A Machine Learning Application for Classification of Chemical Spectra

Michael G. Madden[1] **and Tom Howley**[2]

Abstract. This paper presents a software package that allows chemists to analyze spectroscopy data using innovative machine learning (ML) techniques. The package, designed for use in conjunction with lab-based spectroscopic instruments, includes features to encourage its adoption by analytical chemists, such as having an intuitive graphical user interface with a step-by-step 'wizard' for building new ML models, supporting standard file types and data preprocessing, and incorporating well-known standard chemometric analysis techniques as well as new ML techniques for analysis of spectra, so that users can compare their performance. The ML techniques that were developed for this application have been designed based on considerations of the defining characteristics of this problem domain, and combine high accuracy with visualization, so that users are provided with some insight into the basis for classification decisions.

1 Introduction

This work has been motivated by the need for more accurate analysis of spectroscopic data from mixtures of materials. Raman spectroscopy has been chosen as the specific target for this work, though the techniques developed are equally applicable to other forms of molecular spectroscopy.

Molecular spectroscopic techniques such as infra-red (IR), near infra-red (NIR), and Raman spectroscopy are widely used in analytical chemistry to characterise the molecular structure of materials, by measuring the radiant energy absorbed or scattered in response to excitation by an external light source [1]. When monochromatic light illuminates a material, a very small fraction is inelastically scattered at different wavelengths to the incident light; this is Raman scattering, and is due to the interaction of the light with the vibrational and rotational motions of the molecules. Thus, the Raman spectrum can be used as a molecular fingerprint. Ferraro *et al.* [2] provide an overview of Raman spectroscopy.

[1] Dr Michael G. Madden
Analyze IQ Limited, Cahercrin, Athenry, Galway, Ireland *and*
College of Engineering & Informatics, National University of Ireland, Galway, Ireland.
Email: michael.madden@analyzeiq.com; michael.madden@nuigalway.ie
[2] Dr Tom Howley
Analyze IQ Limited, Cahercrin, Athenry, Galway, Ireland. Email: tom.howley@analyzeiq.com

At the outset of this work, consideration was given to the key defining characteristics of this problem domain, which would inform the development of ML algorithms and associated software, as listed below.

- The dimensionality of the data is high, with typically 500-2000 data points per sample and the data is easily obtained. However, carefully-curated training samples are typically more difficult and expensive to obtain, so often one may have fewer than 100 samples available for training/calibration, particularly if the samples are mixtures of materials.
- A substance may give rise to multiple peaks along a spectrum, with local correlations in the data along the spectrum. Thus, it is best to avoid assuming that data points are independent attributes.
- In a mixture, peaks from the component substances may overlap or mask each other.
- Some materials fluoresce under the external illumination, resulting in a baseline that grows steadily along the spectrum rather than being level.
- The intensity (Y-axis) is arbitrary, depending on equipment and experimental settings, so some form of normalisation may be necessary.
- There may be noise in the signal due to instrumentation, external interference, or inaccuracies in the recorded composition of materials.

Some of these may be observed in Figure 1, which shows the Raman spectra of three pure substances: caffeine, glucose and cocaine.

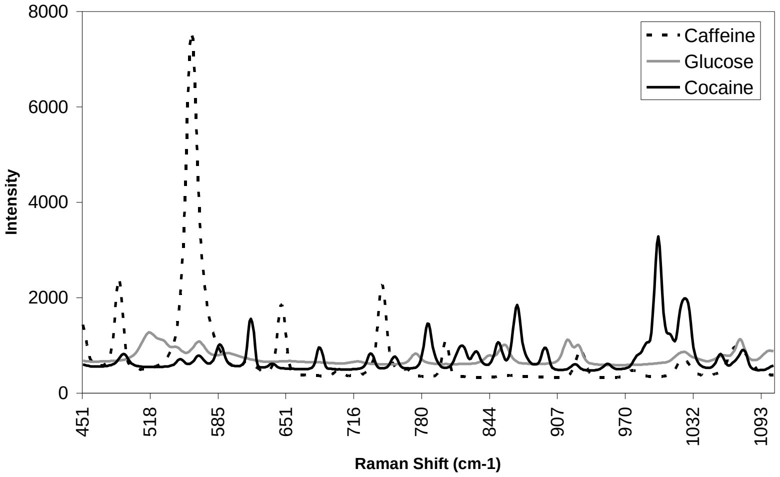


Figure 1. Raman spectra of three substances.

Previous publications have described some of the specific ML techniques that have been developed for this work [4, 5, 6, 7]. This paper focuses on the system's architecture and design, paying attention to features that facilitate its use by the target population of end users, who are typically analytical chemists without prior experience of using machine learning.

2 Review

In the domain of analytical chemistry, software tools for advanced spectral analysis typically implement techniques from a field known as *chemometrics.* Chemometric techniques are based on the use of the statistical transformation technique of Principal Component Analysis (PCA), a classical statistical method for transforming attributes of a dataset into a new set of uncorrelated attributes called principal components (PCs). The key idea is that this allows reduction of the data to a smaller number of dimensions, with low information loss, simply by discarding some PCs. Each PC is a linear combination of the original inputs and each PC is orthogonal, which therefore eliminates the problem of collinearity. SIMCA (Soft Independent Modeling of Class Analogy) is the most widely used chemometric classification technique [8]. In binary classification, SIMCA generates a separate PCA model for the set of samples of both classes. In prediction, the distance of a test sample to either model is calculated. Statistical tests are then used to determine if the test sample belongs to either class.

