Improved Minimum Squared Error Algorithm with Applications to Face Recognition

Qi Zhu[1,2,3], Zhengming Li[1,3,4], Jinxing Liu[5], Zizhu Fan[1,6], Lei Yu[7], Yan Chen[8]*

1 Bio-Computing Center, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China, 2 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China, 3 Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, China, 4 Guangdong Industrial Training Center, Guangdong Polytechnic Normal University, Guangzhou, China, 5 College of Information and Communication Technology, Qufu Normal University, Rizhao, China, 6 School of Basic Science, East China Jiaotong University, Nanchang, China, 7 School of Urban Planning and Management, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China, 8 Shenzhen Sunwin Intelligent Co., Ltd., Shenzhen, China

Abstract

Minimum squared error based classification (MSEC) method establishes a unique classification model for all the test samples. However, this classification model may be not optimal for each test sample. This paper proposes an improved MSEC (IMSEC) method, which is tailored for each test sample. The proposed method first roughly identifies the possible classes of the test sample, and then establishes a minimum squared error (MSE) model based on the training samples from these possible classes of the test sample. We apply our method to face recognition. The experimental results on several datasets show that IMSEC outperforms MSEC and the other state-of-the-art methods in terms of accuracy.

Introduction

The minimum squared error based classification (MSEC) is sound in theory and is able to achieve a high accuracy [1,2]. It has been proven that for two-class classification MSEC is identical to linear discriminant analysis (LDA) under the condition that the number of training samples approximates the infinity [1,2]. In addition, MSEC can be applied to multi-class classification by using a special class label matrix [3]. Various improvements to MSEC such as orthogonal MSEC [4] and kernel MSEC [5–8] have been proposed. The MSEC has been applied to a number of problems such as imbalanced classification [7], palm-print verification [9], low-rank representation [10,11], super-resolution learning [12], image restoration [13], and manifold learning [14].

In recent years, representation based classification (RC) method [15–18] has attracted increasing attention in pattern recognition. The main difference between RC and MSEC is that RC tries to use the weighted sum of all the training samples to represent the test sample, whereas MSEC aims to map the training samples to their class labels. RC can be categorized into two types. The first type is the so-called sparse representation method (SRM) such as the methods proposed in [19,20]. The goal of SRM is to simultaneously minimize the L_1 norm of the weight vector and the representation error that is the deviation between constructed sample and test sample. The second type is the so-called non-sparse representation method such as the methods proposed in [21–26]. The goal of the non-sparse representation method is to simultaneously minimize the L_2 norm of the weight vector and the representation error. The non-sparse representation method has a closed-solution and is usually more computationally efficient than SRM [21].

In this paper, we focus on multi-class classification problem and propose an improved minimum squared error based classification (IMSEC) method. The basic idea of IMSEC is to select a subset of training samples that are similar to the test sample and then build the MSE model based on them. The advantage of the IMSEC is that it seeks the optimal classifier for each test sample. However, MSEC categorizes all the test samples based on a unique classifier. Therefore IMSEC has better performance than MSEC.

The minimum Squared Error Based Classification for Multi-class Problems

Suppose that there are N training samples from c classes. Let the p-dimensional row vector x_i denote the i-th training sample, where $i=1,...,N$. We use a c-dimensional row vector g_i to represent the class label of the i-th training sample. If this sample is from class k, the k-th entry in g_i is one and the other entries are all zeroes.

If a mapping Y can approximately transform each training sample into its class label, we have

$$XY = G. \qquad (1)$$

where $X = \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_N \end{bmatrix}, G = \begin{bmatrix} g_1 \\ . \\ . \\ . \\ g_N \end{bmatrix}$. Clearly, X is an $N \times p$ matrix, G is an $N \times c$ matrix, and Y is a $p \times c$ matrix that is to be solved. As Eq. (1) cannot be directly solved, we convert it into the following

Figure 1. The face images of one subject in the ORL database.
doi:10.1371/journal.pone.0070370.g001

equation:

$$X^T X Y = X^T G \tag{2}$$

If $X^T X$ is non-singular, Y can be solved by

$$Y = (X^T X)^{-1} X^T G. \tag{3}$$

In general, we use $Y = (X^T X + \gamma I)^{-1} X^T G$ to obtain a stable numerical solution, where γ and I denote a small positive constant and the identity matrix, respectively.

