LETTER

Semi-Supervised Nonparametric Discriminant Analysis*

Xianglei XING[†a)], ***Nonmember***, **Sidan DU**[†b)], ***Member***, ***and*** **Hua JIANG**[†], ***Nonmember***

SUMMARY We extend the Nonparametric Discriminant Analysis (NDA) algorithm to a semi-supervised dimensionality reduction technique, called Semi-supervised Nonparametric Discriminant Analysis (SNDA). SNDA preserves the inherent advantages of NDA, that is, relaxing the Gaussian assumption required for the traditional LDA-based methods. SNDA takes advantage of both the discriminating power provided by the NDA method and the locality-preserving power provided by the manifold learning. Specifically, the labeled data points are used to maximize the separability between different classes and both the labeled and unlabeled data points are used to build a graph incorporating neighborhood information of the data set. Experiments on synthetic as well as real datasets demonstrate the effectiveness of the proposed approach.

key words: *semi-supervised learning, nonparametric discriminant analysis, manifold learning*

1. Introduction

Dimensionality reduction plays an important role in information processing, pattern recognition and machine learning. Over the years, many dimensionality reduction techniques [1]–[3] have been proposed. From the perspective of machine learning, these dimensionality reduction methods can be classified into three categories: unsupervised learning, supervised learning, and semi-supervised learning.

Principal component analysis (PCA) [4], as a classical unsupervised algorithm, seeks the directions of maximum variance for optimal reconstruction. If the data is embedded in the linear subspace, PCA is guaranteed to discover the dimensionality of the subspace and produce a compact representation in the form of an orthonormal basis. However, for the data on a nonlinear embedded subspace, PCA has difficulty in discovering the underlying manifold structure. To discover the intrinsic manifold structure of the data, nonlinear dimension reduction algorithms such as locally linear embedding (LLE) [5] and Laplacian eigenmap (LE) [6] were developed. However, they are defined only on the training data points and they do not yield a method for mapping new test points. Locality preserving projections (LPP) [7] was developed to solve this problem. LPP utilizes linear projection function whose properties are similar to the nonlinear maps to project new data points.

Manuscript received August 20, 2012.
Manuscript revised November 6, 2012.

[†]The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing, 210093, China.

[*]This work was supported by the National Science Foundation of China (NSFC No.61271231).

a) E-mail: xingxianglei@gmail.com
b) E-mail: coff128@nju.edu.cn (Corresponding author)

DOI: 10.1587/transinf.E96.D.375

Linear discriminant analysis (LDA) [8], as a supervised algorithm, aims to find the most discriminative features that simultaneously maximize the between-class dissimilarity and minimize the within-class dissimilarity to increase class separability. When sufficient label information is available, LDA can achieve significantly better performance than PCA. However, it suffers a fundamental limitation originating from the assumption that the sample vectors of each class are generated from underlying multivariate Normal distributions of common covariance matrix but different means. Thus, the performance of LDA notably degrades when the actual distribution is non-Gaussian. To address this problem, nonparametric discriminant analysis (NDA) [9], [10] was developed to overcome the problem by introducing a new definition for the between-class scatter matrix, which explicitly emphasizes the samples near the boundary and utilizes the whole training set, instead of merely the class centers. However, when the number of training samples is much smaller than the dimensionality of the feature space, NDA and LDA both suffer the small sample size (SSS) problem due to severe under-sampling of the underlying data distribution. As a result, the generalization capability on testing samples can not be guaranteed.

Semi-supervised Discriminant Analysis (SDA) [11] which makes use of both labeled and unlabeled samples is a reasonable solution to deal with the problem of insufficient training (labeled) samples. However, like other LDA-based methods, SDA still assumes the samples in each class satisfy the Gaussian distribution. Thus, it suffers performance degradation when the intrinsic geometrical structure is generated from non-Gaussian distribution.

We proposed a semi-supervised dimensionality reduction algorithm, called Semi-supervised Nonparametric Discriminant Analysis (SNDA). SNDA aims to best preserve the discriminative information as well as the intrinsic geometric structure in data. Specifically, we construct a nearest neighbor graph to discretely model the manifold structure. Using graph Laplacian, we incorporate the manifold structure into the objective function of the standard NDA as a regularization term. SNDA preserves the inherent advantages of NDA that the Normal assumption is relaxed. SNDA use both labeled and unlabeled samples to estimate the manifold structure of the data. Therefore, it overcomes the SSS problem in the NDA algorithm. Moreover, SNDA shares many of the data representation properties of nonlinear techniques such as LLE and LE, while it yields linear projective maps making it fast and suitable for practical application.