For quantitative analysis, Partial Least Squares (PLS) is a widely used chemometric technique [8]. PLS is a two-step multivariate regression method, which first reduces the data using PCA (using concentration information to extract the PC scores) and then performs linear regression on the PC scores.

Both SIMCA and PLS are industry standards for advanced spectral analysis. However, some researchers have applied machine learning methods to this domain, for example: decision trees [10]; Naïve Bayes [11]; Artificial Neural Network (ANN) [12]; and Support Vector Machine (SVM) [13]. Some machine learning methods appear to be unsuited to dealing with spectral data, including Naïve Bayes, because of its independence assumption, and k-NN, which does not work as well in high-dimensional spaces. These methods require some form of data transformation or reduction in order to be useful in the spectral domain.

The most commonly-used machine learning technique for spectral analysis is the ANN, another technique for which data reduction is advised. The ANN has been shown to give better results than PLS and PCR in regression analysis and it is particularly regarded for its ability to model non-linear relationships in the data [12]. However, chemometric techniques have not always been found to be inferior to the ANN [13]. A key problem with the ANN is that their use is considered by many to be a ‘black art’: finding the right network structure (number of hidden nodes, type of threshold function) and the selection of initial connection weights can be a problem in the generation of an ANN model, all of which has a direct impact on the performance achieved. The ANN is therefore less suitable for use by non-experts in its use, which may often be the case in the analytical chemistry domain. Another often-cited failing of the ANN is that it does not lend itself easily to human interpretation and does not provide any added insight into the data, a key requirement for spectral analysis. In contrast, PCA can be used to generate 2D or 3D views of a spectral dataset, by which clusters or outliers may be discerned.

A final key issue with the use of machine learning methods for spectral analysis is that none of these techniques have been designed or tailored specifically for spectral applications. PCA-based techniques, which have been used in the field of spectral analysis since the early 1970s, are still the most widely used techniques in commercial applications. However, if the above issues were addressed, the potential of machine learning for dealing with the challenges faced in spectroscopic analysis could be realized. The provision of innovative machine learning techniques, which are developed specifically for spectral analysis and can improve on the performance of existing chemometric techniques, is one of the key motivations behind the software package described in this paper. One of these new techniques is based on the SVM, a machine learning method that has been more recently identified as being suited to the analysis of high-dimensional spectral data [13]. However, SVMs have not been extensively tested on Raman spectral data, prior to the work that is the subject of this paper.

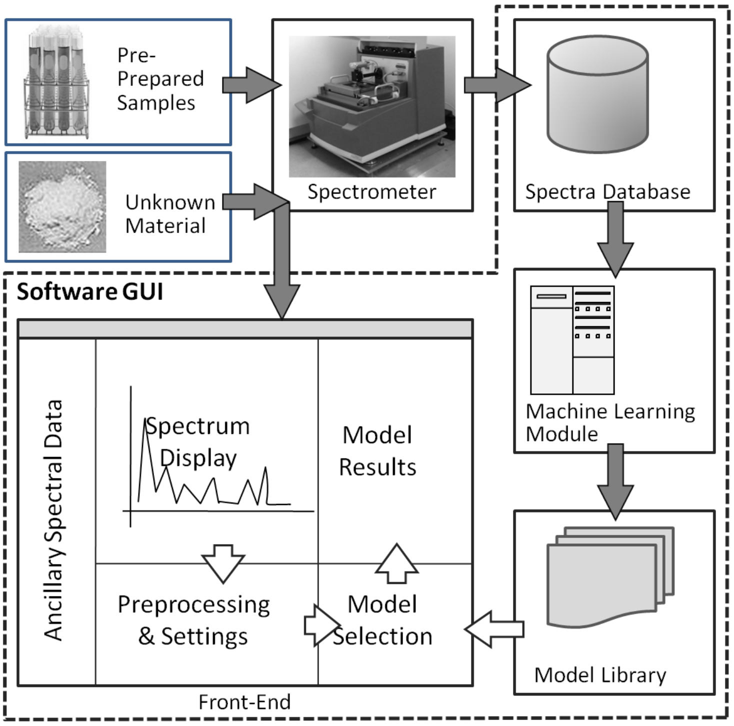


Figure 2. Overview of system architecture.

3 Architecture & User Interface

The architecture of the system is shown in Figure 2. It has four main components:

1. A Spectra Database: this stores spectra from samples of known composition are stored
2. A Machine Learning Module: using this, the user may select an algorithm, configure its parameters and build a model by applying the algorithm to data selected from the database
3. A Model Library, storing previously-built models
4. A GUI (Graphical User Interface): this allows data from a spectroscopy instrument to be retrieved and displayed, and analysed using one or more models from the Model Library.

These components are described in the sub-sections that follow.

