Class-Preserving Manifold Learning for Detection and Classification

Puoya Tabaghi
Department of Electrical and Computer Engineering
Colorado State University
Email: tabaghi@engr.colostate.edu

Mahmood R. Azimi-Sadjadi
Department of Electrical and Computer Engineering
Colorado State University
Email: azimi@engr.colostate.edu

Abstract—This paper proposes a supervised approach for analysis of high-dimensional data using low-dimensional submanifolds. This method offers many useful properties. Using first order approximation for the given nonlinear mapping, we introduce a locally linear model. This model is such that it minimizes the local approximation error resulted by mapping to a local subspace during the learning. Additionally, the proposed method preserves local data energy to conserve local topology. Finally, this method guarantees the separability of the mapped data for different data classes. Two different approaches used for this aim, Linear Discriminant Analysis (LDA) and Regularized Maximum Margin Criterion (RMMC). Having those local feature-domain data, the whole feature domain data can be estimated in MMSE sense. The performance of this method is demonstrated on a sonar imagery dataset for classification of underwater objects.

I. INTRODUCTION

Automatic detection and classification of underwater targets from sonar imagery is a challenging task due to many factors such as: (a) variation of the target returns and features with respect to aspect angle, target range, vehicle elevation and grazing angle; (b) diverse sizes, shapes, compositions and scattering properties of the targets; (c) competing natural and man-made clutter, and (d) highly variable environmental and operating conditions. Thus, it is desirable to devise automatic target detection and classification methodologies that remain robust to such changes in the sonar data. This paper explores the idea of developing novel nonlinear dimensionality reduction algorithms to map the high-dimensional synthetic aperture sonar (SAS) image snippets of different underwater objects into low-dimensional submanifold features that preserve the class separability of the objects in question.

There exist many standard manifold learning methods such as Isometric Feature Mapping (ISOMAP) [1], Locally Linear Embedding (LLE) [2], Laplacian Eigenmaps [3] and Maximum Variance Unfolding (MVU) [4] that attempt to preserve certain neighboring properties in both input and feature domains. However, most of these methods require a large number of data points to form the submanifold and are not necessarily class-preserving. On the other hand, the local manifold learning methods, such as Locally Smooth Manifold Learning (LSML), attempt to parameterize the underlying low-dimensional submanifold based on locally linear approximation [5],[6]. Supervised manifold learning methods, such as supervised LLE (SLLE) [7] and those based on supervised ISOMAP [1], on the other hand, attempt to overcome some of the common issues with the standard manifold learning methods. This paper introduces a new approach to analyze high-dimensional sonar data using low-dimensional submanifold with the goal of preserving the object class. The supervised method uses a linear approximation of its neighboring points while increasing the separation of the different classes by maximizing the between-class scattering matrix of the mapped data while minimizing the inter-class scattering. If this is done properly, all the variations (e.g., w.r.t to pose, grazing angle, etc.) of a particular target type would lead to an orbit (or a trace) in the submanifold domain representing that object. Each target type would have its unique orbit in the submanifold feature space. The class of a new object is then determined based upon closeness to a particular point of an orbit. This manifold-based inference approach can also, in future, provide a mechanism to transverse between the input space and feature space and vice versa which helps us to classify the out-of-sample data points. We shall demonstrate the usefulness of the proposed method on a sonar dataset consisting of synthetically generated targets superimposed on a real sonar background.

II. FORMULATION OF THE PROBLEM

Assume we have a smooth $d$-dimensional manifold that lies in a $D$-dimensional input data space ($D \gg d$). The input data matrix is $X = [x_1 \cdots x_N]$, where $x_i \in \mathbb{C}^D$ and $d < N \ll D$. Our goal is to find a mapping of each point on the manifold with the property that each point can generate its neighbors by a linear transformation. Similar to [5], we denote $W(x_i, \epsilon_i)$ as a...
transformation of \( x_t \) to its neighboring data point in the input space. The vector \( \mathbf{e}_t \in \mathbb{C}^d \) acts as radius of freedom of the transformation in the feature space according to the formula

\[
M : M(y_t + \mathbf{e}_t) = W(x_t, \mathbf{e}_t), \quad \text{where } y_t = M^{-1}(x_t).
\]

