Title	Machine learning methods for quantitative analysis of Raman spectroscopy data
Author(s)	Madden, Michael G.; Ryder, Alan G.
Publication Date	2003-08-27
Publication Information	Michael G. Madden ; Alan G. Ryder; Machine learning methods for quantitative analysis of Raman spectroscopy data. Proc. SPIE 4876, Opto-Ireland 2002: Optics and Photonics Technologies and Applications, 1130 (March 17, 2003); doi:10.1117/12.464039
Publisher	Society of Photo-optical Instrumentation Engineers (SPIE)
Link to publisher's version	http:/dx.doi.org/10.1117/12.464039
Item record	http://hdl.handle.net/10379/6182
DOI	http://dx.doi.org/10.1117/12.464039

Machine Learning Methods for Quantitative Analysis of Raman Spectroscopy Data

Michael G. Madden*[a] and Alan G. Ryder[b]

[a] Department of Information Technology, NUI-Galway, Ireland.

[b] Department of Physics, NUI-Galway, Ireland.

ABSTRACT

The automated identification and quantification of illicit materials using Raman spectroscopy is of significant importance for law enforcement agencies. This paper explores the use of Machine Learning (ML) methods in comparison with standard statistical regression techniques for developing automated identification methods. In this work, the ML task is broken into two sub-tasks, data reduction and prediction. In well-conditioned data, the number of samples should be much larger than the number of attributes per sample, to limit the degrees of freedom in predictive models. In this spectroscopy data, the opposite is normally true. Predictive models based on such data have a high number of degrees of freedom, which increases the risk of models over-fitting to the sample data and having poor predictive power. In the work described here, an approach to data reduction based on Genetic Algorithms is described. For the prediction sub-task, the objective is to estimate the concentration of a component in a mixture, based on its Raman spectrum and the known concentrations of previously seen mixtures. Here, Neural Networks and k-Nearest Neighbours are used for prediction. Preliminary results are presented for the problem of estimating the concentration of cocaine in solid mixtures, and compared with previously published results in which statistical analysis of the same dataset was performed. Finally, this paper demonstrates how more accurate results may be achieved by using an ensemble of prediction techniques.

Keywords: Forensic science; Narcotics; Regression; Raman; Spectroscopy; Machine Learning; Ensemble; Genetic Algorithm; Neural Network.

1. INTRODUCTION

Raman spectroscopy is being used in forensic science research for the identification and analysis of narcotics, explosives, polymers, and other materials.[1] Raman spectra are unique and are based on the vibrational motions of molecules, which provides a chemical fingerprint suitable for identification and discrimination of a wide range of materials.[2] Furthermore, the development of fiber optic Raman probes will allow the implementation of portable devices for *in-situ* examination of suspect materials including narcotic.[3] Examples of illicit narcotics analysed by Raman spectroscopy in the laboratory include cocaine,[4] heroin,[5] and amphetamines in both solid and solution.[6, 7] In reality, however, the composition of seized drug samples can very enormously and it is unlikely that suspect materials will contain only one or two pure diluents. The vast range of possible diluents and impurities that may be present pose several problems for the use of Raman spectroscopy for the quantitative and qualitative analysis of illicit drugs. Difficulties include the presence of fluorescent materials, which obscure Raman peaks, overlap of diluent with narcotic Raman peaks, and variations in signal quality.

To help overcome these problems many investigators a have turned to advanced computational methods to improve the quantitative and qualitative capability of Raman spectroscopy. For quantitative measurements, the use of chemometrics (multivariate analysis) has been demonstrated in the prediction of fuel composition,[8] metabolites in urine,[9] and cocaine.[10, 11] In our previous work, we have employed traditional chemometric methods (Partial Least Squares) for the development of quantitative models for predicting cocaine concentration in solid mixtures.[10, 11] Unfortunately, as the mixtures become more

complex, the computational efficiency and the accuracy decreases. In an order to overcome these problems, this study examines the use of Machine Learning (ML) methods to develop more accurate quantitative methods. In addition, this paper demonstrates how higher accuracies may be achieved by using an *ensemble* method that combines the predictions from multiple methods.

