ELSEVIER

Journal of Chromatography A, 775 (1997) 231–242

JOURNAL OF CHROMATOGRAPHY A

Chromatography pattern recognition of Aroclors using iterative probabilistic neural networks

Glenn R. Magelssen, John W. Elling*

Los Alamos National Laboratory, Mail Stop J580, Los Alamos, NM 87545, USA

Received 21 August 1996; revised 23 December 1996; accepted 2 January 1997

Abstract

The purpose of this article is to demonstrate the application of an iterative probabilistic neural network (PNN) as a classification tool in the analysis of multicomponent environmental samples of Aroclors. The PNN is a neural network implementation of a Bayes classifier. This network is incorporated into an iterative method for classifying Aroclor samples. The performance of the method is demonstrated using experimental gas chromatograms of Aroclors, Aroclor mixtures, and random noise. This technique is compared with standard chromatography data processing procedures and linear regression pattern recognition and found to be more accurate and more sensitive for component identification. The method is appropriate for use in routine environmental screening applications in which the presence or absence of one or more Aroclors must be determined in the presence of interfering signals. © 1997 Elsevier Science B.V.

Keywords: Probabilistic neural networks; Neural networks; Chemometrics; Aroclors; Polychlorinated biphenyls

1. Introduction

Aroclor is the trade name given to complex mixtures of polychlorinated biphenyls (PCBs) that were manufactured in the USA by Monsanto between 1929 and 1978. Aroclors are complex mixtures of chlorinated biphenyls because each of the 10 positions on the biphenyl molecule may be substituted with either chlorine or hydrogen. Theoretically, 209 different chlorinated biphenyls are possible. Typically, 30 to 50 of the 209 possible congeners are major components of each Aroclor. The various Aroclors produced differ in the mixture of congeners. Most Aroclors are given a numerical designation beginning with 12, denoting the 12 carbon biphenyl ring, and ending with two digits expressing the percentage by weight of chlorine in the Aroclor product. Thus, 42% of the average molecular mass of the PCBs in Aroclor 1242 is chlorine and the mixture averages 3.2 chlorine atoms per molecule. Aroclors 1254 and 1260 average approximately five and six chlorine atoms per biphenyl, respectively.

The relative abundance of the chlorinated biphenyls in an Aroclor mixture give rise to a unique signature (fingerprint) of peak areas and retention times (peak profile) in the chromatogram of a sample of that Aroclor. Classifying the Aroclor present in an environmental sample from this fingerprint chromatographic peak profile poses a number of challenges. Fingerprint identification is complicated by sample contamination which causes elevated

*Corresponding author. Tel.: (+1-505) 665-3047; fax: (+1-505) 665-3911.

0021-9673/97/$17.00
PII S0021-9673(97)00052-6

baselines and interference with individual chlorinated biphenyl peaks by co-eluting compounds in the sample. Sample classification is also impeded by Aroclor weathering, which changes the relative abundance of the chlorinated biphenyls in the mixture and so distorts the characteristic fingerprint.

Classifying each Aroclor in environmental samples contaminated with multiple Aroclors is a more difficult problem than identifying a single Aroclor contaminant. Because Aroclor mixtures contain many of the same chlorinated biphenyl congeners, the areas of the peaks from these components cannot be uniquely assigned to any Aroclor. Identification of each component becomes particularly difficult when there is a large (five-fold or more) difference in the concentrations of each Aroclor.

Currently, the manual analysis of environmental Aroclor samples is carried out with a two-step process involving the initial subjective identification of the sample from its chromatographic peak profile followed by the quantization of the Aroclors using the area of isolated, characteristic peaks [1]. This manual process depends on the skill and experience of the analyst, and results are highly variable and error prone [2]. Pattern recognition analytical techniques have the potential to facilitate routine environmental analyses for multicomponent target materials like Aroclor samples by identifying the fingerprints of the target materials with less subjectivity. Furthermore, automating the multicomponent identification in the data interpretation process requires pattern recognition methods that emulate the intuitive analyses performed by experienced technicians [3].

