ELSEVIER

Analytica Chimica Acta 437 (2001) 233–246

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Genetic algorithm for fuel spill identification

B.K. Lavine [a,*], D. Brzozowski [a], A.J. Moores [a], C.E. Davidson [a], H.T. Mayfield [b]

[a] *Department of Chemistry, Clarkson University, Potsdam, NY 13699-5810, USA*
[b] *AL/EQ, 139 Barnes Drive, Suite 2, Tyndall AFB, FL 32403-5323, USA*

Received 29 August 2000; received in revised form 30 January 2001; accepted 26 February 2001

Abstract

Gas chromatography is frequently used to fingerprint fuel spills, with the gas chromatograms of the spill sample and the different candidate fuels compared visually in order to seek a best match. However, visual analysis of gas chromatograms is subjective and is not always persuasive in a court of law. Pattern recognition methods offer a better approach to the problem of matching gas chromatograms of weathered fuels. Pattern recognition methods involve less subjectivity in the interpretation of the data and are capable of identifying fingerprint patterns within gas chromatographic (GC) data characteristic of fuel-type, even if the fuel samples comprising the training set have been subjected to a variety of conditions. In this paper, we report on the development of a genetic algorithm (GA) for pattern recognition analysis of GC fuel spill data. The pattern recognition GA incorporates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection. Its efficacy is demonstrated by way of two studies recently completed in our laboratory on fuel spill identification.

Keywords: Fuel spill identification; Genetic algorithm; Pattern recognition analysis; Feature selection; Machine learning; Fingerprint data

1. Introduction

Water from underground wells or aquifers is an important natural resource, supplementing or replacing surface water supplies in many households and communities in the Southeastern US. The possible contamination of this natural resource by jet fuels stored in leaking underground tanks and pipelines has prompted the US Air Force to develop new methods to identify fuel materials recovered from subsurface environments at or near military airfields. Burgeoning interest in techniques that can establish the type of jet fuel responsible for the contamination of an underground well or aquifer is motivated in large measure by the cleanup costs, legal fees, and fines incurred by the polluter.

Water samples from underground wells or aquifers contaminated by leaking fuels exist in one of two forms. Either the water sample collected from the well has a layer of floating fuel or the sample contains dissolved hydrocarbons from the leaking fuel. In the worse case scenario, that of a leaking fuel, the fuel layer is collected and analyzed by capillary column gas chromatography. The technique is easy to use, sample preparation is minimal, and the instrumentation required for the analysis is inexpensive and readily available. Although optical techniques, e.g. near infrared, mid infrared, and fluorescence spectroscopy, have also been used to characterize oil and fuel spills [1], they do not possess sufficient discriminatory power to type jet fuels because the chemical composition of a jet fuel is primarily middle distillates, i.e.

* Corresponding author. Tel.: +1-315-268-2394; fax: +1-315-268-6610.
E-mail address: bklab@clarkson.edu (B.K. Lavine).

0003-2670/01/$ – see front matter
PII: S0003-2670(01)00946-1

C-9 to C-18 alkanes and alkenes. (Aromatics are only minor constituents.) By comparison, gas chromatography is able to classify volatile and highly complex mixtures of hydrocarbons because of the high resolution of capillary columns and the high sensitivity and linearity of flame ionization detectors towards middle distillates.

Typically, the gas chromatogram of a fuel spill and a number of suspected hydrocarbon sources are compared visually in order to obtain a match. However, this approach to fuel spill identification is subjective and is not always persuasive in a court of law. Furthermore, visual analysis suffers from the drawback that it cannot always take into account the influence of weathering on the overall GC profile of the spill. Small variations in the GC operating conditions (e.g. small changes in the temperature programming rate of the GC column) can also be a problem complicating the identification of leaking fuel [2].

Due to the complexity of the hydrocarbon mixture that constitutes a processed fuel, a systematic comparison of gas chromatograms is necessary to ensure that differences between GC profiles of various fuel-types are significant. Therefore, pattern recognition methods [3,4] offer a better approach to the problem of matching gas chromatograms of hydrocarbon fuels. Pattern recognition methods involve less subjectivity in the interpretation of GC data and are capable of identifying fingerprint patterns within GC data characteristic of fuel-type, even if the fuel samples comprising the training set have been subjected to a variety of conditions. Thus, discriminants can be developed that are less sensitive to changes in the overall GC profile of the fuel due to contamination, weathering or analytical error.

In this paper, the development of a genetic algorithm (GA) for pattern recognition analysis of fuel spill data is reported. The pattern recognition GA [5–7] selects features (i.e. GC peaks) that optimize the separation of the fuel classes in a plot of the two or three largest principal components of the data. A good principal component plot can only be generated using features whose variance or information is primarily about differences between the fuel classes. This fitness criterion dramatically reduces the size of the search space since it limits the search to these types of feature subsets. In addition, the GA focuses on those classes and/or samples that are difficult to classify as it trains by boosting the relative importance of classes and samples that consistently score poorly. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The fuel spill GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection.

Two studies demonstrating the efficacy of the fuel spill identification GA are discussed at length. In the first study, classifiers developed from the gas chromatograms of 284 neat jet fuels were used to predict the fuel-type of 31 jet fuels recovered from a subsurface environment. The features used to develop the classifier were identified by the pattern recognition GA. The second study focused on water samples that contained dissolved hydrocarbons, i.e. water contaminated by aviation turbine fuels. Each water sample was prepared by equilibrating a neat jet fuel with water in a specially designed reaction vessel designed to maximize the contract surface area between the two phases. Pattern recognition analysis of the 133 GC profiles of the dissolved hydrocarbons revealed the existence of fingerprint patterns within the data characteristic of fuel-type. The ease of classifying these highly complex mixtures by selective fractionation prior to gas chromatography becomes apparent when taking into account the fact that an equilibration time of only 3 h is necessary to obtain a reproducible profile of the water-soluble components of a jet fuel.

2. Experimental

Neat samples of JP-4, Jet-A, JP-7, JPTS, JP-5, JP-8, and 100/130-octane aviation gasoline (AVGAS) were obtained from Wright Patterson and/or Mukilteo Energy Management Laboratories. These fuel samples were splits from regular quality control standards used by the two laboratories to verify the authenticity of manufactures claims. The control standards were collected by the two laboratories over a 5-year period and constituted a representative sampling of the fuels.