2. Multiclass Nonparametric Discriminant Analysis

Suppose we have a set of N samples $x_1, x_2, \cdots, x_N \in \mathbb{R}^D$, belonging to C class. In LDA, the data are projected from the original D dimensional space to a $C-1$ dimensional subspace through an optimal linear transformation matrix, such that ratio of the determinant of between-class matrix to that of the within-class matrix is maximized. There are three disadvantages in LDA. First, the rank of the between class matrix is at most $C-1$, so the number of the final LDA feature has an upper limit $C-1$. However, it is often insufficient to separate the classes well with only $C-1$ features, especially when $C \ll D$. Second, the boundary structure of classes is not taken into account in computing between-class scatter matrix, which has been shown to be essential in classification. Third, LDA cannot perform well in the cases of non-Gaussian distribution because it is based on the assumption that all classes share the Gaussian distribution with the same covariance matrix.

NDA has been proposed to solve the aforementioned problems. Nonparametric between-class scatter matrix and within-class scatter matrix are defined as:

$$S_b = \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j\neq i}}^{C} \sum_{p=1}^{k} \sum_{l=1}^{N_i} w(i,j,p,l)(x_l^i - N_p(x_l^i, j))(x_l^i - N_p(x_l^i, j))^T \tag{1}$$

$$S_w = \sum_{i=1}^{C} \sum_{p=1}^{k} \sum_{l=1}^{N_i} (x_l^i - N_p(x_l^i, i))(x_l^i - N_p(x_l^i, i))^T \tag{2}$$

where x_l^i denotes the lth samples from class i, $N_p(x_l^i, j)$ is the pth nearest neighbor from class j to the face vector x_l^i. In our experiments, k is chosen as the median of the training (labeled) sample number for each class as recommended in [10]. The weighting function $w(i, j, p, l)$ is defined as:

$$w(i,j,p,l) = \frac{min\{d^o(x_l^i, N_p(x_l^i, i)), d^o(x_l^i, N_p(x_l^i, j))\}}{d^o(x_l^i, N_p(x_l^i, i)) + d^o(x_l^i, N_p(x_l^i, j))} \tag{3}$$

where o is a control parameter between zero and infinity, and $d(x_l^i, N_p(x_l^i, i))$ is the Euclidean distance between two vectors. The weighting function has the property that near the classification boundary it takes on values close to 0.5 and drops off to zeros if the samples are far away from the classification boundary [9]. This weighting function is used to emphasize the boundary information. The optimal transformation matrix ($A = [\alpha_1, \alpha_2, \cdots, \alpha_d]$) is defined as:

$$A_{opt} = \underset{A}{\operatorname{argmax}} \frac{|A^T S_b A|}{|A^T S_w A|} \tag{4}$$

The optimal projection matrix is formed by the eigenvectors corresponding to the non-zero eigenvalues of a generalized eigenvalue problem.

3. Semi-Supervised Nonparametric Discriminant Analysis

In this section, we try to extend the NDA algorithm to incorporate the manifold structure illustrated by both labeled and unlabeled data. The traditional NDA is a supervised-learning algorithm in nature. Therefore, only the labeled data can be utilized by NDA. In practice, the labeled samples are expensive to obtain. When there are no sufficient training samples, NDA often suffers from the small sample size problem and an extremely degenerated S_w is generated. Regularized Discriminant Analysis (RDA) [12] was developed to solve the above problem. The objective function of NDA with a regularization term can be written as follows:

$$\underset{\alpha}{\operatorname{argmax}} \frac{\alpha^T S_b \alpha}{\alpha^T S_w \alpha + \gamma \alpha^T \alpha} \tag{5}$$

where $\alpha^T \alpha$ is the Tikhonov regularizer, and γ is the parameter controlling the balance between the model complexity and the empirical loss.