Finally, we classify a test sample t as follows: the class label of t is predicted using tY, and then the Euclidean distance between tY and the class label of each class is calculated, respectively. The class label of the j-th class is a row vector whose j-th element is one and whose other elements are all zeros ($j = 1,2...,c$). Among the c classes, if tY is closest to the k-th class, then x is classified into the k-th class.

The Algorithm of Improved Minimum Squared Error Based Classification

Suppose the j-th class has n_j training samples. Let z_j^k be the k-th training sample of the j-th class, where $k = 1,...,n_j, j = 1,...,c$. The algorithm of IMSEC has the following three steps.

Step 1. Determine L possible classes of the test sample, where $L < c$. First, the test sample t is represented as a weighted sum of the training samples of each class, respectively. For the j-th class, it is assumed that $t = \sum_{k=1}^{n} w_j^k z_j^k$ is approximately satisfied.

$t = \sum_{k=1}^{n} w_j^k z_j^k$ can be rewritten as $t = Z_j W_j$, where $W_j = [w_j^1...w_j^n]^T$ and $Z_j = [z_j^1...z_j^n]$. Then we have $\bar{W}_j = (Z_j^T Z_j + \gamma I)^{-1} Z_j y$. $\|t - Z_j \bar{W}_j\|$ is the representation error between the training samples of j-th class and the test sample. The L classes that have the smallest L representation errors are determined, and they are referred to as base classes.

Step 2. Use the base classes to establish the following MSE model

$$\hat{X} \hat{Y} = G_z, \tag{4}$$

where $\hat{X}$ is composed of all the training samples of the base classes, and G_z is composed of the class labels of these training samples. $\hat{Y}$ is computed using $\hat{Y} = (\hat{X}^T \hat{X} + \mu I)^{-1} \hat{X}^T G_z$. μ and I are a small positive constant and identity matrix, respectively.

Step 3. Exploit $\hat{Y}$ and G_z to classify the test sample t. The class label of this test sample can be predicted by using $t\hat{Y}$. Calculate the Euclidean distance between $t\hat{Y}$ and the class label of each base class, respectively. Let dif_k denote the distance of the between g_t and the class label of the k-th class. If $h = \arg\min_k dif_k$, then test sample t is assigned into the h-th class.

Analysis of the Proposed Method

The proposed method and the MSEC have the following differences. MSEC attempts to obtain a unique model for all the test samples, whereas the proposed method constructs a special MSE for each test sample. MSEC tries to minimize the mean square error between the predicted class labels and the true class labels of the training samples. That means MSEC is capable of

Figure 2. The face images of one subject in the AR database.
doi:10.1371/journal.pone.0070370.g002

Figure 3. Some face images from the FERET database.
doi:10.1371/journal.pone.0070370.g003

mapping the training samples to the correct class labels. However, this does not imply that the model of MSEC can map the test sample to the correct class label accurately. Since the test sample and the training samples that are "close" to the test sample have the similar MSE models, it can be expected that IMSEC performs better in mapping the test sample to the correct class label than CMSE.

The proposed method works in the way of coarse-to-fine classification. In detail, step 1 of the proposed method indeed roughly identifies the possible classes of the test sample. Step 2 of the proposed method assigns the test sample into one of the possible classes. For the complicated classification problem, the way of coarse-to-fine classification is usually more effective than the way in one step [27–29].

It is worth pointing out that the proposed method is different from CRC [21] and linear regression based classification (LRC) [30]. The proposed method tries to establish a model to map the training samples to their true class labels, whereas CRC uses the weighted combination of all the training samples to represent the test sample, and LRC uses the class-specific training samples to represent the test sample. Moreover, when classifying a test sample, the proposed method and LRC need to solve one and C MSE models, respectively, where C is the number of the classes. As a result, the proposed method is more efficient than LRC.

