

Speech Recognition Using Locality Preserving Projection Based on Multi Kernel Learning Supervision

Dayong Zou^{1, a}, Juan Wang^{2, b}

*School of Computer and Information Engineering, Henan University,
Kaifeng, 475004, China.*

*School of Computer and Information Engineering, Henan University,
Kaifeng, 475004, China.*

aemail: zd_yong@henu.edu.cn, bemail: wj2001@henu.edu.cn

Abstract

The multi kernel supervising locality preserving projective speech recognition method is proposed. It utilizes a multi-kernel learning method to project speech feature vectors to high dimensional linear space, and then uses linear mapping by locality preserving projective algorithm. It utilizes Locality Preserving Projection to lower the dimension of samples and meanwhile keep the approximate separability of different classes in feature space. Considering the different mapping vector to keep the local structure important degree is different, multiple-kernel mapping feature vectors are given different weight coefficients. Experiments show that the method has achieved dimensionality reduction, improved the retrieval speed of eigenvectors and reduced the amount of computation, and gained a good recognition effect.

Key words: Feature Extraction, Multiple-kernel learning, Supervised learning, Locality Preserving Projection, Speech recognition

Introduction

Man-machine speech communication is not completely realized at present, like the online search engines we are familiar with such as Google, Baidu and so on. Mainly because it is crucial to establish highly efficient index structure if to improve the speed of retrieving speech files and the retrieval performance. But we faced with the high-dimensional eigenvectors with the tens or hundreds of dimensions that extracted from speech frame, it will greatly reduce the performance of traditional indexing mechanism that make the index speed even less than a sequential scan which bring about the dimensional disaster. Therefore, how to dig up the features which users are interested from the high-dimensional data is a hot research topic in recent years.

Researches show that when the high-dimensional data approximates to the linear subspace distributed in the high-dimensional vector space, the traditional linear dimension reduction methods, such as the Principal Component Analysis(PCA) [1] and multi dimensional scaling (MDS)[2] and so on, can effectively learn the data of lower dimensional linear structure; But when the sample characteristics containing heterogeneous information, a larger sample scale, irregularity of the multidimensional data or uneven distribution of multidimensional sample, the linear classifier has been no longer qualified to solve the complex classification problems. Faced with the complication and variability of problems in reality, scholars raise a series of dimensionality reduction methods which are manifold learning methods, such as Locality Preserving Projection [3], Isometric Mapping [4], Locally Linear Embedding [5] and Laplacian Eigenmap [6], etc. These manifold methods have a common framework, namely to figure out the local linear feature of data for each neighborhood and map these linear characteristic to a global low-dimensional space, so as to classify the data better. Traditional manifold learning method, as an unsupervised method, doesn't effectively utilize existing information of categories to improve the efficiency of classification.

These algorithms are based on extraction vector method of one-dimensional feature; multi-dimensional to one-dimensional conversion will result in the loss of local spatial information, and bring difficulty to subsequent operations. Kernel learning method is to project the original data item to a vector space called characteristic space, where the linear method is applied. But there are large amounts of calculation in the process of fusing multiple kernel function, Based on the consideration above, using random projection is able to reduce the sample dimensions while keep the approximate separability of different sorts in feature space. By combining the study of literature [7][8], this paper proposed the Local Preserving Projection (MKL-SLPP) Speech recognition method based on the multi-kernel supervision, making the space dimension lower after the projecting and reducing the complexity of calculation.

1 Basic Theory

1.1 Locality Preserving Projection

Supposed the inputting sample is $X = (x_1, x_2, \dots, x_n) \in R^{d \times n}$, so the key to LPP is to find the same structure of neighbor $Y = V^T X = (y_1, y_2, \dots, y_n) \in R^{l \times n}$; $l \ll d$ with the inputting sample X , V is the transform matrix, the objective optimization equation is:

$$\arg \min \sum_{ij} \|y_i - y_j\|^2 W_{ij}$$

(1)

Weight matrix W representing the similarity among samples is defined as follows:

If x_i and x_j are K -nearest neighbor, defined $W_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{\delta})$; otherwise equal to zero. Because:

$$\begin{aligned} \frac{1}{2} \sum_{ij} \|y_i - y_j\|^2 W_{ij} &= \frac{1}{2} \text{tr} \left(\sum_{ij} (V^T x_i - V^T x_j)(V^T x_i - V^T x_j)^T W_{ij} \right) \\ &= \text{tr}(V^T X(D - W)X^T V) = \text{tr}((V^T X L X^T V)) \end{aligned} \quad (2)$$

In the formula(2) matrix D is the constraint matrix, $D_{ii} = \sum_j W_{ij}$, $\forall i$ the matrix D_{ii} is the diagonal matrix of similarity matrix W , $L = D - W$ is the Laplace Matrix. The constraint of the objective function is $V^T X D X^T V = V^T Y V = I$, where $X L X^T$ and $X D X^T$ are symmetric and positive semi-definite. The minimization the objective function can be transformed into the following form:

$$\begin{aligned} \arg \min \text{tr}(V^T X L X^T V) \\ \text{s.t. } V^T X D X^T V = I \end{aligned} \quad (3)$$

The solution of this optimization problem is the problem of generalized eigenvalue decomposition, as follows.

$$X L X^T v = \lambda X D X^T v \quad (4)$$

1.2 Single kernel learning

Given a supervised machine learning problem: assumed $X = (x_1, x_2, \dots, x_N)$, $x_i \in R^d$ is the input space, the output space is $Y = (y_1, y_2, \dots, y_i) \subseteq R$ (regression) or $Y = \{-1, +1\}$ (two types of classification problem). The non-linear mapping $x \mapsto \phi$ needs to be found out in advance, when the samples are in situation of non-linear classification. The training sample can be mapped to the linear high-dimensional Eigenspace by introducing kernel function $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. So the results of Single kernel learning method can be indicated by linear combination methods:

$$f(x) = \sum_{i=1}^N \alpha_i y_i k(x_i, x) + b \quad (5)$$

In formula (5) α_i is the Lagrange multiplier which to obtain through solving a quadratic optimization problem, b is the threshold of classification. $k(x_i, x_j)$

reflects the similarity of samples x_i and x_j . Under the condition of satisfy the Mercer kernel, the kernel function can take many forms to choose from. Frequently used kernel functions are: Ploynomial kernel, Sigmoid kernel and Radial basis function etc.

1.3 Multiple-kernel learning

Assumed we have a group of dataset $\Omega = \{x_i\}_{i=1}^N$ in which containing N kinds of samples, as well as the description of this dataset to be selected the M kinds of features, supposed the set x_i corresponding kernel-set $\{K_m\}$, every kernel K_m Corresponds to a Hilbert space, The new composed kernel matrix is:

$$K = \sum_{m=1}^M \beta_m K_m \quad ; \quad \beta_m \geq 0 \quad (6)$$

$$K^{(i)} = \begin{bmatrix} k_1(1,i) & \cdots & k_M(1,i) \\ \vdots & \ddots & \vdots \\ k_1(N,i) & \cdots & k_M(N,i) \end{bmatrix} \in R^{N \times M}$$

Where β_m is the weight coefficient of combination kernels in Formulas (6), so the classifying tasks of multi-kernle objective function could be denoted by formula(7):

$$f(x) = \sum_{i=1}^N \alpha_i y_i \sum_{m=1}^M \beta_m k_m(x_i, x) + b \quad (7)$$

Lagrange multiplier $\{\alpha_i\}_{i=1}^N$ and weight coefficient $\{\beta_m\}_{m=1}^M$ in the Formula (8) are the main parameters that need to be optimized solution which in the process of the structuring multi-kernel objective function

2 Based on the MKL-SLPP speech recognition

2.1 Multi-kernel Supervision extension of LPP

Assumed speech signal $X = (x_1, x_2, \cdots, x_N), x_i \in R^d$; $x_i \mapsto \phi(x_i), (i = 1, 2, \cdots, N)$ indicate the feature of the implicit mapping by multi-kernel, the representation of vector x_i mapped to feature space can be expressed by $V^T \phi(x_i)$ for uniqueness. Replace the weight coefficient

matrix V of the Formula (2) by $V = \sum_{n=1}^N \alpha_n \phi(x_n)$. The Kernel form of $V^T \phi(x_i)$ can be indicated as the following formula:

$$V^T \phi(x_i) = \sum_{n=1}^N \sum_{m=1}^M \alpha_n \beta_m k_m(x_n, x_i) = \alpha^T K^{(i)} \beta$$