2. EXPERIMENTAL

2.1 Apparatus and materials

The modular instrument used for near-IR Raman spectroscopy has been described previously.[10] All spectra were recorded at a set interval of 450-1100 cm^{-1} (510 data points) and a resolution of ~ 4 cm^{-1}. The exposure time was set at 30 seconds for all samples. Three Raman spectra at different surface locations were recorded for each sample to minimize the effect of sample morphology. Features in the spectra caused by cosmic rays incident on the detector were manually removed using the EasyPlot software package (ver. 3.00-7, Spiral Software+MIT). The three spectra were co-added, averaged, and smoothed over a five-point range. These spectra were then divided by a normalized white light spectrum to correct for detector response.

Anhydrous D-glucose (BDH), Cocaine hydrochloride and caffeine (Sigma-Aldrich) were reagent grade and were used as received. The sample set (see Table 1) covered a representative range of concentrations and mixtures with the sample mixtures (10-30 mg total weight) being made up by mixing known weights of drug and diluent, followed by grinding in an agate mortar and pestle to ensure sample homogeneity by thorough mixing of components. The mixtures were transferred to clean stainless steel hexagonal sample holders with an internal diameter of ~ 2 mm and tamped into place.

2.2 Software and Hardware

Chemometric analysis was performed using the Unscrambler (V7.5, CAMO ASA, Trondheim, Norway) multivariate analysis software package. Neural Network analyses were performed using the Stuttgart Neural Network Simulator[14] and Genetic Algorithm populations were bred using the GA Playground software[12]. Other software, including the implementation of the k-NN algorithm (Sec. 3.2), was developed for this project by the authors. Machine Learning analyses were carried out on a desktop PC and on NUI Galway's Origin high-performance multi-processor computer.

3. MACHINE LEARNING ANALYSES

3.1 Overview

In this study, the ML analyses have focused on predicting the concentration of cocaine in a sample containing a mixture of components, by examination of the sample's Raman spectrum. As mentioned in the Introduction, the analyses involved two phases: data reduction and prediction. Data reduction involves simplifying the data to improve the accuracy of prediction and reduce computational effort, as is discussed in detail in Sec. 3.3. Statistical and other transforms may be used for data reduction, but in this work feature selection was used, which is a simple form of data reduction whereby some of the input features are selected for use in prediction and the rest are ignored.

Prediction involves building a model of how Raman spectra relate to cocaine concentration, and then using this to predict the concentration of new samples. Details of the prediction methods are presented in Sec. 3.2. Although prediction and feature selection are discussed separately below, the two sub-tasks were actually interlinked, as the feature selection was

optimised to improve predictive performance, as discussed in Sec. 3.3. All analyses were based on the 36 samples used were those listed in Table 1, with 510 data points per sample.

% Cocaine	% Caffeine	% Glucose	% Cocaine	% Caffeine	% Glucose
54.24	23.14	22.62	29.47	9.43	61.1
80.58	8.86	10.56	33.46	20.89	45.65
70.35	17.24	12.41	28.06	8.52	63.42
71.10	9.15	19.75	13.39	15.45	71.16
56.26	10.62	33.12	20.0	10.34	69.66
61.92	19.23	18.85	29.96	29.32	40.72
74.25	19.37	6.38	25.60	18.40	56.0
61.48	29.11	9.41	21.98	20.32	57.70
49.96	10.74	39.30	22.32	28.46	49.22
50.57	19.39	30.04	11.75	30.53	57.72
48.34	41.65	10.01	31.11	38.93	29.96
42.88	10.02	47.10	19.46	39.80	40.74
40.30	19.36	37.34	29.90	50.33	19.77
39.33	30.46	30.21	21.19	29.0	49.81
40.11	39.74	20.15	9.92	38.64	51.44
40.74	49.34	9.92	9.84	50.49	39.67
10.85	10.95	78.20	100.00	0.00	0.00
0.00	100.00	0.00	0.00	0.00	100.00

Table 1: Chemical composition of samples used in the study.

