HYPERSPHERE DISTRIBUTION DISCRIMINANT ANALYSIS

Yi-I Chiu[1], *Chun-Rong Huang*[3,4], *Pau-Choo Chung*[1,2], *Ching-Hsing Luo*[1]

[1] Department of Electrical Engineering, Nation Cheng Kung University, Taiwan
[2] Institute of Computer and Communication Engineering, National Cheng Kung University, Taiwan
[3] Institute of Networking and Multimedia, National Chung Hsing University, Taiwan
[4] Department of Computer Science and Engineering, National Chung Hsing University, Taiwan

ABSTRACT

Current graph embedding frameworks of supervised dimensionality reduction often preserve the intraclass local structures and maximize the interclass variance. However, this strategy fails to provide adequate results when strict within-class multimodalities contradict between-class separations. In this paper, we propose Hypersphere Distribution Discriminant Analysis (HDDA), which determines the affinity by considering not only within-class local structure but also the heteropoint distribution in the neighborhood space. If the heteropoint distribution is relatively high in the feature space, this pair should be mapped apart to avoid mixing problems. By taking both the distribution of heteropoints and the distance into account, HDDA shows more effective results compared to the state-of-the-art methods.

Index Terms— Dimensionality Reduction

1. INTRODUCTION

The concept of *graph embedding* [1] has been utilized in many dimensionality reduction methods, such as principal component analysis (PCA), linear discriminant analysis (LDA) [2], locality preserving projection (LPP) [3], linear discriminant embedding (LDE) [4], marginal Fisher analysis (MFA) [1], and local Fisher dicriminant analysis (LFDA) [5]. Among these methods, many supervised approaches [1][4][5] assign larger values for stronger relations between intraclass nearest-neighbors in order to preserve the local structure in the original space. The strict within-class multimodal strategy works well in many cases. Nevertheless, when multimodality contradicts between-class separation, the latter should be considered as the priority. As shown in Fig. 1, between-class discrimination can be achieved by mapping far-apart within-class samples together, while strictly preserving intraclass multimodalities will result in mixing problems.

To solve this problem, we propose **Hypersphere Distribution Discriminant Analysis** (HDDA), which automatically determines whether or not within-class multimodality should be preserved based on the heteropoints distribution in the interference space of every within-class pair. The *interference space* of a pair is defined by constructing two hyperspheres with radii that are equal to the distance between the pair. If the number of heteropoints in the interference space is relatively large, we should strictly map this within-class pair separately to avoid possible mixing problems. The distance between the pair is also considered since two samples are more similar if they are closer to each other. HDDA inherits the advantages from other linear graph-embedded dimensionality reduction methods. It can automatically balance between within-class multimodality preservation (as LFDA) and between-class separation (as LDA) by considering the heteropoint distribution. With all these characteristics, HDDA arrives to the more effective dimensionality reduction results for supervised tasks compared to the state-of-the-art methods.

The work is supported in part by the National Science Council of Taiwan, R.O.C., under Grant NSC100-2221-E-005-085 and NSC97-2221-E-006-147-MY3.

2. METHOD

2.1. The linear dimensionality reduction problem

Given a set of n data points $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ ($\mathbf{x}_i \in \mathbb{R}^d$) with corresponding class labels $\{y_i\}_{i=1}^n = \{1, \cdots, m\}$. Our goal is to find an appropriate transformation matrix $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_l] \in \mathbb{R}^{d\times l}$ which maps $\mathbf{X}$ to the lower dimensional space $\mathbb{R}^l$ ($l < d$). The data points resulted from mapping can be denoted by $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$ where $\mathbf{Z} = \mathbf{W}^T\mathbf{X}$.

The following equations state the pointwise form of the between (S^b) and the within (S^w) scatter matrices in Sugiyama's work [5].

$$S^w = \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}^w(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{1}$$

$$S^b = \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}^b(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{2}$$

where

$$A_{ij}^w = \begin{cases} A_{ij}/n_c, & \text{if } y_i = y_j = c \\ 0, & \text{if } y_i \neq y_j \end{cases} \tag{3}$$

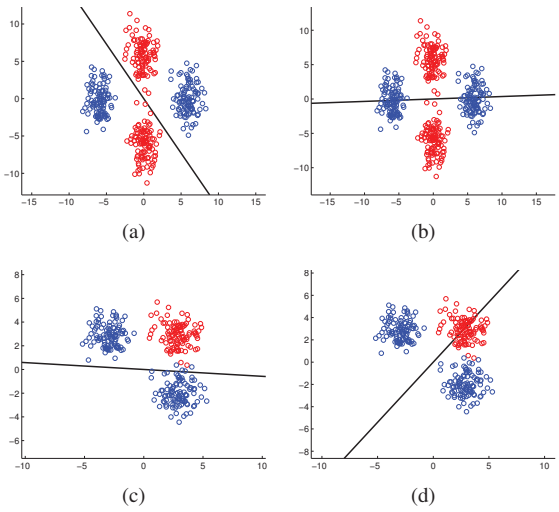


Fig. 1. Fig 1(a) and 1(c) show the results of LFDA. Fig 1(b) and 1(d) show the results of HDDA. The black lines indicate the submanifold after projection. These are the examples that strictly preserving the within-class multimodality as LFDA does will lead to the mixing problem.

