Neurocomputing 135 (2014) 253–261

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Modified minimum squared error algorithm for robust classification and face recognition experiments

Yong Xu [a,b], Xiaozhao Fang [a], Qi Zhu [a], Yan Chen [a,c,*], Jane You [d], Hong Liu [e]

[a] Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[b] Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, China
[c] Shenzhen Sunwin Intelligent Corporation, Shenzhen, China
[d] Biometrics Researcher Centre, Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[e] Engineering Laboratory on Intelligent Perception for Internet of Things, Shenzhen Graduate School, Peking University, Shenzhen, China

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form
25 September 2013
Accepted 25 November 2013
Communicated by D. Tao
Available online 13 February 2014

Keywords:
Minimum squared error (MSE)
Pattern recognition
Face recognition

ABSTRACT

In this paper, we improve the minimum squared error (MSE) algorithm for classification by modifying its classification rule. Differing from the conventional MSE algorithm which first obtains the mapping that can best transform the training sample into its class label and then exploits the obtained mapping to predict the class label of the test sample, the modified minimum squared error classification (MMSEC) algorithm simultaneously predicts the class labels of the test sample and the training samples nearest to it and combines the predicted results to ultimately classify the test sample. Besides this paper, for the first time, proposes the idea to take advantage of the predicted class labels of the training samples for classification of the test sample, it devises a weighted fusion scheme to fuse the predicted class labels of the training sample and test sample. The paper also interprets the rationale of MMSEC. As MMSEC generalizes better than conventional MSE, it can lead to more robust classification decisions. The face recognition experiments show that MMSEC does obtain very promising performance.

1. Introduction

The minimum squared error algorithm has been widely used for pattern classification. The minimum squared error classification (MSEC) takes the sample and its class label as the input and output respectively, and tries to obtain the mapping that can best transform the input into the corresponding output. MSEC first uses the training samples to perform training and then exploits the obtained mapping to predict the class label of the test sample. Finally, MSEC assigns the test sample into the class whose class label is most similar to the predicted class label of the test sample.

MSEC not only can achieve high accuracy but also holds good properties. For example, 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]. LDA and its variants have been widely used [3]. Moreover, if a special class indicator matrix is used, MSEC and LDA are also equivalent for multi-class classification [4]. LDA has also been shown to be equivalent to canonical correlation analysis (CCA) for multi-class classification [5]. As a result, MSEC will perform very similarly as CCA in multi-class classification [6].

Besides MSEC has been extended to multi-class classification, a well-known nonlinear extension of MSEC, kernel MSE (KMSE), has been proposed. KMSE performs very well in the field of pattern recognition too [2,7,8]. Other various improvements to the MSE methodology have also been devised. For example, "Lasso" based MSE (LBMSE) was recently proposed for classification [9–11]. LBMSE tries to obtain good generalization performance by minimizing the l_1 norm of the solution vector and can be viewed as an extension of conventional MSEC. Differing from conventional MSEC, LBMSE takes the training sample and the test sample themselves as the input and the output, respectively. After the mapping between the input and output is constructed, LBMSE also uses a way different from that of MSEC to perform classification. As shown in Refs. [12–14], we can also modify MSEC to a classification algorithm that is similar to LBMSE but subject to the constraint of minimizing the l_2 norm of the solution vector. This algorithm will be computationally more efficient than LBMSE and has comparable classification performance. Linear regression classification (LRC) proposed in Ref. [15] is a typical example of this kind of algorithm. The MSEC algorithms with the constraints of minimizing the l_1 or l_2 norm can also be referred to as

* Corresponding author at: Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. Tel.: +86 755 86169696; fax: +86 755 86169393.
E-mail address: jadechenyan@gmail.com (Y. Chen).

http://dx.doi.org/10.1016/j.neucom.2013.11.025
0925-2312

penalized MSECs [16] or representation-based classification (RBC) algorithms.

Besides the inputs and outputs of the method proposed in our paper are different from those of RBC, it also differs from RBC as follows. The proposed method should solve only one equation and exploit it to predict the class label of all the test samples. However, RBC must solve at least one equation for classifying a test sample. In particular, RBC proposed in Refs. [12–14] should solve and exploit one equation for classifying a test sample. LRC should depend on the solutions of c equations to classify a test sample. c is the number of the classes. As a result, our proposed method is usually computationally more efficient than RBC.