Taking the first order approximation of \( W \) in the neighborhood of \( x_t \), we'll have

\[
W(x_t, \mathbf{e}_t) = M(y_t + \mathbf{e}_t) \approx M(y_t) + \frac{\partial}{\partial y} M(y_t) \mathbf{e}_t = x_t + H_t \mathbf{e}_t,
\]

where each column of the matrix \( H_t \) is the partial derivative of \( M \) with respect to \( y_t \). Thus, our goal would be learning the set of matrices \( H_t \in \mathbb{C}^{d \times d} \) for each data point \( x_t \) and its neighboring points, \( x_j \)'s, such that

\[
x_t + H_t \mathbf{e}_{tj} \approx x_j, \quad \text{in the least } l_2 \text{ norm square sense, or equivalently } H_t \mathbf{e}_{tj} = x_j - x_t, \quad \text{where the neighborhood radius } \mathbf{e}_{tj} = y_j - y_t.
\]

Thus, the problem is to reduce

\[
\hat{x}_j - x_j = H_t(y_j - y_t) - (x_j - x_t), \quad \forall x_j \in N_k(x_t)
\]

in mean-squared sense, i.e. minimize

\[
J(t) = \sum_{x_j \in N_k(x_t)} \|H_t(y_j - y_t) - (x_j - x_t)\|_2^2
\]

(1)

where \( N_k(x_t) \) is the set all \( K \)-nearest neighbors of \( x_t \). In order to weight each term in this objective function according to the distance of \( x_t \) to its neighbors \( x_j \)'s, we can modify the objective as,

\[
J(t) = \sum_{j=1}^{N} \|k(x_t, x_j)(H_t(y_j - y_t) - (x_j - x_t))\|_2^2
\]

(2)

where \( k(x_t, x_j) \) is a radial-basis weighting function with the property,

\[
\left\{ \begin{array}{ll}
  k(x_t, x_j) & \approx 1 \quad \text{for } x_j \in N_k(x_t) \\
  k(x_t, x_j) & \approx 0 \quad \text{for } x_j \notin N_k(x_t)
\end{array} \right.
\]

There are various choices for this weighting function, though here we choose the commonly used Gaussian function,

\[
k(x_t, x_j) = e^{-\frac{\|x_t - x_j\|^2}{\sigma^2}}
\]

where \( \sigma \) must be found experimentally.

III. MODELING THE SMOOTH LOCAL MANIFOLD

It would be beneficial to reformulate the objective function in (2) in terms of centered data matrices of neighboring points of \( x_t \) and \( y_t \). More specifically, we define

\[
X_t = (X - x_t x_t^T)K(x_t)
\]

\[
Y_t = (Y - y_t y_t^T)K(x_t)
\]

where \( K(x_t) = diag(k(x_t, x_1), ..., k(x_t, x_N)) \), 1 represents a one-vector, and \( (.)^T \) denotes the transpose operation. Thus, by using these notations, for the data point \( x_t \) the error function in (2) becomes

\[
J(t) = \text{tr}((H_t Y_t - X_t)^H (H_t Y_t - X_t))
\]

where \((.)^H\) denotes the Hermitian operator and \(\text{tr}(.)\) is the trace operator. Using pseudo-inverse of \( H_t \) the best estimate, in MMSE sense, of the \( y_t \) can be calculated as follow,

\[
Y_t = (H_t^* H_t)^{-1} H_t^* X_t
\]

Alternatively,

\[
J(t) = \text{tr}(X_t^H (I - P_{H_t}) X_t)
\]

where \( P_{H_t} \) is the projection matrix into subspace spanned by the columns of \( H_t \), i.e. \( P_{H_t} = H_t (H_t^* H_t)^{-1} H_t^* \).

