Linear Discriminant Analysis Cauchy Estimator for Single Chinese Character Font Recognition

Yanan Guo

School of Mathematics and Statistics, Yunnan University, Kunming, China.
Email: yanan.guo.ynu@qq.com

Abstract. In machine learning area, the rapid development of optical character recognition (OCR) has prompted interest in Chinese character font recognition (CCFR), especially single Chinese character font recognition. However, when pre-processing pollutes Chinese font images with noise, traditional font recognition algorithms tend not to be suitably discriminant. In this paper, and based on linear discriminant analysis and Cauchy estimator theory, we propose a novel feature selection algorithm called linear discriminant analysis Cauchy estimator (LDACE) for single Chinese character font recognition. LDACE aims to: (1) consider both between-class and within-class geometry in the low-dimensional space, and (2) preserve recognition when input samples are polluted with noise. Experiments with the frequently used FUCCFR dataset demonstrate LDACE’s effectiveness.

1. Introduction
Font recognition, such as Arabic font recognition [1] and Chinese character font recognition, is an important research area for intelligent applications based on optical character recognition (OCR). Font categorization can be thought of as a special image classification scenario, and it has uses in many real-world applications such as high-performance document recovery to obtain the document content and typeface.

Font recognition systems generally follow three steps: feature extraction, feature representation or selection, and classification. Pyramid histogram of orientation gradients (PHOG) feature [2] is the most popular low feature extraction method. PHOG considers the local pyramid histogram in blocks using gradients, it express local features well and can be used to overcome the effects of noise. Locality-constrained linear coding (LLC) [3] feature is the widely used feature representation methods. LLC is applied to the locally linear structure of the input sample and it is highly effective for image classification. However, single Chinese font recognition is characterized by many samples and high-dimensional features after feature extraction, and it can incur high computational costs with immediate classification; therefore, feature selection with dimension reduction is important in single Chinese font recognition. Many feature selection algorithms have been proposed that have greatly improved the quality of feature extraction. However, when Chinese font images are polluted with noise during pre-processing, these feature selection algorithms are less discriminant.

Here we propose a new feature selection algorithm that we call the linear discriminant analysis Cauchy estimator (LDACE). The main contribution has two folds: (1) LDACE considers both the between-class and within-class geometric structure in the low-dimensional space, and (2) by utilizing our single Chinese character font recognition process, recognition is improved when input samples are noisy. Specifically, our single Chinese character font recognition process has four steps: (1) selecting 12 single Chinese character sample classes from the FUCCFR dataset [4], where the size of each original Chinese character is $48 \times 48$; (2) using PHOG to extract features; (3) training the LDACE
2. The Linear Discriminant Analysis Cauchy Estimator

To overcome the limitations of existing techniques, we propose a novel feature selection method that we call the Linear Discriminant Analysis Cauchy estimator (LDACE) for single Chinese character font recognition. In the single Chinese character font recognition problem, a set of training feature vectors $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{D \times N}$ belonging to $C$ are sorted as $C = [c_1, c_2, \ldots, c_k]$, where $x_i$ is a D-dimensional feature vector and $N$ is the number of training feature vectors $X$. $N_j$ training samples are contained in the $j$th class. Applying the new LDACE subspace selection method, a projection matrix $U = [u_1, u_2, \ldots, u_d] \in \mathbb{R}^{D \times d}$ is found, where $d < D$. The new subspace is obtained by $Y = U^T X = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{d \times N}$, with the new succinct and representative feature $y_i$ benefiting subsequent single Chinese character font recognition.

