APPLYING OPTIMAL ALGORITHM TO DATA-DEPENDENT KERNEL FOR HYPERSPECTRAL IMAGE CLASSIFICATION

I-Ling Chen[1] esther.x10@gmail.com
Cheng-Hsuan Li [1,2] ChengHsuanLi@gmail.com
Bor-Chen Kuo [1] kbc@mail.ntcu.edu.tw
Hsiao-Yun Huang[3] stat2021@mail.fju.edu.tw

[1] Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taichung, Taiwan, R.O.C.
[2] Department of Electrical and Control Engineering, National Chiao Tung University, Taiwan, R.O.C.
[3] Department of Statistics and Information Science, Fu Jen Catholic University, Taipei, Taiwan, R.O.C.

1. INTRODUCTION

In the kernel methods, it is very important to choose a proper kernel function to avoid overlapping data. Many data-dependent kernel functions have been developed, and a unified kernel optimization framework proposed by [1] is suitable for classifying low-dimensional data. In this paper, we have two objectives. One objective is to apply the above framework on the hyperspectral image, and the other is to use a Fisher criterion with nonparametric weighted feature extraction (NWFE) [2] to reformulate in the pairwise manner [3] as the objective functions under the kernel optimization framework.

2. METHODOLOGY

2.1. Data-dependent kernel

Given a training set $\{x_1, x_2, \cdots, x_n\} \in R^d$. Data-dependent kernel function [5] is defined as

$$k(x, y) = q(x)q(y)k_0(x, y)$$

where $x, y \in R^d$, $k_0(x, z)$ is called basic kernel, and $q(\cdot)$ is a factor function defined as

$$q(x) = \alpha_0 + \sum_{i=1}^{m} \alpha_i k_1(x, a_i)$$

where $k_1(x,a_i)=e^{-\gamma_1\|x-a_i\|^2}$, $\{a_i \in R^d, i=1,\cdots,m\}$ is the empirical cores, and α_i is the combination coefficients. $\boldsymbol{K}$ and $\boldsymbol{K}_0$ are the kernel matrices corresponding to the kernel function $k(x,y)$ and $k_0(x,y)$, respectively. The kernel matrix from the data-dependent kernel function can be represented as

$$\boldsymbol{K}=\boldsymbol{Q}\boldsymbol{K}_0\boldsymbol{Q}, \text{ where } \quad \boldsymbol{Q}=\begin{bmatrix} q(x_1) & 0 & \cdots & 0 \\ 0 & q(x_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & q(x_n) \end{bmatrix}.$$

$$\mathbf{q}=\begin{bmatrix} q(x_1) \\ q(x_2) \\ \vdots \\ q(x_n) \end{bmatrix}=\begin{bmatrix} 1 & k_1(x_1,a_1) & \cdots & k_1(x_1,a_m) \\ 1 & k_1(x_2,a_1) & \cdots & k_1(x_2,a_{m1}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k_1(x_n,a_1) & \cdots & k_1(x_n,a_m) \end{bmatrix}\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}=\boldsymbol{K}_1\boldsymbol{\alpha}.$$

We denote the vectors $[q(x_1),q(x_2),\cdots,q(x_n)]^{\mathrm{T}}$ and $[\alpha_0,\alpha_1,\cdots,\alpha_m]^{\mathrm{T}}$ by $\mathbf{q}$ and $\boldsymbol{\alpha}$, respectively.

2.2 Kernel optimization algorithm

This optimization task is based on the following Fisher criterion,

$$\mathrm{J}(\alpha)=\frac{tr(\boldsymbol{S}^b(\alpha))}{tr(\boldsymbol{S}^w(\alpha))},$$

where

$$tr(\boldsymbol{S}^w(\alpha))=\alpha^T\boldsymbol{K}_1^T(\boldsymbol{L}^{(w)}.*\boldsymbol{K}_0)\boldsymbol{K}_1\alpha \text{ and } tr(\boldsymbol{S}^b(\alpha))=\alpha^T\boldsymbol{K}_1^T(\boldsymbol{L}^{(b)}.*\boldsymbol{K}_0)\boldsymbol{K}_1\alpha,$$

where $\boldsymbol{L}^{(w)}=\boldsymbol{D}^{(w)}-\boldsymbol{A}^{(w)}$ and $\boldsymbol{L}^{(b)}=\boldsymbol{D}^{(b)}-\boldsymbol{A}^{(b)}$, and $\boldsymbol{D}$ is a diagonal matrix with $D_{i,i}=\sum_j A_{i,j}$. α is a vector of the combination coefficients under constraint $\|\alpha\|=1$. $\boldsymbol{A}.*\boldsymbol{B}$ represents the entry-by-entry product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. $\boldsymbol{A}^{(w)}$ and $\boldsymbol{A}^{(b)}$ are the affinity matrices corresponding to within-class-scatter matrix and between-class-scatter matrix, respectively.

$$A_{i,j}^{(w)}=\begin{cases} \dfrac{1}{n_c} & if \quad y_i=y_j=c \\ 0 & if \quad y_i\neq y_j \end{cases}, \quad A_{i,j}^{(b)}=\begin{cases} \dfrac{1}{n}-\dfrac{1}{n_c} & if \quad y_i=y_j=c \\ \dfrac{1}{n} & if \quad y_i\neq y_j \end{cases},$$

where y_i is the class of the *i*-th sample.