Pattern recognition analysis of chromatograms is an effective tool in the analysis of complex biological and environmental samples. Typical pattern recognition applications in chromatographic data interpretation involve sample classification problems. Multivariate analysis techniques have been used to chromatographically classify fuel samples [4–6], fish oils [7], essential oils [8,9], orange juice [10], and whiskeys [11], to name only a few applications. In these applications, the empirical relationships in a data set of chromatograms from samples of known classification (a training set) are determined with either an unsupervised learning or a supervised learning process. The identified relationships are then used to classify an unknown sample as a member of one or more of the groups identified in the initial chromatogram training set. In these applications, experimental and systematic variations between the training data set and the unknown sample must be avoided to minimize confounding of the desired grouping information.

To be a useful tool in routine environmental testing of Aroclors, pattern recognition applications must work within the restrictions of available data and expected distortions. One restriction is that in the environmental analysis for Aroclor contamination, the assumption that the unknown sample can be classified as a member of a single group in the training data cannot be made. This assumption, however, is implicit in many chromatography pattern recognition applications that seek to identify the origin of a sample [5,7–9]. In routine environmental analysis, Aroclors can be absent from the sample or the sample can contain several Aroclors. In the case that the Aroclor is not present in the sample, the absence must be determined unambiguously by the pattern recognition application. In the case of multiple Aroclors, the fingerprint from each Aroclor in the gas chromatogram must be recognized and the sample must be classified accordingly. Statistical discriminant analysis methods and principal component analysis methods have been developed to recognize individual multicomponent target materials in a mixture of multiple targets [4,12,13]. These methods, by design, categorize the sample as a member of one or more target groups, relying on error estimates in the results to identify erroneous classifications when the target material is not present in the sample.

A second difficulty in the use of pattern recognition for routine environmental screening analyses is the need for an extensive training data set from which the classification relationships are determined [2]. Obtaining a large training set each time an instrument is calibrated is a time-consuming process that is not feasible for routine use in environmental testing laboratories, in which the cost per unknown analysis is of primary concern. From a practicality standpoint, pattern recognition methods must be trainable using no more than the standard single-component calibration data sets [1].

We have developed an iterative PNN pattern recognition method within these restrictions that is amenable to routine use in environmental testing of Aroclor samples. The method is designed for screening analyses of unknown samples containing none, one, or multiple target groups (demonstrated in this work with up to three different Aroclors) in the presence of a chromatographic response from environmental background. The technique uses as its training data set only the standard single-component chromatograms collected in the process of building calibration curves for conventional, manual analysis.

The purpose of this paper is to demonstrate the use of the iterative PNN as a tool for the classification of chromatographic Aroclor data. We show that the method effectively identifies Aroclors and Aroclor mixtures in environmental samples in the presence of peak overlap, missing peaks, and minor contaminants. The PNN method performance on weathered Aroclor samples is not addressed in this work.

The PNN architecture is distinct from the standard back-propagation neural network architecture and typically provides superior performance in classification applications [5,14,15]. A description of probabilistic neural networks is given in the next section. Section 3 describes the production of the experimental chromatograms used to train and test the network as well as the production of the random noise test data. In Section 4 the method used to create the inputs to the PNN from the peak tables generated from the gas chromatograms is described. Also in this section the iterative PNN application is described that addresses the problems associated with Aroclor mixture analysis that causes difficulties with the simple PNN technique. The results of PNN, the iterative PNN analyses, the linear regression analysis, and the standard analysis techniques are presented in Section 5. The final section provides a summary and conclusions.

2. Probabilistic neural networks

The PNN provides a general technique for solving pattern classification problems. In mathematical terms, an input vector, often referred to as a feature vector, is used to determine a category. For example, the spectral energy values from a sonar system can be represented as a feature vector, and based on these values a prediction can be made as to whether a signal is from a ship, submarine, or another source. Neural net classifiers are trained by being shown data of known classifications. The PNN uses the training data to develop distribution functions that are in turn used to estimate the likelihood of a feature vector being within the given categories.

Optionally, this can be combined with the a priori probability of each category to determine the most likely category for a given feature vector. If the relative frequency of the categories is unknown, then all categories can be assumed to be equally likely, and the determination of a category is then solely based on the closeness of the feature vector to the distribution function of a category.

Specht developed the PNN, and has described it in Refs. [14,15]. The PNN represents a neural implementation of a Bayes classifier, where the class-dependent probability density functions are approximated using a Parzen estimator. Since a Bayes classifier provides an optimum approach to pattern classification in terms of minimizing the expected risk, and since Parzen estimators asymptotically approach the true underlying class density functions as the number of vectors increases, PNN provides a very general and powerful classification paradigm when there is adequate data of known classification.