$$A_{ij}^{b} = \begin{cases} A_{ij}(1/n - 1/n_c), & \text{if } y_i = y_j = c \\ 1/n, & \text{if } y_i \neq y_j \end{cases} \quad (4)$$

with A_{ij} indicating the affinity of the within-class samples and n_c denoting the number of samples in class c. In LDA, $A_{ij} = 1$ if $y_i = y_j$, since LDA defines that every within-class pair has equal chance to be mapped together. In LFDA, A_{ij} is defined as the heat kernel in order to preserve the local structure. In our method, a new affinity matrix is generated by considering not only the distance between every within-class pair but also the distribution of the heteropoints in their interference space.

2.2. The heteropoint distribution and affinity weight

A pair of within-class samples is more likely to be mapped closer on the new manifold if the affinity between them is higher. Using the heat kernel [5] as the affinity weight can ensure that multimodality is preserved. However, in some cases (such as those shown in Fig. 1), mapping far-apart within-class samples together can prevent them from mixing with heteropoints. Thus, there is a trade-off between preserving the within-class multimodality and reaching a better between-class separation. As shown in Fig. 1(a) and 1(c), LFDA fails to find appropriate projections in these scenarios.

Given a pair of samples $\mathbf{x}_i$ and $\mathbf{x}_j$ in class c, there is a line that passes through these two points as shown in Fig. 2. If there is a submanifold $\mathcal{M}$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped close, any data points close to the pair and the line between them will be mapped close as well. As a result, heteropoints residing close to this line would also be mapped close to this

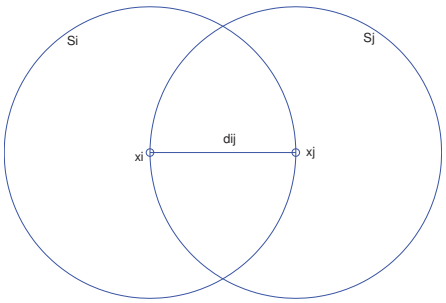


Fig. 2. With these two hyperspheres, the space can be separated into three parts

pair and lead to mixing problems. In contrast, if no such heteropoints exist, mapping this pair together would have less influence on the between-class separation.

In order to determine whether a data point is close to the within-class pair, we construct two hyperspheres S_i and S_j. S_i is centered at $\mathbf{x}_i$, with radius d_{ij}, where d_{ij} is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. S_j is centered at $\mathbf{x}_j$, with the radius d_{ij}. The space covered by these two hyperspheres is defined as the *interference space*.

With more heteropoints located in the interference space, mapping this within-class pair together would lead to a greater chance for mixing problem to occur. If only few or no heteropoints reside in the interference space, there is a smaller chance for the mapping to violate the between-class separation rule. However, the distribution of heteropoints is also related to the size of the hyperspheres, which is determined by the distance between the within-class pair. Therefore, we assign the weights of the affinity matrix based on both the distance of the within-class pair and the heteropoint distribution rather than considering the number of heteropoints alone.

2.3. Hypersphere Distribution Discriminant Analysis

The procedure of hypersphere distribution discriminant analysis (HDDA) is described in this section. We firstly compute the distance matrix $\mathbf{D}$ by finding the Euclidean norm between each pair of samples. For $\mathbf{x}_i$ and $\mathbf{x}_j$ with $y_i = y_j = c$, the number of heteropoints resided in hypersphere S_i is considered by using all the $\mathbf{x}_k$ in different classes as follows:

$$n_i = \sum_{k=1}^{n} \{(1_{\{\mathbf{x}_k | d_{ik} < d_{ij}\}})(1_{\{\mathbf{x}_k | y_k \neq c\}})\} \quad (5)$$

the heteropoint distribution is then defined as follows:

$$h_{ij} = \max\{n_i, n_j\}. \quad (6)$$

Using the heteropoint distribution and the distance between the samples, we can build the new affinity matrix as follows:

$$A_{ij} = \left(\frac{1}{1 + \frac{d_{ij}}{\sum_{\forall i,j} d_{ij}}}\right)^{\log_{n'} h_{ij}} \quad (7)$$

where n' is the square root of the total number of heteropoints, $\sqrt{n-n_c}$. The weighting function is further discussed in Section 3.