3.1 Spectra Database

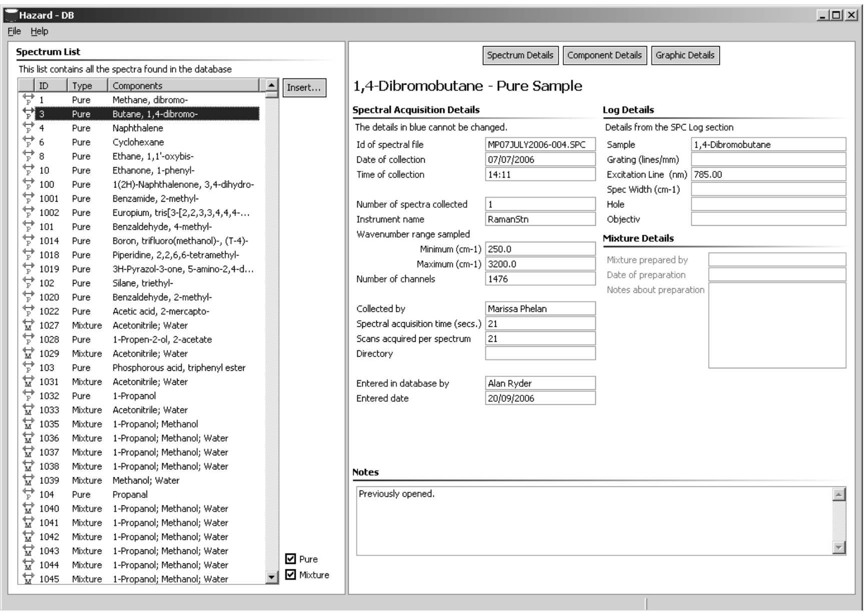
The software package includes a database structure and front-end, designed to store both pure materials and mixtures of materials; mixtures are distinguished by being defined in terms of the proportions of each of their components. A view of the database front-end is shown in Figure 3.

For all materials, the Chemical Abstracts Service (CAS) registry number is stored; this is a unique identifier that links the substance to a wide variety of information about it[3]. Other information about its manufacturer, its common names and its data collection process is also stored. The database front-end allows searching by name or CAS number and includes graphical display of spectra.

3.2 Model Library & Machine Learning Module

The Model Library (shown schematically in Figure 2) has a ‘plug-in’ architecture: at startup, the package scans a specific directory for models, which are loaded dynamically. Thus, in addition to models supplied with the software, new models can be provided to a user or built using the software. The reason for having multiple models is that different models may work with different targets (e.g. distinguish between different white powders; predict the concentration of cocaine in a sample) or may have been built using different analytical techniques (e.g. linear regression; support vector machine).

[3] http://www.cas.org/expertise/cascontent/registry/regsys.html



Spectrum List

This list contains all the spectra found in the database

ID	Type	Components
1	Pure	Methane, dibromo-
3	Pure	Butane, 1,4-dibromo-
4	Pure	Naphthalene
6	Pure	Cyclohexane
8	Pure	Ethane, 1,1'-oxybis-
10	Pure	Ethanone, 1-phenyl-
100	Pure	1(2H)-Naphthalenone, 3,4-dihydro-
1001	Pure	Benzamide, 2-methyl-
1002	Pure	Europium, tris[3-[2,2,3,3,4,4,4-...
101	Pure	Benzaldehyde, 4-methyl-
1014	Pure	Boron, trifluoro(methanol)-, (T-4)-
1018	Pure	Piperidine, 2,2,6,6-tetramethyl-
1019	Pure	3H-Pyrazol-3-one, 5-amino-2,4-d...
102	Pure	Silane, triethyl-
1020	Pure	Benzaldehyde, 2-methyl-
1022	Pure	Acetic acid, 2-mercapto-
1027	Mixture	Acetonitrile; Water
1028	Pure	1-Propen-2-ol, 2-acetate
1029	Mixture	Acetonitrile; Water
103	Pure	Phosphorous acid, triphenyl ester
1031	Mixture	Acetonitrile; Water
1032	Pure	1-Propanol
1033	Mixture	Acetonitrile; Water
1035	Mixture	1-Propanol; Methanol
1036	Mixture	1-Propanol; Methanol; Water
1037	Mixture	1-Propanol; Methanol; Water
1038	Mixture	1-Propanol; Methanol; Water
1039	Mixture	Methanol; Water
104	Pure	Propanal
1040	Mixture	1-Propanol; Methanol; Water
1041	Mixture	1-Propanol; Methanol; Water
1042	Mixture	1-Propanol; Methanol; Water
1043	Mixture	1-Propanol; Methanol; Water
1044	Mixture	1-Propanol; Methanol; Water
1045	Mixture	1-Propanol; Methanol; Water

☑ Pure
☑ Mixture

1,4-Dibromobutane - Pure Sample

Spectral Acquisition Details

The details in blue cannot be changed.