(8)

So each mapping was decided by a certain Lagrange multiplier α and weight vector of kernel β . We can obtain the Minimum optimization objective equation which by using MKL-SLPP when the sample mapped to one-dimensional as follows:

$$\begin{aligned} \min_{\alpha, \beta} \sum_{i,j=1}^N \left\| \alpha^T K^{(i)} \beta - \alpha^T K^{(j)} \beta \right\|^2 w_{ij} \\ s.t. \quad \sum_{i,j=1}^N \left\| \alpha^T K^{(i)} \beta - \alpha^T K^{(j)} \beta \right\|^2 w_{ij}' = 1, \quad \beta_m \geq 0, \quad m = 1, 2, \dots, M \end{aligned}$$

(9)

Extended to P dimensions, it needs the number of P Lagrange multiplier consisting of coefficient matrix $A = [\alpha_1, \alpha_2 \dots, \alpha_P]$. By value A and β can mapped to p-dimensional to obtain $V = [v_1, v_2 \dots, v_P]$. The speech vector x_i is mapped to P dimensional feature space which of expression is: $V^T \phi(x_i) = \alpha^T K^{(i)} \beta \in R^P$.

The Minimum optimization objective equation as follows:

$$\begin{aligned} \min_{A, \beta} \sum_{i,j=1}^N \left\| A^T K^{(i)} \beta - A^T K^{(j)} \beta \right\|^2 W_{ij} \\ s.t. \quad \sum_{i,j=1}^N \left\| A^T K^{(i)} \beta - A^T K^{(j)} \beta \right\|^2 W_{ij}' = 1, \quad \beta_m \geq 0, \quad m = 1, 2, \dots, M \end{aligned}$$

(10)

It can be known from Formula (2) that Formula (10) could be converted into the form as follows:

$$\frac{1}{2} \sum_{i,j=1}^N \left\| A^T K^{(i)} \beta - A^T K^{(j)} \beta \right\|^2 W_{ij} = \text{tr}(A^T S_w^\beta A)$$

(11)

In the formula:

$$S_W^\beta = \sum_{ij=1}^N w_{ij} (K^{(i)} - K^{(j)}) \beta \beta^T (K^{(i)} - K^{(j)})^T$$

$$S_{W'}^\beta = \sum_{ij=1}^N w'_{ij} (K^{(i)} - K^{(j)}) \beta \beta^T (K^{(i)} - K^{(j)})^T$$

If the sample x_i and x_j belong to the same type, structure the nearest neighbor graph and calculate similarity weight. The similarity matrix is:

$$W_{ij} = \begin{cases} \frac{\sqrt{k(x_i, x_j)}}{\sqrt{k(x_i, x_i)} \sqrt{k(x_j, x_j)}}; & \text{If } x_i \text{ and } x_j \text{ are the same samples} \\ 0 & \text{Otherwise} \end{cases}$$

(12)

Basic

kernel

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad \sigma > 0$$

function:

The optimized objective function is converted to:

$$\min_{A, \beta} \text{tr}(A^T S_W^\beta A)$$

$$\text{s.t.} \quad \text{tr}(A^T S_{W'}^\beta A) = 1, \quad \beta_m \geq 0, \quad m = 1, 2, \dots, M$$

(13)

The problem has turned into the settlement of Lagrangian multiplier coefficient $A = \{\alpha_i\}_{i=1}^p$ and kernel weight $\beta = \{\beta_m\}_{m=1}^M$.

2.2 Optimization of parameters

To solve directly the Formula (13) was difficult, we adopted a two-step solution, in the solving process, the first to fix one parameter, then to solve another parameter. Go through the loop, until solving the optimization equation of convergence.

2.3 The algorithm steps

(1) Calculating linear representation of samples of transformation by using the kernel analysis method and selecting similarity matrices W and W' ;

(2) Selecting Basic kernel function and generating the corresponding basic kernel matrix $\{K_m\}_{m=1}^M$;

(3) Calculating the sample coefficient vector $A = [\alpha_1, \alpha_2, \dots, \alpha_p]$ and weight vector β according to the two steps;

(4) Feature vectors of the samples are $Z = A^T K \beta$;

(5) Identifying the classification of the samples waiting for recognition according to the optimization targeted equation of Formula (13).

