ELSEVIER

Chemometrics and Intelligent Laboratory Systems 60 (2002) 161–171

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Innovative genetic algorithms for chemoinformatics

B.K. Lavine *, C.E. Davidson, A.J. Moores

Department of Chemistry, Clarkson University, Box 5810, Potsdam, NY 13699-5810, USA

Abstract

In this paper, we report on the development of a genetic algorithm (GA) for pattern recognition analysis of multivariate chemical data. The GA identifies feature subsets that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by the selected features is about differences between classes in the data set. The principal component (PC) plot function as embedded information filter. Sets of features are selected based on their principal component plots, with a good principal component plot generated by features whose variance or information is primarily about differences between classes in the data set. This limits the GA to search for these types of feature subsets, significantly reducing the size of the search space. In addition, the pattern recognition GA focuses on those classes and/or samples that are difficult to classify by boosting their weights over successive generation using a perceptron to learn the class and sample weights. Samples that consistently classify correctly are not as heavily weighted in the analysis as samples that are difficult to classify. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a ''smart'' one-pass procedure for feature selection. The efficacy and efficiency of the pattern recognition GA is demonstrated via problems from chemical communication and environmental analysis.

Keywords: Genetic algorithms; Chemoinformatics; Perceptron

1. Introduction

Many relationships in chemical data cannot be expressed in quantitative terms. These relationships are better expressed in terms of similarity and dissimilarity among groups of multivariate data. The task confronting the chemist when investigating these types of relationships is two-fold: (1) Can the data be divided into categories for the prediction of some property?, and (2) Can the features necessary for differentiating the classes in the data be identified? Pattern recognition techniques are well suited for tackling both these tasks since they can display variability between a large number of samples and show the major clustering trends in large data sets [1–3].

Pattern recognition methods were originally developed to solve the class membership problem. In a typical pattern recognition study, samples are classified according to a specific property using measurements indirectly related to that property. An empirical relationship or classification rule is developed from a set of samples for which the property of interest and the measurements are known. The classification rule is then used to predict the property in samples that are not part of the original training set. The property in question may be the type of fuel responsible for a spill, and the measurements are the areas of selected gas chromatographic peaks. Classification is synonymous

* Corresponding author. Tel.: +1-315-268-2394; fax: +1-315-268-6610.

E-mail address: bklab@clarkson.edu (B.K. Lavine).

0169-7439/02/$ - see front matter
PII: S0169-7439(01)00193-9

with pattern recognition and scientist have turned to it to analyze the large data sets generated in studies that involve environmental or biological samples.

Problems can arise when applying pattern recognition techniques to chemical data. Classification success rates vary with the pattern recognition method employed. Unfavorable classification results are obtained for the prediction set despite a linearly separable training set. Automation of these techniques for the solution of a general class of pattern recognition problems is often difficult.

The basic premise underlying the research described in this paper is that all classification methods will work well when the problem is simple. By identifying the appropriate features, a "hard" problem is reduced to "simple" one. Thus, our goal is feature selection, in order to increase the signal to noise ratio of the data by discarding measurements on components that are not characteristic of the source profile of the classes in the data set. To ensure identification of all relevant features, it is best that a multivariate approach to feature selection be employed. The approach should also take into account the existence of redundancies in the data.

In this paper, we report on the development of a genetic algorithm (GA) for pattern recognition analysis of multivariate chemical data [4–7]. The GA identifies a set of features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Since principal components maximize variance, the bulk of the information encoded by the selected features is about differences between the classes in the data set. The principal component plot used by the fitness function acts as an embedded information filter. Sets of features are selected based on their principal component plots, with a good principal component plot generated by features whose variance information is primarily about differences between the classes. This limits the search to these types of feature subsets, thereby significantly reduces the size of the search space. In addition, the GA can focus on those classes and/or samples that are difficult to classify by boosting their weights over successive generations using a perceptron to learn the class and sample weights. Samples that consistently classify correctly are not as heavily weighted in the analysis as samples that are difficult to classify. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection. The efficacy and efficiency of the pattern recognition GA is demonstrated via problems from chemical communication and environmental analysis.

2. Pattern recognition GA

A block diagram of the pattern recognition GA is shown in Fig. 1. The GA builds a population of binary strings of fixed length, each of which represents a potential solution to the pattern recognition problem. For a feature to be included in the subset, it is necessary for the corresponding bit in the string to be set at 1. If the bit is set to 0, the corresponding feature is not included.