Id of spectral file	MP07JULY2006-004.SPC
Date of collection	07/07/2006
Time of collection	14:11
Number of spectra collected	1
Instrument name	RamanStn
Wavenumber range sampled	
Minimum (cm-1)	250.0
Maximum (cm-1)	3200.0
Number of channels	1476
Collected by	Marissa Phelan
Spectral acquisition time (secs.)	21
Scans acquired per spectrum	21
Directory	
Entered in database by	Alan Ryder
Entered date	20/09/2006

Log Details

Details from the SPC Log section

Sample	1,4-Dibromobutane
Grating (lines/mm)	
Excitation Line (nm)	785.00
Spec Width (cm-1)	
Hole	
Objectiv	

Mixture Details

Mixture prepared by	
Date of preparation	
Notes about preparation	

Notes

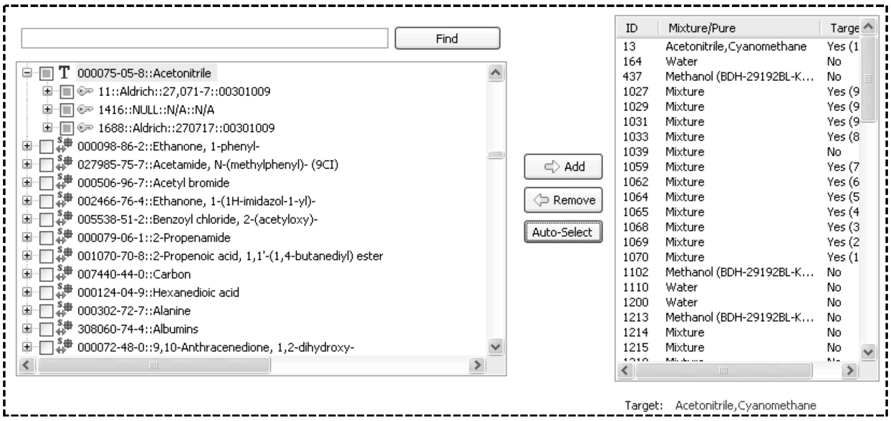
Previously opened.

Figure 3. Screen-shot of Spectra Database.

The Machine Learning Module provides functionality for those users who wish to build their own models. A conventional supervised learning approach is followed. In a laboratory setting, samples are prepared with carefully controlled composition. Their spectra are then collected; these form a labelled training set from which a classification or regression model may be built using any appropriate machine learning technique. The model's accuracy may be optionally evaluated by holding out a specified percentage of the training data, loading a test set, or using repeated cross-validation runs. The model is then added automatically to the Model Library, so that it may be used to analyse new samples.

To facilitate its use, the Machine Learning Module uses a step-by-step 'wizard' approach. As shown in Figure 4, the user begins by selecting a target substance for the model from the database, then selecting the rest of the data to that will be included in the training set. Functionality is provided to 'auto-select' the training set, by scanning the databases for all samples that include the target material, and then all other samples that contain materials that occur in mixtures with the target material. After that, the user may choose to apply pre-processing operations such as normalisation, smoothing using an *n*-point moving average, or *k*-th order Savitzky-Golay derivative [3].

Next, as shown in Figure 5, the user may chose an analysis method from a tree-structured list of ML methods specific to this work as well as some standard ML methods and chemometric methods. Depending on the method chosen, parameters may be specified; sensible defaults are provided.

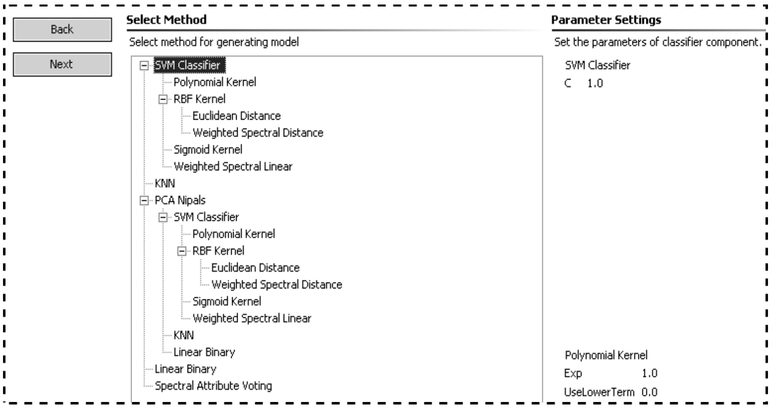


- 000075-05-8::Acetonitrile
  - 11::Aldrich::27,071-7::00301009
  - 1416::NULL::N/A::N/A
  - 1688::Aldrich::270717::00301009
- 000098-86-2::Ethanone, 1-phenyl-
- 027985-75-7::Acetamide, N-(methylphenyl)- (9CI)
- 000506-96-7::Acetyl bromide
- 002466-76-4::Ethanone, 1-(1H-imidazol-1-yl)-
- 005538-51-2::Benzoyl chloride, 2-(acetyloxy)-
- 000079-06-1::2-Propenamide
- 001070-70-8::2-Propenoic acid, 1,1'-(1,4-butanediyl) ester
- 007440-44-0::Carbon
- 000124-04-9::Hexanedioic acid
- 000302-72-7::Alanine
- 308060-74-4::Albumins
- 000072-48-0::9,10-Anthracenedione, 1,2-dihydroxy-

ID	Mixture/Pure	Targe
13	Acetonitrile,Cyanomethane	Yes (1
164	Water	No
437	Methanol (BDH-29192BL-K...	No
1027	Mixture	Yes (9
1029	Mixture	Yes (9
1031	Mixture	Yes (9
1033	Mixture	Yes (8
1039	Mixture	No
1059	Mixture	Yes (7
1062	Mixture	Yes (6
1064	Mixture	Yes (5
1065	Mixture	Yes (4
1068	Mixture	Yes (3
1069	Mixture	Yes (2
1070	Mixture	Yes (1
1102	Methanol (BDH-29192BL-K...	No
1110	Water	No
1200	Water	No
1213	Methanol (BDH-29192BL-K...	No
1214	Mixture	No
1215	Mixture	No

Target: Acetonitrile,Cyanomethane

Figure 4. Selecting the target material and training data.



