Expert Systems With Applications 91 (2018) 252–262

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Linear dimensionality reduction for classification via a sequential Bayes error minimisation with an application to flow meter diagnostics

Kojo Sarfo Gyamfi*, James Brusey, Andrew Hunt, Elena Gaura

Faculty of Engineering and Computing, Coventry University, Coventry CV1 5FB, United Kingdom

ARTICLE INFO

Article history:
Received 4 July 2017
Revised 6 September 2017
Accepted 9 September 2017
Available online 9 September 2017

Keywords:
Linear dimensionality reduction
LDA
Heteroscedasticity
Bayes error
Flow meter diagnostics

ABSTRACT

Supervised linear dimensionality reduction (LDR) performed prior to classification often improves the accuracy of classification by reducing overfitting and removing multicollinearity. If a Bayes classifier is to be used, then reduction to a dimensionality of $K-1$ is necessary and sufficient to preserve the classification information in the original feature space for the K-class problem. However, most of the existing algorithms provide no optimal dimensionality to which to reduce the data, thus classification information can be lost in the reduced space if $K-1$ dimensions are used. In this paper, we present a novel LDR technique to reduce the dimensionality of the original data to $K-1$, such that it is well-primed for Bayesian classification. This is done by sequentially constructing linear classifiers that minimise the Bayes error via a gradient descent procedure, under an assumption of normality. We experimentally validate the proposed algorithm on 10 UCI datasets. Our algorithm is shown to be superior in terms of the classification accuracy when compared to existing algorithms including LDR based on Fisher's criterion and the Chernoff criterion. The applicability of our algorithm is then demonstrated by employing it in diagnosing the health states of 2 ultrasonic flow meters. As with the UCI datasets, the proposed algorithm is found to have superior performance to the existing algorithms, achieving classification accuracies of 99.4% and 97.5% on the two flow meters. Such high classification accuracies on the flow meters promise significant cost benefits in oil and gas operations.

1. Introduction

Linearly reducing the dimensionality of a dataset is an important preprocessing step in machine learning for a number of reasons. For one thing, linear dimensionality reduction (LDR) enables easy visualisation of data when the data is reduced to two or three dimensions. For another, performing LDR prior to learning can reduce model complexity while alleviating the small sample size problem in algorithms such as Fisher's linear discriminant, where a very large dimensionality and much smaller training data cause the scatter matrix to be singular (Lu, Plataniotis, & Venetsanopoulos, 2003; Sharma & Paliwal, 2015). More importantly, however, LDR often improves learning in the low-dimensional manifold in which the data is reduced to lie (Brunzell & Eriksson, 2000; Duin & Loog, 2004). This is usually due to the fact that LDR results in useful feature extraction from a dataset, thus reducing overfitting (Bermingham et al., 2015; James, Witten, Hastie, & Tibshirani, 2013). In algorithms such as k-Nearest Neighbours (kNN), the performance improvement obtained from LDR is also attributable to the fact that LDR mitigates the effects of the so-called curse of dimensionality (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999).

LDR has been applied to several problems such as medical diagnosis e.g. Sharma and Paliwal (2008),Coomans, Jonckheer, Massart, Broeckaert, and Blockx (1978),Sengur (2008) and Polat, Güneş, and Arslan (2008), face and object recognition e.g. Song, Zhang, Wang, Liu, and Tao (2007),Chen, Liao, Ko, Lin, and Yu (2000),Liu, Chen, Tan, and Zhang (2007) and Yu and Yang (2001) and credit card fraud prediction e.g. Mahmoudi and Duman (2015) to reduce the dimensionality of very high-dimensional feature spaces. Indeed, there are several other emerging application areas where dimensionality reduction can be employed to improve learning. One such area is flow meter diagnostics which is described in Section 4.