The total least squares (TLSs) [17,18] is another well-known improvement to the MSE. TLS assumes that both the input and output are corrupted and each of them can be expressed as the sum of the corresponding "true data" and "measurement noise". Differing from TLS, conventional MSE methods just assumes that the output is corrupted but the input is not. Based on TLS, researchers also proposed the weighted and structured total least squares (WSTLSs) [17–20]. WSTLSs are usually numerically solved by using local optimization methods [17]. In addition, recursive least-squares methods were proposed as reinforcement learning algorithms [21]. Two-stage least squares (2SLS) was proposed for latent variable models [22]. Bayesian minimum mean-square error was also proposed to explore the theoretical issue in pattern classification such as to estimate the classification error [23–25]. In addition, some means such as the regularized term was also used to improve the numerical stability of MSE [26]. The means of regularization is indeed widely used and Hessian regularization proposed in Ref. [27] obtained very good performance in image annotation. Orthogonal MSE [28] and computationally more efficient MSE algorithm [29–31] were also devised. Besides pattern classification [32], the minimum squared error algorithms have been applied to other fields such as density estimation, clustering, feature extraction, data fitting and regression as well as image coding [7,17,30,31,33–36]. We also note that MSE has been widely used in the field of signal processing for resolving some important problems such as direction estimation, estimation of deterministic parameters with noise covariance uncertainties, optimization of the downlink multiuser MIMO systems and multipath channel estimations [37–39]. The MSE algorithm was also used for other issues such as Kalman filters and probabilistic principal component analysis [40]. The naïve MSE algorithm and its variants have been also widely used in regression [41,42].

Researchers have also paid much attention to improve the generalization performance of the classification algorithm. For MSEC, a conventional and important way to improve the generalization performance is to impose the constraint of minimizing the norm especially the l_2 norm of the solution vector on it. Of course, this way is very useful for avoiding the case where the predicted class label of the test sample corrupted by little noise greatly deviates from its true class label. However, the above way still cannot perform well in the case where the test sample is corrupted by great noise. For example, in real-world face recognition applications the test sample might be very different from the training sample from the same subject owing to varying expression, pose and illumination [43–45]. Consequently, the predicted class label of the test sample might have large deviation from its true class label. However, we see that the predicted class label of the training sample is always very close to its true class label. This somewhat means that the MSEC algorithm has great confidence in predicting the class label of the training sample but has less confidence in predicting the class label of the test sample. As a result, if a training sample is very near to the test sample, it is reasonable to integrate the predicted class labels of this training sample and the test sample to classify the test sample.

In this paper, in order to obtain more robust MSEC algorithm, we improve the MSEC algorithm by modifying its classification rule. We establish the same equation as that of the conventional MSEC and also solve it in the same way. Then we exploit the obtained solution to simultaneously predict the class labels of the test sample and the training samples nearest to it and combine the predicted results to ultimately classify the test sample. We use a weighted fusion scheme to combine the predicted class labels of the test sample and the training samples. The weight of the test sample is assigned a larger value in comparison with those of the training samples. When more than one training sample are exploited, we also assign a larger coefficient to the training sample that is closer to the test sample. The experiments also show that MMSEC does obtain much higher classification accuracy than conventional MSEC. This paper has the following noticeable contributions. First, it for the first time proposes the idea to take advantage of the predicted class labels of the training samples to classify the test sample. It also carefully demonstrates the underlying rationale of MMSEC. Second, it devises a weighted fusion scheme to fuse the predicted class labels of the training sample and test sample.

2. The minimum squared error classification (MSEC)

In this section we take the multi-class problem as an example to describe MSEC. Suppose that there are c classes. We assign a class label to each class. If a mapping is able to transform a sample into its class label and we can get this mapping by learning, then we can exploit the learned mapping to predict the class label of each test sample. Let x_i be a p-dimensional row vector and denote the ith training sample, $i=1,\ldots,N$. N is the total number of the training samples. We use a c-dimensional vector to represent the class label. If a sample is from the first class, we take $g=[1 \quad 0 \quad . \quad . \quad . \quad 0]$ as its class label. If a sample is from the cth class, we take $g=[0 \quad . \quad . \quad . \quad 0 \quad 1]$ as its class label. In other words, if a sample is from the kth class, then the kth element of its class label is one and the other elements are all zeroes. This class label is also referred to as the class label of the kth class.

Assuming that matrix Y can approximately transform each training sample into its class label, MSEC has the following equation:

$$XY=G \tag{1}$$

where

$$X=\begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_N \end{bmatrix},\ G=\begin{bmatrix} g_1 \\ . \\ . \\ . \\ g_N \end{bmatrix}$$

It is clear that X is an $N\times p$ matrix, G is an $N\times c$ matrix, and Y is a $p\times c$ matrix. We refer to Y as transform matrix. g_i is the class label of the ith training sample.

As Eq. (1) cannot be directly solved, we convert it into the following equation:

$$X^TXY=X^TG \tag{2}$$

We can obtain Y using

$$\overline{Y}=(X^TX+\gamma I)^{-1}X^TG \tag{3}$$

where γ and I denote a small positive constant and the identity matrix, respectively. MSEC classifies a test sample x in the form of row vector as follows: the class label of x is first predicted using $g_x=x\overline{Y}$. Then the distances between g_x and the class labels of all the c classes are calculated. As shown above, the class label of the