Select Method

Select method for generating model

- SVM Classifier
  - Polynomial Kernel
  - RBF Kernel
    - Euclidean Distance
    - Weighted Spectral Distance
  - Sigmoid Kernel
  - Weighted Spectral Linear
- KNN
- PCA Nipals
  - SVM Classifier
    - Polynomial Kernel
    - RBF Kernel
      - Euclidean Distance
      - Weighted Spectral Distance
    - Sigmoid Kernel
    - Weighted Spectral Linear
  - KNN
  - Linear Binary
- Linear Binary
- Spectral Attribute Voting

Parameter Settings

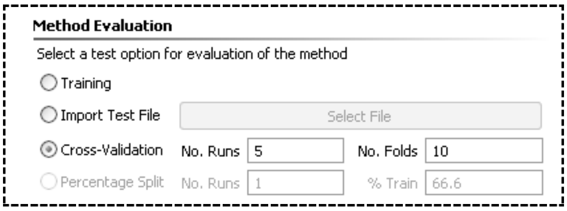
Set the parameters of classifier component.

SVM Classifier
C 1.0

Polynomial Kernel
Exp 1.0
UseLowerTerm 0.0

Figure 5. Selecting a model generation method and specifying parameters.

Finally, the user chooses what form of evaluation to use to measure the performance of the model being built, using the options shown in Figure 6.



Method Evaluation

Select a test option for evaluation of the method

○ Training

○ Import Test File Select File

● Cross-Validation No. Runs 5 No. Folds 10

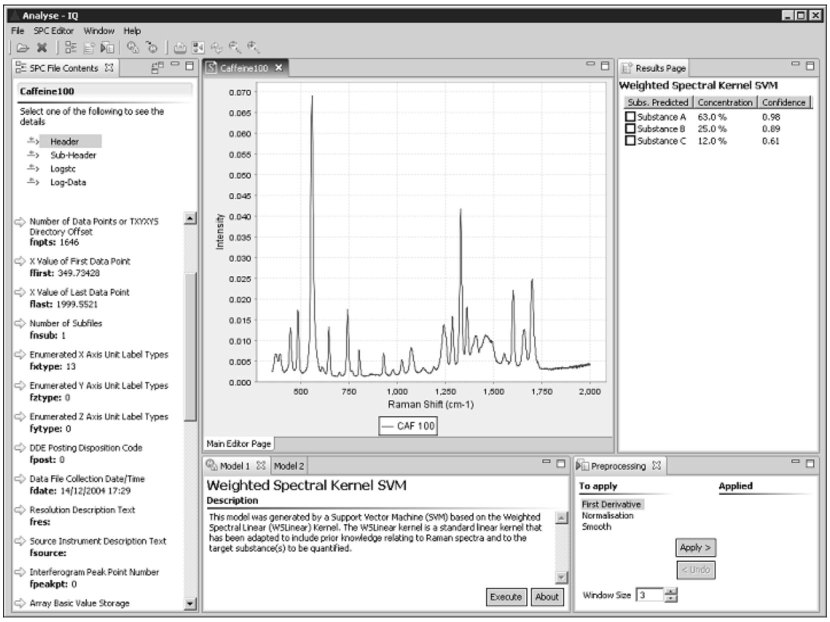
○ Percentage Split No. Runs 1 % Train 66.6

Figure 6. Selecting a method for model evaluation.

Having specified this information, the model is built automatically, and the evaluation results are displayed for the user to review. The model is automatically added to the Model Library so that it is available for further use.

3.3 Graphical User Interface

The GUI allows users to work with the library of machine learning models in an intuitive way. When a spectrum file is loaded, it is displayed in the main graphics pane, with ancillary data relating to it on the left pane, as shown in Figure 7.



Caffeine100

Select one of the following to see the details

- Header
- Sub-Header
- Logstc
- Log-Data

Number of Data Points or TXYXYS Directory Offset
fnpts: 1646

X Value of First Data Point
ffirst: 349.73428

X Value of Last Data Point
flast: 1999.5521

Number of Subfiles
fnsub: 1

Enumerated X Axis Unit Label Types
fxtype: 13

Enumerated Y Axis Unit Label Types
fztype: 0

Enumerated Z Axis Unit Label Types
fytype: 0

DDE Posting Disposition Code
fpost: 0

Data File Collection Date/Time
fdate: 14/12/2004 17:29

Resolution Description Text
fres:

Source Instrument Description Text
fsource:

Interferogram Peak Point Number
fpeakpt: 0

Array Basic Value Storage

Weighted Spectral Kernel SVM

Subs. Predicted	Concentration	Confidence
□ Substance A	63.0 %	0.98
□ Substance B	25.0 %	0.89
□ Substance C	12.0 %	0.61

Weighted Spectral Kernel SVM

Description

This model was generated by a Support Vector Machine (SVM) based on the Weighted Spectral Linear (WSLinear) Kernel. The WSLinear kernel is a standard linear kernel that has been adapted to include prior knowledge relating to Raman spectra and to the target substance(s) to be quantified.