One of the most popular LDR techniques is Principal Components Analysis (PCA) (Barber, 2012), which works by projecting the

* Corresponding author.
E-mail addresses: gyamfik@uni.coventry.ac.uk (K.S. Gyamfi), j.brusey@coventry.ac.uk (J. Brusey), ab8187@coventry.ac.uk (A. Hunt), csx216@coventry.ac.uk (E. Gaura).

http://dx.doi.org/10.1016/j.eswa.2017.09.010
0957-4174/

original data onto a subspace where the variance of the data is maximised in each dimension. However, when statistical classification is desired after dimensionality reduction, PCA may lose the class-discriminatory information in the data, as the directions of maximum variance does not necessarily coincide with the most class-discriminative directions.

In order to maximise the class-discriminatory information while linearly reducing the dimensionality, LDR aimed for classification makes use of class labels to inform the choice of the transformation matrix **M**. In this case, the optimum objective function to minimise is the Bayes error in the linearly reduced space (Buturovic, 1994; Fukunaga, 2013). However, as an analytic expression for the Bayes error is hard to obtain for any arbitrary probability distribution, several approximations have been made (Buturovic, 1994; Duda, Hart, & Stork, 2012; Fukunaga, 2013), leading to several supervised dimension reduction techniques (Barber, 2012; Brunzell & Eriksson, 2000; Cunningham & Ghahramani, 2015; Duin & Loog, 2004). Notable among these techniques is Linear Discriminant Analysis (LDA) (Barber, 2012; Fisher, 1936; Fukunaga, 2013; Izenman, 2009). At its core, LDA is built on the assumption that the data is normally distributed in each class, with the covariance matrices of the classes being equal (an assumption known as homoscedasticity). Consequently, Fisher's LDA maximises Fisher's criterion (Barber, 2012; Duin & Loog, 2004; Fukunaga, 2013) as a measure of class separability, by taking only the differences in the projected class means into account, ignoring any differences in covariance matrices that might be present among the various classes in the data (Duin & Loog, 2004).

However, experimental results have shown that if one accounts for the violation of the assumptions in the original procedure, the performance of LDA can be improved (Hastie & Tibshirani, 1996; Marks & Dunn, 1974; Mika, Ratsch, Weston, Scholkopf, & Mullers, 1999; Zhao, Sun, Yu, Liu, & Ye, 2009). Along this line, our previous work describe an iterative procedure to obtain a one-dimensional subspace where the Bayes error is minimised in the two-class problem under the normality assumption in LDA, while accounting for heteroscedasticity (Gyamfi, Brusey, Hunt, & Gaura, 2017).

In this paper, we present a novel technique to LDR, that projects the original data onto a $(K-1)$-dimensional subspace for the K-class problem. We do this by sequentially creating linear classifiers that minimise the Bayes error under assumptions of normality and heteroscedasticity via a gradient descent procedure. This procedure is described in Section 3. Though iterative, the proposed algorithm is fast, and it remains unaffected by the number of training examples. In Section 4, we describe the applicability of LDR to flow meter diagnostics. In Section 5, we experimentally validate the proposed algorithm on 10 University of California, Irvine (UCI) datasets, as well as in the diagnosis of the health states of two ultrasonic flow meters, using datasets provided by the National Engineering Laboratory (NEL), United Kingdom.

2. Background and related work

Consider a dataset $\mathcal{D} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ with n examples and a dimensionality of d. The dataset is assumed to be labelled and made up of K classes, i.e., $\mathcal{D} = [\mathcal{C}_1, \ldots, \mathcal{C}_K]$. We aim at finding a linear transformation $\mathbf{T}$ such that $\tilde{\mathcal{D}} = \mathbf{T}^\top \mathcal{D}$ has a dimensionality of q, i.e., $\mathbf{T} \in \mathbb{R}^{d\times q}$, where $q < d$, while maximising the class-discriminatory information.

Let $\bar{\mathbf{x}}_k$, $\mathbf{S}_k$ and $\pi_k = p(\mathcal{C}_k)$ respectively be the mean, covariance and prior probability of the kth class, for $k \in \{1, \ldots, K\}$. Also, let $\bar{\mathbf{x}}$ be the mean of the overall dataset $\mathcal{D}$.