3 Experiment and analysis

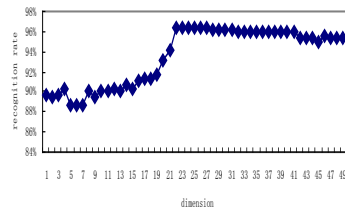
Experiment sample adopts total speech files of 100 which were recorded under the quiet laboratory environment, among which there are 50 men and 50 women, and the file type is WAV, the length of files are all about 1s, with Mono and sampling frequency of 8 KHz and quantization accuracy of 16 bits. A series of preprocessing is needed, such as pre-emphasis, frame, windowing and etc. After the preprocessing procedure, we can take out 50 speech frames effectively from each speech file, frame by frame calculating 12 dimensional MFCC parameters, each speech files are composed by a 50×12 dimensional matrix, these files were not marked. Take out 80 files for template training and 20 files for identification training from them.

Take x_i at random, and use MKL-SLPP to calculate values with other 79 samples x_j , then the training templates could be obtained, if x_j and x_i in accordance with the conditions of Formula (11), they are classified as one category; and then use the remaining samples to do recognition training, the recognition rate is calculated as follows:

$$\text{Recognition rate} = \text{number of successful identification} / 100 \quad (14)$$

Fig.1 displays recognition rate of speech in different mapping dimensional P , from the figure we can see, when the value P is smaller, the recognition rate is lower, indicating that the mapping to the training template part information missing; When $P \geq 22$, recognition success rate is the highest. When values P continue to increase the recognition rate did not continue to raise, Fig.2 shows that training time is also multiplied.

The relationship between the dimensions with the recognition rate



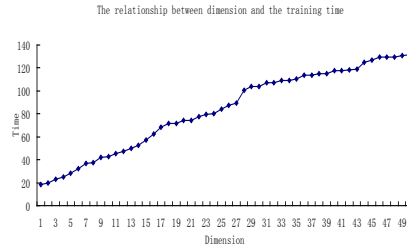


Fig.1 Relationship between recognition rate and dimension. Fig. 2
Relationship between dimension and training time

4 Conclusions

This paper uses the locality preserving projection method of multiple kernel learning supervision, obtaining the speech cepstral features 50×12 matrix through the Speech framing and feature extraction. During the period of multi-kernel learning stage, it uses multi kernel combination mechanism by weighted multiplying to combine the basic kernels together, and adopts the two-step optimization method to learn the kernel parameters and training classification model; through LPP learning, the feature space dimensions are reduced after mapping, the computational complexity is lowered and better recognition effect is gained.

References

- [1] JOLLIFFIT. Principal component analysis[M]. Berlin: Springer-Verlag, 1986.384 -389.
- [2] COX M A A, COX T F. Multidimensional scaling [M]. London: Chapman & Hall,2001. 316-341.
- [3] HE XIAOFEL,NIYOGI P. Locality preserving projections[C]. Proceedings of the 17th Annual Conference of on Neural Information Processing System.Vancouver: MIT press,2004. 153-160.
- [4] JOSHUA B T, VIN DE S, LANG-FORD J C. A global geometric framework for nonlinear dimensionality reduction [J] . Science,2000, 290(5500) 2319-2323.
- [5] SAM T R, LAWRENCE K S. Nonlinear dimensionality reduction by locally linear embedding[J]. Science,2000,290(5000) 2323-2326.
- [6]MIKHAIL B, NIYOGI P. Laplacian Eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003,15(6) 1373-1396.
- [7]Jun-Bao Li, Zhi-Ming Yang, Yang Yu, Zhen Sun.Semi-supervised Kernel Learning Based Optical Image Recognition[J]. Optical Communications. 2012,

Vol.285(18)3697-3703.

- [8]Jun-Bao Li,Jeng-Shyang,Pan,Shyi-Ming,Chen.Kernel.Self-Optimized
Locality Preserving Discriminant Analysis for Feature Extraction and
Recognition[J]. Neurocomputing. 2011.Vol.74(17)3019-3027.