To apply

First Derivative
Normalisation
Smooth

Applied

Window Size 3

Figure 7. Main user interface screen with a spectrum loaded.

To assist in examining data, the user may choose to apply the pre-processing operations described previously, and see their effects immediately. The user can then select a model from the model library and apply it to the spectrum; the model output is displayed in the right pane. If a classification model has been chosen, the output will list the material(s) predicted to be present. In the case of a regression model, the concentrations of materials will be listed. Depending on the model, a confidence factor associated with predictions may also be displayed. The user may select materials from the model predictions, to have their spectra overlaid on the original spectrum.

Significant effort has been put into the design of the system, to minimise risk of user error, bearing in mind that typical users may not be experienced in working with machine learning. For example, the models include information about what pre-processing steps (e.g. normalisation, smoothing or derivatives) were applied to its training data, as well as the spectrum range and sampling frequency of the training data. When a new spectrum is being analysed, it is automatically pre-processed and resampled if required.

4 Evaluation

This section evaluates the analytical chemistry software package in two ways, firstly considering the new machine learning techniques it provides, and then considering its broader functionality, including the insight it provides for users.

4.1 Evaluation of Performance of Machine Learning Techniques

The new ML techniques in the software package are based on the principle of incorporating specific spectral domain knowledge into their design. They include:

1. Spectral Attribute Voting (SAV): an ensemble-based approach that generates models for each point on a spectrum and also uses the shape of the spectral profile (i.e. peaks and troughs) in the generation of these models.
2. Weighted Spectral Kernel (WS Kernel): this approach uses an SVM with a custom kernel that is designed to take the shape of the spectral profile into account when comparing spectra. The WS Kernel also incorporates the pure spectrum of a target that is to be analysed. Modifications of the standard Linear and RBF kernels have been developed using this approach.
3. Improved Genetic Programming (GP): this technique uses a fitness function designed to optimise the assurance levels associated with discovered rules, so as to reduce the likelihood of misclassification of future samples [5].

Initial research into the above techniques focussed on direct comparisons with closely related ML techniques in addition to comparisons with chemometric techniques. The SAV method was shown to significantly outperform a number of ensemble techniques (e.g. AdaBoost and Bagging) in the classification of substances based on Raman spectra [7]. Similarly, previous experiments have shown that SVMs using WS kernels significantly improve on the performance of standard kernel SVMs, especially in spectral classification tasks [6].

The evaluation presented here examines the performance of the new ML techniques that were developed specifically for this application, along with standard chemometric techniques. The performance evaluation considers two separate tasks: the classification and quantification of materials.

For classification, WSLinear SVM (the WS Kernel approach combined with a standard Linear kernel), SAV and the standard SIMCA technique are applied to the classification of acetonitrile, an industrial solvent, based on Raman spectra. The dataset used for this set of experiments comprises 74 samples: 53 samples containing acetonitrile and 21 without acetonitrile. Stratified cross-validation was used to ensure that the same distribution of acetonitrile and non-acetonitrile samples was present in the training and test sets. The dataset was also normalised prior to this analysis, so that the intensity values of each sample ranged from a minimum of 0 to a maximum of 1. Table 1(a) reports the average error achieved by each technique in a 5x10-fold cross-validation test; the standard deviation of this error is also included. The average error is determined as follows: calculate the error rate over each run and then calculate average error over the five runs. Similarly, the standard deviation is based on the five test run averages.

The SIMCA analysis was undertaken using the Unscrambler multivariate analysis software package (V8.0, CAMO AS, Trondheim, Norway). For SIMCA based classification, separate models were constructed for both classes (samples with target and samples without target). A significance level of 5% was used for each classification step. The WSLinear Kernel SVM used an internal 3-fold cross-validation on the training set to determine the optimal settings for SVM and kernel. The SAV method used default settings with a decision tree as the base classifier. The results of Table 1(a) clearly show that the new ML techniques improve on the classification performance of SIMCA.

Table 1. Comparing standard chemometric methods with new Machine Learning methods for: (a) identification of Acetonitrile; (b) quantification of Cocaine concentration

Acetonitrile Classification	**%Error**
Chemometrics: SIMCA	8.65±2.23
New ML Techniques	
WSLinear Kernel SVM	2.16±1.54
Spectral Attribute Voting	1.08±1.13

Cocaine Quantification	**%Error**
Chemometrics: PLS	5.225
New ML Techniques	
WSLinear Kernel SVM	4.433
WSRBF Kernel SVM	3.704

The second phase of this evaluation compares the performance of two variants of the WS Kernel SVM technique, WSLinear and WSRBF, against PLS for the quantification of cocaine. The cocaine dataset comprises the spectra of 36 samples containing varying amounts of cocaine. More details of this dataset are given in Madden & Ryder [14]. Table 1(b) shows the average root mean squared error of predication (RMSEP) achieved by each method in the quantification of cocaine. This is computed using leave-one-out cross-validation: for each sample in turn, that sample was removed from the dataset and the remainder was used to build a model, which was then used to predict the concentration of cocaine in the sample that had been removed. This type of test was chosen due to the small dataset size. The PLS performance is taken from Madden & Ryder [14], which used the same cocaine dataset. As with the classification experiments, these results show that the new ML techniques have better accuracy than the standard PLS technique.