The transformation matrix $\mathbf{W}$ can be computed by solving the following optimization problem:

$$\mathbf{W} = \arg\max[((\mathbf{W}^T S^b \mathbf{W})^{-1}(\mathbf{W}^T S^w \mathbf{W}))] \quad (8)$$

To find the transformation matrix $\mathbf{W}$, we can apply the Spectral Graph Theory [6] and compute the eigenvectors and eigenvalues for the following eigenvalue problem

$$XL^bX^T\mathbf{w} = \lambda XL^wX^T\mathbf{w}. \quad (9)$$

Here, $L^b = D^b - A^b$ and $L^w = D^w - A^w$ where D^b and D^w are diagonal matrices with entries are column (row) sums of A^b and A^w, i.e. $D^b_{ii} = \sum_j A^b_{ji}$ and $D^w_{ii} = \sum_j A^w_{ji}$.

With the optimal transformation matrix $\mathbf{W}$,the linear embedding is stated as follows:

$$\mathbf{x}_i \rightarrow \mathbf{z}_i = \mathbf{W}^T\mathbf{x}_i \quad (10)$$

where $\mathbf{z}_i$ is a l-dimensional vector and $\mathbf{W}$ is an $n \times l$ matrix.

3. CHOICE OF AFFINITY WEIGHT

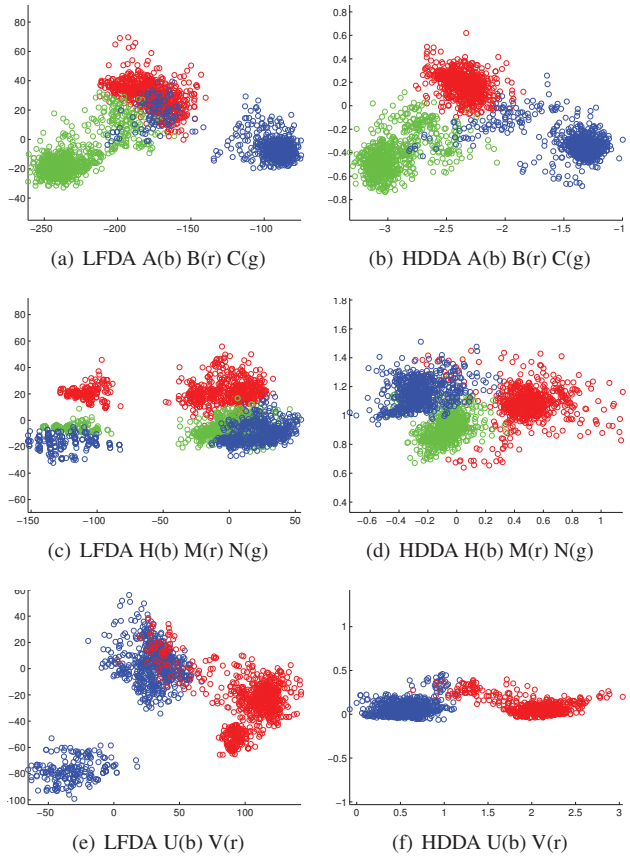
The choice of the within-class affinity weight depends on two variables h_{ij} (the heteropoint distribution) and d_{ij} (the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$). The higher the number of heteropoints residing in the interference space (higher h_{ij}), the lower the affinity between the within-class pair, since mapping the pair of points close to each other would likely cause a mixing problem. Also, a pair is considered more similar if they are closer in the original space (lower d_{ij}).

We set a threshold for the number of "tolerable" heteropoints located in the interference space. If the number of the heteropoints exceeds the threshold n', the pair of within-class samples should not be mapped together in order to avoid mixing with heteropoints. n' is defined as the square root of the total number of heteropoints, $\sqrt{n-n_c}$. The weight in the affinity matrix drops rapidly if $h_{ij} > n'$. On the other hand, if the number of heteropoints in interference space is smaller than the threshold value, affinity would be enhanced since mapping these two data points together would not cause mixing. Therefore, for within-class samples x_i and x_j, their affinity is defined as (7) in Section 2.3. The function of d_{ij} is appropriately chosen as a monotone decreasing function with its value between $(0, 1]$, which causes the affinity value to drop down when the number of heteropoints exceeds the threshold. We do not consider the heat kernel for its strengthening the strict within-class multimodality.

4. EXPERIMENTS

4.1. 2D Data Visualization

We apply the UCI Letter Recognition Dataset [7] for our 2D data visualization experiments. This dataset consists of samples with 16 features in 26 classes ('A'-'Z'). In the experiment, three groups of testing samples, including some easily confused combinations, were chosen. These combinations are {A,B,C}, {H,M,N} and {U,V}. Many hand-written letters have in-class multimodality. However, arbitrarily preserving the multimodality can sometimes lead to undesired results. The results in Fig. 3 show that HDDA performs better than LFDA when applied to 2D data visualization problems.