2.1. Weathered jet fuel dataset

For the first study, each jet fuel sample was stored in a sealed container at 20°C. Prior to GC analysis, each fuel sample was diluted with methylene

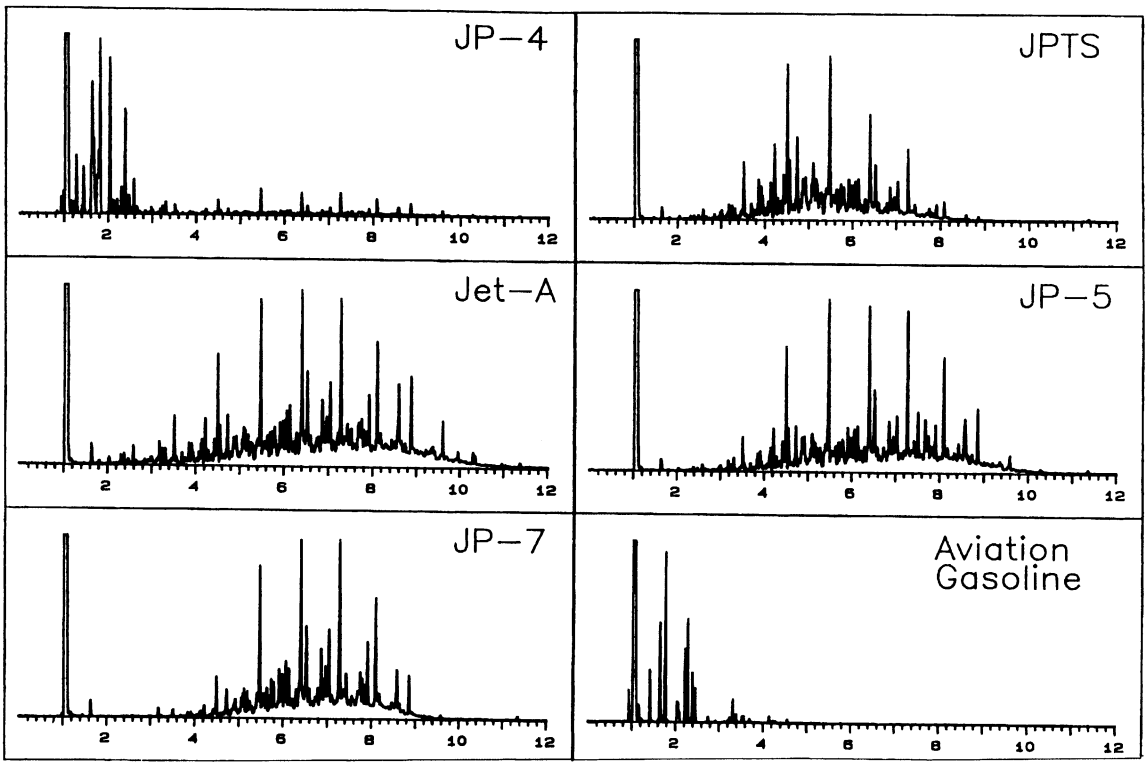

Fig. 1. High-speed gas chromatograms of JP-4, Jet-A, JP-7, JPTS, JP-5, and AVGAS.

chloride. The fuel samples were injected onto a capillary column using a split column technique. The high efficiency fused silica capillary column (10 m in length with an i.d. of 0.10 mm and coated with 0.34 mm of a bonded and cross-linked 5% phenyl-substituted polymethylsiloxane stationary phase) was temperature programmed from 60 to 270°C at 18°C/min. The resulting high-speed gas chromatograms were digitized using an HP-3357 laboratory automation system. High-speed gas chromatograms representative of JP-4, Jet-A, JP-7, JPTS, JP-5, and AVGAS are shown in Fig. 1. The gas chromatograms of the neat jet fuels constituted the training set (see Table 1). The prediction set consisted of 31 gas chromatograms of weathered jet fuels (see Table 2). Seventeen of the 31 weathered fuels were collected from sampling wells as a neat oily phase found floating on top of the water; 11 of the 31 fuels were extracted from the soil near various fuel spills; and the other three fuels had been subjected to weathering in a laboratory.

Table 1
Training set

Fuel-type	Number of samples
JP-4	54
Jet-A	66
JP-7	28
JPTS	34
JP-5	44
AVGAS	18
JP-8	40
Total	284

2.2. Dissolved hydrocarbon dataset

The water-soluble fraction was obtained by equilibrating 2 ml of a neat jet fuel with 250 ml of water while stirring gently for 12 h in a vessel designed by McIntyre and Burris [8] to maximize surface contact between fuel and water while avoiding

Table 2
Prediction set

Fuel-type	Sample number
JP-4	2650-T9W007[a]
JP-4	2660-T9W008
JP-4	2670-T9W009
JP-4	2680-T9W010
JP-4	2690-T9W011
JP-4	2700-T9W012
JP-4	2710-T9W013
JP-4	2720-KSE1M2[b]
JP-4	2730-KSE2M2
JP-4	2740-KSE3M2
JP-4	2750-KSE4M2
JP-4	2760-KSE5M2
JP-4	2770-KSE6M2
JP-4	2780-KSE7M2
JP-4	27908/9/90SMP1[c]
JP-4	28008/9/90SMP2
JP-4	28108/9/90SMP3
JP-4	28208/9/90SMP4
JP-4	2830STALE-1[d]
JP-4	2840STALE-2
JP-4	2850STALE-3
JP-5	2860PIT1UNK[e]
JP-5	2870PIT1UNK
JP-5	2880PIT2UNK
JP-5	2890PIT2UNK
JP-4	0010TYNDL-1[f]
JP-5	2910PIT2UNK[e]
JPTS	2600PAT[g]
JPTS	2610PAT[g]
AVGAS	3100Richmnd[h]
AVGAS	3200Richmnd

[a] Sampling well at Tyndall. The sampling well was near a previously functioning storage depot. Each well sample was collected on a different day.

[b] Soil extract near sampling well at Tyndall. Dug with a hand auger at various depths. Distance between sampling well and soil extract was approximately 80 yards.

[c] JP-4 diluted with methylene chloride was added to sand and later re-extracted (simulated soil extract).

[d] Weathered in laboratory.

[e] Sampling pit at Keywest Naval Air Station. Two pits were dug near a seawall to investigate a suspected JP-5 fuel leak.

[f] Recovered from a previously functioning storage facility at Tyndall.

[g] Recovered from the subsurface environment at Patrick Air Force Base.