To minimize the reconstruction error w.r.t. \( H_t \), we apply singular value decomposition (SVD) to \( X_t \),

\[
X_t = U_t \Sigma_t V_t^H
\]

where \( U_t \in \mathbb{C}^{d \times d} \) and \( V_t \in \mathbb{C}^{N \times N} \) contain the left and right singular vectors of \( X_t \) and \( \Sigma_t \in \mathbb{C}^{d \times d} \) is a rectangular diagonal matrix of its singular values. Similarly, if we apply SVD to \( H_t \) we have,

\[
H_t = R_t \Lambda_t E_t^H
\]

with the corresponding singular vector matrices \( R_t \in \mathbb{C}^{d \times d} \) and \( E_t \in \mathbb{C}^{d \times d} \) and rectangular diagonal matrix of the singular values \( \Lambda_t \in \mathbb{C}^{d \times d} \). It can readily be shown that at the optimum solution \( U_t = R_t \) and the minimum error would be,

\[
J^*(t) = \text{tr}(\Sigma_t^H \Sigma_t (I - I_N^d))
\]

(3)

and hence

\[
P_{H_t} = U_t I_d \Sigma_t V_t^H
\]

(4)

Additionally, \( Y_t \) can be given by

\[
Y_t = E_t (\Lambda_t^H \Lambda_t)^{-1} \Lambda_t^H \Sigma_t V_t^H
\]

(6)

which implies that \( \Lambda_t \) and \( E_t \) are yet to be found in order to compute \( Y_t \).

IV. PRESERVING LOCAL DATA ENERGY

Due to the unit variance constraint in some of nonlinear dimensionality reduction methods, e.g., LLE, the global metric information is completely lost \([13]\). Retaining the global metric in the feature space is a necessity. Preserving the local energy leads to local topology conservation. Using this idea we try to find \( \Lambda_t \) by forcing our mapping to preserve the local data energy by the following constraint:

\[
\|Y_t\|_F^2 = \|X_t\|_F^2,
\]

where \( \|\cdot\|_F \) is the Frobenius norm. Using cyclic property of trace operator, we will have

\[
\text{tr}(\Sigma_t^H (\Lambda_t \Lambda_t^H)^{-1} \Sigma_t) = \text{tr}((\Sigma_t^H \Sigma_t),
\]

where \((\Lambda_t \Lambda_t^H)^+\) is the Moore–Penrose pseudo-inverse of the matrix \( \Lambda_t \Lambda_t^H \). One solution leads to \( \Lambda_t = \lambda_t I_{d \times d} \), where \( \lambda_t \) is given by

\[
\lambda_t = \frac{\|I_{d \times d} \Sigma_t\|_F}{\|\Sigma_t\|_F}
\]

(7)

Thus, so far we have \( Y_t \) and \( H_t^* \) as follows:

\[
Y_t = E_t \lambda_t^{-1} I_{d \times d} \Sigma_t V_t^H
\]

(8)

\[
H_t^* = U_t \lambda_t I_{d \times d} E_t^H
\]

(9)

We can interpret the objective function in the following form:

\[
\sum_{i\neq t} k^2(x_t, x_i)\|x_i - x_t\|_2^2 = \sum_{i\neq t} k^2(x_t, x_i)\|y_i - y_t\|_2^2
\]

which is similar to the idea of Curvilinear Component Analysis \([14]\) for preserving local topology of data.
V. CLASS-PRESERVING FORMULATION

So far, we have developed a model which has the minimum approximation/reconstruction error and preserves the local Gram matrix of each data point in $X_t$. Now, $E_t$ is the only unknown parameter for each data point $Y_t$ that needs to be identified. $E_t$ can be considered as a dictionary matrix that generates the data points in $Y_t$, if one can find a global dictionary, say $E$, that can generate any data matrix $Y_t$, $t = 1, ..., N$, then

$$Y_t = E \lambda_t^{-1} I_{d \times D} \Sigma_t V_t^H.$$  (10)