During each generation, the strings are decoded yielding the actual feature subset that is sent to the fitness function for evaluation. Each string is assigned a value by the fitness function, which is a measure of the degree of separation between the classes in a principal component map of the data defined by the extracted feature subset. The fitness (i.e., the quality

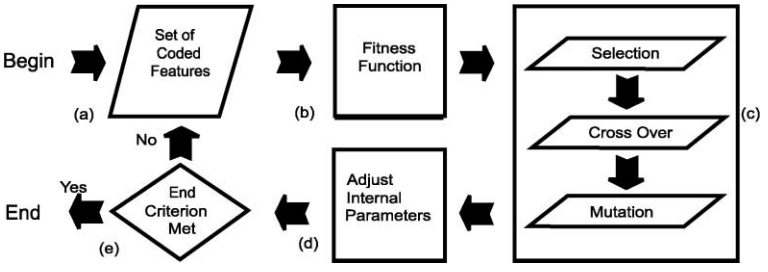

Fig. 1. A block diagram of the pattern recognition GA.

of the proposed feature subset for the pattern recognition problem) is used to select potential solutions for recombination, which produces a new population of strings. The power of the GA arises from recombination [8,9], which causes a structured yet randomized exchange of information between solutions, with the expectation that good solutions can generate even better ones. In addition, some of the binary strings may undergo mutation, where one of the bits is randomly changed.

The aforementioned process (evaluation, selection, crossover, reproduction, and adjustment of internal parameters) is repeated until a specified number of generations or a feasible solution has been found. The pattern recognition GA for chemoinformatics differs from conventional genetic algorithms in the types of operators that it employs. The operators unique to the pattern recognition GA are described below.

2.1. Evaluation

The pattern recognition GA uses machine emulation of human pattern recognition to score the principal component plots. To facilitate the tracking and scoring of principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see Eqs. (1) and (2)). The class weights sum to 100; the sample weights for samples constituting a particular class sum to a value equal to the corresponding class weight.

$$\mathrm{CW}(c) = 100\frac{\mathrm{CW}(c)}{\sum_{c}\mathrm{CW}(c)}, \tag{1}$$

$$\mathrm{SW}_c(s) = \mathrm{CW}(c)\frac{\mathrm{SW}_c(s)}{\sum_{s\in c}\mathrm{SW}_c(s)}. \tag{2}$$

Each principal component plot generated for each chromosome after the subset of features in the chromosome has been extracted is scored using the K-nearest neighbor (K-NN) classification algorithm [10]. For a given data point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to largest. A poll is then taken of the point's k-nearest neighbors. For the most rigorous classification, k equals the number of samples in the class to which the point belongs. The number of k-nearest neighbors with the same

$$F(d) = \sum_{c}\sum_{s\in c}\frac{1}{K_c}\times \mathrm{SHC}(s)\times \mathrm{SW}(s), \tag{3}$$

class label as the sample point in question, the so-called sample hit count (SHC), is computed ($0 \le \mathrm{SHC}(s) \le K_c$). It then becomes a simple matter to score each principal component plot (see Eq. (3)).

To better understand the scoring of the principal component plots, consider a data set with two classes, which have been assigned equal weights. Class 1 has 20 samples, and class 2 has 50 samples. At generation 0, all samples in a given class will have the same weight. Thus, each sample in class 1 has a sample weight of 2.5, whereas each sample in class 2 has a weight of 1. Suppose a sample from class 1 has as its 20 nearest neighbors 14 class one samples. Hence, $\mathrm{SHC}/K = 0.7$, and $(\mathrm{SHC}/K)\times \mathrm{SW} = 0.7\times 2.5$, which equals 1.75. By using $(\mathrm{SHC}/K_c)\times \mathrm{SW}$ for each sample, the principal component plot can be scored.

2.2. Adjusting internal parameters

The GA is able to focus on samples and classes that are difficult to classify by boosting their weights over successive generations (see Fig. 2). In order to boost the weights, it is necessary to first compute the sample hit rate, SHR(s), which is the mean value of SHC/K_c over all feature subsets in a particular generation. SHR(s), which is a measure of the difficulty of classifying a particular sample. If a sample is difficult to classify, it has a low sample hit rate since it has a low SHC/K_c value in most feature subsets of the population. If a sample is easy to classify, it has a high sample hit rate since it has a high SHC/K_c value in most feature subsets of the population:

$$\mathrm{SHR}(s) = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC}_i(s)}{K}, \tag{4}$$

$$\mathrm{CHR}_g(c) = \mathrm{AVG}(\mathrm{SHR}_g(s) : \forall_{s\in c}). \tag{5}$$

Next, the class hit rate (see Eq. (5)), which is the average sample hit rate for all of the samples in a class,