[h] Subsurface fuel spill from Richmond Airport.

mixing. Following equilibration, several milliliters of water were discharged from the vessel to ensure the delivery tube was clear of fuel, and two 25 ml aliquots of the water phase were delivered into gas tight syringes equipped with Luer-lock open shut valves. Solid phase extraction (SPE)/gas chromatography was used to characterize the 25 ml water samples containing the dissolved hydrocarbons. For the SPE procedure, each 25 ml aliquot was forced through a C-18 Sep-pak (Millipore Corporation) SPE cartridge. The Sep-Pak was partially dried with a 5 ml slug of air and was extracted with 1 ml of carbon disulfide. A 1 ml aliquot of each carbon disulfide extract was injected directly onto a fused silica capillary column (0.25 mm bonded polyethylene glycol stationary phase), which was temperature programmed from 40 to 200°C at 5°C/min with an initial isothermal hold of 4 min. Gas chromatograms of the carbon disulfide extract were obtained using a Hewlett Packard 5880 GC equipped with a mass selective detector. Fig. 2 shows GC profiles (i.e. total ion chromatograms) representative of the solid phase extracts of JP-4, Jet-A, JP-7, JPTS, JP-5, and AVGAS. The SPE dataset, which consisted of 133 gas chromatograms, is described in Table 3. Further information about this dataset can be found elsewhere [9].

3. Data preprocessing

The GC data were digitized and stored using an HP-3357 laboratory automation system implemented on an HP-1000-F minicomputer. A FORTRAN program was used to translate the integration reports into ASCII files that were formatted for entry into SETUP [10], a computer program for peak-matching. SETUP matches peaks by first: (1) computing the Kovat's retention index for the compounds eluting off the GC column or (2) dividing each chromatogram into intervals defined by major peaks that are always present and linearly scaling the retention times of the peaks within the intervals for best fit with respect to a reference chromatogram. For the neat jet fuels, the *n*-alkane peaks were the most prominent features present [11], so it was a simple matter to compute Kovat's retention indices for the GC peak in the weathered jet fuel dataset. For the SPE gas chromatograms, there were a number of peaks that are present in all the gas chromatograms so developing a retention scale rotted on the majors was feasible. The peak-matching program then analyzed the GC data in three distinct steps. First, a template of peaks

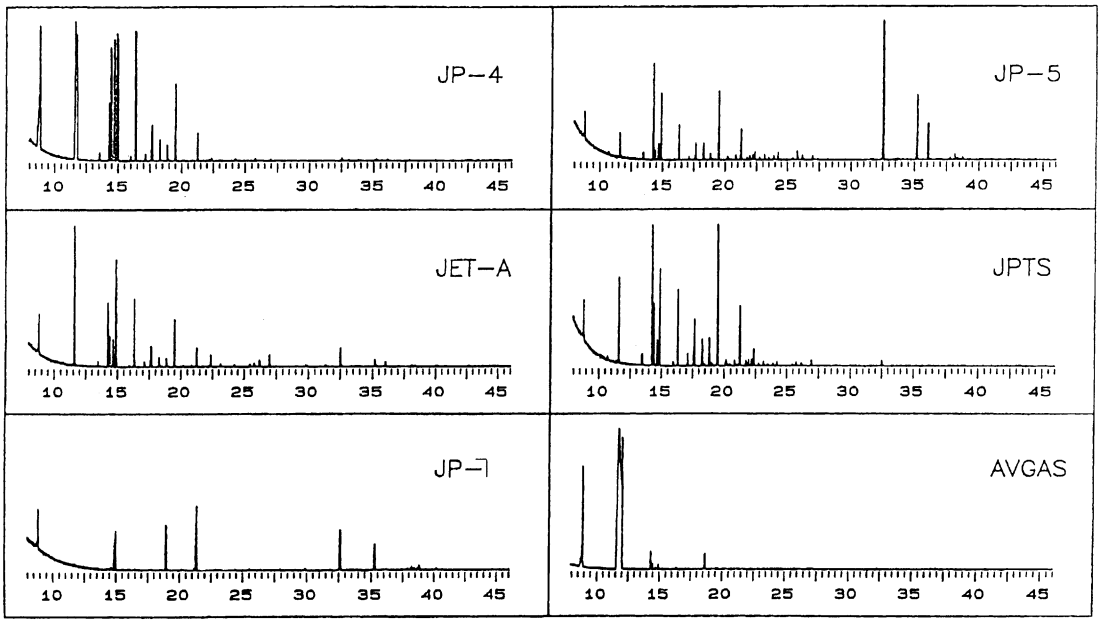

Fig. 2. Reproduced total ion chromatograms representative of the solid phase extracts of JP-4, Jet-A, JP-7, JPTS, JP-5, and AVGAS.

was developed by examining integration reports, and adding features to the template, which did not match the retention indices of previously observed features. (The integration reports list the integrated areas of the chromatographic peaks in each gas chromatogram.) Second, a preliminary data vector was produced for each gas chromatogram by matching the retention indices of GC peaks with the retention indices of the features in the template. (The template lists the standardized retention indices of the different peaks encountered in the gas chromatograms of the dataset being investigated.) A feature is assigned a value corresponding to the normalized area of the GC peak in the chromatogram. Unmatched peaks are zeroed, whereas poorly resolved and tailing peaks are excluded from the analysis. (A peak is matched provided that differences in adjusted retention times, e.g. KI values for the neat jet fuels or retention times rotted on the majors for SPE, fall within the user specified tolerance window for a given peak pair.) Third, the frequency of each feature was computed, i.e. the number of times a particular feature is found to have a nonzero value is calculated, and features below a user specified number of nonzero occurrences (which is set equal to 10% of the total number of fuel samples in the training set) are deleted from the dataset, whereas features that passed the nonzero frequency criterion are retained. The peak-matching software yielded a final cumulative reference file containing 85 peaks for the weathered jet fuel dataset and 48 peaks for the SPE dataset. Hence, for pattern recognition analysis, each neat jet fuel gas chromatogram was initially represented as an 85 dimensional data vector, $x = (x_1, x_2, x_3, \ldots, x_j, \ldots, x_{85})$, and each SPE gas chromatogram was initially represented as a 48-dimensional data vector $x = (x_1, x_2, x_3, \ldots, x_j, \ldots, x_{48})$ where x_j is the area of

Table 3
Solid phase extraction dataset

Fuel-type	Number of fuel samples	Number of chromatograms
JP-4	20	27
Jet-A	27	54
JP-7	4	8
JPTS	10	20
JP-5	9	18
AVGAS	6	6
Total	76	133