Motivated by the success of RDA, a semi-supervised NDA (SNDA) is developed here by incorporating the manifold structure when a set of unlabeled samples available. The objective function of SNDA is defined as follows:

$$\underset{\alpha}{\operatorname{argmax}} \frac{\alpha^T S_b \alpha}{\alpha^T S_w \alpha + \gamma_1 \alpha^T \alpha + \gamma_2 J_{MR}(\alpha)} \tag{6}$$

where J_{MR} is a data-dependent manifold regularizer which plays a central role in preserving the manifold structure in data. The manifold assumption [13], which states that the target function varies smoothly along the manifold, has played a key role in manifold regularization. Specifically, if two points are close on the manifold, they are likely to have the same label. However, the data manifold is usually unknown in practice. To reflect the geometric structure of the manifold, we can construct a nearest neighbor graph on a scatter of data points. Consider a graph G with N vertices where each vertex corresponds to a data point. Define the corresponding edge weight matrix W as follows:

$$W_{ij} = \begin{cases} 1 & if\ x_i \in N_p(x_j)\ or\ x_j \in N_p(x_i) \\ 0 & otherwise. \end{cases} \tag{7}$$

where $N_p(x_i)$ denotes the set of p nearest neighbors of x_i.

According to the spectral graph theory [14], the criterion used to measure the smoothness after the mapping can be defined as follows:

$$\begin{aligned} J_{MR}(\alpha) &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha^T x_i - \alpha^T x_j)^2 W_{ij} \\ &= \sum_{i=1}^{N} \alpha^T x_i D_{ii} x_i^T \alpha - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^T x_i W_{ii} x_i^T \alpha \\ &= \alpha^T X L X^T \alpha \end{aligned} \tag{8}$$

where D is a diagonal matrix with $D_{ii} = \sum_j W(i, j)$, $L = D - W$ is the Laplacian matrix of graph G, and $X = [x_1, x_2, \cdots, x_N]$. Observe that if x_i and x_j are linked by an edge ($W_{ij} = 1$), it incurs a heavy penalty in the cost function when the respective $\alpha^T x_i$, $\alpha^T x_j$ are far apart in the

Table 1 SNDA algorithm.

Input:	Data set $X = \{X_l, X_u\}$, balance parameters λ_1, λ_2, control parameter o, number of nearest neighbors p
Output:	A low-dimensional representation of x with enhanced discriminatory power. $y = A^T x$

Algorithm:
1: Calculate S_b and S_w based on the labeled training samples in X_l using Eqs. (1) and (2);
2: Construct a p-nearest neighbor graph matrix W based on all training samples in X using Eq. (7) and calculate the graph Laplacian matrix $L = D - W$;
3: Calculate the optimal projections by the generalized eigenvalue problem (9), and the projection matrix $A = [\alpha_1, \alpha_2, \cdots, \alpha_d]$ where α_i's are the eigenvectors corresponding to the largest d eigenvalues.
4: Transform original samples into the embedded subspace by $y = A^T x$

low-dimensional subspace. In addition, if the points are not neighbors, they do not affect the minimization because their respective weights are zeros. The cost function is minimized when the mapping function varies smoothly on the graph.

The optimal projective vector α's are the eigenvectors corresponding to the maximum eigenvalues of the following generalized eigenvalue problem:

$$S_b\alpha = \eta(S_w + \lambda_1 I + \lambda_2 XLX^T)\alpha \tag{9}$$

Our SNDA algorithm is summarized in Table 1.

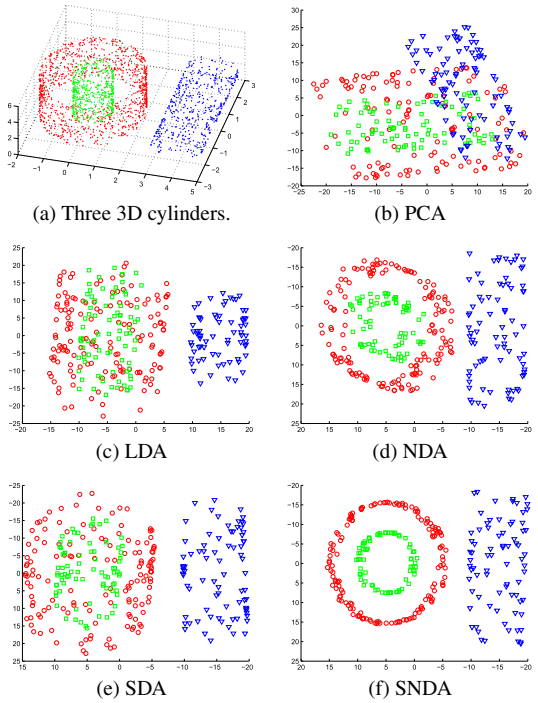
4. Experiments

Several experiments were performed to investigate the performance of the proposed SNDA. We begin with a synthetic example to emphasize the ideas behind the proposed method. Then we proceed with experiments on real world data in order to compare our method to other approaches.