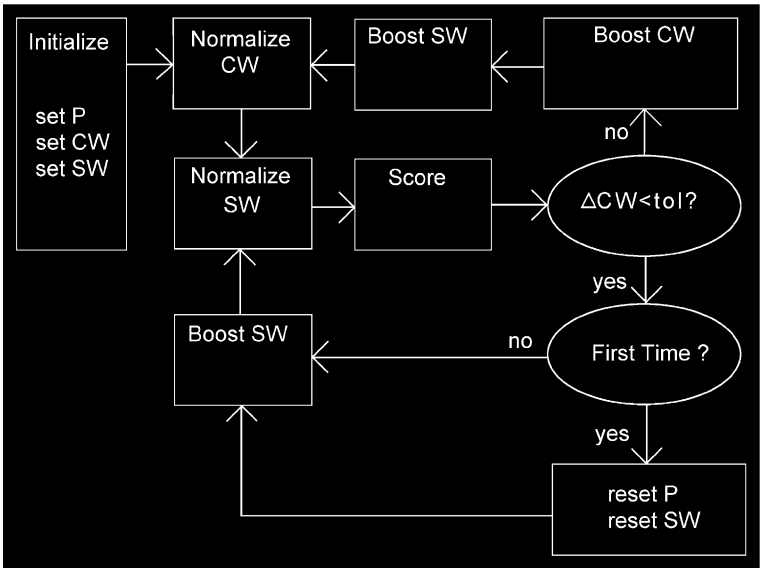

Fig. 2. Block diagram of the boosting algorithm used to adjust the weights of difficult classes and/or samples.

is computed. The class and sample weights are then adjusted using a perceptron (see Eqs. (6) and (7)). Classes with a low class hit rate and samples with a low sample hit rate are weighted more heavily than classes or samples that score well. The user must set the momentum, *P*. The value of *P* should be high enough to facilitate learning while ensuring that a particular sample or class does not dominate the calculation, which would result in other samples and/or classes not contributing to the fitness function. After a certain number of generations, the class weights do not change. Eq. (6) is then turned off and the GA focuses exclusively on the troublesome samples via Eq. (7). During each generation, class and sample weights are updated (i.e., boosted) using the class and sample hit rates from the previous generation ($g+1$ is the current generation, whereas g is the previous generation.) Boosting of sample and class weights is crucial because it modifies the fitness landscape, as the population evolves, potentially mitigating the problem of convergence to a local optimum.

$$\mathrm{CW}_{g+1}(s) = \mathrm{CW}_g(s) + P(1 - \mathrm{CHR}_g(s)), \quad (6)$$

$$\mathrm{SW}_{g+1}(s) = \mathrm{SW}_g(s) + P(1 - \mathrm{SHR}_g(s)). \quad (7)$$

2.3. Reproduction

Selection, crossover, and mutation operators are applied to the chromosomes to develop new and potentially better solutions. The selection operator used by the pattern recognition GA is implemented by ordering the population of strings, i.e., the potential solutions, from best to worst by their fitness while simultaneously generating a copy of the same population and randomizing the order of the strings in this copy with respect to their fitness. A fraction of the population is then selected as per the selection pressure, which is usually set at 0.5. The top half of the ordered population is mated with strings from the top half of the random population, guaranteeing the best 50% are selected for reproduction, while every string in the randomized copy has a uniform chance of being selected due to the randomized selection criterion imposed on the strings from this population.

For each pair of strings selected for mating, two new strings are generated using a variation of three-point crossover (see Fig. 3). As in the case of simple three-point crossover, the length of each new string or solution is the same as the dimensionality of the data. However, the crossover operator used by the pattern recognition GA is not compelled to preserve order

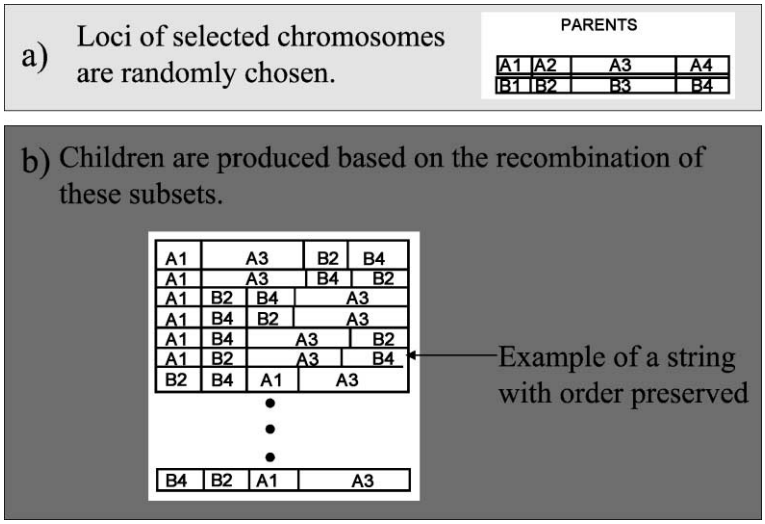

Fig. 3. Three-point crossover with a reordering algorithm embedded in it. Instead of swapping alleles and simultaneously preserving their position, four chromosome fragments are distributed and recombined at random with four factorial unique possibilities.

among exchanged string fragments. This safeguards the loss of information or features in the population. Furthermore, this variation of three-point crossover may be useful in searching for good string arrangements. If the current population has bad ordering, where features with a high synergism are spaced apart at great distances, simple crossover would probably destroy these important allele packets. On the other hand, there is a chance to obtain good allele ordering, by using a crossover operator with a reordering algorithm embedded in it.