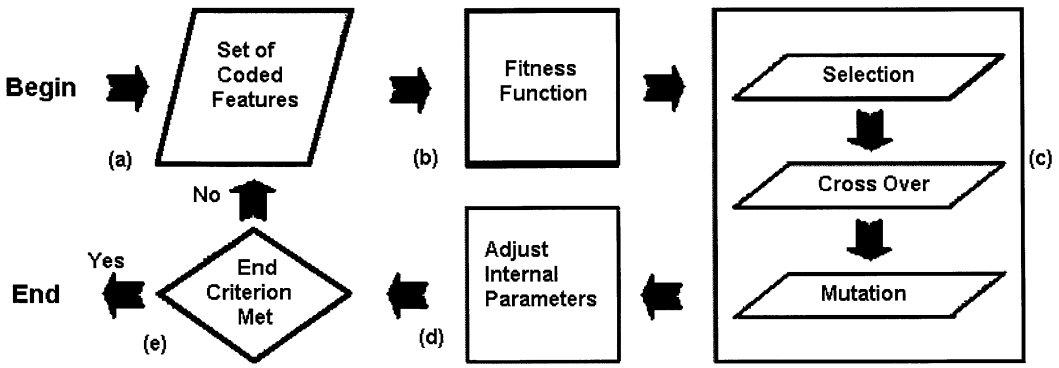

Fig. 3. A block diagram of the pattern recognition GA used for fuel spill identification.

the jth peak. The data vectors were normalized to constant sum, i.e. each x_j was divided by the total integrated peak area.

3.1. Pattern recognition analysis

A block diagram of the pattern recognition GA for fuel spill identification is shown in Fig. 3. The GA [12,13] builds a population of binary strings, each of which represents a possible solution, i.e. a unique subset of the GC peaks. 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 GC peak is not included in the feature subset. During each generation, the strings are decoded yielding the actual parameter set, which is sent to the fitness function for evaluation. The fitness function determines the string's relative importance to other members of the population. Fit solutions are selected for crossover, that is, fit feature subsets are broken up, swapped, and recombined creating new subsets of features, which are introduced into the population of potential solutions. This process is repeated until a specified number of generations are executed or a feasible solution is found.

The pattern recognition GA for fuel spill identification differs from conventional GAs in the types of operators that it utilizes. The fitness function, which is graphically based, actually emulates human pattern recognition through machine learning to identify a set of features (i.e. GC peaks) that optimize the separation of the fuel classes in a plot of the two largest principal components of the data. To track and score the principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see Eqs. (1) and (2)) where CW(c) is the weight of class c with c varying from 1 to the total number of classes in the dataset. $\mathrm{SW}_c(s)$ is the weight of sample s in class c. The class weights sum to 100, and the sample weights for the objects comprising a particular class sum to a value that is equal to the class weight of the class in question.

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

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

Each principal component plot generated for each feature subset after it has been extracted from its chromosomes is scores using the K-nearest neighbor classification algorithm [14]. 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 taken of the point's K_c-nearest neighbors. For the most rigorous classification (which was also the case for the two studies described in this paper), K_c equals the number of samples in the class to which the point belongs. (Thus, K_c usually has a different value for each class.) The number of K_c-nearest neighbors with the same class label as the sample point in question, the so-called sample-hit count, SHC(s), is computed ($0 < \mathrm{SHC}(s) < K_c$) for each

sample. It is then a simple matter to score a principal component plot (see Eq. (3)). First, the contribution to the overall fitness by each sample in class 1 is computed, with the scores of the samples comprising the class summed to yield the contribution by this class to the overall fitness. This simple calculation is again repeated for classes 2, 3, etc., with the scores from each class summed to yield the overall fitness, $F(d)$.

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

To better understand how principal component plots are scored, consider a dataset with two classes, which have been assigned equal weights. Class 1 has 50 samples, and class 2 has 10 samples. At generation 0, the samples in a given class have the same weight. Thus, each sample in class 1 has a sample weight of 1, whereas each sample in class 2 has a weight of 5. Suppose a sample from class 2 has as its nearest neighbors 8 class one samples. Hence, $\mathrm{SHC}/K = 0.8$, and $(\mathrm{SHC}/K) \times \mathrm{SW} = 0.8 \times 5$, which equals 4. By summing $(\mathrm{SHC}/K_c) \times \mathrm{SW}$ for each sample, each principal component plot can be scored. One advantage of using this procedure to score the principal component plots is that a class with a large number of samples will not dominate the analysis due to the class weights.

The fitness function of the GA is able to focus on samples and classes that are difficult to classify by boosting their weights over successive generations. (Boosting the weights is referred to as adjusting the internal parameters in the block diagram of the pattern recognition GA.) In order to boost, it is necessary to compute both the sample-hit rate (SHR), which is the mean value of SHC/K_c over all feature subsets produced in a particular generation (see Eq. (4)), and the class-hit rate (CHR), which is the mean SHR of all samples in a class (see Eq. (5)). ϕ in Eq. (4) is the number of chromosomes in the population, and AVG in Eq. (5) refers to the average or mean value. During each generation, class and sample weights are adjusted by a perceptron (see Eqs. (6) and (7)) with the momentum, P, set by the user. ($g + 1$ is the current generation, whereas g is the previous generation.) Classes with a lower CHR are boosted more heavily than those classes that score well.

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

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

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

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

Boosting is crucial for the successful operation of the fuel spill identification GA because it modifies the fitness landscape by adjusting the values of the class and sample weights. This helps to minimize the problem of convergence to a local optimum. Hence, the fitness function of the GA changes as the population evolves towards a solution.

The selection operator of the pattern recognition GA utilizes both the adults and children to develop new solutions. Potential solutions are placed in two columns. In the first column, the solutions are ordered from best to worst with respect to their fitness. In the second column, a copy of the same population is randomly ordered with respect to the fitness of the chromosomes. The first row of the first column is combined with the first row of the second column using a set of rules encoded in the crossover operator to yield new and potentially better solutions for the fuel spill identification problem. Typically, the selection pressure is set at 0.5 so, the top half of the ordered population is mated with strings or chromosomes from the top half of the random population. For each pair of strings selected for mating, two new strings are generated. Because the best feature subsets are always being used, each new population is expected to yield better results than the previous generation. However, each chromosome (i.e. potential solution) has a chance of being selected (because of the second column) ensuring that a significant degree of diversity is maintained during the search for the best solution. It is unlikely that any individual feature will be zeroed out of the analysis when this selection procedure is used.