(a) LFDA A(b) B(r) C(g) (b) HDDA A(b) B(r) C(g)

(c) LFDA H(b) M(r) N(g) (d) HDDA H(b) M(r) N(g)

(e) LFDA U(b) V(r) (f) HDDA U(b) V(r)

Fig. 3. The results of 2D data visualization of the letter recognition dataset applying HDDA and LFDA. These results show that with the justification of the multimodality by detecting heteropoints in the interference space; a better between-class separation can be achieved.

4.2. Classification results of IDA datasets

We also compare the classification results of the IDA data sets [8] of HDDA and those in Sugiyama's work [5]. Table 1 shows the average error rates of HDDA and of the state-of-the-art methods when using 1NN classifier. The embedding dimensionality of HDDA is chosen by 5-fold cross-validation. We divide these datasets into three groups based on the performance of LDA and LPP.

Group 1 includes the datasets in which LDA outperforms

LPP, indicating that arbitrarily preservation of the multimodality of these datasets does not provide a better result. HDDA outperforms LFDA and shows comparable or slightly better results to LDA. The result indicates that HDDA inherits the characteristic of LDA, which appropriately maps within-class samples together to achieve better recognition rates. LFDA, on the other hand, strictly preserves the within-class multimodality and shows less accurate results in those cases.

Group 2 includes the datasets in which LPP performs better; however, the datasets do not contain strong multimodality. For these datasets, HDDA performs better results than LDA does, since it gives the priority to detecting appropriate multimodalities over mapping all the within-class data together. HDDA also shows comparable (if not better) results compared to LFDA in most of datasets in this group, as it applies the between-class separation concepts proposed in LDA.

Group 3 contains the datasets with the strong within-class multimodality. HDDA does not perform as well as LFDA within these data sets, as these datasets already have very strong multimodality. Thus, arbitrarily preserving multimodality would be the best choice for seeking the best projection space. In the thyroid datasets, the direct inhibition of the heteropoints between the within-class pairs exists only in a small portion of those pairs.

Unlike LDA and LPP, HDDA neither strictly preserves nor rejects the within-class multimodality. Instead, HDDA judges when to apply the within-class multimodality to achieve better classification results. In unsupervised scenarios, LPP offers adequate results since no class information is provided. However, this projection rule cannot be strictly followed in supervised scenarios. With the exception for datasets with strong multimodality and imbalanced number of data in different classes, HDDA shows better results in most cases than does the state of art methods.

5. CONCLUSIONS

In this paper, we proposed a novel linear supervised dimensionality reduction method called Hypersphere Distribution Discriminant Analysis (HDDA). The major advantage of our approach is that the within-class multimodality is preserved when it can lead to better between-class separation, instead of applying it strictly. By constructing the distribution matrix of the heteropoints using hyperspheres, we can determine if mapping a within-class pair together would cause possible mixing problems. The new within-class affinity is assigned based on the heteropoint distribution and the distance between the pair. Performance improvement of HDDA over the state-of-the-art methods is demonstrated through several experiments.

Table 1. Means of the error rates applying HDDA, LFDA, LPP, PCA and LDA to IDA datasets. The superscript following the dataset name indicates the the group that the dataset belongs to.

Data set	HDDA	LFDA	LPP	PCA	LDA
breast-cancer[1]	**32.3**	34.7	33.5	34.5	32.9
diabetes[1]	**30.3**	32.0	31.5	31.2	30.6
flare-solar[1]	**38.6**	39.2	39.2	39.1	39.0
german[1]	**29.6**	29.9	30.7	30.2	30.5
heart[2]	**20.8**	21.9	23.3	24.3	24.0
image[2]	4.7	**3.2**	3.6	3.4	6.5
ringnorm[2]	**17.9**	21.1	20.6	21.6	31.2
splice[2]	**16.3**	16.9	23.2	22.6	33.7
thyroid[3]	5.7	4.6	**4.2**	4.9	5.3
twonorm[2]	**3.4**	3.5	3.7	3.6	5.0
waveform[3]	12.8	12.5	**12.4**	12.7	17.6

6. REFERENCES

[1] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern analysis and machine intelligence*, vol. 29, no. 1, pp. 40 – 51, 2007.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Inc., Boston, ii edition, 1990.

[3] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems 16*, 2004.

[4] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Computer Vision and Pattern Recognition*, 2005.

[5] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research 8*, pp. 1027–1061, 2007.

[6] F. Chung, "Spectral graph theory," *Regional Conf. Series in Math.*, , no. 92, 1997.

[7] A. Frank and A. Asuncion, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`, 2010.

[8] G. Ratsch, T. Onoda, and K.-R Muller, "Soft margins for adaboost.," *Machine Learning*, vol. 42, no. 287-320, 2001.