Experiments

A. Ethics Statement

Some face datasets were used in this paper to verify the performance of our method. These face datasets are publicly available for face recognition research, and the consent was not needed. The face images and the experimental results are reported in this paper without any commercial purpose.

B. Experimental Results

Face recognition has become a popular pattern classification task. We perform the experiments on ORL, FERET and AR face databases. Our method, CMSE, CRC, SRC, Eigenface [31], Fisherface [32], Nearest Neighbor Classifier (1-NN), 2DPCA [33], Alternative-2DPCA [34], 2DLDA [35], Alternative-2DLDA [36] and 2DPCA+2DLDA [37] were tested in the experiments. Before implementing each method, we converted every face image into a unit vector with the norm of 1. When CRC was implemented, the regular parameter was set to 0.001. In Eigenface method, we used the first 50, 100…, 400 Eigenfaces for feature extraction, respectively, and reported the lowest error rate. In the 2D based subspace methods, including 2DPCA, Alternative-2DPCA, 2DLDA, Alternative-2DLDA and 2DPCA+2DLDA, the number of the projection axes was set to 1,2,…,5, and the lowest error rate was reported.

In the ORL database, there are 40 subjects and each subject has 10 different images. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no

Table 1. Rates of classification errors of the methods on the ORL database (%).

Number of the original training samples per class	3	4	5	6
The proposed method	11.07	5.83	4.50	1.87
CMSE	13.93	7.92	7.50	3.75
CRC	15.36	9.17	8.00	5.63
SRC	19.29	15.00	14.50	11.87
Eigenface	26.07	20.00	14.00	10.00
Fisherface	23.01	22.64	23.29	9.08
1-NN	20.36	15.00	14.00	8.75
2DPCA	14.29	11.25	9.50	3.75
Alternative-2DPCA	13.93	10.42	8.50	3.75
2DLDA	11.79	7.92	9.50	4.37
Alternative-2DLDA	17.50	13.75	13.50	4.37
2DPCA+2DLDA	16.07	12.50	10.00	4.37

doi:10.1371/journal.pone.0070370.t001

Table 2. Rates of classification errors of the methods on the AR database (%).

Number of the original training samples per class	4	5	6	7
The proposed method	25.27	23.29	22.96	20.92
CMSE	27.92	24.88	25.87	25.48
CRC	29.89	28.02	29.71	28.90
SRC	41.97	43.41	34.04	29.78
Eigenface	41.78	47.66	24.79	26.05
Fisherface	44.17	40.71	25.13	23.89
1-NN	37.69	39.40	25.87	25.04
2DPCA	40.38	41.87	30.83	31.36
Alternative-2DPCA	40.23	40.87	30.63	31.54
2DLDA	50.68	52.22	35.33	33.25
Alternative-2DLDA	54.09	55.83	41.96	36.40
2DPCA+2DLDA	35.53	37.90	26.42	28.03

doi:10.1371/journal.pone.0070370.t002

Table 3. Rates of classification errors of the methods on the FERET database (%).

Number of the original training samples per class	5	6
The proposed method	19.00	3.50
CMSE	29.25	7.50
CRC	38.75	29.50
Eigenface	37.00	36.00
Fisherface	47.50	61.00
SRC	59.25	52.50
1-NN	28.50	27.00
2DPCA	35.25	49.50
Alternative-2DPCA	36.00	50.00
2DLDA	29.25	25.25
Alternative-2DLDA	29.25	31.50
2DPCA+2DLDA	30.50	35.00

doi:10.1371/journal.pone.0070370.t003

glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). Each face image contains 92×112 pixels, with 256 grey levels per pixel [38]. We resized each face image into a 46 by 56 matrix. Figure 1 shows the face images of one subject in the ORL database. We took the first three, four, five and six face images of each subject as training images and treated the others as test images, respectively. In our method, L was set to $0.3 \times c$.

