Available online at www.sciencedirect.com
ScienceDirect
Knowledge-Based Systems 19 (2006) 363–370
Knowledge-Based SYSTEMS
www.elsevier.com/locate/knosys

The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data

Tom Howley [a], Michael G. Madden [a,*], Marie-Louise O'Connell [b], Alan G. Ryder [b]

[a] *Department of Information Technology, National University of Ireland, University Road, Galway, Ireland*
[b] *National Centre for Biomedical Engineering Science, National University of Ireland, University Road, Galway, Ireland*

Received 28 October 2005; accepted 28 November 2005
Available online 8 February 2006

Abstract

This paper presents the results of an investigation into the use of machine learning methods for the identification of narcotics from Raman spectra. The classification of spectral data and other high-dimensional data, such as images, gene-expression data and spectral data, poses an interesting challenge to machine learning, as the presence of high numbers of redundant or highly correlated attributes can seriously degrade classification accuracy. This paper investigates the use of principal component analysis (PCA) to reduce high-dimensional spectral data and to improve the predictive performance of some well-known machine learning methods. Experiments are carried out on a high-dimensional spectral dataset. These experiments employ the NIPALS (Non-Linear Iterative Partial Least Squares) PCA method, a method that has been used in the field of chemometrics for spectral classification, and is a more efficient alternative than the widely used eigenvector decomposition approach. The experiments show that the use of this PCA method can improve the performance of machine learning in the classification of high-dimensional data.

Keywords: Machine learning; High-dimensional data; Principal component analysis; NIPALS; Spectroscopy

1. Introduction

The automatic identification of illicit materials using Raman spectroscopy is of significant importance for law enforcement agencies. High-dimensional spectral data can pose problems for machine learning as predictive models based on such data run the risk of overfitting. Furthermore, many of the attributes may be redundant or highly correlated, which can also lead to a degradation of prediction accuracy.

This problem is equally relevant to many other application domains, such as the classification of gene-expression microarray data [1], image data [2] and text [3]. In the classification task considered in this paper, Raman spectra are used for the identification of acetaminophen, a pain-relieving drug that is found in many over-the-counter medications, within different mixtures. In the physical sciences, a statistical approach to classification is normally taken (Chemometrics) [4], and these methods use PCA to handle the high-dimensional spectra. PCA is a classical statistical method for transforming attributes of a dataset into a new set of uncorrelated attributes called principal components (PCs). PCA can be used to reduce the dimensionality of a dataset, while still retaining as much of the *variability* of the dataset as possible. The goal of this research is to determine if PCA can be used to improve the performance of machine learning methods in the classification of such high-dimensional data.

In the first set of experiments presented in this paper, the performance of five well-known machine learning techniques (support vector machines, *k*-nearest neighbours, C4.5 decision tree, RIPPER and Naive Bayes) along with classification by linear regression are compared by testing

* Corresponding author.
E-mail addresses: thowley@vega.it.nuigalway.ie (T. Howley), michael.madden@nuigalway.ie (M.G. Madden), ML.OConnell@nuigalway.ie (M.-L. O'Connell), alan.ryder@nuigalway.ie (A.G. Ryder).

0950-7051/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2005.11.014

them on a Raman spectral dataset. A number of pre-processing techniques such as normalisation and first derivative are applied to the data to determine if they can improve the classification accuracy of these methods. A second set of experiments is carried out in which PCA and machine learning (and the various pre-processing methods) are used in combination. This set of PCA experiments also facilitates a comparison of machine learning with the popular chemometric technique of principal component regression (PCR), which combines PCA and linear regression.

The main contributions of this research are as follows:

(1) It presents a promising approach for the classification of substances within complex mixtures based on Raman spectra, an application that has not been widely considered in the machine learning community. This approach could also be applied to other high-dimensional classification problems.
(2) It proposes the use of NIPALS PCA for data reduction, a method that is much more efficient than the widely used eigenvector decomposition method.
(3) It demonstrates the usefulness of PCA for reducing dimensionality and improving the performance of a variety of machine learning methods. Previous work has tended to focus on a single machine learning method. It also demonstrates the effect of reducing data to different numbers of principal components.

The paper is organised as follows. Section 2 will give a brief description of Raman spectroscopy and outline the characteristics of the data it produces. Section 3 describes PCA, the NIPALS algorithm for PCA that is used here and the PCR method that incorporates PCA into it. Section 4 provides a brief description of each machine learning technique used in this investigation. Experimental results along with a discussion are presented in Section 5. Section 6 describes related research and Section 7 presents the conclusion of this study.

2. Raman spectroscopy

Raman spectroscopy is the measurement of the wavelength and intensity of light that has been scattered inelastically by a sample, known as the Raman effect [5]. This Raman scattering provides information on the vibrational motions of molecules in the sample compound, which in turn provides a molecular fingerprint. Every compound has its own unique Raman spectrum that can be used for sample identification. Each point of a spectrum represents the intensity recorded at a particular wavelength. A Raman dataset therefore has one attribute for each point on its constituent spectra. Raman spectra can be used for the identification of materials such as narcotics [4], hazardous waste [6] and explosives [7].

Raman spectra are a good example of high-dimensional data; a Raman spectrum is typically made up of 500–3000 data points, and many datasets may only contain 20–200 samples. However, there are other characteristics of Raman spectra that can be problematic for machine learning:

- *Collinearity*: many of the attributes (spectral data points) are highly correlated to each other which can lead to a degradation of the prediction accuracy.
- *Noise*: particularly prevalent in spectra of complex mixtures. Predictive models that are fitted to noise in a dataset will not perform well on other test datasets.
- *Fluorescence*: the presence of fluorescent materials in a sample can obscure the Raman signal and therefore make classification more difficult [4].
- *Variance of Intensity*: a wide variance in spectral intensity occurs between different sample measurements [8].

3. Principal component analysis

In the following description, the dataset is represented by the matrix X, where X is a $N \times p$ matrix. For spectral applications, each row of X, the p-vector x_i contains the intensities at each wavelength of the spectrum sample i. Each column, X_j contains all the observations of one attribute. PCA is used to overcome the previously mentioned problems of high-dimensionality and collinearity by reducing the number of predictor attributes. PCA transforms the set of inputs $X_1, X_2, \ldots, X_n$ into another set of column vectors $T_1, T_2, \ldots, T_N$ where the T's have the property that most of the original data's information content (or most of its variance) is stored in the first few T's (the principal component scores). The number of PCs that account for a portion of the total variance of the original data (i.e., their associated eigenvalues are non-zero) is equal to either $N-1$ or p,which ever is the smaller. With PCA the data can be reduced to a smaller number of dimensions, with low information loss, simply by discarding the higher numbered PCs and retaining the first set of PCs that account for most of the original data's total variance. Each PC is a linear combination of the original inputs and each PC is orthogonal, which therefore eliminates the problem of collinearity. This linear transformation of the matrix X is specified by a $p \times p$ matrix P so that the transformed variables T are given by:

$T = XP$ or alternatively X is decomposed as follows:

$$X = TP^T, \tag{1}$$

where P is known as the *loadings matrix*. The columns loadings matrix P can be calculated as the eigenvectors of the matrix X^TX [9], a calculation which can be computationally intensive when dealing with datasets of 500–3000 attributes. A much quicker alternative is the NIPALS method. The NIPALS method does not calculate all the PCs at once as is done in the eigenvector approach. Instead, it calculates the first PC by getting the first PC score,