2.1. Fisher's criterion

Fisher's LDA aims to maximise Fisher's criterion as given by:

$$J_F = \mathrm{trace}\left((\mathbf{T}^\top \mathbf{S}_W \mathbf{T})^{-1} (\mathbf{T}^\top \mathbf{S}_B \mathbf{T})\right) \tag{1}$$

where $\mathbf{S}_W$, the within-class scatter matrix and $\mathbf{S}_B$, the between-class scatter matrix are both given by

$$\mathbf{S}_W = \sum_{k=1}^{K} \pi_k \mathbf{S}_k \quad \text{and} \quad \mathbf{S}_B = \sum_{k=1}^{K} \pi_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^\top. \tag{2}$$

In the two-class case, where reduction to only one dimension is possible, maximising Fisher's criterion tends to minimise the Bayes error in the one-dimensional subspace onto which the data is projected, when the normality and homoscedasticity assumptions hold (Hamsici & Martinez, 2008; Izenman, 2009).

2.2. Mahalanobis criterion

For the K-class case (where $K > 2$), however, maximisation of Fisher's criterion does not guarantee the minimisation of the Bayes error, even when the assumptions of homoscedasticity and normality are satisfied. To get around this problem, an upper bound on the Bayes error based on the Mahalanobis distance has been employed for LDR in the multi-class scenario (Brunzell & Eriksson, 2000). The Mahalanobis-based LDR seeks to preserve the separation given by

$$J_M = \prod_{1 \le i < j \le K} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^\top (\mathbf{S}_i + \mathbf{S}_j)^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) \tag{3}$$

in the linearly reduced space.

However, the Mahalanobis distance, just like Fisher's criterion, does not take the difference in covariance matrices into account.

2.3. Chernoff criterion

To account for the difference in covariance matrices among the classes, a heteroscedastic extension of the Mahalanobis distance based on the Bhattacharya distance has been proposed for LDR (Decell Jr & Marani, 1976). Following this, there has been the use of a wider class of Bregman divergences, notably, the Kullback–Leibler divergence (Decell & Mayekar, 1977) for heteroscedastic LDR. Yet, while the Bhattacharya distance provides a good enough bound on the Bayes error, it has been shown that the Chernoff bound provides a slightly tighter bound than the Bhattacharya distance (Duda et al., 2012; Nielsen, 2014). Thus, a directed distance matrix (DDM) based on the Chernoff criterion has been developed for dimensionality reduction in the two-class case (Loog & Duin, 2002), as well as in the multi-class setting (Duin & Loog, 2004). Specifically, based on this DDM, the following Chernoff criterion is derived:

$$\begin{aligned} J_C = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \pi_i \pi_j \mathrm{trace}\Big[& (\mathbf{T}^\top \mathbf{S}_W \mathbf{T})^{-1} \mathbf{T}^\top \mathbf{S}_W^{\frac{1}{2}} \Big((\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} \mathbf{S}_W^{-\frac{1}{2}} \\ & \times (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^\top \mathbf{S}_W^{-\frac{1}{2}} (\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} + \frac{1}{\tau_i \tau_j} (\log \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}} \\ & - \tau_i \log \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_i \mathbf{S}_W^{-\frac{1}{2}} - \tau_j \log \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_j \mathbf{S}_W^{-\frac{1}{2}}) \Big) \mathbf{S}_W^{\frac{1}{2}} \mathbf{T} \Big], \end{aligned} \tag{4}$$

with

$$\tau_i = \frac{\pi_i}{\pi_i + \pi_j}, \quad \tau_j = \frac{\pi_j}{\pi_i + \pi_j} \quad \text{and} \quad \mathbf{S}_{ij} = \pi_i \mathbf{S}_i + \pi_j \mathbf{S}_j, \tag{5}$$

which is maximised to obtain an optimum linear transformation (Duin & Loog, 2004).

However, while the original LDA procedure provides reduction to at most $K-1$ dimensions, the LDR approaches described do