For AR face database, we used 3120 gray face images from 120 subjects, each providing 26 images [39]. These images were taken in two sessions and show faces with different facial expressions, in varying lighting conditions and occluded in several ways. Figure 2

shows the 26 face images of one subject in the AR database. We took the first four, five, six, seven and eight face images of each subject as training images and treated the others as test images, respectively. In our method, L was set to $0.15 \times c$.

A subset of the FERET face database is used to test our method. This subset includes 200 subjects, and each subject has 7 images. It is composed of the images whose names are marked with two-character strings: 'ba', 'bj', 'bk', 'be', 'bf', 'bd', and 'bg'. This subset involves variations in facial expression, illumination, and pose [40]. The facial portion of each original image was cropped to form a 40×40 image. Figure 3 shows some face images from the FERET database. We took the first five and six face images of each subject as training images and treated the others as test images, respectively. In our method, L was set to $0.15 \times c$.

Tables 1, 2 and 3 show the classification error rates of the methods on the ORL, AR and FERET databases, respectively. We can observe that our method always obtains the lowest classification error rate. In other words, our method can achieve the desirable classification result.

Conclusions

The proposed method, i.e. IMSEC, establishes a special MSE model for each test sample. When building the classification model, IMSEC uses only the training samples that are close to the test sample. Theoretical analyses were presented to explore the properties of IMSEC. Compared with MSEC that classifies all the test samples based on a unique model, IMSEC can perform better in classifying the test samples. We tested the proposed method on three face datasets. The experimental results clearly demonstrated that IMSEC outperforms MSEC and the other state-of-the-art methods.

Author Contributions

Conceived and designed the experiments: QZ ZML. Performed the experiments: QZ JXL ZZF. Analyzed the data: QZ ZZF. Contributed reagents/materials/analysis tools: QZ LY. Wrote the paper: QZ ZML YC.

References

1. Duda RO, Hart PE, Stork DG (2001) Pattern Classification(2nd ed.). Wiley-Interscience Publication.
2. Xu J, Zhang X, Li Y (2001) Kernel MSEC algorithm: a unified framework for KFD, LS-SVM and KRR. International Joint Conference on Neural Networks: 1486–1491.
3. Ye J (2007) Least Squares Linear Discriminant Analysis. Proc. Int'l Conf.Machine Learning: 1087–1094.
4. Chen S, Hong X, Luk BL, Harris CJ (2009) Orthogonal-least-squares regression: A unified approach for data modeling. Neurocomputing 72(10–12): 2670–2681.
5. Xu Y, Zhang D, Jin Z, Li M, Yang JY (2006) A fast kernel-based nonlinear discriminant analysis for multi-class problems. Pattern Recognition 39(6): 1026–1033.
6. Zhu Q (2010) Reformative nonlinear feature extraction using kernel MSE. Neurocomputing 73(16–18): 3334–3337.
7. Wang J, You J, Li Q, Xu Y (2012) Extract minimum positive and maximum negative features for imbalanced binary classification. Pattern Recognition 45: 1136–1145.
8. Xu Y (2009) A new kernel MSE algorithm for constructing efficient classification procedure. International Journal of Innovative Computing, Information and Control 5(8): 2439–2447.
9. Zuo W, Lin Z, Guo Z, Zhang D (2010) The multiscale competitive code via sparse representation for palmprint verification. IEEE Conference on Computer Vision and Pattern Recognition: 2265–2272.
10. Liu G, Lin Z, Yu Y (2010) Robust subspace segmentation by low-rank representation. International Conference on Machine Learning: 663–670.
11. Liu R, Lin Z, Torre FD, Su Z (2012) Fixed-rank representation for unsupervised visual learning. IEEE Conference on Computer Vision and Pattern Recognition: 598–605.
12. Lin Z, He J, Tang X, Tang CK (2008) Limits of learning-based superresolution algorithms. International Journal of Computer Vision 80(3): 406–420.
13. Zuo W, Lin Z (2011) A generalized accelerated proximal gradient approach for total-variation-based image restoration. IEEE Transactions on Image Processing 20(10): 2748–2759.
14. Lai Z, Wong WK, Jin Z, Yang J, Xu Y (2012) Sparse approximation to the Eigensubspace for discrimination. IEEE Trans. Neural Netw. Learning Syst. 23(12): 1948–1960.
15. Wagner A, Wright J, Ganesh A, Zhou ZH, Ma Y (2009) Towards a practical face recognition system: robust registration and illumination by sparse representation. IEEE Conference on Computer Vision and Pattern Recognition.
16. Zhu Q, Sun CL (2013) Image-based face verification and experiments. Neural Computing and Applications. doi:10.1007/s00521-012-1019-x.
17. Xu Y, Zhang D, Yang J, Yang JY (2011) A two-phase test sample sparse representation method for use with face recognition. IEEE Transactions on Circuits and Systems for Video Technology 21(9): 1255–1262.
18. Xu Y, Zhu Q (2013) A simple and fast representation-based face recognition method. Neural Computing and Applications 22(7): 1543–1549.
19. Yang M, Zhang L, Yang J, Zhang D (2011) Robust sparse coding for face recognition. IEEE Conference on Computer Vision and Pattern Recognition.
20. Wright J, Yang A, Ganesh A, Shankar S, Ma Y (2009) Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2): 210–227.
21. Zhang L, Yang M, Feng XC (2011) Sparse Representation or Collaborative Representation: Which Helps Face Recognition. International Conference on Computer Vision.
22. Xu Y, Zhu Q, Zhang D, Yang JY (2011) Combine crossing matching scores with conventional matching scores for bimodal biometrics and face and palmprint recognition experiments. Neurocomputing 74: 3946–3952.
23. Xu Y, Fan Z, Zhu Q (2012) Feature space-based human face image representation and recognition. Optical Engineering 51(1).
24. Xu Y, Zhong A, Yang J, Zhang D (2011) Bimodal biometrics based on a representation and recognition approach. Optical Engineering 50(3).