The resulting population of strings, both parents and children, are sorted by their fitness and the top ϕ strings are retained for the next generation. The new population is expected to perform better on average than its predecessor because the selection criterion used for reproduction exhibits bias for the higher-ranking strings. However, the aforementioned reproduction operators also assure a significant degree of diversity in the population, since the crossover points and reordering of exchanged string fragments of each chromosome pair is selected at random.

3. Chemical communication

The first data set used to evaluate the efficiency and efficacy of the pattern recognition GA consisted of gas chromatograms of the post-pharygeneal gland (PPG) hydrocarbons extracts of the ant *Cataglyphis niger*. Previous workers [11–13] have shown that both cuticular and PPG hydrocarbons of ants are important in nestmate recognition, the process by which ants recognize both colony and social caste of conspecifics. Because the queen plays a central role in the ant colony, it is logical to assume that she influences the nestmate recognition cues used by individual ants.

To assess this hypothesis, a subset of ants from a laboratory colony were isolated and furnished with a new queen, which also came from the same laboratory colony. (The new queen was a reserve queen in the laboratory colony.) The hydrocarbon profiles of the ants in this sub-colony were monitored over time: 0, 1, 2, and 3 months. The entomologists who performed this experiment wanted to answer the following question: Do the PPG hydrocarbons of the ants in the sub-colony change systematically over time to reflect the profile of their new queen?

The following experimental protocol was used to generate the hydrocarbon profile data. Random samples of 10 ants were sacrificed at specific time intervals (0, 1, 2, and 3 months) and their PPG hydrocarbons were quantified by gas chromatography using eicosane (C_{20}) as the interval standard. To assure equal treatment of the data, 23 peaks out of a total of 72 were chosen for pattern recognition analysis.

The 23 peaks selected could be accurately and reliably quantified by GC/MS. For pattern recognition analysis, each gas chromatogram was represented by a data vector, $\boldsymbol{x} = (x_1, x_2, x_3 \ldots x_j \ldots x_{23})$, where x_j is the area of the jth peak normalized using the total integrated peak area so each peak was expressed as a fraction of the total. The data were auto-scaled to ensure that each peak had equal weight in the analysis.

The first step in the study was to apply principal component analysis [14] to the data. Principal component analysis is the most widely used multivariate analysis technique in science and engineering. It is a method for transforming the original measurement variables into new, uncorrelated variables called principal components. Each principal component is a linear combination of the original measurement variables. Using this method is analogous to finding a new coordinate system better at conveying information present in the data than axes defined by the original measurement variables. This new coordinate system is linked to variation in the data. The basis vectors of this new coordinate system are the principal components. Often, only the two or three largest principal components are necessary to explain all of the information present in a data set if the data contains a large number of interrelated measurement variables. Using principal component analysis, dimensionality reduction, classification of samples, and identification of outliers in high dimensional data is possible.

Fig. 4 shows a plot of the two largest principal components of the 40 *C. niger* ant samples. Each ant sample is represented as a point on the PC map. 1 represents 0 month ants, 2 represents 1 month ants, 3 represents 2 month ants, and 4 represents 3 month ants. When the gas chromatogram of the queen was projected onto the principal component (PC) map defined by the 23 gas chromatographic peaks and the 40 ant samples, it was not evident whether any of the groups possessed a hydrocarbon profile similar to the queen.

A GA for pattern recognition analysis was used to uncover features characteristic of the gas chromatographic profile of each group. The GA identified