3.2   Prediction Methods

Two regression methods have been evaluated in this study:

1   k-Nearest Neighbours
2   Feed-Forward Neural Networks

Neural Networks are a popular ML technique for non-linear mapping of inputs to outputs. Based on highly simplified models of the operation of the brain, each neuron in the network has a set of inputs that are weighted and summed, and a non-linear threshold function is applied to the result to produce an output value. In a feed-forward network, the neurons are arranged in layers, with each layer's outputs providing the inputs for the following layer. In this study, the inputs to the first layer are the Raman spectral data points and the final output is an estimate of cocaine concentration. The number of hidden (i.e. intermediate) layers and number of neurons per hidden layer have been varied in experiments. Training a neural network involves adjusting the weights on each neuron's inputs until the outputs are as close as possible to their expected values. In this work, the Resilient Backpropogation (RProp) algorithm[13] has been used for training the network, as preliminary experiments indicated that it achieved relatively fast convergence and was not overly sensitive to parameter settings. The Stuttgart Neural Network Simulator[14] was used for this work.

The k-Nearest Neighbours algorithm, as described by Mitchell[15], works by comparing a sample to be classified with known samples, identifying the k samples that are nearest to the new one, and estimating the cocaine concentration of the new sample by averaging them. In this work, cosine and Euclidian distance measures were used, and the averages were weighted by the inverse of distance. In preliminary experiments, various values of k (number of neighbours) were tried, and $k=3$ was selected for the main experiments.

These two methods were selected because they contrast strongly with each other: kNN is an instance-based learner, is fast, is deterministic and is particularly sensitive to irrelevant and correlated attributes. Neural networks are model-based, are slower, may converge to local maxima, and are less sensitive to irrelevant/correlated attributes although they also benefit from feature selection. However, neural networks are able to represent complex non-linear relationships in data, giving good performance on a wide range of prediction tasks.

3.3 Feature Selection Methods

In a 'typical' ML application with well-conditioned data, the number of cases is much larger than the number of attributes per case. In this study, however, there are 36 cases with 510 attributes per case. For many ML techniques, this causes problems. For example, a neural network with 510 inputs, 10 neurons in a single hidden layer, and 1 output layer would have over 5000 degrees of freedom, which with just 36 cases will most often result in convergence to a poor local minimum. Accordingly, feature selection has been used to reduce the number of attributes per case. The following approaches to feature selection have been assessed:

1   Local Maxima
2   Optimal Search using a Genetic Algorithm

In each case, the impact of the feature selection methods on the performance of the ML algorithms has been assessed, as discussed in Section 4 below. For purposes of comparison, the ML algorithms have also been applied without any feature selection.

The Local Maxima approach involved selecting all points that appeared as peaks (defined as having at least two points on each side with lower values) on the spectrum of the 100% cocaine sample, but excluding minor peaks below the average value. This resulted in the selection of 17 points.

The Optimal Search approach was more sophisticated. The objective here was to find a set of attributes that was as small as possible, with as high accuracy as possible measured relative to a learning algorithm. The optimal search was performed using a Genetic Algorithm (GA), which is a technique based on the paradigm of biological evolution[16]. The essential idea is that a population of potential solutions to a problem is created, and the fitness of each individual is assessed by calculating the how appropriate it is for the problem. The fittest individuals are carried forward to the next generation, and the new generation's population is augmented by crossover (producing new individuals that combine features of existing individuals) and mutation (random changes to an individual). The initial population is created at random, and bred for a large number of generations until it reaches steady state. This is termed a *wrapper* approach to feature selection, as the GA is wrapped around the learning algorithm.

In the first set of GA experiments, the learning algorithm chosen as the target for the accuracy measure was k-Nearest Neighbours (kNN), as it runs fast and is known to be sensitive to cross-correlated attributes. Each individual consisted of a string of 510 binary digits, representing a configuration with some attributes selected and others not, depending on whether

the corresponding digit in the bit-string was 1 or 0. The fitness of each individual was assessed by measuring the performance of kNN with the indicated set of attributes selected, and giving a penalty based on the number of attribute selected, to encourage the development of individuals that selected the best set of attributes and to favour the selection of as few attributes as possible. Populations were bred using the GA Playground software[12], a general-purpose Genetic Algorithm toolkit implemented in Java, which the authors interfaced to their own kNN software. Repeated runs were carried out with various population sizes, crossover rates, mutation rates, and penalties.

