Pattern Recognition 48 (2015) 1863–1877

Contents lists available at ScienceDirect

Pattern Recognition

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

Pareto models for discriminative multiclass linear dimensionality reduction

Karim T. Abou-Moustafa [a,*], Fernando De La Torre [b], Frank P. Ferrie [c]

[a] Dept. of Computing Science, ATH 3-55, University of Alberta, Edmonton, AB, Canada T6G 2E8
[b] Robotics Institute, Smith Hall 211, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[c] Dept. of Electrical & Computer Engineering and Centre of Intelligent Machines, McGill University, McConnell Engineering Building, Room 441, 3480 University Street, Montreal, QC, Canada H3A 2E9

ARTICLE INFO

Article history:
Received 10 January 2014
Received in revised form
13 August 2014
Accepted 8 November 2014
Available online 20 November 2014

Keywords:
Fisher discriminant analysis
Supervised linear dimensionality reduction
Feature transformation
Metric learning
Subspace learning
Multiobjective optimization
Pareto optimality
Kullback–Leibler divergence

ABSTRACT

We address the class masking problem in multiclass linear discriminant analysis (LDA). In the multiclass setting, LDA does not maximize each pairwise distance between classes, but rather maximizes the sum of all pairwise distances. This results in serious overlaps between classes that are close to each other in the input space, and degrades classification performance. Our research proposes Pareto Discriminant Analysis (PARDA); an approach for multiclass discriminative analysis that builds over multiobjective optimizing models. PARDA decomposes the multiclass problem to a set of objective functions, each representing the distance between every pair of classes. Unlike existing LDA extensions that maximize the sum of all distances, PARDA maximizes each pairwise distance to maximally separate all class means, while minimizing the class overlap in the lower dimensional space. Experimental results on various data sets show consistent and promising performance of PARDA when compared with well-known multiclass LDA extensions.

1. Introduction

Fisher Discriminant Analysis (FDA) originally developed by Fisher in 1936 [1] is a technique for supervised linear dimensionality reduction that is optimal for classification under two assumptions: (i) the number of classes c is exactly two, and (ii) the samples in each class are assumed to be generated from a multivariate Gaussian distribution with different means and equal covariance matrices (homoscedastic data) [2]. In this context, FDA is guaranteed to find a one dimensional subspace that will classify the samples with the optimal error rate, *Bayes error*, and the subspace is known to be *Bayes optimal* [2]. Rao [3] extended this approach to the multiclass homoscedastic case ($c > 2$), under the condition that the data features $d \geq c$ (and assuming the number of samples $n > d$). The resultant $c-1$ dimensional subspace is also guaranteed to be Bayes optimal, and the technique has become known as Linear Discriminant Analysis (LDA). Rao also noted that in the homoscedastic case, if the lower dimensional subspace has dimensionality $d_0 < c-1$, the resultant subspace will not be Bayes optimal. It is only recently that Hamsici and Martinez [4] pushed the homoscedastic case further and derived a Bayes optimal one dimensional subspace when $c > 2$.

When the equal covariance assumption does not hold for $c \geq 2$ (heteroscedastic data), Rao proposed to approximate the heteroscedastic problem with a homoscedastic setting and solve the approximated problem instead. His approximated problem considered that all classes have different means but share a common covariance matrix which is a weighted average of all the covariance matrices of the original problem. This approximation matrix became known as the pooled sample covariance matrix, or the average within-class scatter matrix $\mathbf{S}_w$. Rao's final solution became the well known formulation based on the Rayleigh quotient of the between-class scatter matrix $\mathbf{S}_b$ and $\mathbf{S}_w$. The obtained subspace, however, is not Bayes optimal for the original heteroscedastic problem.

Several researchers, backed by theoretical justifications, have scrutinized the limitations and non-optimality (in terms of Bayes error) of LDA when its strong assumptions do not hold and proposed extensions derived from Gaussian assumptions [5–8] and kernel methods [9,10] to generalize LDA to the multiclass heteroscedastic case. The result was a plethora of algorithms that have been reported to perform well in a variety of application

* Corresponding author. Tel.: +1 780 655 2685; fax: +1 780 492 1071.
E-mail addresses: aboumous@cs.ualberta.ca (K.T. Abou-Moustafa), ftorre@cs.cmu.edu (F. De La Torre), ferrie@cim.mcgill.ca (F.P. Ferrie).