In traditional feature selection, the least squares method is often used to weight the variance between samples; however, this results in poor recognition when images or signal are polluted by noise. More recently, Cauchy estimator theory [6] has been applied to many fields to deal with noise. We also use Cauchy estimator theory to replace the least squares method and reduce noise effects. The Cauchy estimator can be represented as:

$$H(x) = \log \left( 1 + \left( \frac{x}{c} \right)^2 \right)$$ (1)

We next explain how this theory overcomes noise effects. First, $H(x)$ increases slower than the square function, which can be thought of as follows. $h(x)$ is the corresponding influence function of $H(x)$:

$$h(x) = \frac{d \left( \phi(x) \right)}{dx} = \frac{2x}{c^2 + x^2}$$ (2)

In equation (2), $h(x)$ is the measure of the rate of change of $H(x)$, and the larger the absolute value of $h(x)$, the faster the change in $H(x)$. Second, $H(x)$ limits growth; from equation (2), $h(x)$ must be less than the constant $c$, so the rate of change of $H(x)$ is smaller than that of the least squares method. According to this function’s properties, even if the dataset samples are contaminated with noise, the Cauchy estimator will decrease the error rate.

2.1. Linear Discriminant Analysis

LDA was first proposed by Fisher [7], who considered that subspace samples in different classes can be well distinguished. Subspace selection with LDA makes within-class distances as close as possible and between-class distances as large as possible. The within-class scatter matrix $S_w$ and the between-class scatter matrix $S_b$ are defined as:

$$S_w = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (x_j - \bar{x}_j)(x_j - \bar{x}_j)^T,$$

and $S_b = \sum_{j=1}^{C} N_j (\mu_j - \bar{x})(\mu_j - \bar{x})^T$ (3)

where $\bar{x}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_j$ is the mean of the $j$th class in high-dimensional space and $\bar{x} = \frac{1}{N} \sum_{j=1}^{C} \sum_{i=1}^{N_j} x_j$ is the mean of all samples in high-dimensional space.

The LDA model maximizes the trace of the between-class scatter matrix $S_b$ and simultaneously minimizes the trace of the within-class scatter matrix $S_w$. 

feature selection matrix using training samples; and (4) classifying projected LDACE samples using the LIBSVM [5] classifier.
\[
\arg\max_{U} \left( \text{tr} \left( U^{T} S_{b} U \right) - \beta \text{tr} \left( U^{T} S_{w} U \right) \right)
\]  \hspace{1cm} (4)

According to matrix trace rules, the objective function can be transformed as follows:
\[
\arg\min_{U} \left( \text{tr} \left( U^{T} \left( S_{w} - \gamma S_{b} \right) U \right) \right) = \arg\min_{U} \left( \text{tr} \left( U^{T} L_{p} U \right) \right)
\]  \hspace{1cm} (5)

where \( L_{p} = S_{w} - \gamma S_{b} \).

2.2. The Cauchy Estimator
In general, LDA performs well on samples without noise, with performance decreasing dramatically when samples are contaminated with noise. Applying the Cauchy estimator to equation (5), we can rewrite it as
\[
\arg\min_{U} \log \left( 1 + \frac{\text{tr} \left( U^{T} L_{p} U \right)}{c^{2}} \right)
\]  \hspace{1cm} (6)

2.3. Margin Maximization
Due to the training samples being too close in proximity for accurate determination in the high-dimensional space, we want to separate training samples as much as possible in the low-dimensional subspace.

For each training sample in the low-dimensional subspace, we maximize the sum of the distance between training samples and the mean of all the training samples:
\[
\arg\max_{y} \sum_{i=1}^{N} \left\| y_{i} - \bar{y} \right\|^{2}
\]  \hspace{1cm} (7)

where \( \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_{i} \) is the mean of all of the training samples in the low-dimensional subspace.

