Globality-Locality Consistent Discriminant Analysis for Phone Classification

Heyun Huang[1], Yang Liu[2], Jort Gemmeke[1], Louis ten Bosch[1], Bert Cranen[1], Lou Boves[1]

[1]Department of Linguistics, Radboud University Nijmegen,
Erasmuslaan 1, 6525 HT, Nijmegen, the Netherlands
[2]Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong

{h.huang,j.gemmeke,l.tenbosch,b.cranen,l.boves}@let.ru.nl, csygliu@comp.polyu.edu.hk

Abstract

Concatenating sequences of feature vectors helps to capture essential information about articulatory dynamics, at the cost of increasing the number of dimensions in the feature space, which may be characterized by the presence of manifolds. Existing supervised dimensionality reduction methods such as Linear Discriminant Analysis may destroy part of that manifold structure. In this paper, we propose a novel supervised dimensionality reduction algorithm, called Globality-Locality Consistent Discriminant Analysis (GLCDA), which aims to preserve global and local discriminant information simultaneously. Because it allows finding the optimal trade-off between global and local structure of data sets, GLCDA can provide a more faithful compact representation of high-dimensional observations than entirely global approaches or heuristic approaches aimed to preserve local information. Experimental results on the TIMIT phone classification task show the effectiveness of the proposed algorithm.

Index Terms: Dimensionality Reduction; Discriminant Analysis; Global and Local Data Structure; TIMIT; Affinity Matrix

1. Introduction

MFCCs and PLP coefficients computed every 10 ms over 25 ms windows are able to characterize the short-term characteristics of speech signals. However, speech is generated by continuous movements of the articulators and these short-term features do not capture the underlying articulatory dynamics completely. In typical speech recognition systems, the delta and delta-delta coefficients of adjacent frames are used to model the temporal dynamics, but this is not enough for representing dynamic changes at the level of syllables. To better capture the articulatory dynamics, at least three different research directions have been proposed. First, features that combine variable-length windows can be used, including the TRAP features proposed in [1]. Second, the temporal dynamics observed in successive frames of MFCCs/PLPs can be characterized by polynomial regression functions (e.g. [2, 3]). The third direction, which is taken in this paper, is using sequences of short-term feature vectors as units for subsequent processing.

If we want to represent the articulatory dynamics at the level of syllables (which on average have a duration of about 250 ms) we need to concatenate up to 25 short-term feature frames. With the popular 13 element MFCC vectors this will result in feature spaces of over 300 (arguably non-orthogonal) dimensions. Because it is extremely difficult to define meaningful distance measures in very high dimensional spaces, there is a need for dimensionality reduction, preferably one that enhances the similarity of observations in the same class and simultaneously maximizes the distance between different classes. One way for doing this is Fisher Discriminant Analysis (FDA) [4]. However, FDA assumes that all classes are characterized by unimodal (Gaussian) distributions. Therefore, FDA is not able to represent complex local variations and subtleties of the data structure, such as concavities and protrusions. Yet, phone classes might well exhibit such complex distributions, if only because of the co-articulation effects.

In order to enhance the power of Discriminant Analysis models, several algorithms that introduce the idea of "locality" [5] have been proposed recently, such as local FDA (LFDA) [6, 7], locality sensitive discriminant analysis (LSDA) [8], local discriminant embedding (LDE) [9], and marginal Fisher analysis [10]. All these local algorithms are based on the same idea: they seek to maximize the discriminant information among nearby data points by constructing a neighborhood relationship graph in terms of an Affinity Matrix. Compared with the original FDA, these algorithms perform better in describing the local structure of a data set. However, one common drawback of these local algorithms is that they are not designed to describe the global distribution of the data set, which is also of great importance in the learning procedure.

To obtain an optimal trade-off between global distribution and local information in the dimensionality reduction process, this paper proposes a novel algorithm called Globality-Locality Consistent Discriminant Analysis (GLCDA), which provides a flexible affinity matrix that can adapt to the unknown data structure. By exploring both global and local discriminant information in a unified framework, GLCDA can provide a more faithful projection of the high-dimensional observations than approaches that are fully global or treat local information heuristically.

The rest of the paper is organized as follows. In Section 2, we recall how FDA can be reformulated in terms of affinity matrices. Then we generalize the affinity matrices to integrate both global and local discriminant information and propose the new algorithm GLCDA. In Section 3 and Section 4, the data and approaches to evaluate our algorithm are introduced and the results of different dimensionality reduction algorithms on TIMIT phone classification task [11] are reported, respectively. General discussions and conclusion are given in Section 5.

