Neurocomputing 237 (2017) 114–132

Contents lists available at ScienceDirect

Neurocomputing

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

Moments discriminant analysis for supervised dimensionality reduction

K. Ramachandra Murthy*, Ashish Ghosh

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Communicated by Dacheng Tao

Keywords:
Dimensionality reduction
Subspace learning
Projection space
Statistical moments
Kernel
Image retrieval

ABSTRACT

Most of the well-known supervised dimensionality reduction methods assume unimodal or Gaussian likelihoods, which may not be appropriate in the real life applications. In this manuscript, we introduce a novel supervised dimensionality reduction approach, moments discriminant analysis, which models linear relationships between the high-dimensional input space and a low-dimensional space by maximizing the discrimination between second order raw moments of different classes to improve the generalization capability of a classifier. Unlike the state-of-the-art methods, moments discriminant analysis is intended to accommodate data distributions that may be multimodal and non-Gaussian. Initially, experiments using synthetic random data (generated from different probability distributions) are performed to prove the efficiency of the proposed method for multimodal and non-Gaussian data with the help of five separability measures. Also, extensive experimental results on UCI machine learning repository and image retrieval on WANG and MIT (Oliva and Torralba) databases are carried out in order to exhibit the effectiveness of moments discriminant analysis over the state-of-the-art methods.

1. Introduction

Any progresses in efficient data processing and storage capacities need control on the number of useful variables/features/attributes. Real world applications, such as e-science, medical image processing, video processing, speech signal analysis, bio-informatics, biometrics, document classification, etc., deal with the data of very high dimensionality [1–4]. High dimensional data is a big challenge for the learning problems because of the difficulty in modelling the precise relationships between the large number of features and the class variables. In such cases, it may be desirable to reduce the dimensionality in order to improve the accuracy and performance of a classifier. The goal of Dimensionality Reduction (DR) is to embed high dimensional data samples in a low-dimensional space while most of 'intrinsic information' contained in the data is preserved. Moreover, the reduced feature set of the data can have better interpretability than the original ones [5–8].

Statistical moments play an important role among DR methods like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) [9], Maximum Margin Criterion (MMC) [10,11], Angle Linear Discriminant Embedding (ALDE) [12], Two-Stage ALDE (TSALDE), Linear Boundary Discriminant Analysis (LBDA) [13], Local Fisher's Discriminant Analysis (LFDA) [14], Exponential Discriminant Analysis (EDA) [15], etc. For example, PCA tries to maximize the variational (second order central moments) information of features, and LDA, LBDA, LFDA, EDA and MMC maximize distance between the means (first order raw moments) of the classes and minimize the within-class scatters (variances, i.e., second order central moments). Whereas, ALDE and TSALDE use CO-Angle to model between-class (with the help of different class means) and within-class scatter matrices. Most of the situations, these methods generate a projection space which trend towards means (centers) of the classes by assuming Gaussian likelihoods on the data. For example, from Fig. 1(a), they assume Gaussian distribution even on non-Gaussian (uniform) data and generate a mean biased projection space (W). That means, W maximizes the distance $|\mu_1 - \mu_2|$, and doesn't bother about discrimination of σ_1 and σ_2 (i.e., $|\sigma_1 - \sigma_2|$ is very small), where $|.|$ denotes the modulus of a real number. Whereas, W_{opt} tries to maximize the discrimination between both means (μ) and variances (σ) to preserve the separability between classes. That means, both the distances $|\mu_1 - \mu_2|$ and $|\sigma_1 - \sigma_2|$ are maximized on W_{opt}. Thus, W_{opt} balances the discrimination between means and variances. Also, if the means of different classes are overlapped (in case of multimodal data) then the between class scatter of these methods will vanish. This may force them to depend on within class scatter only and lead to overlapping of classes. Fig. 1(b) narrates two multimodal classes for which means are overlapped at a point. In this scenario, the above methods generate a projection space (W) through minimizing within class scatter only and ignores between class scatter. But, W_{opt} produces non-overlapping classes by preserving the discrimination (i.e., $|\sigma_1 - \sigma_2|$) between class variances (σ_1 and σ_2). The

* Corresponding author.
E-mail addresses: k.ramachandra@isical.ac.in, kramachandramurthy@gmail.com (K.R. Murthy).

http://dx.doi.org/10.1016/j.neucom.2016.09.048
Received 26 May 2016; Received in revised form 28 July 2016; Accepted 22 September 2016
Available online 30 September 2016

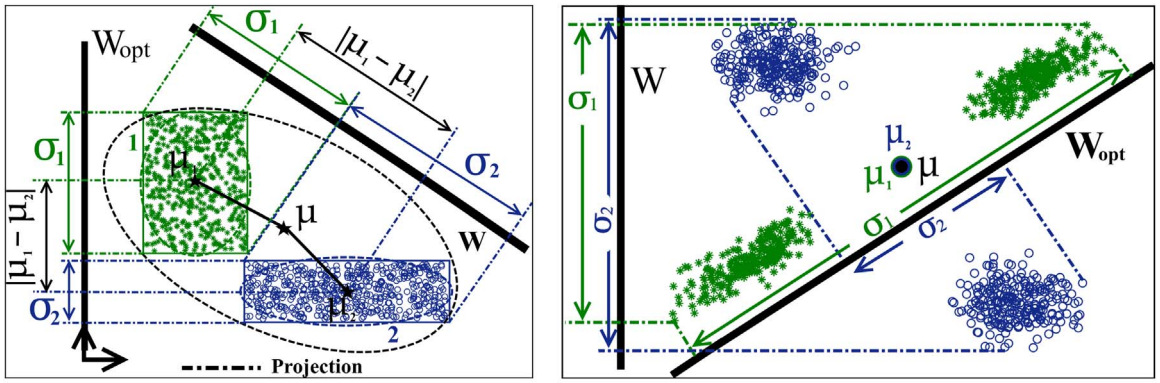

Fig. 1. μ_l and σ_l are mean and variance of the class l, respectively, l=1, 2. Here *(Green) and o (blue) patterns represent class 1 and 2, respectively. (a) Non-Gaussian (b) Multimodal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

main aim of this article is to provide such a projection space (W_{opt}).