Combine \( Y = U^{T} X \), we can rewrite equation (7) as
\[
\arg\max_{y} \sum_{i=1}^{N} \left\| y_{i} - \bar{y} \right\|^{2} = \arg\max_{y} \sum_{i=1}^{N} \text{tr} \left[ (y_{i} - \bar{y})(y_{i} - \bar{y})^{T} \right] \\
= \arg\max_{U} \text{tr} \left[ \sum_{i=1}^{N} \left[ U^{T} \left( x_{i} - \bar{x} \right) \left( x_{i} - \bar{x} \right)^{T} U \right] \right] = \arg\max_{U} \text{tr} \left( U^{T} MU \right)
\]  \hspace{1cm} (8)

where \( M = \sum_{i=1}^{N} (x_{i} - \bar{x})(x_{i} - \bar{x})^{T} \).

2.4. L2 Penalty Term
To avoid over fitting, we apply the \( l_{2} \)-norm to our objective function. Combining equation (8) and (10), the objective function becomes:
\[
\arg\min_{U} \log \left( 1 + \frac{\text{tr} \left( U^{T} L_{p} U \right)}{c^{2}} \right) - \alpha_{1} \text{tr} \left( U^{T} S U \right) + \alpha_{2} \left\| U \right\|^{2}, \quad \text{s.t.} \ U^{T} U = I
\]  \hspace{1cm} (9)

Utilizing Lagrange’s multiplier, equation (9) can be simplified and transformed into an eigenvalue problem. We can obtain the projection matrix \( U \) given by \( d \) eigenvectors ordered by the \( d \) smallest eigenvalues.
2.5. Optimization
Applying Lagrange’s multiplier, we simplify equation. (9) to
\[
\arg \min_u \log \left( 1 + \frac{\text{tr}(U^T L_u U)}{c^2} \right) - \alpha \text{tr}(U^T MU) + \alpha_2 \|U\|^2 + \lambda (U^T U - I) \text{, s.t. } U^T U = I
\]
(10)
Setting the gradient of above function with respect to \(U\) as 0,
\[
\frac{L_p U}{c^2 + \text{tr}(U^T L_p U)} - \alpha_t MU + 2 \alpha_2 U + 2 + \lambda U = 0
\]
(11)
This equation can be rewritten as
\[
\left( \frac{L_p}{c^2 + \text{tr}(U^T L_p U)} - \alpha_t M + 2 \alpha_2 I \right) U = \lambda U
\]
(12)
Setting \(E_p = \text{tr}(U^T L_p U)\), immediately solving the eigenvectors of \(\frac{L_p}{c^2 + E_p} - \alpha_t M + 2 \alpha_2 I\) is difficult because the projection matrix \(U\) is unknown. To obtain the eigenvectors of \(\frac{L_p}{c^2 + E_p} - \alpha_t M + 2 \alpha_2 I\), we randomly initialize the values of projection matrix \(U\). Then, projection matrix \(U\) is updated iteratively using equation. (12) until \(U\) converges.

3. Experiments
We next conducted comparative experiments on the FUCCFR dataset to illustrate LDACE’s performance. The FUCCFR dataset contains seven Chinese typefaces combined with four font styles. It comprises 28 classes of Chinese character samples, and each font set consists of over three thousand different simplified Chinese characters. To demonstrate the advantages of LDACE compared to other feature selection algorithms on noisy data, white Gaussian noise was added to all FUCCFR samples, specifically, we selected 12 Chinese character sample classes including four Chinese typefaces (Song, Kai, Lishu, and Xingkai) combined with three font styles (normal, bold, and italic), the signal-to-noise ratio (SNR) was set to 10. Details of the experimental setup are presented below.

In the FUCCFR dataset, each font set consisted of 3744 different simplified Chinese character samples. We randomly selected 2000 training samples from each class and tested with the remaining samples. The training set was used to calculate the projection matrix, and the test set was used to examine recognition results. The LIBSVM classifier was used for recognition and to compute the recognition rate from the 2-dimensional to the 50-dimensional in increments of 2. Experiments were randomly repeated 5 times for fairness.