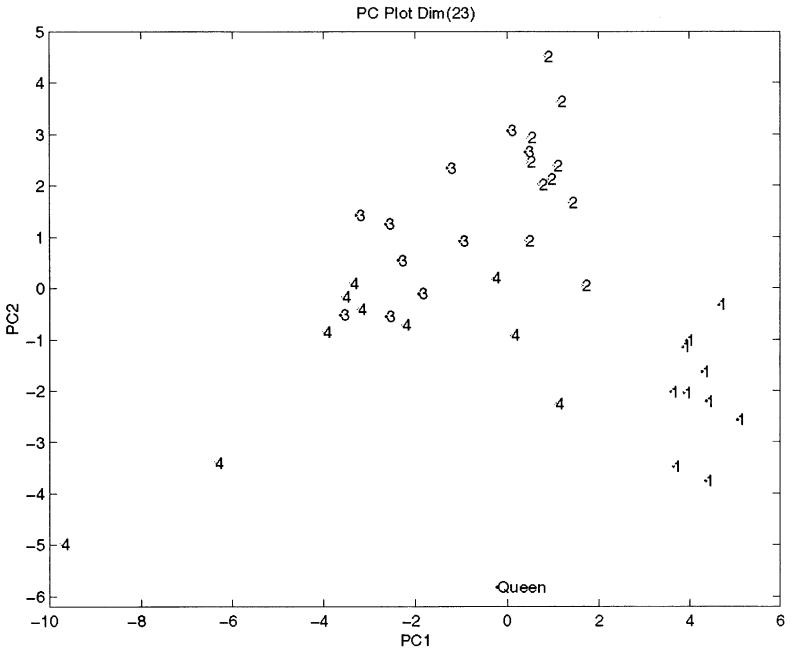

Fig. 4. A plot of the two largest principal components of the 40 *C. niger* ant samples developed from the 23 gas chromatographic peaks. Each ant sample is represented as a point in the principal component map (1 = 0 month ants, 2 = 1 month ants, 3 = 2 month ants, and 4 = 3 month ants). The queen has been projected onto this PC map.

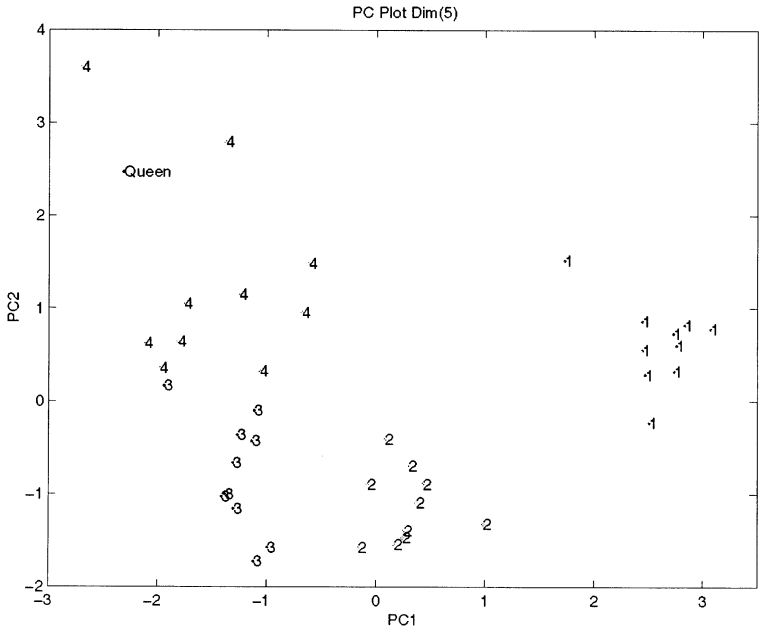

Fig. 5. A plot of the two largest principal components of the 40 *C. niger* ant samples developed from the five gas chromatographic peaks identified by the pattern recognition GA. Each ant sample is represented by a point in the principal component plot (1 = 0 month ants, 2 = 1 month ants, 3 = 2 month ants, and 4 = 3 month ants). The queen has been projected onto this PC map.

features by sampling key feature subsets, scoring their PC plots, and tracking those classes and/or samples that were difficult to classify. A boosting routine was used to steer the population to an optimal solution. After 100 generations, the GA identified five gas chromatographic peaks whose PC plot showed clustering of the ants on the basis of time period (see Fig. 5). When the gas chromatogram of the queen was projected onto the principal component (PC) map defined by the five gas chromatographic peaks and the 40 ant samples, it was evident that 3 month ants possessed a hydrocarbon profile similar to the queen. This result further reinforced the hypothesis formulated by these workers that PPG hydrocarbons play an important role in nestmate recognition.

4. Post-consumer identification of plastics

The second data set used to evaluate the performance of the pattern recognition GA consisted of 188 Raman spectra of six common household plastics: High density polyethylene (HDPE), low density polyethylene (LDPE), polyethylene terephthalate (PET), polypropylene (PP), polystyrene (PS), and polyvinylchloride (PVC). The overall goal of this study was to develop a potential method to differentiate common household plastics by type using Raman spectroscopy. Since the most valuable reprocessed plastics are prepared from pure polymer streams, sorting of plastics by type is crucial to ensure the economic viability of recycling.