In the second set of GA experiments, a neural network was the target learning algorithm. Because of the relative slowness of neural network training, the fitness function was based on the sum of the squared error in training, since full cross-validation for the 36 samples would have taken 36 times longer. Likewise, the 510 attributes per sample were reduced to a representative set of less than 10%, comprising local maxima, local minima and some intermediate points. In addition, because comparisons across networks with different sets of input attributes were being performed, a fixed number of hidden nodes was used for all networks and all networks were trained for the name number of epochs. Having a penalty term to encourage solutions with fewer attributes was not found to be beneficial in the neural network case.

3.4 Ensembles

Ensemble methods are currently an active area of research within machine learning. An ensemble is a learning algorithm that uses a set of predictors and combines their predictions through a voting scheme to arrive at a decision. Ensembles produce more accurate results than their individual members provided that they are accurate (i.e. their performance is better than random) and diverse (i.e. different members make different errors on new data)[17]. A good overview is given by Dietterich [18]. Ensembles are usually constructed by producing variations on a single classifier, for example by training several neural networks using different feature subsets for each. In this study, however, completely different prediction methods are used for constructing the members of the ensemble.

4. RESULTS & DISCUSSION

4.1 Results of Feature Selection

Figure 2 illustrates the effect of the feature selection features on the Raman spectrum of a pure cocaine sample. Without any feature selection, the full spectrum of 510 points is used. If the maxima are used, this reduces the data set to 17 points per sample, corresponding to the local maxima of the pure cocaine sample (even though these may not be maxima in other samples). The plot also shows the results of the GA optimisation when the kNN algorithm and the NN algorithm are each used as targets for the optimisation procedure. Lines connect the points to make them easier to identify. These two results provide an interesting contrast with each other: for kNN, the optimal solution is a set of just four points, three of which are clustered around the largest peak at 996 cm^{-1}. For the neural network, on the other hand, the optimal solution does not use points on that peak at all, instead being based on other local maxima and local minima of the spectrum.
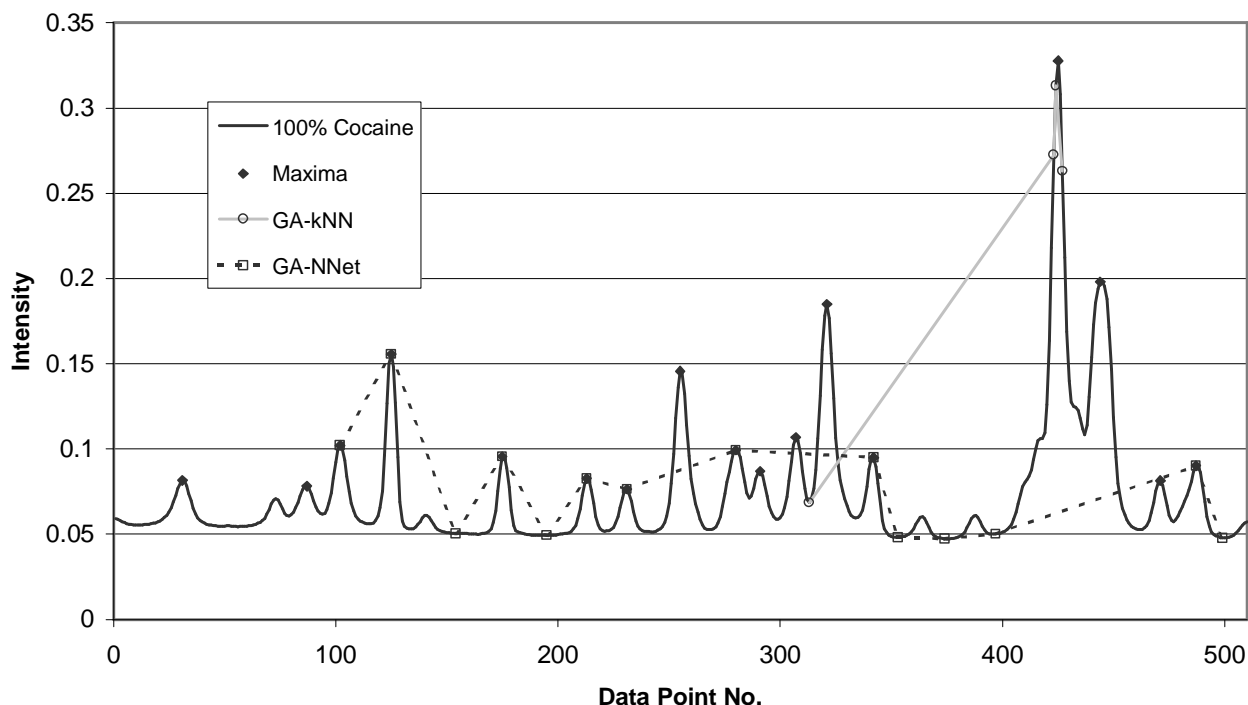
Figure 1: Results of attribute selection procedures.

