Pattern Recognition 47 (2014) 806–819

Contents lists available at ScienceDirect

Pattern Recognition

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

Regularized discriminant entropy analysis

Haitao Zhao [a,b], W.K. Wong [a,*]

[a] Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong
[b] Automation Department, East China University of Science and Technology, Shanghai, China

ARTICLE INFO

Article history:
Received 3 September 2012
Received in revised form
4 August 2013
Accepted 23 August 2013
Available online 5 September 2013

Keywords:
Regularized discriminant entropy
Entropy-based learning
Discriminant entropy analysis

ABSTRACT

In this paper, we propose the regularized discriminant entropy (RDE) which considers both class information and scatter information on original data. Based on the results of maximizing the RDE, we develop a supervised feature extraction algorithm called regularized discriminant entropy analysis (RDEA). RDEA is quite simple and requires no approximation in theoretical derivation. The experiments with several publicly available data sets show the feasibility and effectiveness of the proposed algorithm with encouraging results.

1. Introduction

Feature extraction from original data is an important step in pattern recognition, often dictated by practical feasibility. It is also an essential process in exploratory data analysis, where the goal is to map input data onto a feature space which reflects the inherent structure of original data. We are interested in methods that reveal or enhance the class structure of data, which is an essential procedure for a given classification problem.

For unsupervised learning's sake, feature extraction is performed by constructing the most representative features out of original data. For pattern classification, however, the purpose of feature extraction is to map original data onto a discriminative feature space in which samples from different classes are clearly separated. Many approaches have been proposed for feature extraction, such as principal component analysis (PCA) [1], linear discriminant analysis (LDA) [1,2], isometric feature mapping (ISOMAP) [3], local linear embedding (LLE) [4], locality preserving projection (LPP) [5] and graph embedding [6].

As feature extraction methods are applied to realistic problems, where dimensionality is high or the amount of training data is very large, it is impractical to manually process the data. Therefore, feature extraction methods that can robustly obtain a low-dimensional subspace are of particular interest in practice. Many robust algorithms, such as robust PCA [7], robust ISOMAP [8,9], robust LLE [10] and robust LDA [11], have been proposed in the past few years.

Recently, many feature extraction algorithms addressing the robust classification problem have involved the use of information theoretic learning (ITL) techniques [12–14]. The maximum likelihood principle [15], entropy [16–18], Kullback Leibler (KL) divergence [19,20], Bhattacharyya distance [21], and Chernoff distance [22] are often selected to develop feature extraction algorithms. In these algorithms, however, the handling of robust feature extraction is achieved at the cost of increased computational complexity. The projection matrix or the projection procedure of these algorithms is often solved by iterative optimization which has relatively high computational complexity. Due to the often used non-convex constraints, these algorithms are also prone to the local minimum (or maximum) problem.

In order to preserve computational simplicity and the characteristics of eigenvalue-based techniques, such as LDA, He et al. [11] proposed the maximum entropy robust discriminant analysis (MaxEnt-RDA) algorithm. Renyi's quadratic entropy was used in MaxEnt-RDA as a class-separability measure. The MaxEnt distribution was estimated by the nonparametric Parzen window density estimator with a Gaussian kernel. Due to the first-order Taylor expansion of each Gaussian kernel term, MaxEnt-RDA is an approximate algorithm which avoids iterative calculation of entropy. MaxEnt-RDA is proposed for discriminant feature extraction. It can effectively overcome the limitations of traditional LDA algorithms in a data distribution assumption and is robust against noisy data [11].

MaxEnt-RDA tried to obtain the projection matrix U by solving the following constraint MaxEnt problem:

$$\max_U H(U^TX) \quad \text{s.t.} \quad H(U^TX|C)=c_1, \quad U^TU=I, \tag{1}$$

where

$$H(U^TX)=-\ln\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}G(U^Tx_j-U^Tx_i,\sigma)\right)$$

* Corresponding author. Tel.: +852 27666471; fax: 852 27731432.
E-mail address: calvin.wong@polyu.edu.hk (W.K. Wong).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.08.020

is the esimate of Renyi's entropy by Parzen method, $G(U^Tx_j-U^Tx_i,\sigma)=(1/\sqrt{2\pi}\sigma)\exp(-\|U^Tx_j-U^Tx_i\|^2/2\sigma^2)$. And $H(U^TX|C)=\sum_{j=1}^{c}p(C_j)H(U^TX|C=C_j)$, where C_j is the label of the jth class. The closed form solution of the constraint MaxEnt problem is hard to find, because the problem is nonlinear. To solve this problem, He et al. [11] used the first-order Taylor expansion of each Gaussian kernel:

$$G(U^Tx_j-U^Tx_i,\sigma)\approx -G(x_j-x_i,\sigma)\|U^Tx_j-U^Tx_i\|^2+\text{const.} \tag{2}$$

Substituting Eq. (2) into (1), MaxEnt-RDA is reduced to a constraint graph embedding problem:

$$\max \mathrm{Tr}(U^TXL_tXU) \quad \text{s.t.} \quad \mathrm{Tr}(U^TXL_wXU)=c_1, \quad U^TU=I,$$

where

$$L_t=D^t-W^t,$$

$$L_w=D^w-W^w,$$

$$W_{ij}^t=\frac{2G(x_i-x_j,\sigma)}{\sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n}G(x_i-x_j,\sigma)},$$

$$W_{ij}^w=I_{(C_i=C_j)}p(C_i)\frac{2G(x_i-x_j,\sigma)}{\sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n}G(x_i-x_j,\sigma)},$$

$D_{ii}^t=\sum_{j=1}^{n}W_{ij}^t$, $D_{ii}^w=\sum_{j=1}^{n}W_{ij}^w$, D^t and D^w are diagonal matrices. For more details, refer to [11].