The reproduction operator of the pattern recognition GA generates new solutions employing an unusual variation of three-point crossover. As in the case of simple crossover, the length of each new string is the same as the dimensionality of the data.

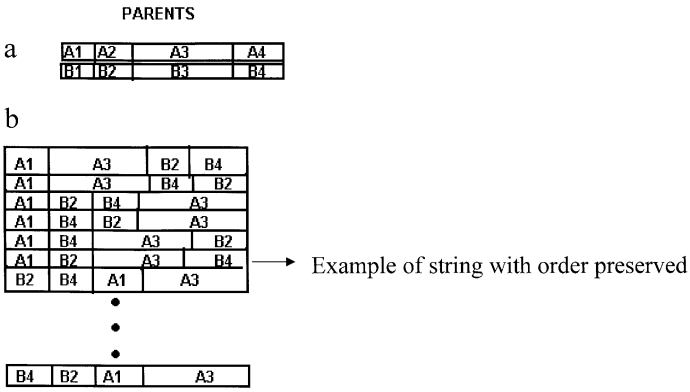

Fig. 4. Crossover operator with a reordering algorithm embedded in it. (a) Loci of selected chromosomes are randomly chosen, and (b) children are produced based on the recombination of chromosomal fragments. The preservation of original fragment order is not required.

Unlike simple three-point crossover, the crossover operator used by the pattern recognition GA is not compelled to preserve order among exchanged string fragments, which safeguards the loss of information or features in the population (see Fig. 4). As a result, it becomes less likely for the population variability to fall below a critical value due to the additional degree of freedom provided by the reordering. This variation of three-point crossover is also useful in searching for good string arrangements. For example, consider a population with bad ordering, i.e. where great distances separate good features. Simple crossover would probably destroy these important feature packets. In this situation, there is a chance to obtain good ordering, if a crossover operator is used with a reordering algorithm embedded in it.

In the last step of reproduction, a mutation operator is applied to the new strings. The mutation probability of the operator is set at 0.01, so 1% of the feature subsets are selected at random for mutation. A chromosome marked for mutation has a single bit flipped. This allows the GA to explore other regions of the parameter space. If the GA finds a better point in the solution space, the chromosome representing this point will invade the population, allowing optimization to continue in a new direction.

The pattern recognition GA was coded using Matlab 5.3. All calculations in this study were performed on a 166 MHz Pentium computer with 128 Mb of EDO RAM running under Windows 95. Fitness evaluation was the step with the highest computational load. The 3–4 h were typical run times for the pattern recognition GA on this platform.

4. Weathered jet fuels

The first step in any fuel spill identification problem is to apply principal component analysis (PCA) to the data. PCA is a method of transforming the original measurement variables into new, uncorrelated variables called principal components. Each principal component is a linear combination of the original measurement variables. The largest principal component is formed by determining the direction of largest variation in the data and modeling it by a line that passes through the center of the data. The second largest principal component lies in the direction of next largest variation; it passes through the center of the data and is orthogonal to the largest principal component. The third largest principal component lies in the direction of next largest variation; it passes through the center of the data and is orthogonal to the first and second largest principal components, and so forth. By using PCA, the original measurement variables, which constitute a correlated axis system, can be converted into an orthogonal system that removes correlations by forcing the new axes to be independent. This requirement dramatically reduces the dimensionality of the data because only a few independent axes are necessary to describe the data. PCA is routinely applied to high dimensional data to affect dimensionality reduction, classify samples, and/or identify outliers.

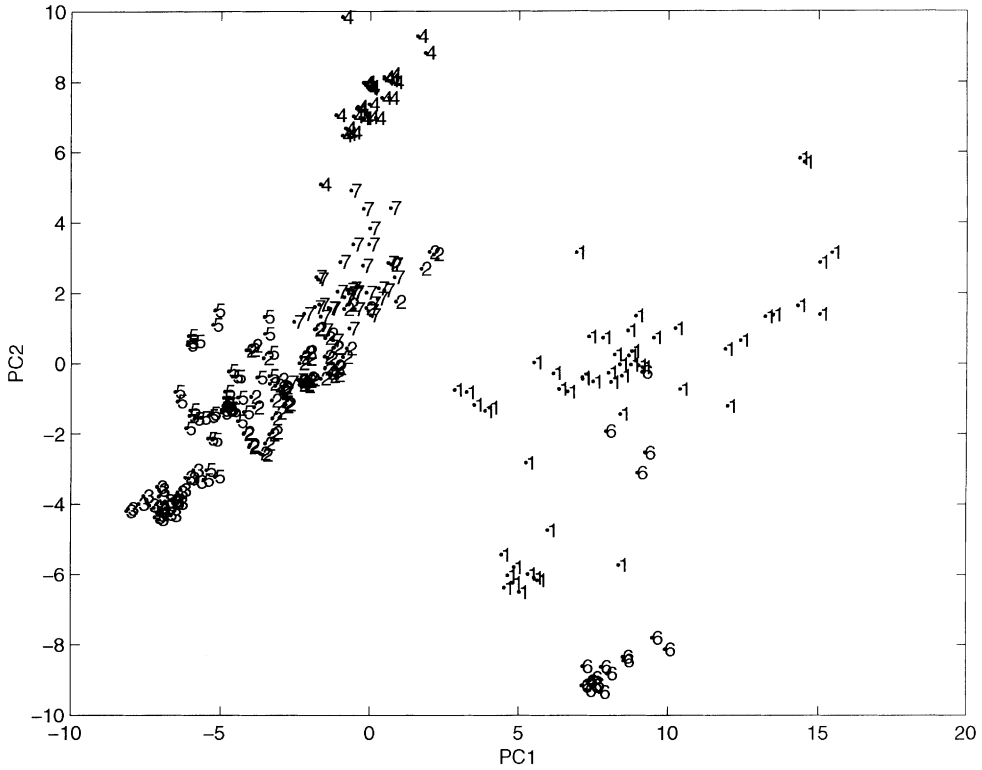

Fig. 5. A plot of the two largest principal components of the 85 GC peaks obtained from the 284 neat jet fuel gas chromatograms. Each fuel sample or gas chromatogram is represented as a point in the map (1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5, 6 = AVGAS, and 7 = JP-8). The two largest principal components explain 65% of the total cumulative variance.