4.2 Results of Analyses

Table 2 summarises the performance of the kNN and neural network methods, when combined with the different attribute selection strategies, and also includes the performance of the Partial Least Squares technique as applied to this dataset previously by the authors[11]. For each combination of prediction method and attribute selection technique, the table lists the number of attributes selected, the root mean squared error of prediction (RMSEP) and the absolute maximum error in prediction (MaxErrP). In all cases, the prediction procedure was standard leave-one-out cross-validation: for each sample in turn, that sample was removed from the set 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. For the NN and PLS methods, calibration statistics (RMSEC and MaxErrC) are also listed: these are found from using all samples together to build a model, and then using the model to predict the concentration of each sample in turn. In the NN case, this is essentially the training error of the method. For the kNN algorithm, calibration statistics are not meaningful – for the 1-neighbour case, if the sample to be predicted is available, the kNN algorithm will have zero error. As is typical of neural networks, it can be seen in rows 5-7 that calibration statistics are much better than prediction statistics, because of the way the network can represent nonlinear data accurately.

Examining Table 2, it is seen that the kNN and NN methods both benefit from careful selection of attributes, as reflected in reduced values for RMSEP and MaxErrP when GA-based selection is used compared to when no selection or simple maxima-based selection is used. In the best case, the NN outperforms the PLS method both in terms of root mean squared error and maximum error. However, the performance of the kNN method is not quite as good. Overall, the best-case performance of all methods are quite similar; it appears that all methods are fundamentally limited by a lack of information

stemming from having such a limited dataset. Figure 2 shows, for each sample, its measured concentration of cocaine plotted against its value as predicted by the best-case neural network with GA feature selection. While there is some scatter around the dashed centreline, the plot shows a strong correlation between predicted and actual values, as indicated by the low RMSEP and MaxErrP values reported in Table 2.

Prediction Method	Attribute Selection	No. of Data Points	RMSEC %	MaxErrC %	RMSEP %	MaxErrP %
k-Nearest Neighbours	None	510	—	—	8.911	24.467
	Maxima	17	—	—	7.674	24.690
	GA	4	—	—	5.837	24.690
Neural Network	None	510	1.324	3.942	7.727	25.038
	Maxima	17	3.922	12.003	6.108	19.351
	GA	15	2.583	5.642	5.206	11.628
Partial Least Squares		510	4.655	13.522	5.225	17.286

