Information Sciences 382–383 (2017) 1–14

Contents lists available at ScienceDirect

Information Sciences

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

Building a discriminatively ordered subspace on the generating matrix to classify high-dimensional spectral data

Rui Zhu[a], Kazuhiro Fukui[b], Jing-Hao Xue[a,*]

[a] Department of Statistical Science, University College London, London WC1E 6BT, UK
[b] Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan

ARTICLE INFO

Article history:
Received 22 July 2016
Revised 28 November 2016
Accepted 3 December 2016
Available online 5 December 2016

Keywords:
Discriminatively ordered subspace
Generalised difference subspace
Generating matrix
SIMCA
Spectral data classification
Subspace method

ABSTRACT

Soft independent modelling of class analogy (SIMCA) is a widely-used subspace method for spectral data classification. However, since the class subspaces are built independently in SIMCA, the discriminative between-class information is neglected. An appealing remedy is to first project the original data to a more discriminative subspace. For this, generalised difference subspace (GDS) that explores the information between class subspaces in the generating matrix can be a strong candidate. However, due to the difference between a class subspace (of infinite scale) and a class (of finite scale), the eigenvectors selected by GDS may not also be discriminative for classifying samples of classes. Therefore in this paper, we propose a discriminatively ordered subspace (DOS): different from GDS, our DOS selects the eigenvectors with high discriminative ability between classes rather than between class subspaces. The experiments on three real spectral datasets demonstrate that applying DOS before SIMCA outperforms its counterparts.

1. Introduction

High-dimensional spectral data, such as near infrared (NIR) spectroscopic data and mass spectrometry (MS) data, are widely used in a variety of fields, for example chemometrics, bioinformatics and hyperspectral image analysis. In the analysis of spectral data, classification is an omnipresent task [2,4,7,9,10,13], which enables us to distinguish different species, identify the geographical origins of the products, or predict molecular substructure, to name a few.

Fig. 1 shows an example for NIR spectroscopic data of two classes, the chicken meat samples and the turkey meat samples. Each curve depicts the spectrum of a sample, which is usually represented by a high-dimensional feature vector. A classification task is to classify the spectra of new samples into the two classes based on the information provided by some labelled training spectra. In this paper, we focus on two-class classification. Based on the two-class classification results, multi-class classification can be readily obtained by using the one-vs-one or one-vs-all strategy [3].

Soft independent modelling of class analogy (SIMCA) [12] is a subspace-based classification method that is widely used in the two-class classification of high-dimensional spectral data in chemometrics [2,4,10]. When SIMCA is used for two-class classification, firstly two class subspaces are built for the two classes separately through using principal component analysis (PCA). Then an F-test, which tests whether the residual standard deviation of a new sample from the subspace of a class is

[*] Corresponding author.
E-mail address: jinghao.xue@ucl.ac.uk (J.-H. Xue).

http://dx.doi.org/10.1016/j.ins.2016.12.001
0020-0255/© 2016 Elsevier Inc. All rights reserved.

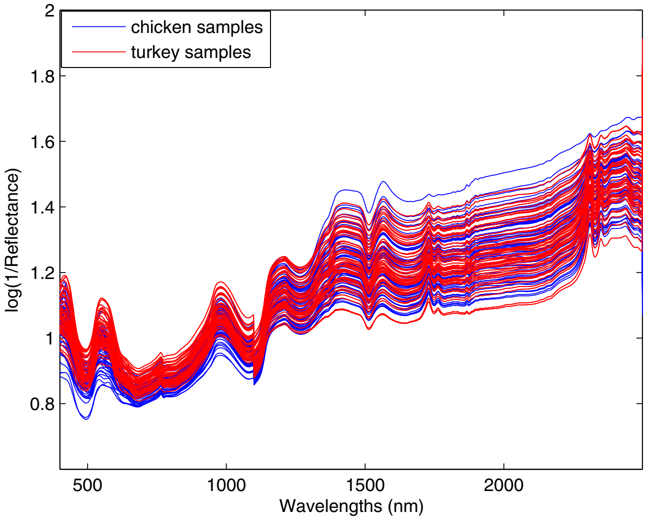
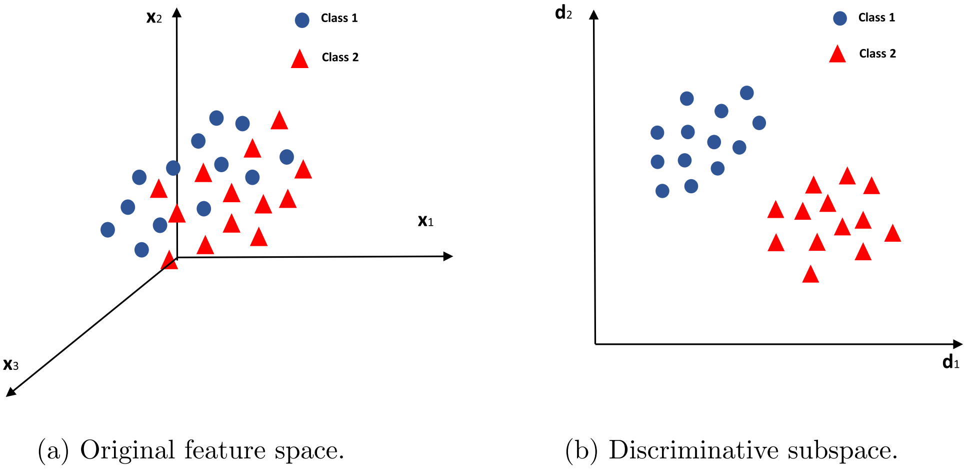


Fig. 1. Spectra of meat samples from two classes: chicken and turkey.



(a) Original feature space.

(b) Discriminative subspace.

Fig. 2. (a) Two classes of samples are mixed together in the original 3-dimensional feature space. (b) The same groups of samples can be well separated when they are projected to a discriminative 2-dimensional subspace.

statistically significantly different from the residual standard deviation of the training set of that class, is used to determine the class membership of the new sample. The PC-subspace is considered as a good class model for high-dimensional data because it extracts the most variable information in the data to few PCs and gets rid of a large amount of redundant information in the original feature dimensions. SIMCA is originally designed for both outlier detection and classification. In this paper, we treat SIMCA as a simple classification method that assign a new sample to the class with the smallest *F*-value as suggested in [8].

In spite of its wide use, SIMCA suffers from the problem that the class subspaces are built independently without considering between-class information. Therefore the *F*-value calculated independently for each class may not be discriminative enough to classify a new sample.

An appealing solution to this problem is to find a more discriminative subspace than the original feature space and project the data to this subspace before applying SIMCA. The projections of the samples to this discriminative subspace are expected to be more separated and can be more easily classified than those in the original feature space, as illustrated in Fig. 2. Also, as the new subspace contains more discriminative information for classification, the *F*-value calculated in this subspace is expected to be more discriminative. It is therefore the objective of our work in this paper to find such a discriminative subspace.

Recently, Fukui and Maki [6] propose the generalised difference subspace (GDS) projection as a preprocessing method to improve a popular subspace-based classifier called mutual subspace method (MSM) in image set-based object recognition. GDS aims to tackle an issue of MSM: the class subspaces are independently generated by PCA in a class-by-class manner, and