Fig. 5 shows a plot of the two largest principal components of the 85 GC peaks obtained from the 284 neat jet fuel gas chromatograms. Each fuel sample or gas chromatogram is represented as a point in the map. The high degree of overlap of Jet-A, JP-5, and JP-8 fuel samples in the principal component map of the data is not altogether surprising. JP-5 and JP-8 are kerosene-based jet fuels that are similar in composition to Jet-A, the fuel used by civilian airliners. Mayfield and Henley [15] observed that Jet-A and JP-5 fuels are more difficult to classify than other types of jet fuels because of the similarity in their overall hydrocarbon composition. Lavine et al. [16,17] showed that statistical discriminant analysis could be used to differentiate Jet-A, JP-5, JP-4, JPTS, and JP-7 fuels. However, when this approach was applied to a set of chromatograms that included JP-8, the statistical discriminant could not differentiate Jet-A from JP-5 and JP-8.

The next step was feature selection. It is important to delete uninformative features to ensure that discriminatory information about fuel-type is the major source of variation in the data. If noisy features are not removed from the data, their presence can be detrimental to the performance of pattern recognition techniques such as linear and quadratic discriminant analysis [14] or SIMCA [18], since information characteristic of fuel-type will be swamped out by the large amount of qualitative and quantitative data due to experimental conditions [19]. Furthermore, many pattern recognition techniques, e.g. linear and quadratic discriminant analysis, do not perform well in small sample/high dimensional settings requiring the user to select an optimal set of features for dis-

criminant analysis [20]. Methods that estimate the inverse of the covariance matrix for each class will often not do well in these settings because of the difficulties in computing the inverse of the covariance matrix for each class due to the problem of collinearity, which arises from having more features than samples. It is the smaller of the two (features or samples) that define the number of independent axes needed to describe the data. Collinearity will inflate the size of the larger eigenvalues at the expense of the smaller eigenvalues. Since the inverse of the covariance matrix is determined by the smaller eigenvalues, this deflation resulting in values near the noise level in the data is a serious problem for unregularized methods such as quadratic or linear discriminant analysis [21].

Feature selection can also transform a difficult problem pattern recognition problem into a simple one. If separation by sample type is evident in a principal component score plot of the selected features, the probability of successfully developing a classifier from these features is high since the between group differences are large compared to among group differences.

The pattern recognition GA for fuel spill identification was used to uncover features characteristic of the GC profile of each fuel class. For this study, the GA was configured in the following manner. The number of chromosomes or binary strings in each population, ϕ, was set to 100, whereas the length of each binary string was 85. The momentum, P, was 0.8, and K_c for each class equaled the number of samples in the class. The selection pressure was 0.5 and the mutation rate was 0.01. Maximum number of iterations for a run was set to 100.

The pattern recognition GA sampled key feature subsets, scored their principal component plots, and tracked those samples and/or classes that were most

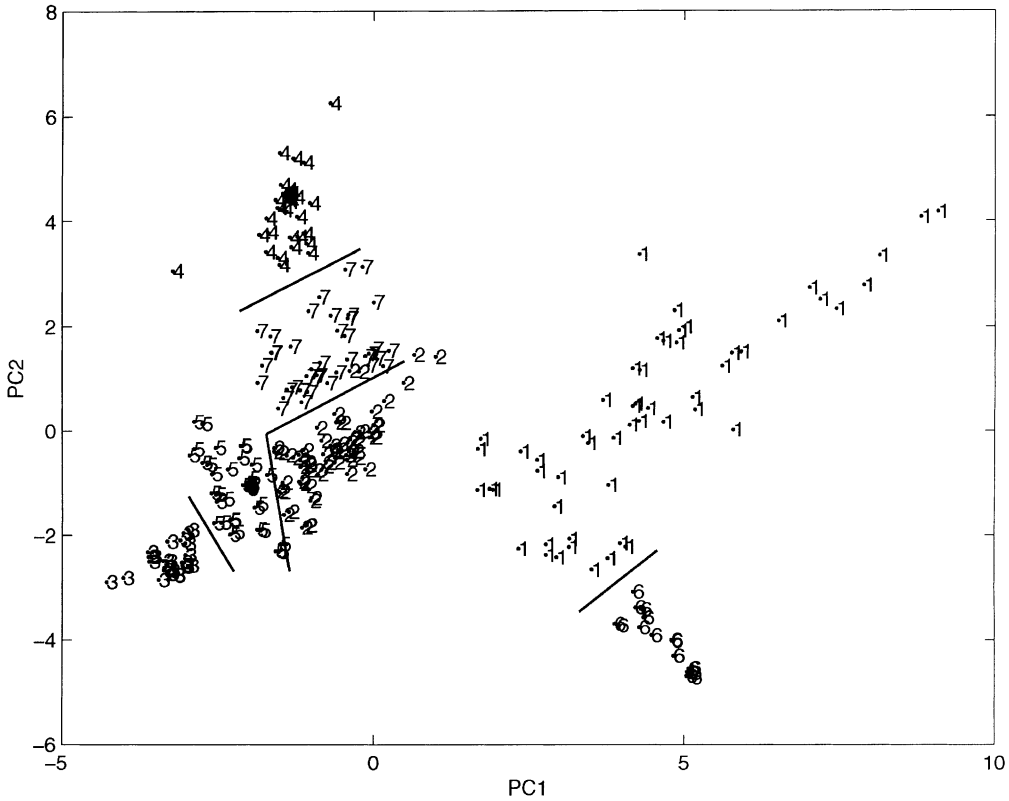

Fig. 6. A plot of the two largest principal components of the 22 GC peaks selected by the pattern recognition GA for fuel spill identification. Each fuel sample or gas chromatogram is represented as a point in the map (1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5, 6 = AVGAS, and 7 = JP-8). The two largest principal components explain 60% of the total cumulative variance.

difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 100 generations, the GA identified 22 standardized retention time windows whose principal component plot showed clustering of the fuel samples on the basis of fuel-type (see Fig. 6). This suggests that information about fuel-type is contained within the gas chromatograms of the neat jet fuels.

The 22 GC peaks identified by the GA were used as the starting point for a seven-way classification study involving JP-4, Jet-A, JP-7, JPTS, JP-5, AVGAS, and JP-8 fuels. This classification study, which is a logical extension of an earlier effort [16,17], was undertaken because of the change from JP-4 to JP-8 as the principal US Air Force fuel. The difficulty of identifying JP-8 fuels from GC data has been previously reported [16].

A classification rule was developed from the 22 GC peaks using regularized discriminant analysis (RDA) [21]. RDA is similar to SIMCA but employs a more complex scheme to obtain a biased estimate of the inverse of the class covariance matrix. Optimum values for the shrinking parameters used in RDA are computed for a given dataset by cross validating on the total number of misclassifications. (In other words, a vector of misclassifications as a function of the shrinkage parameter is generated, with the value of the evaluated parameter corresponding to the lowest error rate selected.).