Table 2: Results of analyses using various prediction methods and attribute selection schemes. Note that correlation is not meaningful for the kNN algorithm. Figures for the PLS method are included for comparison.
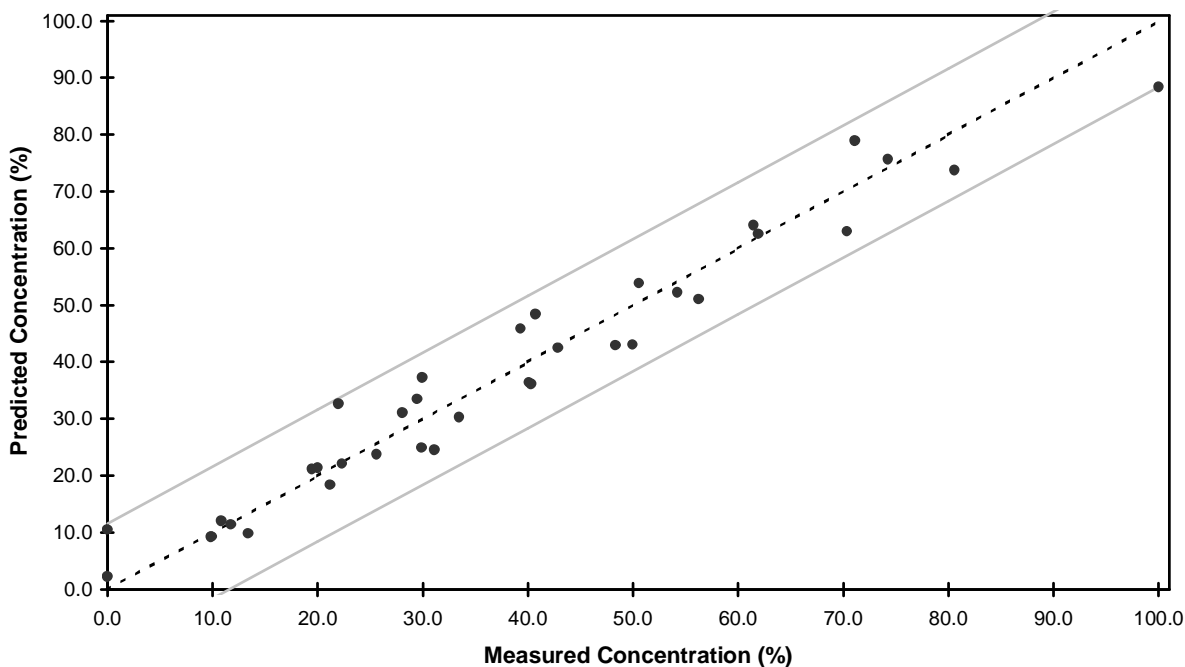

Figure 2: Predicted versus measured concentration of cocaine using GA with Neural Network.

4.3 Combining Predictors in an Ensemble

As was explained in Sec. 3.4, an ensemble is a committee of predictors, where the opinion of each member of the ensemble is sought when performing predictions. In this case, it is proposed to use the kNN, NN and PLS algorithms as members of the ensemble, and to simply average the predictions from each to arrive at a final prediction. (It would be possible to use a weighted average if independent experiments had established the relative accuracy of each, but lack of data prevents that from being done in this case.)