Statistical moments are numerical characteristics of a probability distribution and are applicable to many different aspects of pattern recognition, image processing, machine learning, bio-informatics, etc. [16,17]. When applied to pattern classification, they describe the discrimination information of classes (or distributions) [17]. For a bounded probability distribution, the collection of raw moments uniquely determines the underlying distribution. Any probability density can be modelled with the help of it's population parameters. In statistics, 'the method of moments' is a procedure to estimate the population parameters with the help of 'sample' raw moments [18]. One starts with deriving equations, that relate the 'population' moments (i.e., the expected values of powers of the random variable under consideration) to the parameters of the distribution. Then a sample is drawn and 'population' raw moments are estimated from that 'sample'. The equations are then solved for the parameters of the distribution, using the 'sample' raw moments in place of (unknown) 'population' raw moments. Moreover, the 'sample' raw moments are unbiased estimates of 'population' raw moments. On the other hand, the 'sample' central moments are not unbiased because their computation uses up a degree of freedom by using the 'sample' mean. For the higher order central moments, the unbiased estimates of the 'population' central moments will become more and more complex. Also central moments may have disturbances because the moments are biased towards mean of the data [18]. Thus, the projected directions of DR methods which are modelled on central moments, explicitly, may be biased towards means of the classes (Fig. 1). Therefore, collection of 'sample' raw moments are good enough to model the underlying probability density and henceforth, MDA will use 'sample' raw moments to reduce the dimensionality.

In this work, we propose a novel supervised DR approach, Moments Discriminant Analysis (MDA), that models linear relationship between feature vectors and class labels by maximizing the discrimination between second order class raw moments. In order to capture the non-linear relationships within the features and class variable, kernel version of MDA (KMDA) has been developed. MDA possesses the following advantages over the existing state-of-the-art DR methods.

(i) MDA is applicable to the data with distributions that may be non-Gaussian.
(ii) Even for multimodal distribution data, MDA can be used for simplification of decision boundary.

Experimental study has been performed on wide variety for data sets as,

- Initial experiments have been performed on thirty two (32) synthetic data sets[1] to prove the efficiency of MDA to Gaussian, multimodal and non-Gaussian data. The comparisons have given using different state-of-the-methods according to their applicability for
  (i) Gaussian: ALDE [12], EDA [15], MMC [10,11] and LDA [9].
  (ii) Multimodal: Quadratic Mutual Information (QMI)[19], Exponential Local Discriminant Embedding (ELDE) [20], Linear Discriminative Gaussian (LDG) [21], LBDA [13] and LFDA [14].
  (iii) Non-Gaussian: QMI [19], Stable Orthogonal Local Discriminant Embedding (SOLDE) [22] and Exponential Marginal Fisher Analysis (EMFA) [23]. These methods have been evaluated using five different separability measures, namely, volume of overlap region (*overlap*), Thornton's Separability Index (*TSI*), Fraction of points on Boundary (*FB*), volume of local neighborhood (*volume*) and Nonparametric Separability (*NS*) measures [24].
- Next, twelve UCI machine learning [25] data sets have been used to compare the performance of MDA with the state-of-the-art methods like, ALDE, TSALDE, LBDA, LDG, EDA, QMI, ELDE, SOLDE, EMFA, LFDA, MMC, Discriminant Component Analysis (DCA) [26] and LDA. And KMDA's performance is compared with the available kernel versions of the above linear methods like Kernel QMI (KQMI) [19], Kernel LFDA (KLFDA) [14], Kernel MMC (KMMC) [11], Kernel DCA (KDCA) [26], and Kernel Discriminant Analysis (KDA) [27,28] with the help of Nearest Neighborhood (NN) [9] classifier.
- Finally, on WANG [29] and MIT (Oliva and Torralba) [30] image data sets, MDA and KMDA are compared with above mentioned methods in the context of image retrieval.

This paper is organized as follows. Section 2 discusses related work and Section 3 introduces MDA. Non-linear version of MDA has been developed in Section 4 and Asymptotic time complexity analysis of MDA has been discussed in Section 5. Experimental results are presented in Section 6 and the manuscript has been concluded in Section 7.

2. Related work

Linear Dimensionality Reduction (DR) methods can be divided into unsupervised [31] and supervised. One of the popular and well-known unsupervised DR method is Principal Component Analysis (PCA) [9]. PCA tries to maximize data variances captured in the low-dimensional subspace, i.e., PCA minimizes the reconstruction error of the projected data points with the original data. Independent Component Analysis

[1] MDA mex file and synthetic data sets are available at http://www.isical.ac.in/~k.ramachandra/MDA.html.