When λ was set at 0.5 and γ was set at 0, the apparent classification success-rate for the neat jet fuels was 100%. The predictive ability of these descriptors was assessed by first computing the cross-validated and bootstrapped error rate which was approximately 2%. To further test the predictive ability of these GC peaks and the discriminant that they supported, a prediction set of 31 gas chromatograms was employed (see Table 2). 30 of the 31 samples in the prediction

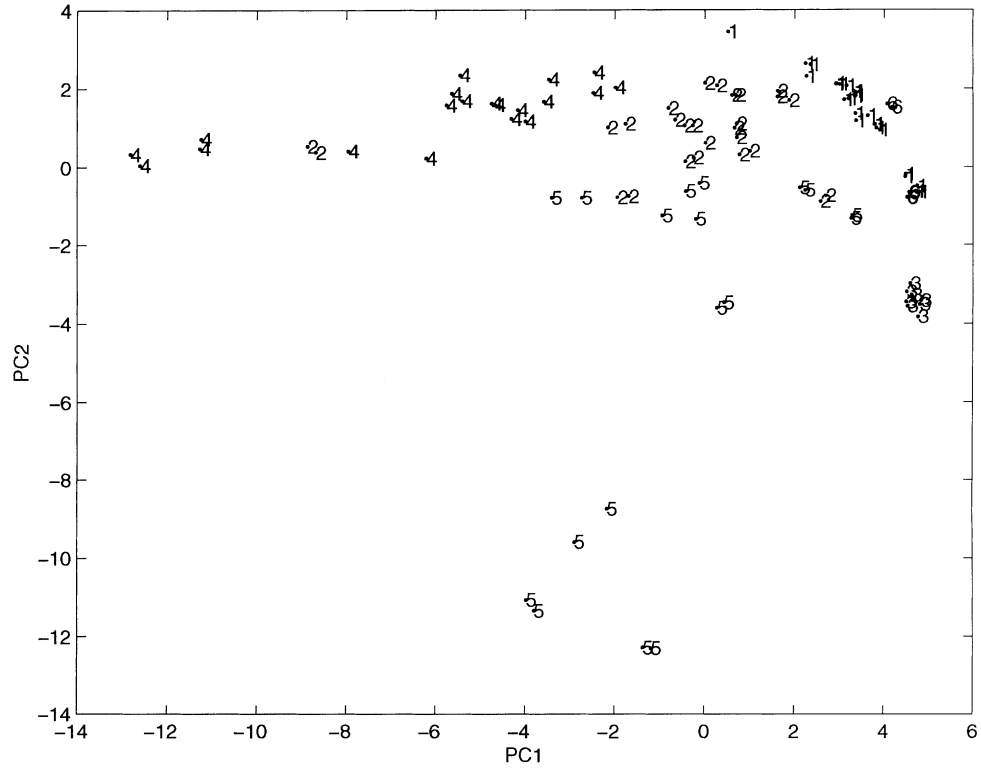

Fig. 7. A plot of the two largest principal components of the 48 GC peaks obtained from the 133 SPE gas chromatograms. Each gas chromatogram is represented as a point in the plot (1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5, and 6 = AVGAS). The two largest principal components explain 70% of the total cumulative variance.

set were correctly classified using the discriminant developed from the gas chromatograms of the neat jet fuels. The only misclassification involved a JP-5 fuel recovered from a monitoring well at Keywest Naval Air Station. The regularized discriminant recognized the fuel as Jet-A. (It classified the fuel sample as Jet-A with a 65% probability; the remaining 35% of the probability was assigned to the JP-5 fuel class.) This result is not altogether surprising because of the similarity in the overall hydrocarbon composition of these two fuel materials. Anecdotal data from our laboratory suggests that some JP-5 fuels are simply Jet-A fuels containing different surfactant additive packages. Clearly, the high classification success-rate obtained for the weathered jet fuels suggests that information about fuel-type is present in the 22 GC peaks identified by the pattern recognition GA. Furthermore, the potential of using gas chromatography to differentiate JP-4 (principal USAF fuel prior to 1990) and JP-8 (currently the principal USAF Fuel) from Jet-A and JP-5 has been demonstrated.

5. Dissolved hydrocarbons

Fig. 7 shows a plot of the two largest principal components of the 48 GC peaks obtained from the 133 SPE gas chromatograms. Each gas chromatogram is represented as a point in the plot (1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5, and 6 = AVGAS). JP-4 and JP-7 yield well define clusters, well separated from the gas chromatograms of the other fuels. The overlap of JP-5, Jet-A, and JPTS fuel samples in the PC plot is not surprising because of the similarity in the physical and chemical properties of these fuels, e.g. flash point, freezing point, vapor pressure, and distillation curve [22]. Mayfield and Henley [15] observed that gas chromatograms of kerosene-based