Here we propose two different methods for finding $E$. Let us start by denoting the selection matrix $S_t$ such that

$$X_t = (X - x_t1^T)K(x_t) = X(I - S_t)K(x_t),$$

that is, $XS_t = x_t1^T$. Clearly, the same matrix can also be used for $Y_t$, i.e.,

$$Y_t = (Y - y_t1^T)K(x_t) = Y(I - S_t)K(x_t).$$

By stacking all $Y_t$, $t \in [1, N]$ into a big matrix,

$$[Y_1, ..., Y_N] = Y([I - S_t]K(x_1), ..., ([I - S_N]K(x_N)),$$

we get,

$$E \Theta = YW,$$  (11)

where $\Theta = [\theta_1, ..., \theta_N]$ for $\theta_t = \lambda_t^{-1} I_{d \times D} \Sigma_t V_t^H$ and $W = ([I - S_t]K(x_1), ..., ([I - S_N]K(x_N))$. To solve (11), we minimize the $l_2$ norm squared error of the mapped low dimensional data,

$$\min_Y \|YW - E \Theta\|_2^2.$$  (12)

It can be easily shown that the optimal solution is of the form

$$Y = E \Theta W^+ = E \Phi$$  (13)

where $W^+$ is the Moore–Penrose pseudo-inverse of the sparse matrix $W$ and $\Phi = \Theta W^+$. The goal of the rest of this section is to find $E$ such that the mapped input data points $x_t$ improve the class separation in the corresponding d-dimensional feature space. To accomplish this goal, we use two different approaches one using the Linear Discriminant Analysis (LDA) and the other using regularizing Maximum Margin Criterion (RMMC). These are explained next.

1) LDA: LDA is a supervised learning method [15], which utilizes the category information associated with each sample. The goal of LDA is to maximize the between-class scatter measure while minimizing the within-class scatter measure. Mathematically speaking, the within-class scatter matrix $S_w$ and the between-class scatter matrix $S_b$ are defined as

$$S_w = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (y_{(i,j)} - \mu_{(j)})(y_{(i,j)} - \mu_{(j)}),$$  (14)

$$S_b = \sum_{j=1}^{C} N_j(\mu_{(j)} - \mu)(\mu_{(j)} - \mu)^H,$$  (15)

where $y_{(i,j)}$ is the i-th data point of the j-th class, $\mu_{(j)}$ is the mean of the j-th class, $\mu$ is the mean of all the data, $N_j$ is the number of data points in j-th class and $C$ is the number of different classes. By stacking different data points of all classes we get

$$Y = [Y^{(1)}, ..., Y^{(C)}] = E[\Phi^{(1)}, ..., \Phi^{(C)}]$$  (16)

where $Y^{(i)}$ is the data points of the i-th class and $\Phi^{(i)}$ is its corresponding parameter matrix. The between-class scattering and the within-class scattering matrices are then given by

$$S_w = ER_wE^H$$

$$S_b = ER_bE^H$$

where $R_w$ and $R_b$ are

$$R_w = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (\Phi^{(j)}(i) - \frac{1}{N_j}\Phi^{(j)}1)(\Phi^{(j)}(i) - \frac{1}{N_j}\Phi^{(j)}1)^H$$

$$R_b = \sum_{j=1}^{C} \sum_{i=1}^{N_j} \Phi^{(j)}1 - \frac{1}{N}\Phi1(\Phi^{(j)}1 - \frac{1}{N}\Phi1)^H$$

where $\Phi^{(j)}(i)$ is the i-th column of the matrix $\Phi^{(j)}$. The goal here is to select mapping matrix $E$ in order to maximize the ratio $\frac{\det S_b}{\det S_w}$. If $R_w$ is nonsingular then this ratio is maximized when the dictionary matrix $E$ consists of $d$ generalized eigenvectors corresponding to the $d$ largest eigenvalues of $R_w^{-1}R_b$. Note that there are at most $C-1$ nonzero generalized eigenvalues, and hence an upper bound on $d$ is $C-1$. That is,

$$\max_E \frac{\det S_b}{\det S_w} = \max_E \frac{\det ER_bE^H}{\det ER_wE^H}$$

leads to an eigenvalue problem in which the columns of $E$ are the eigenvectors of $R_w^{-1}R_b$ matrix.