Each plastic sample was cut from collected containers obtained from residential homes and BFI

Table 1
Training set

Plastic type	Number of spectra
HDPE	33
LDPE	26
PET	35
PP	26
PS	32
PVC	17
Total	169

Table 2
Prediction set

Plastic type	Number of spectra
HDPE	5
LDPE	2
PET	5
PP	2
PS	5
PVC	0
Total	19

Recycling in Pocatello, ID. The sample geometry was chosen based on optimal placement in the sample holder of Raman spectrometer. Raman spectra were measured using a Spex 500 M 1/2 meter Raman spectrometer incorporating a Spex Model 1449 collection optics module, an Omnichrome Model 160 T/B air-cooled Ar^+ laser and liquid nitrogen cooled charged coupled detector device. Further details about the data can be found elsewhere [15].

Each Raman spectrum, an average of 16 one-second scans, was collected over the wave number range 850 to 1800 cm^{-1} to yield 1093 points. The Raman spectra were boxcar averaged every 10 points yielding 218-point spectra, which were baseline corrected for offsets using a linear polynomial. Each spectrum was then normalized to unit length to adjust for variations in the optical path length.

The spectra were divided into a training set of 169 spectra (see Table 1) and a prediction set of 19 spectra (see Table 2). Members of the prediction set were chosen by random lot. The data were auto-scaled to ensure that each wavelength had equal weight in the analysis. For pattern recognition analysis, each plastic sample was represented by a data vector, $\boldsymbol{x} = (x_1, x_2, x_3 \ldots x_j, x_{218})$, where x_j is the Raman intensity of the jth point of the baseline corrected normalized Raman spectrum.

The first step in the study was to apply principal component analysis to the data. Fig. 6 shows a plot of the two largest principal components of the 218-point Raman spectra that comprised the training set. Each spectrum is represented as a point in the principal component plot (1 = HDPE, 2 = LDPE, 3 = PET,

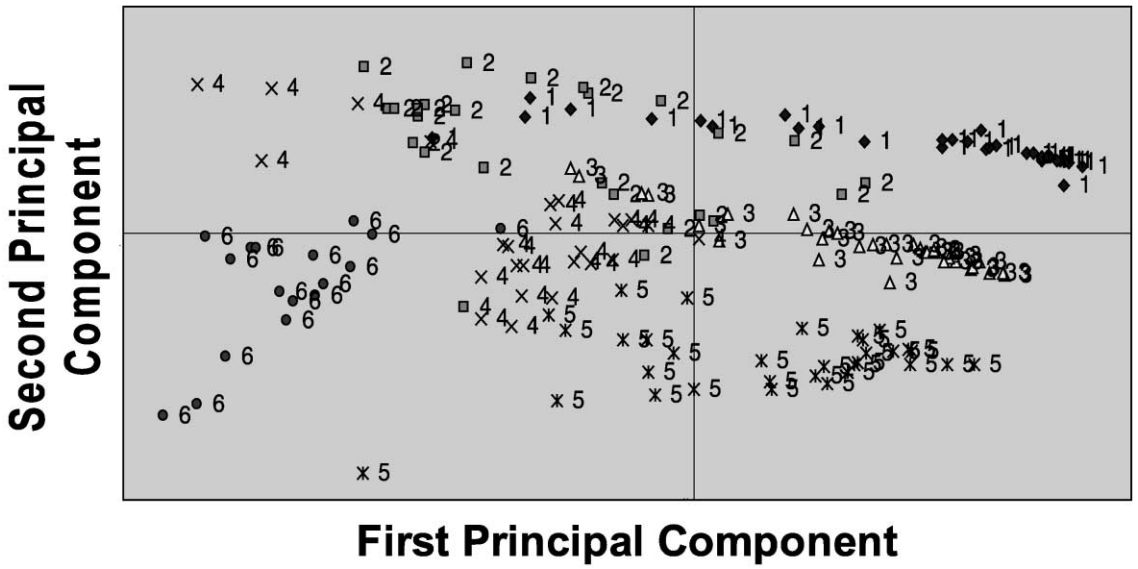

Fig. 6. A plot of the two largest principal components of the 218-point Raman spectra that comprised the training set. Each spectrum is represented as a point in the principal component map (1 = HDPE, 2 = LDPE, 3 = PET, 4 = PP, 5 = PS, and 6 = PVC). Reprinted with the kind permission of SPIES from B. K. Lavine, and A.J. Moores, "Genetic Algorithms for Pattern Recognition Analysis and Fusion of Sensor Data," in Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, K. Siddiqui and D. Eastwood (Eds.), Proceedings of SPIES, 1999, pp. 103–112.