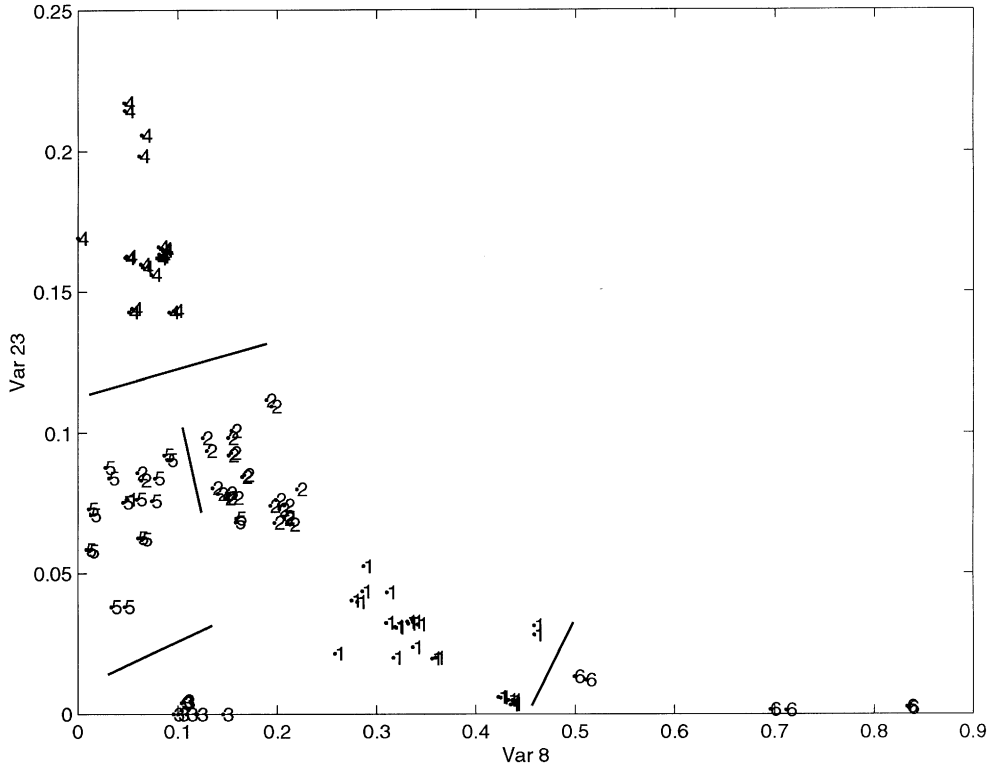

Fig. 8. A plot of the two GC peaks (standardized retention time windows 8 and 23) identified by the pattern recognition GA for fuel spill identification. (1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5, and 6 = AVGAS).

fuels (e.g. Jet-A, JP-5, and JPTS) are more difficult to classify than other types of jet fuels due to the similarity in their overall hydrocarbon composition. Nevertheless, Mayfield and Henley were able to identify fingerprint patterns within the gas chromatograms of kerosene-based jet fuels characteristic of fuel-type, which in turn motivated us to investigate the existence of these types of patterns in SPE data.

The fuel spill identification GA was used to uncover features characteristic of the GC profile of each fuel class. For this study, the GA was configured in the following manner. The number of chromosomes or binary strings in each population, ϕ, was set to 100, whereas the length of each binary string was 48. The momentum, P, was 0.5, and K_c for each class equaled the number of samples in the class. The selection pressure was 0.5 and the mutation rate was 0.1. The maximum number of iterations for a run was set to 100.

After 100 generations, the GA identified two standardized retention time windows (standardized retention time windows 8 and 23) whose plot showed clustering of the fuel samples according to fuel-type (see Fig. 8). Uncovering this solution was possible because the initial population was configured to consist of sparse feature subsets. (The number of features in each feature subset of the initial population can be a critical parameter.) If the feature sets are initially sparse, the probability of including features, which are neither good nor bad, is low since the fitness function does not provide additional points for adding them. On the other hand, the probability of removing these same features as a result of using less sparse feature subsets is also low since there is no advantage to deleting them.

Clearly, information about fuel-type is captured by the gas chromatograms of the water-soluble components of the jet fuels. The ease of classifying jet fuels using selective fractionation becomes apparent when taking into account the fact that an equilibration time of only 3 h is necessary to obtain a reproducible profile of the water-soluble components of a jet fuel [23].

6. Conclusions

The pattern recognition GA, which involves the evaluation, reproduction, and boosting of potential solutions, is well suited for analyzing GC data of fuel spills because of its attributes. First, the GA utilizes a multivariate approach to feature selection ensuring identification of all relevant features. Second, features that contain discriminatory information about a specific pattern recognition problem would be expected to be correlated, which is why feature selection methods based on PCA are preferred. Third, chance classification is not a serious problem since the bulk of the variance or information content of the feature subset selected is about the class membership problem being investigated. Fourth, the PCA 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 are investigated. The fitness function of the GA which combines 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 these different learning approaches.

References

[1] B.K. Lavine, Underground fuel spills, source identification, in: R.A. Meyers (Ed.), Encyclopedia of Environmental Analysis and Remediation, Wiley, New York, 1998, pp. 4923–4938.
[2] B.K. Lavine, Chemolab 15 (1992) 219–230.
[3] M. Sharaf, D. Illman, B.R. Kowalski, Chemometrics, Wiley, New York, 1986.
[4] R.G. Brereton (Ed.), Multivariate Pattern Recognition, Elsevier, Amsterdam, 1992.
[5] B.K. Lavine, A.J. Moores, H. Mayfield, A. Faruque, Microchem. J. 61 (1999) 69–78.
[6] B.K. Lavine, A.J. Moores, J. Bull. Khim. 12 (1997) 73–86.
[7] B.K. Lavine, A.J. Moores, H.T. Mayfield, A. Faruque, Anal. Lett. 31 (15) (1998) 2805–2822.
[8] W.G. MacIntyre, D.R. Burris, Arch. Environ. Contam. Toxicol. 13 (1984) 171–180.
[9] B.K. Lavine, D.M. Brzozowski, J. Ritter, A.J. Moores, H.T. Mayfield, J. Chrom. Sci., submitted for publication.
[10] H.T. Mayfield, W.J. Bertsch, J. Comp. Appl. Lab. 1 (1983) 130–137.
[11] A.J. Roberts, USAF OEHL Report 84-146sZ111CFC, USAF Occupational and Environmental Health Laboratory, May 1984.
[12] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[13] Z. Michalewicz, Genetic Algorithms + Data Structure = Evolution Programs, 3rd Edition, Springer, Berlin, 1996.
[14] M. James, Classification, Wiley, New York, 1985.

[15] H.T. Mayfield, M. Henley, in: J.R. Hall, G.D. Glayson (Eds.), Monitoring Water in the 1990s: Meeting New Challenges, ASTM, Philadelphia, PA, 1991, pp. 578–597.
[16] B.K. Lavine, A. Stine, H.T. Mayfield, Anal. Chim. Acta 277 (1993) 357–367.
[17] B.K. Lavine, H. Mayfield, P.R. Kroman, A. Faruque, Anal. Chem. 67 (1995) 3846–3852.
[18] S. Wold, M. Sjosterom, in: B.R. Kowalski (Ed.), Chemometrics: Theory and Applications, ACS, Washington, DC, 1982, pp. 243–282.
[19] J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine, A.M. Harper, Anal. Chem. 57 (1985) 295–303.
[20] P.J. Gemperline, L.D. Webber, F.O. Cox, Anal. Chem. 61 (1989) 138–144.
[21] I.E. Frank, J.H. Friedman, J. Chemo. 3 (1989) 463–475.
[22] Handbook of Aviation Fuel Properties, Coordinating Research Council, Atlanta, GA, 1983.
[23] B.K. Lavine, Solid phase micro-extraction applied to the problem of fuel spill identification, Summer Faculty Research Program, Air Force Office of Scientific Research, Bollings Air Force Base, Washington, DC, August 1995.