2. Advanced Fisher Discriminant Analysis

2.1. Fisher Discriminant Analysis

Given n training data points $\mathbf{x}_i \in \mathbb{R}^D$ $(i = 1, 2, \ldots, n)$, each data point is associated with a class label $c_{x_i} = c \in$

$\{1, 2, \ldots, C\}$; n_c is the number of data points in class c. Fisher Discriminant Analysis (FDA) [4] aims to find the projection matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$ $(d \leq D)$ such that the low-dimensional representations $\mathbf{z}_i \in \mathbb{R}^d$ can be obtained by $\mathbf{z}_i = \mathbf{W}^T \mathbf{x}_i$. In order to maximize the discrimination power, LDA optimizes the objective function:

$$\underset{\mathbf{W}}{\arg\max}[(\frac{tr(\mathbf{W}^T \mathbf{S}^{(b)} \mathbf{W})}{tr(\mathbf{W}^T \mathbf{S}^{(w)} \mathbf{W})})] \qquad (1)$$

$\mathbf{S}^{(w)}$ and $\mathbf{S}^{(b)}$ are the within-class scatter matrix and the between-class scatter matrix, respectively. They can be formulated in terms of the sum of pairwise scatter matrices: $(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$ in the following way [6, 10]:

$$\mathbf{S}^{(w)} = \frac{1}{2} \sum_{ij} A_{ij}^w (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \qquad (2)$$

$$A_{ij}^w = \begin{cases} 1/n_c & \text{if } c_{x_i} = c_{x_j} = c \\ 0 & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

$$\mathbf{S}^{(b)} = \frac{1}{2} \sum_{ij} A_{ij}^b (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \qquad (3)$$

$$A_{ij}^b = \begin{cases} 1/n - 1/n_c & \text{if } c_{x_i} = c_{x_j} = c \\ 1/n & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

Each pair $(\mathbf{x}_i, \mathbf{x}_j)$ makes a contribution to each of the two scatter matrices which is independent of their distance. Since A_{ij}^w and A_{ij}^b in equations (2) and (3) are independent of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ (and thus ignore any local information), FDA can be regarded as a global method.

2.2. Incorporating Local Information into FDA

The *global* dimensionality reduction method FDA described in Section 2.1 assumes that all classes have the same multivariate normal distribution, apart from the mean vectors. In contrast, "locality" implies that the data structure is characterized by a (large) number of (sub-)manifolds, which cannot be accurately described by multivariate normal distributions. The Locality Preserving Projection (LPP) approach [5] was proposed to describe the manifolds in terms of an adjacency graph. To represent the adjacency relationship explicitly, the affinity matrices introduced in equations (2) and (3) are now redefined by introducing a weight to capture the distance between data pairs $(\mathbf{x}_i, \mathbf{x}_j)$:

$$a_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2) \qquad (4)$$

Based on the definition in (4), LPP is capable of describing the locality information that represents the local manifolds (theoretically) well; therefore, it provides an alternative tool to analyze the data space. However, in real-world applications, including phone classification, the structure of the data in a high-dimensional space is very hard to predict. Consequently, it is not a priori known whether the globality or locality information is more important. Motivated by this consideration, LFDA [6] combines the idea of LPP and FDA:

$$\mathbf{S}^{(w)} = \frac{1}{2} \sum_{ij} A_{ij}^w (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \qquad (5)$$

$$A_{ij}^w = \begin{cases} a_{ij}/n_c & \text{if } c_{x_i} = c_{x_j} = c \\ 0 & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

$$\mathbf{S}^{(b)} = \frac{1}{2} \sum_{ij} A_{ij}^b (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \qquad (6)$$

$$A_{ij}^b = \begin{cases} a_{ij}(1/n - 1/n_c) & \text{if } c_{x_i} = c_{x_j} = c \\ 1/n & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

To better describe the locality information, a locality preserving scaling kernel was adopted in LFDA by [12], by defining the weights as:

$$a_{ij} = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma_i \sigma_j}\right) \qquad (7)$$

where σ_p (where p is i or j) is assigned as:

$$\sigma_p = ||\mathbf{x}_p - \mathbf{x}_p^L|| \qquad (8)$$

in which $\mathbf{x}_p^L$ denotes the Lth nearest neighbour in the same class.