4.1 Synthetic Data

We generated a synthetic 3D cylinder data set, consisting of three cylinders: two concentric cylinders and a vertical cylinder. As shown in Fig. 1 (a), the data set has large within-class variations, thus the in-cluster data distribution is far from Gaussian. About 15% data are kept as test examples, 2% data are used as labeled data and 83% data are used as unlabeled data. We then visually compare our SNDA with several famous dimensionality reduction techniques: PCA, LDA, NDA, and SDA. The test data points are projected from 3D to 2D by the above algorithms. The dimensionality reduction results of our SNDA and other algorithms are shown in the last five sub-figures of Fig. 1. The SNDA's parameters are set to $\lambda_1 = 0$, $\lambda_2 = 0.25$, $o = 8$ and $p = 7$ in the experiment.

As can be observed from Fig. 1, the proposed SNDA performs much better than the other algorithms such that the remaining 2D projection is easy to perceive as a three cluster arrangement and has much more discriminating power than the other four algorithms. Specifically, the PCA algorithm has the smallest discriminating power of all in this example, as can be observed in Fig. 1 (b). This result is because PCA



(a) Three 3D cylinders. (b) PCA (c) LDA (d) NDA (e) SDA (f) SNDA

Fig. 1 Example of dimensionality reduction on a 3D data set.

chooses the directions of maximum variance without paying particular attention to the underlying class structure. The LDA algorithm successfully separates the projective data points of the blue cylinder from that of the two concentric cylinders, see Fig. 1 (c). However, the projective data points of the two concentric cylinders are obviously overlapped. This result is because LDA suffers a fundamental limitation originating from the assumption that the sample vectors of each class satisfy the Gaussian distribution. This restriction also exists in the SDA algorithm, which is in nature the couple of LDA and manifold assumption, as can be observed from Fig. 1 (e). The NDA algorithm, where the Normal assumption is relaxed, has more discriminating power than LDA, see Fig. 1 (d).

4.2 Real World Data

In this subsection, we compare our SNDA algorithm with five popular algorithms in dimensionality reduction: PCA−LDA [8], PCA−LFDA [15], NDA [10], LPP [7] and SDA [11]. After dimensionality reduction has been performed, we apply a simple nearest-neighbor classifier to perform classification in the embedding space. We evaluate these algorithms on five benchmark data sets, including two UCI [16] data sets, a TDT2 [17] document data set, and two image data sets: USPS [18] and COIL20 [19]. See Table 2 for more details. Specifically, for the USPS data set, we performed three experiments with the highly confusing digits: binary classification of digits 4 vs. 9, three-way classification of 1, 7, 9 and four-way classification 1, 4, 7, 9. For the TDT2 document data set, 10 categories from the largest 4th-13th are kept in this experiment. For the high-dimensional

Table 3 Average classification errors for each method on each data set. Each number inside brackets shows the corresponding standard derivation.

Data set	PCA–LDA	PCA–LFDA	NDA	LPP	SDA	SNDA
LSD	0.2198(0.0208)	0.2040(0.0226)	0.2301(0.0193)	0.1978(0.0237)	0.2299(0.0279)	**0.1891(0.0152)**
Vehicle	0.3067(0.0418)	0.2988(0.0364)	**0.2807(0.0307)**	0.3866(0.0463)	0.3827(0.0441)	0.3232(0.0360)
USPS(4 vs. 9)	0.0765(0.0361)	0.0681(0.0344)	0.0757(0.0353)	0.1052(0.0286)	0.0673(0.0258)	**0.0456(0.0147)**
USPS(1, 7, 9)	0.0696(0.0157)	0.0565(0.0173)	0.0669(0.0193)	0.0572(0.0132)	0.0554(0.0181)	**0.0429(0.0111)**
USPS(1, 4, 7, 9)	0.0923(0.0214)	0.0861(0.0213)	0.1229(0.0245)	0.1003(0.0160)	0.1020(0.0189)	**0.0763(0.0177)**
COIL20	0.1109(0.0265)	0.1067(0.0239)	0.1042(0.0189)	0.0919(0.0182)	0.0565(0.0177)	**0.0465(0.0176)**
TDT2	0.2447(0.0431)	0.2294(0.0461)	0.1267(0.0372)	0.1527(0.0552)	0.0939(0.0300)	**0.0698(0.0254)**

Table 2 Statistics of the data sets and the number of labeled data for each data set.

