the classification errors of k-NN and SVM are considerably lower than others.

5 SUMMARY

We have introduced the basic concept of ICA, PCA, Bayes, and SVM, and we have discussed their functions in classification and evaluated their performances for different applications. If global properties are more important, then these features are more easily extracted by PCA than ICA. If the features are more localized, ICA is better than PCA. For small ship classification, where global features are more important than localized spatial features, PCA is good enough for feature extraction. However, for micro-Doppler time-frequency signatures, the localized spatial features are more important, and ICA should be used in feature extraction. From two simulated examples, we see that SVM can find more complicated decision boundaries, and the classification errors of k-NN and SVM are considerably lower than the others.

6 ACKOWLEDGEMENT

This work was supported by ONR, through the NRL base program.

REFERENCES

1. A.K. Jain, R.P.W. Duin and J. Mao, "Statistical pattern recognition: A review", *IEEE Trans. on PAMI*, vol.22, no.1, pp.4-37, Jan. 2000.
2. A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and Applications", *Neural Networks*, 13(4-5), pp.411-430, 2000.
3. A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution", Neural Computation, 7, pp.1129-1159, 1995.
4. P. Comon, "Independent component analysis: A new concept?" Signal Processing, 36(3),pp.287-314, 1994.
5. C.J.C. Burges, "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, 2, pp.121-167, 1998.
6. V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, Inc., New York, 1998.
7. Q. Zhao and J. Principe, "Support vector machines for SAR automatic target recognition", *IEEE Trans. on AES*, vol.37, no.2, pp.643-654, 2001.
8. R.P.W. Duin, *PRTools – A Matlab Toolbox for Pattern Recognition*, http://www.ph.tn.tudelft.nl/prtools.