25. Xu Y, Zuo W, Fan Z (2011) Supervised sparse presentation method with a heuristic strategy and face recognition experiments. Neurocomputing 79: 125–131.
26. Yang M, Zhang L (2010) Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary. European Conference on Computer Vision.
27. Gangaputra S, Geman D (2006) A design principle for coarse-to-fine classification. IEEE Conference on Computer Vision and Pattern Recognition.
28. Amit Y, Geman D, Fan X (2004) A coarse-to-fine strategy for multi-class shape detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(12): 1606–1621.
29. Pham TV, Smeulders AWM (2006) Sparse representation for coarse and fine object recognition. IEEE Transactions on Pattern Analysis and Machine 28(4): 555–567.
30. Naseem I, Togneri R, Bennamoun M (2010) Linear regression for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11): 2106–2112.
31. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cognitive Neurosci 3(1): 71–86.
32. Belhumeur PN, Hespanha J, Kriegman DJ (1997) Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine 19(7): 711–720.
33. Yang J, Zhang D, Frangi AF, Yang J (2004) Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Transactions on Pattern Analysis and Machine 26 (1): 131–137.
34. Zhang D, Zhou ZH (2005) (2D)2PCA: two-directional two-dimensional PCA for efficient face representation. Neurocomputing 69(1–3): 224–231.
35. Xiong H, Swamy MNS, Ahmad MO (2005) Two-dimensional FLD for face recognition. Pattern Recognition. 38: 1121–1124.
36. Zheng WS, Lai JH, Li SZ (2008) 1D-LDA vs. 2D-LDA: when is vector-based linear discriminant analysis better than matrix-based. Pattern Recognition. 41(7): 2156–2172.
37. Qi Y, Zhang J (2009) (2D)2PCALDA: an efficient approach for face recognition. Appl. Math. Comput. 213 (1): 1–7.
38. Samaria F, Harter A (1994) Parameterisation of a stochastic model for human face identification, Proceedings of 2nd IEEE Workshop on Applications of Computer Vision.
39. Yang J, Zhang D, Frangi AF, Yang JY (2004) Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 26(1): 131–137.
40. Yang J, Yang JY, Frangi AF (2003) Combined Fisherfaces framework. Image Vision Comput. 21(2): 1037–1044.