Available online at www.sciencedirect.com
SCIENCE DIRECT
Knowledge-Based Systems 18 (2005) 217–224
Knowledge-Based SYSTEMS
www.elsevier.com/locate/knosys

An improved genetic programming technique for the classification of Raman spectra

Kenneth Hennessy*, Michael G. Madden, Jennifer Conroy, Alan G. Ryder

Department of Information Technology and Department of Chemistry, National University of Ireland, Galway, Ireland

Received 26 October 2004; accepted 30 October 2004
Available online 11 April 2005

Abstract

The aim of this study is to evaluate the effectiveness of genetic programming relative to that of more commonly-used methods for the identification of components within mixtures of materials using Raman spectroscopy. A key contribution of the genetic programming technique proposed in this research is that it explicitly aims to optimise the certainty levels associated with discovered rules, so as to minimize the chance of misclassification of future samples.

Keywords: Machine learning; Genetic programming; Neural networks; Spectroscopy; Raman

1. Introduction

Raman spectroscopy may be described as the measurement of the intensity and wavelength of inelastically scattered light from molecules when they are excited by a monochromatic light source. The Raman scattered light occurs at wavelengths that are shifted from the incident light by the energies of molecular vibrations. The analytical applications of Raman spectroscopy continue to grow; typical applications are in structure determination [1], multi-component qualitative analysis and quantitative analysis [2].

Traditionally, multivariate data analysis techniques such as partial least squares (PLS) and principal component regression (PCR) have been used to identify the presence of specific compounds in mixtures from their Raman spectra [2]. However, Raman spectral elucidation suffers from several problems. The presence of fluorescent compounds, impurities, complex mixtures and other environmental and instrumental factors can greatly add to the difficulty in identifying compounds from their spectra [3]. Increasingly, machine learning techniques are being investigated as a possible solution to these problems, as they have been shown to be successful in conjunction with other spectroscopic techniques, such as the use of neural networks to identify bacteria from their infra-red spectra [4] and the application of neural networks to quantification of Fourier transform infra-red (FTIR) spectroscopy data [5]. Schultz et al. [6] used a neural network and PLS to identify individual components in biological mixtures from their Raman spectra, and Benjathapanum et al. [7] used PCR and neural networks to classify ultraviolet–visible spectroscopic data.

In this paper, neural networks, PLS and PCR are compared with the evolutionary technique of genetic programming for predicting which of four solvents are present in a range of mixtures. Genetic programming offers an advantage over neural networks and chemometric methods in this area as the rules generated are interpretable and may be used in isolation or in conjunction with expert opinion to classify spectra.

In combination with the environmental and instrumental problems outlined above, a significant challenge that also arises in other machine learning problems, is in the high sample dimensionality and low sample number commonly found in this area. In many real laboratory applications, it is required to identify materials based on a small number of reference spectra. While commercial spectral databases

* Corresponding author. Address: Department of Information Technology, National University of Ireland, Galway, Ireland. Tel.: +353 91 524411x2041.
E-mail addresses: hennessy@vega.it.nuigalway.ie (K. Hennessy), michael.madden@nuigalway.ie (M.G. Madden), jennifer.conroy@nuigalway.ie (J. Conroy), alan.ryder@nuigalway.ie (A.G. Ryder).

0950-7051/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2004.10.001

typically contain spectra for some thousands of materials, they are organised into categories and for individual groups of materials such as the solvents considered here, spectra would be provided for only a small number of mixtures, if any. Machine learning models exhibiting poor generalisation and overfitting to the training data are a consequence of this problem.

In response to this, rather than aiming simply to evolve equations that classify the training data correctly, our approach aims to optimise selection of equations so as to minimize the chance of misclassification of future predicted samples and thereby minimize the problems associated with low sample numbers.

Not many research groups have published applications of genetic programming for the interpretation of spectra. Goodacre [8] discusses the application genetic programming to FTIR spectroscopy image analysis. Using the same genetic programming software, Ellis et al. [9] have quantified the spoilage of meat from its FTIR spectra and Taylor et al. [10] have classified *Eubacterium* species based on their pyrolysis mass spectra.

2. Description of task

Raman spectra were recorded on a Labram Infinity (J-Y Horiba) equipped with a liquid nitrogen cooled CCD detector and a 488 nm excitation source. All spectra were recorded at a set interval of $\sim$400–3340 cm^{-1} with a resolution of $\sim$11 cm^{-1}. The liquid samples were held in 1 cm pathlength quartz cuvettes and mounted in a macro sample holder (J-Y Horiba). The macro lens has a focal length of 40 mm, which focuses through the cuvette to the centre of the liquid. The spectral data was not corrected for instrument response. Three spectra were taken for each sample, the raw data for each sample were then averaged and analysed using the Unscrambler chemometrics software package. The solvents (all spectroscopic grade), acetone, acetonitrile, cyclohexane, and toluene were obtained from Sigma-Aldrich and used as received. Solutions of different concentrations (Table 1) were made up by mixing known volumes of each solvent.

The objective is to be able to predict accurately whether or not a specific solvent is present in a mixture of other solvents. The 24 samples contain differing combinations of four solvents, acetone (A), cyclohexane (C), acetonitrile (Acn) and toluene (T), with compositions as listed in Table 1. Identification of each solvent is treated as a separate classification task. For each solvent, the dataset was divided into a training/testing set of 14 samples and a validation set of 10. The validation set in each case contained five positive and five negative samples.

There are two challenging aspects to the dataset. Firstly, as mentioned earlier, the dimensionality of the data is very high, with 1024 points per sample and the number of samples is low. Secondly, for all four solvents, the most intense peaks occur in the same region of the spectra. This may be seen in Figs. 2–4 (Section 4.1), which plot the Raman spectra of each pure solvent. The solvent mixtures detailed in Table 1 are a preliminary dataset produced for this study; the authors are currently collecting a more extensive and diverse dataset for future research.

Table 1
Chemical composition of samples used in this study (acetone (A), cyclohexane (C), acetonitrile (A_{cn}) and toulene (T))

No.	A (%)	C (%)	Acn (%)	T (%)
1	0	100	0	0
2	0	0	0	100
3	100	0	0	0
4	0	0	100	0
5	50	50	0	0
6	50	0	0	50
7	50	0	50	0
8	0	50	0	50
9	0	0	50	50
10	75	25	0	0
11	75	0	0	25
12	75	0	25	0
13	0	75	0	25
14	25	75	0	0
15	25	0	0	75
16	0	25	0	75
17	0	0	25	75
18	25	0	75	0
19	0	0	75	25
20	33	0	33	33
21	33	33	33	0
22	33	33	0	33
23	0	33	33	33
24	25	25	25	25

3. Analysis techniques

3.1. Overview

This section outlines the use of standard chemometric techniques and neural networks to identify components in mixtures from their Raman spectra. It then goes on to describe an alternative technique based on genetic programming.

As mentioned in Section 1, chemometric techniques are widely used for analysing spectra. While there are many such techniques, the two chosen in this study are PCR and PLS, as they are particularly well established for the classification of spectroscopic data [11–13]. Neural networks have been used successfully in conjunction with spectroscopic data in past research for classification purposes [4,6]. Conventional feed-forward neural networks, however, can be hard to configure for a given problem and the means by which they form predictions are not particularly easy to interpret. Hence, they are often viewed as a ‘black box’ technique.

Genetic programming is a well-known and well-documented technique in machine learning [14]. In the approach taken in this paper, we attempt to evolve