http://dx.doi.org/10.1016/j.patcog.2014.11.008
0031-3203/

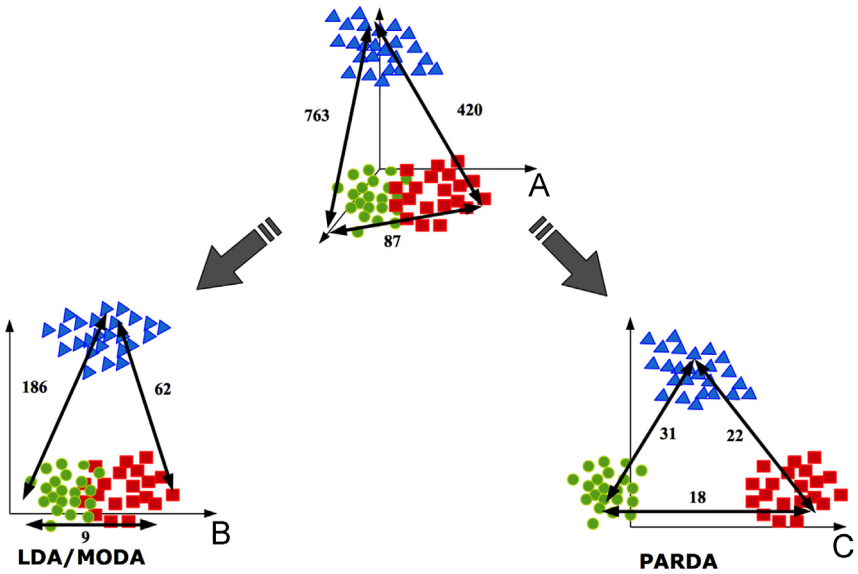

Fig. 1. (A) A Synthetic example of a 3-class problem with three dimensional data; L_1 triangles, L_2 squares, and L_3 circles. The numbers shown on arrows indicate the KL divergence between classes. The contribution of each pairwise divergence to the total divergence is 60%, 33%, and 7% for (L_1,L_2), (L_1,L_3), and (L_2,L_3), respectively. (B) and (C) Projections using MODA and PARDA, respectively, on two-dimensional subspaces. Note that the divergences in the lower dimensional subspaces are always less than the divergences in the original input space. This is due to the information loss incurred from the linear transformation and it shall be explained in Section 4.3. For MODA's projection, the contribution of each pairwise divergence to the total divergence is 72%, 28%, and 3% for the same class ordering. Note that the largest KL divergence in the input space increased in the lower dimensional subspace, and that the other (less) KL divergences became even smaller in the lower dimensional subspace – which is the class masking effect. For PARDA's projection, the contribution of each pairwise divergence to the total divergence is 44%, 31%, and 25% for the same class ordering. Note that, while MODA decreases the separation from 7% to 3% for (L_2,L_3), PARDA increases the separation to 25%.

domains, most notably face recognition (see [4,11–14] for a good review of these methods).

Of particular interest is the extension proposed by De La Torre and Kanade [15], namely Multimodal Oriented Discriminant Analysis (MODA), where it was shown that FDA's objective function is a special case of a more general objective that maximizes the Kullback–Leibler (KL) divergence [16] between two Gaussian densities, when the two Gaussians share the same covariance matrix. Note that the symmetric KL divergence considers the difference in mean locations and the difference in covariance matrices (size and orientation). Therefore, MODA searches for a linear transformation that maximizes the symmetric KL divergence between the two classes in the low dimensional subspace.

To account for the multiclass heteroscedastic case, MODA sums over all KL divergences between every pair of different classes and maximizes that sum in the lower dimensional subspace. This is similar to LDA's objective function which, as shown by Loog et al. [11], maximizes the sum of pairwise FDAs between all pairs of different classes. Hence MODA is a consistent generalization of FDA/LDA to multimodal Gaussian distributions with different means and covariance matrices.

However, as noted by several researchers [11,12,17,18], even if all the homoscedastic assumptions are satisfied, LDA and MODA suffer from the serious problem of merging classes that are close to each other in the original input space, *a.k.a* the class masking problem. This is due to the fact that LDA and MODA shift the 2-class problem to the multiclass setting by maximizing the sum of all KL divergences, which is a suitable objective function when all classes are equally distant from each other in terms of KL divergence.

Fig. 1A depicts a synthetic example for a 3-class problem with three dimensional data. Traditional methods like LDA or MODA find projections that maximize the sum of pairwise Mahalanobis distance (for LDA) or the KL divergence (for MODA) between pairwise classes. Note that the first term in the symmetric KL divergence – for two multivariate Gaussians see Eq. (6) – and the Mahalanobis distance (a special case from the KL divergence) are positive quadratic distance functions. From the optimization of minimax functions [19], it is known that the sum of positive powered functions, $\sum_{j=1}^{m}[f_j]^p$, where $p>1$, is a smooth approximation for $\max_{1\le j\le m}[f_j]^p$, as p is increasing, and hence $\sum_{j=1}^{m}[f_j]^p \approx [f_r]^p$ where $f_r > f_j \ \forall j \neq r$. Using this argument,[1] and for $p=2$, we argue that LDA is in fact maximizing a smooth approximation of the maximum of quadratic distances. Similarly, due to the quadratic distance in the first term of the symmetric KL divergence (in the case of Gaussians), MODA also maximizes a smooth approximation of the maximum divergences between Gaussians. Hence, LDA and MODA intrinsically prefer solutions that encourage maximizing the largest distance in the input space to make it even larger in the lower dimensional subspace, i.e., LDA and MODA put needless effort to maximize already distant classes in the input space. This effect can be seen in Fig. 1B, where MODA's projection gives relatively better increase in terms of KL divergence to the classes that are farther away in the input space, while it only makes a slight effort to separate between classes that are closer to each other in the input space.

1.1. Contribution

We note that the multiclass problem for LDA and MODA defines an independent objective function for each pair of different classes that needs to be optimized, namely maximize the symmetric KL divergence between every pair of different classes. Hence, the set of all pairs of different classes defines an optimization problem with *multiple objective functions* that share one final solution, and if possible, they all need to be *simultaneously optimized*. Given this perspective, maximizing the sum over all pairwise KL divergences (or quadratic distances) does not consider each objective function independently, since as explained above, maximizing that sum approximates a max function that only encourages maximizing the largest KL divergence. In other words, upgrading the problem of learning a discriminant subspace from the 2-class setting to the

[1] This will be explained in more detail in Section 4.