After adding white Gaussian noise to the dataset, we extracted PHOG features to represent the appearance and shape of single Chinese character font images. The parameter setting and basic procedure were as follows: 2 × 2 cells were integrated into a block, and each block contained 24×24 pixels, 8×8 pixels, and 2×2 pixels. Fifty percent of blocks overlapped during sliding. The four-bin PHOG descriptor was extracted from each block to obtain a 16-dimensional representation. All features were concatenated into a single long vector, a 7136-dimensional representation.

LDACE was compared to five typical feature selection algorithms: linear discriminant analysis (LDA), principal component analysis (PCA), supervised locality-preserving projections (SLPP) [8], orthogonal locality-preserving projections (OLPP) [9], and sparse principal component analysis (SPCA) [10]. To accelerate learning, PCA projection was conducted as a pre-processing step prior to applying the supervised algorithms. From figure. 1, LDACE achieve a higher average recognition rate at higher dimensions for datasets polluted with noise level of SNR = 10. LDACE is robust due to the
use of the Cauchy estimator and considers both between-class and within-class geometry in the low-dimensional space; therefore, it outperforms other feature selection algorithms when samples are noisy.

4. Conclusions
CCFR is extremely important for intelligent applications based on Chinese character fonts, and many useful algorithms have been proposed for this purpose. However, in real-word applications, these systems tend to be very complex and samples are easily polluted with noise during pre-processing. Previous algorithms have not adequately considered the effects of noise. In this paper, we propose a novel feature selection method called LDACE to address the above problem. LDACE is robust due to the use of the Cauchy estimator and considers both between-class and within-class geometry in the low-dimensional space; therefore, it outperforms other feature selection algorithms when samples are noisy.

![Graph showing recognition rate vs. subspace dimension on the FUCCFR dataset at a noise level of SNR = 10.]

Figure 1. The average recognition rate vs. the subspace dimension on the FUCCFR dataset at a noise level of SNR = 10.

5. Acknowledgments
This work was supported in part by the Yunnan University’s Research Innovation Fund for Graduate Students, in part by the Yunnan University’s Research Innovation Fund for Graduate Students under Grant YDY17103, and in part by the Yunnan Province’s New Academic Researcher Award for Graduate Students.

6. References
[1] Moussa S, Zahour A, Ben Abdelhafid A, and Alimi A New features using fractal multi-dimensions for generalized Arabic font recognition 2010 Pattern Recognition Lett. 31 5 pp 361 – 371
[2] Bai Y, Guo L, Jin L, and Huang Q A novel feature extraction method using Pyramid Histogram of Orientation Gradients for smile recognition 2009 IEEE International Conference on Image Processing, pp 3305 – 3308
[3] Wang J, Yang J, Yu K, Lv F, Huang T, and Gong Y Locality-constrained linear coding for image classification 2010 IEEE International Conference on Computer Vision and Pattern Recognition pp 3360 – 3367
[4] Tao D, Lin X, Jin L, and Li X Principal Component 2-D Long Short-Term Memory for Font Recognition on Single Chinese Characters 2016 IEEE trans. on cybern. 46 3 pp 756 - 765
[5] Chang C and Lin C LIBSVM: a Library for Support Vector Machines 2006 ACM Trans. on Intelligent Systems & Tech. 2 3 pp 389 – 396
[6] Mizera I and Muller C Breakdown points of cauchy regression-scale estimators 2002 Statistics & probability Lett. 57 1 pp 79 – 89
[7] Fisher R and F R The use of multiple measurements in taxonomic problems 1936 Ann. of Eugen. 7 2 pp 179 – 188
[8] He X and Niyogi P Locality Preserving Projections 2003 Neural Information Processing Systems 16 pp 186 – 197
[9] Cai D, He X, Han J, and Zhang H Orthogonal Laplacianfaces for Face Recognition 2006 IEEE Trans. on Image Process. 15 11 pp 3608 – 14
[10] Zou H, Hastie T, and Tibshirani R Sparse principal component analysis 2006 J. Comput. Graphical Stat. 15 2 pp 262 – 286