4.2 Evaluation of Functionality

In the domain of analytical chemistry, commonly used analysis packages include the Unscrambler (CAMO AS) and Grams/AI (Thermo Scientific). However, the analytical chemistry software described in this paper has several features that are not available in other packages:

- As well as providing standard chemometric analysis techniques, it provides a range of standard ML analysis techniques
- It has new spectral ML techniques, such as Weighted Spectral Kernel SVMs and Spectral Attribute voting, that tend to outperform other techniques and are not available in any other package
- It includes distinctive features to enable analytical chemists to apply ML in their work while minimising risk of errors, such as automatically applying appropriate pre-processing steps when analysing a spectrum with a model
- Its step-by-step model-building wizard leads non-experts in ML through the process of building new ML models
- Unlike most other analytical chemistry packages, it provides functionality for best-practice evaluation of the performance of new models, in particular supporting repeated n-fold cross-validation.

4.3 Evaluation of Insight into Decisions

The new ML techniques described above have been designed to provide novel visualisations that can provide domain experts (e.g. analytical chemists) with added insight into the data under analysis. For example, Figure 8, reproduced from [7], shows a typical visualisation generated by the SAV technique that is part of this software package. This figure highlights the points selected in the classification of 1,1,1-trichloroethane in a mixture of solvents. *Positive* points can be interpreted as providing significant evidence that the target (1,1,1-trichloroethane) **is** contained in the material being analysed, whereas *Negative* points provide evidence that it is **not**. In Figure 8, positive evidence greatly outweighs negative evidence, so the conclusion is that the target is present.

Figure 8 also shows that the presence or absence of 1,1,1-trichloroethane was determined using points principally located on a large peak at 520 cm^{-1} and a smaller peak at 720 cm^{-1}. The 520 cm^{-1} band is the C-Cl stretch vibration and would be expected to be the primary discriminator. The large peak at 3000 cm^{-1} is largely ignored as this area corresponds to the C-H bond region of the spectrum, which is less helpful in classification as all of the solvent in the dataset contained C-H bonds. This correlation between points chosen and chemical structure of the target in question provides the user with a direct way of understanding the basis for decisions, and serves to reassure the user that predictions being made by the software are reasonable.

While Figure 8 shows a visualisation generated from the SAV technique, equivalent visualisations may be generated using the other techniques listed in Section 4.1. (Visualisations for one of these, the improved GP technique, have been presented previously [5].) An interesting aspect of this is that visualisations can be generated even when SAV uses a relatively ‘opaque’ ML technique such as an ANN as its base classifier. Similarly, useful visualisations can be generated from the WS Kernel technique, even in the context of an SVM with RBF kernel.

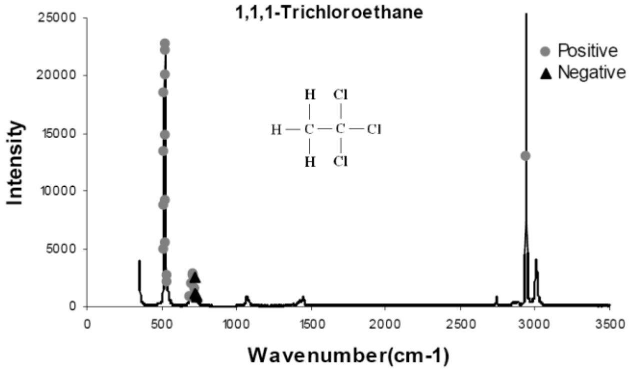


Figure 8. Spectrum of 100% 1,1,1-trichloroethane showing points selected by SAV technique [7]

5 Commercial Benefits

This software package specifically addresses the problem of identifying the components in mixtures, based on their spectra. For a wide range of real-world applications, it is essential to be able to assess the composition of mixtures accurately. For example:

- First Responders: The ML techniques developed for this work can aid the identification of household poisons, unlabelled medicines, and hazardous materials, in combination with portable instrumentation.
- Law Enforcement and Forensics: Drug concentration is useful intelligence that indicates where on the ‘supply chain’ a narcotics seizure has originated. By being able to identify the diluting agents, more intelligence can be provided to law enforcement agencies.
- Pharmaceutical Industry and Process Analytical Technologies: This software is applicable in a diverse range of analytical area from raw materials quality control (QC), formulation QC, tablet production, polymorph analysis, and characterisation of hazardous materials.
- Environment: Raman instruments, in conjunction with the analysis techniques developed for this work, can be used to identify unknown chemicals and other materials, and determine the concentrations of components.

Although much of this work has considered applications to Raman spectroscopy, the software package is equally applicable to other molecular spectroscopy techniques, so it has potential for impact in a wide range of application domains.