2.3. Globality-Locality Consistent Discriminant Analysis

Although LFDA has shown good performance in describing locality information, it has two limitations. First, LFDA only exploits the locality information within each class, but fails to consider the locality information between different classes, which is also of great importance in supervised learning tasks. For instance, if two manifolds from different classes exist, the projection matrix should learn to separate them in the lower dimension space. Second, the affinity matrix of LFDA cannot be adapted to represent the trade-off between local and global information in the data structure. To overcome these limitations, we present an algorithm called Globality-Locality Consistent Discriminant Analysis (GLCDA), which defines the affinity matrices of the within and between-class scatter matrices as follows:

$$\mathbf{S}^{(w)} = \frac{1}{2} \sum_{ij} A_{ij}^w (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \qquad (9)$$

$$A_{ij}^w = \begin{cases} a_{ij}(\gamma_1)/n_c & \text{if } c_{x_i} = c_{x_j} = c \\ 0 & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

$$\mathbf{S}^{(b)} = \frac{1}{2} \sum_{ij} A_{ij}^b (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \qquad (10)$$

$$A_{ij}^b = \begin{cases} 0 & \text{if } c_{x_i} = c_{x_j} = c \\ b_{ij}(\gamma_2)) & \text{if } c_{x_i} \neq c_{x_j} \end{cases}$$

where the adjacency graphs are defined as:

$$a_{ij}(\gamma_1) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{(\sigma_i \sigma_j)^{\gamma_1}}\right) \qquad (11)$$

$$b_{ij}(\gamma_2) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{(\sigma_i \sigma_j)^{\gamma_2}}\right) \qquad (12)$$

The definition of σ_i and σ_j is similar to equation (8). Compared with LFDA, γ_1, γ_2, and b_{ij} are introduced in our algorithm, which will be explained in following subsections. It is worth mentioning that, the elements of the between-class scatter affinity matrix in Eq. (10) when $c_{x_i} = c_{x_j} = c$ are set to zero while the corresponding elements of FDA in Eq. (3) and LFDA in Eq. (6) are non-zero. We do this because the distance between two points from the same class is already accounted for in the affinity matrix of the within-class scatter. Using it again in the affinity matrix of the between-class scatter would mean that the same information is used twice, which should be avoided in dimensionality reduction.

2.3.1. Globality and Locality Consistency by γ_1 and γ_2

Introducing γ_1 and γ_2 provides a way to adapt the dimensionality reduction method to the unknown data structure. By decreasing γ_1 and γ_2, the locality information becomes more important. Thus, tuning of the γs allows us to find the optimal

trade-off between globality information and locality information.

The following considerations further explicate the relationship between GLCDA and the FDA, LPP, and LFDA methods.

- If $\gamma_1 = \gamma_2 = +\infty$, all elements a_{ij} and b_{ij} are equal to 1 (The prerequisite is that all σ should be larger than 1 which is easily guaranteed by re-scaling). This means that: 1) a pair of data points which belong to the same class makes an identical (i.e. distance independent) contribution to the within-class scatter; and 2) a pair of data points which belong to different classes also contributes equally to the between-class scatter. Although not entirely identical to FDA (due to the non-zero A_{ij}^b in Eq. (3) for data points that belong to the same class), the underlying objective of GLCDA for this setting of γ_1 and γ_2 is very similar to that of FDA.

- If $\gamma_1 = 1, \gamma_2 = +\infty$, all elements a_{ij} are equal to those in Eq. (5) while all elements b_{ij} are equal to 1. In other words, for this setting of γ_1 and γ_2, the way in which GLCDA accounts for local and global information is very similar to that of LFDA.

- If $\gamma_1 = \gamma_2 = 0$, the coefficients a_{ij} and b_{ij} are identical to those in LPP [5].

2.3.2. Incorporating Local Discriminant Information by b_{ij}

Besides balancing global and local information, another advantage of the proposed GLCDA over LFDA is the consideration of local discriminant information. If $(\mathbf{x}_i, \mathbf{x}_j)$ are located in two different manifolds, i.e., they are from different classes, maximizing their distance in the lower dimensional space becomes very important. The elements of the affinity matrix b_{ij} will increase the penalty of close pairs from different classes, effectively increasing the distance so that the overlap between local manifolds is reduced [9]. Thus, the parameter γ_2 allows to flexibly adapt the trade-off between local and global discriminant information.