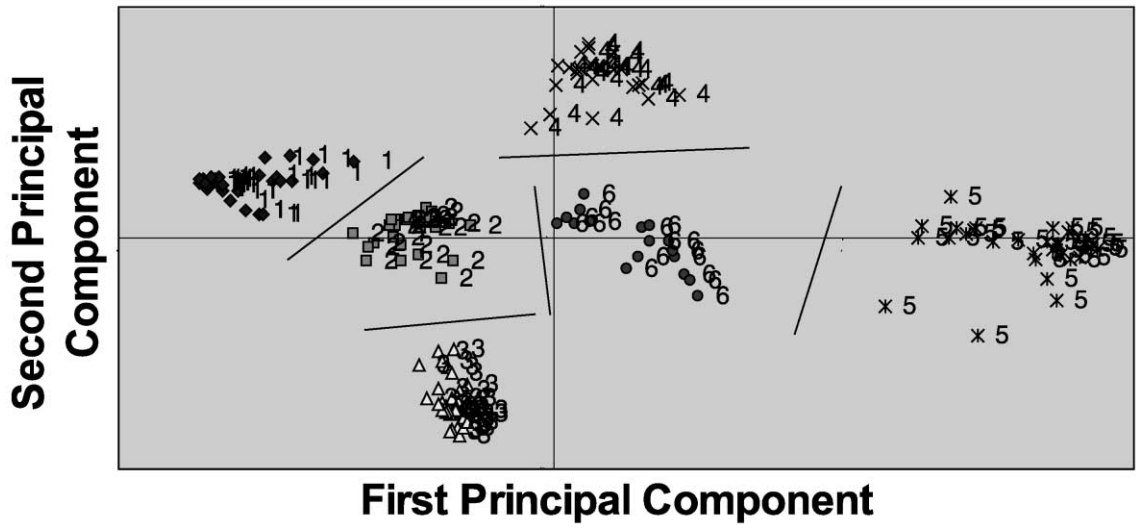

Fig. 7. A plot of the two largest principal components of the 169 Raman spectra that comprise the training set and nine spectral features selected by the GA. Each spectrum is represented as a point in the principal component map (1 = HDPE, 2 = LDPE, 3 = PET, 4 = PP, 5 = PS, and 6 = PVC). Reprinted with the kind permission of SPIES from B. K. Lavine, and A. J. Moores, "Genetic Algorithms for Pattern Recognition Analysis and Fusion of Sensor Data," in Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, K. Siddiqui and D. Eastwood (Eds.), Proceedings of SPIES, 1999, pp. 103–112.

4 = PP, 5 = PS, and 6 = PVC). The overlap of HDPE, LDPE, PP and PS in the PC plot is not surprising in view of the similarity of their Roman spectra.

The pattern recognition GA was used in this study to uncover features characteristic of the Raman profile of each class. Features were identified by sampling key feature subsets, scoring their principal component plots, and tracking classes and/or samples, which were most difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 100 generations, the pattern recognition GA identified nine spectral features whose principal component plot showed clustering of the Raman spectra on the basis of class (see Fig. 7).

A prediction set of 19 Raman spectra was used (see Table 2) to assess the predictive ability of the nine wavelengths identified by the pattern recognition GA. The prediction set samples were projected onto the principal component map developed from the 169 spectra and nine wavelengths. Fig. 8 shows the projection of the prediction set samples onto a principal component map defined by the nine wavelengths selected by the GA. Each projected sample lies in a region of the map occupied by plastic samples possessing the same class label. Evidently, the GA can identify features in the Raman spectra characteristic of the plastic-type. This suggests that Raman spectroscopy can be used to sort plastic containers by type.

5. Conclusion

The advantages of using the pattern recognition GA for feature selection are four-fold. First, chance classification is not a serious problem since the bulk of the variance or information content of the features selected is about the pattern recognition problem of interest. Second, features that contain discriminatory information about a particular classification problem will be correlated, which is why feature selection should be performed using methods based on principal component analysis. Third, the principal component analysis routine of the fitness function is able to dramatically reduce the size of the search space since it can correctly assess the true dimensionality of the data ensuring that only those regions of the solution space with information about the problem of interest