6 Conclusions and Observations

This paper has presented an innovative analytical chemistry software package that enables chemists to use machine learning analysis techniques in their work. As has been described, it has a range of features to ensure a good user experience and to minimise risk of user error.

A key advantage of this new software package is that it allows analytical chemists make use of powerful and novel machine learning techniques, which have been specifically designed to handle the characteristics of spectral data. The use of a step-by-step model-building wizard means that these techniques can be adopted without requiring expert knowledge of machine learning. Furthermore, as demonstrated in Section 4, its classification accuracy is superior to that of techniques conventionally used in this domain. Its model-based approach also has lower CPU and storage requirements than direct matching against a database, so that the approach could potentially be deployed on handheld portable instruments.

This paper will conclude with some broader observations drawn from the experience of developing this software package. Firstly, for a ML practitioner, tackling new domains and working with new datasets are always rewarding, as they motivate the development of new algorithms. A helpful starting point is to consider the characteristics of the data and identify any characteristics that may violate standard assumptions. Secondly, close collaboration with domain experts is invaluable, since they are best positioned to identify appropriate pre-processing techniques, as well as to perform baseline analyses using industry standard techniques. Their role is also important in interpreting results, sanity-checking conclusions and assessing the value of new work. Naturally, domain experts are essential when developing software requirements and designing user interfaces.

Related to this, it must be observed that the benefit of investing effort to develop a good user interface is that it aids end-users in evaluation and encourages adoption of the software. To encourage adoption, however, it is even more important to provide users with insight into the basis for decisions; in this work, it was determined in consultation with chemists that they would find it informative to see what parts of the spectrum most strongly contribute to decisions.

Finally, users are not in general prepared to sacrifice performance accuracy in order to gain insight/visualisation. However, in our experience, when the ML techniques are tailored to the characteristics of the application domain and are designed to enable specific forms of visualisation, this can have a positive effect on classification performance.

Further information and evaluations are available from www.AnalyzeIQ.com.

Acknowledgements

The authors acknowledge the contribution of Dr A. Ryder, School of Chemistry, National University of Ireland, Galway, who was involved in the research work that led to this publication, and who performed the SIMCA analysis for classification of Acetonitrile that is reported in Table 1(a).

References

1 Glossary of Terms Related to Chemical and Instrumental Analysis of Fire Debris. IAAI Forensic Science Committee, http://www.fire.org.uk/glossary.htm (Accessed Jan 2008).

2 Ferraro, J.R., Nakamoto, K. and Brown, C.W. (2003). Introductory Raman Spectroscopy. Academic Press, San Diego, second edition.

3 Savitzky, A. & Golay, M.J.E. (1964). “Smoothing and differentiation of data by simplified least squares procedures.” Analytical Chemistry, 36, 1627–1639.

4 Howley, T., Madden, M.G., O’Connel, M.L., Ryder, A.G. (2006). "The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data". Knowledge Based Systems, Vol. 19, Issue 5.

5 Hennessy, K., Madden, M.G., Conroy, J., Ryder, A.G. (2005). "An Improved Genetic Programming Technique for Identification of Solvents from Raman Spectra", Knowledge Based Systems, Vol. 18, Issue 4-5.

6 Howley, T. (2007). "Kernel Methods for Machine Learning with Applications to the Analysis of Reaman Spectra". PhD Thesis, National University of Ireland, Galway.

7 Hennessy, K. (2007). "Machine Learning Techniques for the Analysis of Raman Spectra". PhD Thesis, National University of Ireland, Galway.

8 Geladi, P. & Kowalski, B.R. (1986). Partial Least Squares: A Tutorial. Analytica Chemica Acta, 185, 1–17.

9 Wold, Svante, and Sjostrom, Michael (1977). SIMCA: A method for analyzing chemical data in terms of similarity and analogy, in Kowalski, B.R., ed., Chemometrics Theory and Application, American Chemical Society Symposium Series 52, Wash., D.C., American Chemical Society, p. 243-282.

10 Markey, M.K., Tourassi, G.D. & Floyd, C.E. (2003). Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. Proteomics, 3, 1678–1679.

11 Liu, H., Li, J. & Wong, L. (2002). A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Informatics, 13, 51–60.

12 Yang, H., Griffiths, P.R. & Tate, J.D. (2003). Comparison of partial least squares regression and multi-layer neural networks for quantification of non-linear systems and application to gas phase fourier transfrom infrared spectra. Analytica Chimica Acta, 489, 125–136.

13 Zou, T., Dou, Y., Mi, H., Ren, Y. & Ren, Y. (2006). Support vector regression for determination of component of compound oxytetracycline powder on near-infrared spectroscopy. Analytical Biochemistry, 355, 1–7.

14 Luinge, H.J., van der Maas, J.H. & Visser, T. (1995). Partial least squares regression as a multivariate tool for the interpretation of infrared spectra. Chemometrics and intelligent laboratory system, 28, 125–138.

15 Madden, M.G. and Ryder A.G. (2002). Machine learning methods for quantitative analysis of Raman Spectroscopy data. In Proceedings of SPIE, Vol. 4876, 1013-1019.