As was mentioned previously, for the accuracy of an ensemble to be greater on average than that of any of its members, it is necessary and sufficient that each member is accurate and that the predictions produced by members are diverse[17]. Figure 3 shows superimposed plots of predicted versus measured concentration of cocaine, similar to the plot of Figure 2. In Figure 3, results are shown for the partial least squares method (hollow square) and kNN algorithm (hollow triangle) as well repeating the neural network result (hollow circle). It is clear from the plot that the three methods are better than random and that their predictions are diverse: there is at least 15% difference in some predictions. Hence, it is reasonable to construct an ensemble from the three algorithms. Accordingly, Figure 3 also shows the performance of the ensemble that results from averaging predictions from the three methods (solid square). Corresponding statistics are presented in Table 3. As Figure 3 shows, performance on predicting concentration of the 100% cocaine is not good, particularly for the kNN method, as this requires significant extrapolation. Since predictions at such high concentrations are of less practical significance than at low concentrations, Table 3 also lists statistics when the 100% cocaine sample is excluded, denoted RMSEP* and MaxErrP*.
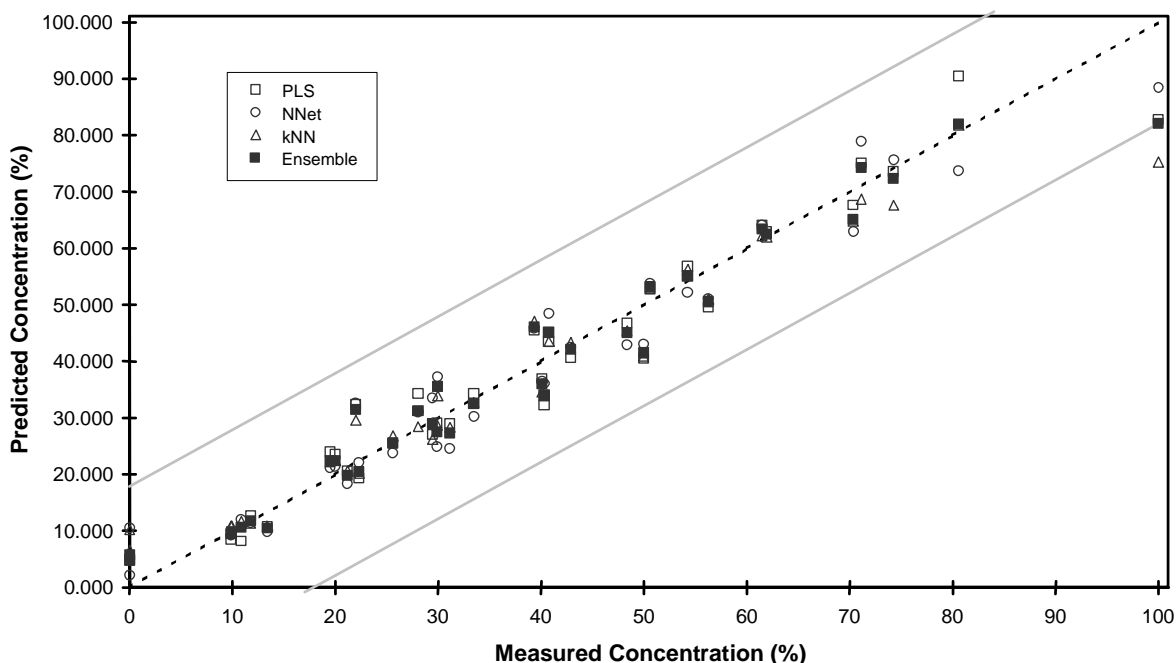

Figure 3: Predicted versus measured concentration of cocaine using each of the three algorithms individually (hollow symbols) and when all three are combined in an ensemble (solid square).

Method	RMSEP %	MaxErrP %	RMSEP* %	MaxErrP* %
Partial Least Squares	5.225	17.286	4.421	10.446
Neural Network	5.206	11.628	4.901	10.634
k-Nearest Neighbours	5.837	24.690	4.199	10.203
Ensemble	4.857	17.868	3.891	9.549

Table 3: Performance of individual prediction methods and of an ensemble that averages predictions
of the three methods.

As Table 3 shows, overall performance of the ensemble is better than that of any of the individual members, as shown by the RMSEP values. Naturally, the MaxErrP of the ensemble cannot be as good as that of the best individual member, since it averages values from all members. Nonetheless, it is seen that ensemble has a significant ameliorating effect on an individual bad value such as the MaxErrP of the kNN algorithm.

5. CONCLUSIONS

This paper has investigated the application of Machine Learning techniques for prediction and data reduction to the task of predicting the concentration of cocaine in solid mixtures, using on Raman spectroscopy. The study has shown that good results are achievable by Neural Networks and k-Nearest Neighbours, provided that data reduction is used to improve the dimensionality of the data. In this study, the data reduction took the form of selecting specific wavelengths and discarding all others. This selection process was optimised by using a Genetic Algorithm, which in combination with the Neural Network method produced greater prediction accuracy than the Partial Least Squares method. The resulting predictors are simple, basing their predictions on a small number (less than 20) of data points. Accordingly, after the models have been built the classifiers operate rapidly, and they could be implemented on hardware for portable probes because they just a small number of simple mathematical operations (addition and multiplication).

This paper has also demonstrated how an ensemble of different predictors can be used to produce predictions that are better than any one of the individual predictors, though naturally this comes at the cost of increased computational effort in constructing multiple predictors.