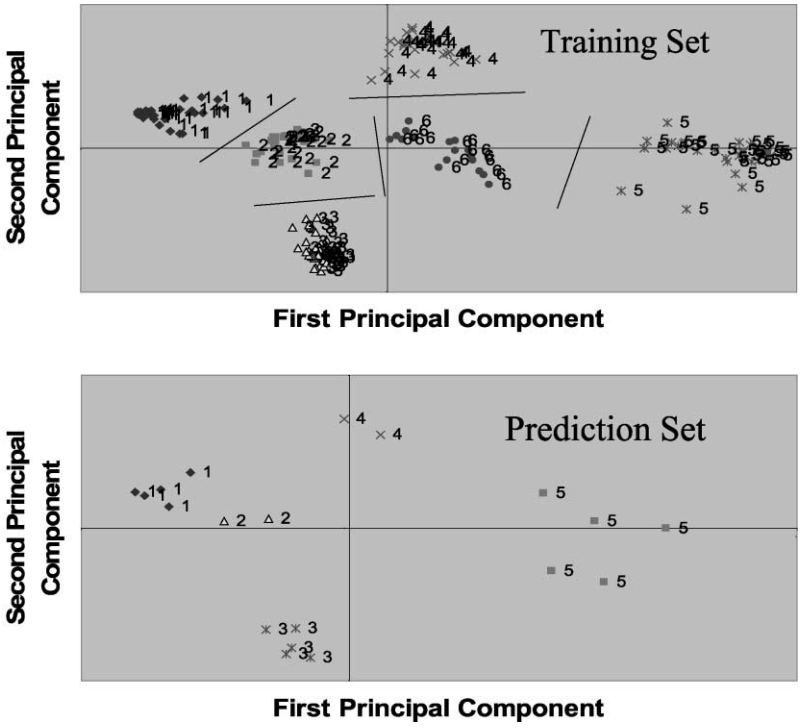

Fig. 8. A projection of the prediction set samples onto a principal component map defined by the training set samples, and the nine spectral features identified by the pattern recognition GA. Each spectrum in the prediction set is represented as a point on the principal component map (1 = HDPE, 2 = LDPE, 3 = PET, 4 = PP, 5 = PS, and 6 = PVC). The projected samples lie in a region of the map occupied by plastic samples possessing the same class label. Reprinted with the kind permission of SPIES from B. K. Lavine, and A. J. Moores, "Genetic Algorithms for Pattern Recognition Analysis and Fusion of Sensor Data," in Pattern Recognition, Chemometric, and Imaging for Optical Environmental Monitoring, K. Siddiqui and D. Eastwood (Eds.), Proceedings of SPIES, 1999, pp. 103–112.

are investigated. Fourth, the pattern recognition GA through the PC plot that it generates allows the user to interpret the meaning of the underlying pattern recognition relationship in the data and understand how the decision for a classification is made from the principal component plot generated.

The approach used by the GA for feature selection and pattern recognition is the same approach used by many chemists for multivariate data analysis. However, the GA has the advantage that it can search a large space in a systematic manner using human pattern recognition. The combination of human pattern recognition and machine learning implemented through the language of reproduction and natural selection produces a learning paradigm superior to that of man or machine alone because of the synergism created by coupling different learning approaches.

References

[1] R.G. Brereton (Ed.), Multivariate Pattern Recognition in Chemometrics—Illustrated by Case Studies, Elsevier, Amsterdam, 1992, pp. 1–45.
[2] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley Interscience, New York, 1992.
[3] K. Fukunaga, Statistical Pattern Recognition, 2nd edn., Academic Press, San Diego, 1990.
[4] B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, H.T. Mayfield, Anal. Chem. 72 (2) (2000) 423–431.
[5] B.K. Lavine, A.J. Moores, H.T. Mayfield, A. Faruque, Genetic algorithms applied to pattern recognition analysis of high speed gas chromatograms of aviation turbine fuels using an integrated Jet-A/JP-8 data base, Microchem. J. 61 (1999) 69–78.
[6] B.K. Lavine, A. Moores, L.K. Helfend, A genetic algorithm for pattern recognition analysis of pyrolysis gas chromatographic data, J. Anal. Appl. Pyrolysis 50 (1999) 47–62.
[7] B.K. Lavine, A. Moores, H.T. Mayfield, A. Faruque, Fuel spill

identification using gas chromatography/genetic algorithm-pattern recognition techniques, Anal. Lett. 31 (15) (1998) 2805–2822.
[8] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, Reading, MA, 1989.
[9] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer Verlag, New York, 1992.
[10] M. James, Classification Algorithms, Wiley Interscience, New York, 1985.
[11] R.K. Vander Meer, D.P. Wojcik, Chemical mimicry in the myrmecaphodius excavaticollis, Science 218 (1982) 806–808.
[12] M.D. Beecher, Signature systems and kin recognition, Am. Zool. 22 (1982) 477–490.
[13] N.F. Carlin, B. Holldobler, Nestmate and kin recognition in interspecific mixed colonies of ants, Science 222 (1983) 1027–1029.
[14] J. Edward Jackson, A User's Guide to Principal Component Analysis, Wiley, New York, 1991.
[15] V. Allen, J.H. Kalivas, R.G. Rodriguez, Post-consumer plastic identification using Raman spectroscopy, Appl. Spectrosc. 53 (6) (1999) 672–681.