3. Experimental Setup

3.1. Feature Extraction and Classification Task

The GLCDA method will be evaluated by means of the TIMIT phone classification task [11]. We use the standard NIST training set, which includes 462 speakers, 3696 utterances, and $139,852$ phones. GLCDA will be used to reduce the dimensionality of the vectors (x_i), and classification will be performed by directly using a weighted k-Nearest Neighbor (WkNN) classifier [13]. The core test set is used to evaluate the classification performance, containing 24 speakers, 192 utterances, and $7,195$ phones. In addition, the development set advised in [14] is used to tune parameters. As usual, the 64 phone labels were reduced to 48 classes according to [15]. The classification error is evaluated by further mapping the 48 classes into 39 classes in line with [15] as well. Glottal stops (q) were excluded in all sets.

For each phone, the feature vector is generated as follows: A short-time Fourier analysis is performed every 10 ms using a 25 milliseconds Hamming window. Thirteen MFCCs are computed for each frame. For each phone the middle frame is selected, and then 11 frames to the left and 11 frames to the right are concatenated. The stack of MFCCs may or may not contain some frames from the neighboring phones, depending on

the length of the phone. Consequently, every phonetic segment is characterized by a $13 \times 23(299)$ feature block, which is then reshaped to a single 299 dimensional feature vector. Next, as is common in discriminant analysis [16], Principal Component Analysis (PCA) is applied to reduce the original 299-dimensional space to a 150-dimensional space, which keeps 97% of the total variance. The reduced feature vectors characterize the phone segments of training, development, and test sets, which act as the input of all different dimensionality reduction algorithms in the remainder of this paper.

3.2. Weighted k-Nearest Neighbour Classifier

We choose a Weighted k-Nearest Neighbor (WkNN) [13] classifier to classify the reduced feature vectors into one of 48 phone classes. Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$ be the k nearest neighbours of a test vector $\mathbf{x}$. The weights of these neighbours are calculated by Eq. (13).

$$w_i = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}||^2}{\tau}\right), i = 1, 2, ..., k \qquad (13)$$

To determine the class label of that test vector, the weights of neighbors that belong to each class are accumulated and the class label $\hat{C}$ associated with the largest sum is selected.

4. Experimental Results

We compare different discriminant analysis methods, including FDA, LPP, LFDA, and the proposed GLCDA. Parameters γ_1 and γ_2 are jointly tuned for the proposed GLCDA. Two different approaches will be adopted to show the results. First, a cross-validation on the development set is performed. The parameters of kNN classifier and dimensionality reduction (if necessary) are optimized using this set. Next, the classification accuracy is measured on the core test set. Also, to investigate the impact of the mismatch between the development and core test set, we will exchange the role of the development and core test set.

Second, to investigate the sensitivity to small changes in the parameters, we also provide results for slightly perturbed values around the optimal values of γ_1 and γ_2. By doing so, we show that GLCDA is robust against small changes in the parameters. To allow for a fair comparison, we will use 47 dimensions (the number of phone classes minus 1) in all dimensionality reduction methods.

4.1. Performance Evaluation by Cross-Validation

The free parameters are jointly tuned on the development set, which means that (k, τ) are optimized for FDA/LPP/LFDA and $(k, \tau, \gamma_1, \gamma_2)$ for the proposed GLCDA. In preliminary experiments, meaningful ranges of these parameters were determined: $15 < k < 60$, $3.5 < \tau < 6.5$, $0 < \gamma_1, \gamma_2 < 1.5$. The tuning method used uniform sampling in these ranges and searching the full grid. The values maximizing the performance on the development set are used to evaluate the core test set.

The results are reported in the upper "Accuracy" row of Table. 1. The purely global method, FDA, has lower accuracy than all methods which take the local information into account. The enhancement over FDA proves that meaningful local structures do exist in the TIMIT feature space. For the three methods with local information, LPP and LFDA have similar performance, but worse than GLCDA. This shows that adjusting the importance of local and global information is meaningful.This also holds when the roles of development and core test set are ex-

Table 1: Performance comparison of FDA, LPP, LFDA, and GLCDA on the core test set (upper row) and development set (lower row) with optimally tuned parameters on the development set and core test set.