2) RMMC: Maximum Margin Criterion (MMC), aims at finding the optimal linear transformation by increasing between-class scattering and decreasing within-class scattering such that the class structure of the original high-dimensional space is preserved in the transformed lower dimensional space [10]. This method can extract the most discriminatory features and does not suffer from the small sample size problem. In addition, it does not need to use inverse operation as with LDA [9]. We use the regularized version of MMC which yields the objective function,

$$\max_E \frac{\det S_b}{\det S_w} = \max_E \frac{\det ER_bE^H}{\det ER_wE^H}$$

leads to an eigenvalue problem, in which the columns of $E$ are the eigenvectors of $R_w^{-1}R_b$ matrix.

VI. CLASSIFICATION RESULTS AND DISCUSSIONS

In this section, we apply the developed manifold learning method in conjunction with a two-class classifier to a sonar dataset for mine-like versus non-mine-like discrimination. This sonar dataset is described next.
A. Data Description

The dataset used in this study is a sonar image database containing actual Synthetic Aperture Sonar (SAS) images of the seafloor with synthetically generated targets inserted into the images. The dataset consists of a high-resolution high frequency (HF) sonar image as well as a broadband (BB) sonar image co-registered over the same region of the seafloor. The HF channel provides higher spatial resolution and better ability to capture target characteristics while BB sonar has the advantage of much better clutter suppression. The images in this dataset consist of complex-valued pixels and are the direct result of the k-space beamformer.

The image database consists of 145 pairs of HF and BB complex valued sonar images. Each HF image of size $1800 \times 1992$ pixels has a corresponding BB image of size $1800 \times 4150$ meaning there are two images for each data collection scenario characterized by its background and object configuration though here only the HF images are used.

Each image pair uses one of 29 different backgrounds that are real (magnitude of the beamformed data) images of the seafloor and contains the synthetically generated signatures of four different objects that are superimposed on the background image according to one of five different configurations that define the orientations and relative positions of the objects. Each pair of HF and BB image snippets, which are of size $120 \times 240$ and $120 \times 480$ respectively, contains one of each of the following four target types: Block, Cone, Sphere, and Cylinder. Image backgrounds range in complexity from very smooth, where object highlights and shadows are easy to visually distinguish from the background, to highly cluttered, where objects may be very easy to confuse with background when visually inspected. The signatures of each object are synthetically generated and are placed at various aspect angles ranging from $25\text{o}$ to $75\text{o}$ and various ranges or cross-track locations from 12m to 40m. A detector was first applied to examine every region of interest (ROI) in the pair of HF and BB images and localize the areas which could contain potential targets. This detector is based upon the broadband coherence method developed in [12]. The goal of our classifier (see the next section) is then to assign mine-like versus non-mine-like labels to those detected image snippets to further reduce the false alarms. For the purpose of this study, the mine-like class consists of Cylinder and Cone while the non-mine-like class consists of Sphere, Block, and any natural clutter (e.g., coral and rock) detected by the aforementioned detector. Figures 2 and 3 give examples of several target and non-target HF snippets, respectively, while Figures 4 and 5 display the corresponding BB snippets. From the images in Figures 2 and 3 one can see that the HF snippets are typically characterized by a highlight region corresponding to a strong sonar return from the object itself immediately followed by a shadow region. Likewise, one can see from Figures 4 and 5 that the BB snippets also exhibit a strong highlight region but generally lack a noticeable shadow structure.

B. Classifier

Here, it is assumed that the distribution of the data in the low-dimensional submanifold domain is multivariate normal under both classes, i.e. $y|_{H_1} \sim N(\mu_i, C_i)$ where $\mu_i$ and $C_i$ are the sample mean vector and sample covariance matrix under hypothesis (or class) $H_1$. As mentioned before, this classifier decides if the data contains Target (Cone or Cylinder) or Non-Target (Sphere, Block, competing Clutter). Thus, the likelihood ratio (LR) for this test is

$$L(y) = \frac{p_Y(y|H_1\{\text{Cylinder, Cone}\})}{p_Y(y|H_0\{\text{Sphere, Block, Clutter}\})} \geq_{NT} \gamma$$