Data set	#Dim (D)	#Inst (n)	#Class (C)	#Labeled (l)
LSD	36	6435	6	20
Vehicle	18	846	4	20
USPS(4 vs. 9)	256	1673	2	20
USPS(1, 7, 9)	256	2882	3	20
USPS(1, 4, 7, 9)	256	3734	4	20
COIL20	1024	1440	20	10
TDT2	36,771	3008	10	10

data set such as COIL20 and TDT2, we first apply PCA to reduce the dimension to 256 for computational efficiency.

For each data set, we randomly select 15% data points as test data and l data points from each class as labeled data. The remaining data points form the unlabeled data. Table 2 shows the number of the labeled points for each data set. There are four important parameters in SNDA algorithm: the balance parameters λ_1, λ_2, the control parameter o, and the number of nearest neighbors p. In our experiments, we empirically set them to 0.01, 0.25, 8, and 7, respectively. We perform 20 random trials and report the mean and standard derivation over the 20 trials. The experimental results are listed in Table 3. For each data set, the lowest classification error is shown in bold. As we can see, the performance of SNDA is better than other methods in most situations. Specifically, for LSD, USPS, COIL20 and TDT2 which may contain manifold structure, the performance of SNDA is the best of all and SNDA offers significant performance gain over the traditional NDA algorithm. For Vehicle which may not contain any manifold structure, the performance of NDA is the best of all and SNDA ranks No.2. Moreover, SNDA outperforms its semi-supervised manifold learning competitor on all the data set.

5. Conclusions

In this study, we have presented a new approach for semi-supervised dimensionality reduction. By making use of both labeled and unlabeled data in learning a linear transformation, this approach overcomes a serious limitation of NDA under situations where labeled data are limited. Inheriting the advantages of NDA, SNDA overcomes the limitation of the LDA-based methods that the samples of each class satisfy the Gaussian distribution assumption. SNDA utilizes the relative benefits of both the discriminant analysis and the manifold learning paradigms. Experimental results have validated the effectiveness of our method.

References

[1] R. Saegusa, H. Sakano, and S. Hashimoto, "A nonlinear principal component analysis of image data," IEICE Trans. Inf. & Syst., vol.E88-D, no.10, pp.2242–2248, Oct. 2005.
[2] M. Zhu and A.M. Martinez, "Subclass discriminant analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.8, pp.1274–1286, 2006.
[3] Y. Zhang and D.Y. Yeung, "Semi-supervised generalized discriminant analysis," IEEE Trans. Neural Netw., vol.22, no.8, pp.1207–1217, 2011.
[4] I.T. Jolliffe, Principal component analysis, Springer-Verlag, 1986.
[5] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, no.5500, pp.2323–2326, 2000.
[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," Proc. NIPS 14, pp.585–591, 2001.
[7] X. He and P. Niyogi, "Locality preserving projections," Proc. NIPS 16, pp.153–160, 2004.
[8] P.N. Belhumeur, J.P. Hepanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.7, pp.711–720, 1997.
[9] K. Fukunaga, "Nonparametric discriminant analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.5, no.6, pp.671–678, 1983.
[10] Z.F. Li, D.H. Lin, and X.O. Tang, "Nonparametric discriminant analysis for face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.4, pp.755–761, 2009.
[11] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," Proc. Int. Conf. Computer Vision (ICCV'07), pp.1–7, 2007.
[12] J.H. Friedman, "Regularized discriminant analysis," J. Amer. Statist. Assoc., vol.84, no.405, pp.165–175, 1989.
[13] M. Belkin, P. Niyogi, V. Sindhwani, and P. Bartlett, "Manifold regularization: A geometric framework for learning from examples," J. Mach. Learn. Res., vol.7, pp.2399–2434, 2006.
[14] F.R.K. Chung, Spectral graph theory, Regional conference series in mathematics, vol.92, AMS, 1997.
[15] M. Ihara, S. Maeda, K. Ikeda, and S. Ishii, "Low-dimensional feature representation for instrument identification," SICE J. Cont. Meas. Syst. Integ., vol.5, pp.249–258, 2012.
[16] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.
[17] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel, "The TDT-2 text and speech corpus," Proc. Darpa Broadcast News Workshop, pp.57–60, 1999.
[18] J.J. Hull, "A database for handwritten text recognition research," IEEE Trans. Pattern Anal. Mach. Intell., vol.16, no.5, pp.550–554, 1994.
[19] S.A. Nene, S.K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," tech. rep., CUCS-005-96, Feb. 1996.