Tune/Test	FDA	LPP	LFDA	GLCDA
Development/Core	73.47	73.99	74.00	**74.64**
Core/Development	74.16	74.31	74.44	**74.88**

Table 2: Average suboptimal performance of GLCDA. The upper data row reports the accuracy on core test set and the lower data row reports that on development set.

Test Set	Optimal	Sub-Opt.(0.1%)	Sub-Opt.(0.2%)
Core	74.64	74.79	74.77
Development	74.88	74.86	74.87

changed (bottom row in Table. 1). In both cases, the performance of FDA, LPP, and LFDA is still worse than GLCDA.

4.2. Robustness of the Proposed Method

In many real-world pattern recognition tasks, the development and core test sets might not match very well. In TIMIT, the core test set contains all possible SX sentences, none of which appears more than once. In the development set the SX sentences are not unique. To prevent the occasional optima, we are concerned with the parameters γ_1 and γ_2 whose performance with best tuned kNN classifier is slightly worse than the (globally) optimal ones. They are called "suboptimal" to evaluate the robustness of the proposed GLCDA. Concretely speaking, the parameters γ_1 and γ_2 are evaluated in the core test set when their classification accuracy are 0.1% and 0.2% lower than the maximum of development set. The accuracy on average is then reported in Table. 2. Similarly, the roles of development and core test sets are exchanged. Still, the average accuracy is (more) higher than FDA, LPP, and LFDA, which means the improvement is robust and not accidental. Note that the optimal (γ_1, γ_2) as found on the development set does not always yield an optimal performance on the core test set (top row).

5. General Discussions and Conclusion

This paper presents a novel dimensionality reduction method for TIMIT phone classification, called Globality and Locality Consistent Discriminant Analysis. The conventional FDA was first reformulated in terms of data-pair scatter matrices. The concept of locality was then introduced together with LPP and the discriminative method, LFDA. Two obvious deficits of LFDA, inability of being adapted to the unknown feature space and lack of local discriminant information, motivated us to propose the affinity matrices defined by Eq. (9) and Eq. (10). GLCDA is capable of adjusting the weights of global and local structures by two parameters: γ_1 and γ_2.

Experimental results on TIMIT substantiated the effectiveness of proposed method and the existence of both global and local structures in the MFCC feature space. Two heuristic local methods, LPP and LFDA, outperformed the conventional FDA, which means that the locality information is worth exploiting for phone modeling for speech recognition. Furthermore, the proposed GLCDA, with optimal (suboptimal) tuned γ_1 and γ_2 in the development set, reached the highest accuracy of 74.64% (74.79% on average). When the roles of development set and core test set were changed, GLCDA still achieved the best performance as 74.88% (74.86% on average) classification accuracy. Compared with FDA, we can draw the conclusion that exploiting the local structure of TIMIT by our proposed GLCDA achieved a significant accuracy gain of 1.17% (1.32% on average) on TIMIT core test set and 0.72% (0.70% on average) on TIMIT development set.

6. Acknowledgement

7. References

[1] H. Hermansky and P. Jain, "Band-independent speech-events categories for TRAP based ASR," in *Proc. Eurospeech*, 2003, pp. 1013 – 1016.

[2] H.Gish and K.Ng, "Parametric trajectory models for speech recognition," in *Proc. ICASSP*, 1996, pp. 466–469.

[3] Y. Han, J. de Veth, and L. Boves, "Trajectory clustering for solving the trajectory folding problem in automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15:4, pp. 1425 – 1434, 2007.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[5] X. He and P. Niyogi, "Locality preserving projections," in *NIPS 16*, 2004.

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

[7] M. Sakai, N. Kitaoka, and K. Takeda, "Feature transformation based on discriminant analysis preserving local structure for speech recognition," in *Proc. ICASSP*, 2009, pp. 3813 – 3816.

[8] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *Proc. 20th IJCAI*, 2007, pp. 708–713.

[9] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proc. CVPR*, 2005, pp. 846–853.

[10] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extension: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 40–51, 2007.

[11] J. S. Garofolo, *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*, National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

[12] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *NIPS 17*, 2004, pp. 1601–1608.

[13] S.Tan, "Neighbor-weighted k-nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, vol. 28, pp. 667–671, 2005.

[14] A. K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," *Ph.D. Thesis, MIT*, 1998.

[15] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hmms," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[16] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 337–350, 2006.