(17)

Using the normality assumption, the Log-Likelihood Ratio (LLR) test becomes

$$l(y) = (y-\mu_0)^H C_0^{-1}(y-\mu_0) - (y-\mu_1)^H C_1^{-1}(y-\mu_1) \geq_{NT} \gamma'$$

(18)

This classifier is then applied to the dataset described earlier. More specifically, the two class-preserving manifold learning methods, i.e. CPML-LDA and CPML-RMMC are applied to the detected image snippets to reduce the dimension of the data to only those features that are critical for this classification problem. The results are given in the next section.

C. Classification Results and Discussions

The training dataset for the classifier consists of 5 image snippets for each object type at different configuration (aspect, range, etc.) and 6 different background conditions (2 easy, 2 medium, and 2 difficult) while the test dataset consists of 5 image snippets for each object type and 12 background cases (4 easy, 4 medium and 4 difficult). For the CPML-LDA, the dimension of the mapped data is upper bounded by $d = 4 < C$ where $C = 5$ the number of different objects encountered; whereas for CPML-RMMC this dimension is not restricted. Thus, we choose $d = 10 > C - 1$ as we want to reduce the reconstruction error mentioned earlier. The remaining issue
with these algorithms is to find the optimum value for $\sigma$. Choosing a large value for $\sigma$ sacrifices the smoothness of the manifold which violates the model formulations. The optimal value of $\sigma^2$ was chosen experimentally. For CPML-RMMC $\sigma^2_{opt}$ was found to be $0.67D$ and $0.2D$ for BB and HF data, respectively where $D$ is the dimension of data points. While for CPML-LDA this was $0.2D$ and $2.2D$ for BB and HF data, respectively.

Figure 6 shows the receiver operating characteristic (ROC) curve for the CPML-LDA method ($d = 4$) which gives correct classification rate $P_{cc}$ versus false alarm rate $P_{fa}$. The classification performance at the knee-point of ROC curve, where $P_{cc} + P_{fa} = 1$, gives $P_{cc} = 85\%$ and $P_{fa} = 15\%$ for BB images and $P_{cc} = 87\%$ and $P_{fa} = 13\%$ for HF images. Figure 7, on the other hand, gives the ROC curve for CPML-RMMC method for $\alpha = 0.2$ and $d = 10$. This ROC curve exhibits almost identical performance at the knee point showing $P_{cc} = 85\%$ and $P_{fa} = 15\%$ for BB images and $P_{cc} = 86.6\%$ and $P_{fa} = 13.4\%$ for HF images while using larger dimension for the mapped data. However, although both methods exhibit somewhat identical classification performance at the knee point of ROC curve, clearly for a fixed $P_{fa}$ (e.g., 20\%) the ROC of the CPML-LDA method on the HF data is much better than that of the CPML-RMMC; whereas for the BB data this is reversed.

**Fig. 6.** ROC curve for CPML-LDA with $d = 4$.

**Fig. 7.** ROC curve for CPML-RMMC with $d = 10$ and $\alpha = 0.2$.

**VII. CONCLUSIONS**

This paper proposes a class-preserving manifold learning method (or CPML) method for data representation and inference. The proposed method achieves the minimum local manifold distortion. This property helps us to overcome the global metric information loss issue that generally exist in almost all spectral embedding methods. Experiments on the synthesized sonar datasets show that the performance of proposed methods for feature extraction and classification purposes directly depends on the local cost function and the dimensionality of the input data. We showed that both CPML-LDA and CPML-RMMC method provides similar results because they both have similar local cost functions. The priority of the CPML-RMMC over CPML-LDA is that it does not use inverse operator and it does not impose any constraint on $d$. This method could easily further be modified for multiple dictionary model situations. Moreover, it can provide a new framework to map out-of-sample data points and transverse between the input and feature space.

**ACKNOWLEDGMENT**

This work was supported by the Office of Naval Research (ONR), Code 3210E under contract #N00014-14-1-0110.

**REFERENCES**