It appears that all prediction methods considered in this study are fundamentally limited by the lack of information inherent in a limited test dataset of just 36 samples of various concentrations of cocaine, glucose and caffeine. To achieve substantially improved prediction accuracies, a much more comprehensive database of samples and their Raman spectra would be required. In addition, further samples would be necessary for independent verification of the performance of the prediction methods. Fortunately, in a real-world application, the number of samples available would be continually expanding as new samples would be analysed on an ongoing basis from law-enforcement seizures and similar sources.

6. ACKNOWLEDGEMENTS

7. REFERENCES

1   A.H. Kuptsov. "Applications of Fourier transform Raman spectroscopy in forensic science." *J. Forensic Sci.* **39**, pp. 305-318, 1994.

2   B.J. Bulkin. The Raman effect: an introduction. In: *Analytical Raman Spectroscopy. Chemical Analysis volume 114.* B.J. Bulkin BJ, and J.G.Grasselli, editors, pp. 1-19, New York: John Wiley and Sons, Inc, 1991.

3   S.M. Angel, J.C. Carter, D.N. Stratis, B.J. Marquardt, and W.E. Brewer. "Some new uses for filtered fiber-optic Raman probes: In situ drug identification and in situ and remote Raman imaging." *J. Raman Spectrosc.* **30**, 795-805, 1999.

4   J.C. Carter, W.E. Brewer, and S.M. Angel. "Raman spectroscopy for the in situ identification of cocaine and selected adulterants," *Appl. Spectrosc.* **54**, pp. 1876-1881, 2000.

5   J. Akhavan, and C.M Hodges. "The use of Fourier transform Raman spectroscopy in the forensic identification of illicit drugs and explosives." *Spectrochim. Acta*. **46A**, pp. 303-307, 1990.

6   S.E.J. Bell, D.T. Burns, A.C. Dennis, and J.S. Speers. "Rapid analysis of ecstasy and related phenethylamines in seized tablets by Raman spectroscopy." *Analyst* **125**, pp. 541-544, 2000.

7   H. Tsuchihashi, M. Katagi, M. Nishikawa, M. Tatsuno, H. Nishioka, A. Nara, et al. "Determination of methamphetamine and its related compounds using Fourier transform Raman spectroscopy." *Appl. Spectrosc.* **51**, pp. 1796-1799, 1997.

8   J.B. Cooper, K.L. Wise, W.T. Welch, M.B. Sumner, B.K. Wilt, and R.R. Bledsoe. "Comparison of near-IR, Raman and mid-IR spectroscopies for the determination of btex in petroleum fuels," *Appl. Spectrosc.* **51**, pp. 1613-1620, 1997.

9   X. Dou, Y. Yamaguchi, H. Yamamoto, S. Doi, and Y. Ozaki. "Quantitative analysis of metabolites in urine using a highly precise, compact near-infrared Raman spectrometer," *Vib. Spectrosc.* **13**, pp. 83-89, 1996.

10 A.G. Ryder, G.M. O'Connor, and T.J. Glynn. "Identifications and Quantitative Measurement of Narcotics in Solid Mixtures Using Near-IR Raman Spectroscopy and Multivariate Analysis." *J. Forensic Sci.* **44**, pp.1013-1019, 1999.

11 A.G. Ryder, G.M. O'Connor, and T.J. Glynn. "Quantitative analysis of cocaine in solid mixtures using Raman spectroscopy and chemometric methods," *J. Raman Spectrosc.* **31**, pp. 221-227, 2000.

12 A. Dolan, "The GA Playground", http://www.aridolan.com/ga/gaa/gaa.html, 1998.

13 M. Riedmiller, and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm", *Proc. International Conference on Neural Networks*, 1993.

14 Stuttgart Neural Network Simulator, © University of Stuttgart and University of Tübingen, http://www-ra.informatik.uni-tuebingen.de/SNNS/, 1990-2002.

15 T.M. Mitchell, *Machine Learning*, McGraw Hill, 1997.

16 M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, 1996.

17 L. Hanson and P. Salamon, "Neural Network Ensembles", *IEEE Trans. Pattern Analysis and Machine Intelligence,* **2**, pp. 993-1001.

18 T.G. Dietterich, "Ensemble Methods in Machine Learning", *Lecture Notes in Computer Science*, **1857,** pp. 1-15, 2000.