Das and Nenadic [23] proposed a discriminant feature extraction method based on information discriminant analysis (IDA) [24] and showed how a feature extraction matrix can be obtained through eigenvalue decomposition, reminiscent of LDA.

IDA is based on maximization of an information-theoretic objective function referred to as a $\mu-$measure, which enjoys many properties of mutual information and the Bayes error [24,23]. For a continuous random variable $X\in\mathbb{R}^n$ and class variable $\{C_1,C_2,\ldots,C_c\}$, the simplified form of the $\mu-$measure is given by

$$\mu(X)=\frac{1}{2}\left(\ln(|S_T|)-\sum_{i=1}^{c}p(C_i)\ln(|S_i|)\right),$$

where S_T is the total (unconditional) covariance matrix, S_i is the class-conditional covariance matrix. The optimal feature extraction matrix U is found by the maximization of $\mu(Z)$, where $Z=U^TX$ is the feature random variable. While both the gradient and the Hessian of $\mu(Z)$ (with respect to U) can be found analytically, the maximization of μ must be performed numerically.

Das and Nenadic's method was referred to as approximate information discriminant analysis (AIDA) [23]. Das and Nenadic [23] showed that their criterion is related to the $\mu-$measure in an approximate way, which made their algorithm simple and computationally efficient. Assume $Q\in\mathbb{R}^{n\times n}$ be a symmetric positive definite matrix with eigenvalue and eigenvector matrices, $\Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ and V, respectively. Let $xQ=V\ \mathrm{diag}(\ln\lambda_1,\ldots,\ln\lambda_n)V^T$. AIDA evaluates the matrix

$$S_{AIDA}=\ln S_T-\sum_{i=1}^{c}p(C_i)\ln S_i$$

in the transformed domain (note that $S_T\to S_W^{-1/2}S_TS_W^{-1/2}$, $S_i\to S_W^{-1/2}S_iS_W^{-1/2}$ and $S_W=\sum_{i=1}^{c}p(C_i)S_i$). The eigenvectors corresponding to the largest m eigenvalues of S_{AIDA} are taken to form the projection matrix U. For more details, refer to [23].

Motivated by information theoretic learning, we propose the regularized discriminant entropy (RDE), and then introduce a novel supervised feature extraction algorithm, called regularized discriminant entropy analysis (RDEA), which preserves computational simplicity and characteristics of eigenvalue-based techniques. Several interesting perspectives should be addressed:

1. The RDE is simple and easy to understand. It considers both class information and scatter information on original data. Section 2 shows that maximization of the RDE has connections to the mean shift algorithm and the pre-image reconstruction. However, the RDE is designed for supervised learning and is based on the within-class entropy and scatter information on data. The regularization parameter used in the RDE is proposed for the first time ever and does not appear in the mean shift algorithm and the pre-image reconstruction.
2. RDEA utilizes the results of the RDE, which has a clear theoretical foundation. RDEA is reduced to a simple constrained optimization problem, which can be obtained by eigen decomposition. For more details, refer to Section 3. RDEA is a direct solution without approximation. This is quite different form MaxEnt-RDA and AIDA. In order to use eigenvalue-based techniques, both MaxEnt-RDA and AIDA use approximation in their theoretical derivation.
3. RDEA can be regarded as a framework for supervised feature extraction. Firstly, we use the most widely used Shannon's entropy to design the RDE. If other generalized entropies are used, one may obtain other forms of RDE. Secondly, we put orthogonal constraints on basis functions to obtain projection matrix. The RDE and the likelihood values can be preserved under orthogonal projection (refer to Section 3 for details). Other constraints can be used here to substitute orthogonal constraints. The combination of different constraints and the maximization of the RDE can derive different feature extraction algorithms.
4. In Section 5, we compare our method with several other second-order feature extraction techniques using real datasets. The results show that RDEA compares favorably with other methods. We conclude that RDEA should be considered as an alternative to the prevalent feature extraction techniques.

2. Maximum entropy principle and regularized discriminant entropy

2.1. Maximum entropy principle

Let $P=(p_1,p_2,\ldots,p_N)$ be a probability distribution for N variates $x_1,x_2,\ldots,x_N$, and then there is uncertainty about the outcomes. Shannon used the measure

$$T(P)=-\sum_{i=1}^{N}p_i\ln p_i \tag{3}$$

to measure this uncertainty and called it the entropy of the probability distribution P [25]. It can be regarded as a measure of equality of $p_1,p_2,\ldots,p_N$ among themselves. Renyi, Havrda and Charvat, Kapur, Sharma-Taneja and others proposed other functions [25–27] of $p_1,p_2,\ldots,p_N$ to measure this uncertainty and called these functions as generalized measures of entropies or simply generalized entropies, such as Burg's entropy:

$$T^B(P)=\sum_{i=1}^{N}\ln p_i \tag{4}$$

and Kapur's entropy:

$$T^K(P)=-\sum_{i=1}^{N}p_i\ln p_i-\sum_{i=1}^{N}(1-p_i)\ln(1-p_i). \tag{5}$$