We employ the standard gradient method to iterate and update the combination coefficients $\boldsymbol{\alpha}$ with an appropriate stepsize η

$$\alpha_{(t+1)} = \alpha_{(t)} + \eta_{(t)} \frac{\partial J(\alpha_{(t)})}{\partial \alpha_{(t)}},$$

where $\eta_{(t)} = \eta_0(1 - \frac{t}{T})$, η_0 is the initial stepsize rate, T is a pre-specified total number of iterations, and t stands for the current number of iterations.

2.3 Optimizing Kernel-based NWFE

The between-class scatter matrix $\boldsymbol{S}_b^{KNW}$ and the within-class scatter matrix $\boldsymbol{S}_w^{KNW}$ of KNWFE in the feature space H are

$$\boldsymbol{S}_b^{KNW} = \sum_{i=1}^{L} P_i \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{\ell=1}^{N_i} \frac{\lambda_\ell^{(i,j)}}{N_i} (\phi(x_\ell^{(i)}) - M_j(\phi(x_\ell^{(i)})))(\phi(x_\ell^{(i)}) - M_j(\phi(x_\ell^{(i)})))^T$$

and

$$\boldsymbol{S}_w^{KNW} = \sum_{i=1}^{L} P_i \sum_{\ell=1}^{N_i} \frac{\lambda_\ell^{(i,i)}}{N_i} (\phi(x_\ell^{(i)}) - M_i(\phi(x_\ell^{(i)})))(\phi(x_\ell^{(i)}) - M_i(\phi(x_\ell^{(i)})))^T$$

where the scatter matrix weight $\lambda_\ell^{(i,j)}$ is defined by

$$\lambda_\ell^{(i,j)} = \frac{\mathrm{dist}(\phi(x_\ell^{(i)}), M_j(\phi(x_\ell^{(i)})))^{-1}}{\sum_{t=1}^{N_i} \mathrm{dist}(\phi(x_t^{(i)}), M_j(\phi(x_t^{(i)})))^{-1}},$$

$M_j(\phi(x_\ell^{(i)})) = \sum_{k=1}^{N_j} w_{\ell k}^{(i,j)} \phi(x_k^{(j)})$ denotes the weighted mean with respect to $\phi(x_\ell^{(i)})$ in class j and

$$w_{\ell k}^{(i,j)} = \frac{\mathrm{dist}(\phi(x_\ell^{(i)}), \phi(x_k^{(j)}))^{-1}}{\sum_{t=1}^{N_j} \mathrm{dist}(\phi(x_\ell^{(i)}), \phi(x_t^{(j)}))^{-1}}.$$

Note that

$$tr(\boldsymbol{S}_b^{KNW}) = \sum_{i=1}^{L} P_i \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{\ell=1}^{N_i} \frac{\lambda_\ell^{(i,j)}}{N_i} (\phi(x_\ell^{(i)}) - M_j(\phi(x_\ell^{(i)})))^T (\phi(x_\ell^{(i)}) - M_j(\phi(x_\ell^{(i)}))) = \boldsymbol{\alpha}^T \boldsymbol{K}_1^T \boldsymbol{B} \boldsymbol{K}_1 \boldsymbol{\alpha}$$ and

$$tr(\boldsymbol{S}_w^{KNW}) = \sum_{i=1}^{L} P_i \sum_{\ell=1}^{N_i} \frac{\lambda_\ell^{(i,i)}}{N_i} (\phi(x_\ell^{(i)}) - M_i(\phi(x_\ell^{(i)})))^T (\phi(x_\ell^{(i)}) - M_i(\phi(x_\ell^{(i)}))) = \boldsymbol{\alpha}^T \boldsymbol{K}_1^T \boldsymbol{W} \boldsymbol{K}_1 \boldsymbol{\alpha}.$$

The optimizer can be obtained by the gradient method described in the section 2.2.

3. SOME EXPERIMENTAL RESULTS

For investigating the influences of small training sample sizes to the dimension, three distinct cases, N_u=20 (case 1), N_u=40 (case 2) and N_u=300 (case 3), are discussed, and for investigating the influences of different kernel functions to the effect of classification, we apply these kernel functions to support vector machine. Three common hyperspectral image sources are used in this paper, and they are Indian Pine Site, Washington DC Mall and Kennedy Space Center. The experimental results display the superiority of the optimizing kernel function over the RBF kernel function with 5-fold cross-validation method, especially, in the Indian Pine Site Image.

4. REFERENCES

[1] Bo Chen, Hongwei Liu, and Zheng Bao, "Optimizing the Data-dependent Kernel under A Unified Kernel Optimization Framework," *Pattern Recognition*, Vol. 41,No. 6, pp. 2107-2119,2008.

[2] 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 Anal. Mach. Intell.*, Vol. 29 No. 1, pp. 40–51, 2007.

[3] B. C. Kuo and D. A. Landgrebe, "Nonparametric Weighted Feature Extraction for Classification," *IEEE Trans. Geosci. Remote Sens.*, Vol.42, No. 5, pp. 1096-1105, 2000.

[4] B. C. Kuo, C. H. Li, J. M. Yang, '' Kernel Nonparametric Weighted Feature Extraction for Hyperspectral Image Classification,'' *IEEE Trans. Geosci. Remote Sens.*, Vol. 47, No. 4, pp. 1139-1155, 2009.

[5] Huilin Xiong, and M. Omair Ahmad, "Optimizing the Kernel in the Empirical Feature Space," *IEEE Trans. Neural Network*, Vol. 16, No. 2, pp. 460-474,2005.

[6] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed.New York: Springer-Verlag, 2001.

[7] Fukunaga K., *Introduction to Statistical Pattern Recognition*, CA: Academic, San Diego, 1990.