Suppose a classification problem has k classes, and suppose that the data on which decisions are based is represented by an N-dimensional vector

$$\vec{v} = (\nu_1, \nu_2, \dots \nu_N).$$

Let

$$f_1(\vec{v}), f_2(\vec{v}), \dots, f_k(\vec{v})$$

be the probability density functions (PDFs) of the class populations and let

$$\omega_1, \omega_2, \dots, \omega_k$$

be the a priori probabilities that a vector will lie in a given class. Then the Bayes decision rule compares the k values

$$\omega_1 f_1(\vec{v}), \omega_2 f_2(\vec{v}), \dots, \omega_k f_k(\vec{v})$$

and chooses the class corresponding to the highest value. Loss functions can also be factored into the calculations, but the crux of the decision rule is to evaluate the multivariate class PDFs at the given vector, weight them, and compare them.

This decision rule depends on knowing the class PDFs. The Parzen estimator is a non-parametric method of estimating PDFs which makes no assumption about the nature of the distribution. The Parzen estimator is built up from a set of smaller parametric functions, typically Gaussian multivariate functions. In essence, a small Gaussian curve is found for each training vector, then the curves are added together and smoothed.

The Parzen estimator used in PNN is of the form [16]

$$f_k(v) = \left(\frac{1.0}{2\pi\sigma^2}\right)^{(N+1)/2} \cdot \frac{1}{T_k}\sum_{j=1}^{T_k}\{\exp\left(\frac{-D_j^2}{2\sigma^2}\right)\} \qquad (1)$$

where

$$D_j = \|(v - V^j)\| = \sqrt{\sum_{i=1}^{N}(v_i - V_i^j)^2}$$

$$V^j = (V_1^j,...,V_N^j)$$

is the jth vector in class k, T_k is the number of training records in class k, and $\sigma = \sigma(T_k)$ is a smoothing parameter which must satisfy

$$\lim_{T_k \to \infty} \sigma(T_k) = 0$$

and

$$\lim_{T_k \to \infty} (T_k\sigma(T_k)) = \infty$$

One way to satisfy this is to define

$$\sigma = \frac{S}{T_k^{E/N}}$$

where E is a constant varying between 0.0 and 1.0 and S, the sigma scale, is optimized when σ is optimized in the training process. Typically, E is set to 0.5, which is the value used in our networks. The summation terms in Eq. (1) are referred to as Parzen kernels, and each kernel is implemented as a principle element in the pattern layer of PNN. In the standard pattern unit, D_j is defined to be the Euclidean distance between the input vector and the stored center V^j.

The NeuralWork (Neuralware, Pittsburgh, PA, USA) neural network software toolkit is used to develop this PNN application. The PNN architecture developed in this work has an input layer with nodes for the retention times of two peaks and the ratio of their peak areas (described below), a pattern layer with clustering, a summation layer that sums the Parzen kernels for each class, and an output layer that calculates the normalized probabilities from the values in the summation layer, as shown in Fig. 1. The output layer of this PNN provides the average value of the positive response of the connected summation nodes over a series of vectors that are applied to the input layer.

Ordinarily, the number of weights in the pattern layer is equal to the number of training vectors. As a result, if the number of training vectors is large, the network could become both computational- and memory-intensive. Clustering is a method for reducing the number of weights [16]. The clustering procedure works in the following way. One starts constructing the set of Gaussian kernels of Eq. (1) by introducing training vectors to the network. After the first few kernels have been set up, the kernel created by an incoming training vector is compared to those already established in the same class. The comparison is made by finding its distance to the closest center of a previously created kernel in the same class. If this distance is less than a radius of

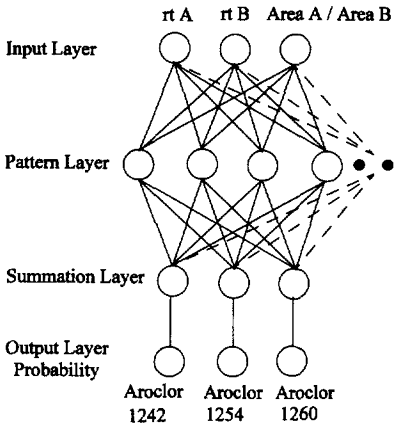

Fig. 1. Neural network architecture. The number of nodes in the pattern layer is optimized in the training process. Two peak retention times and the ratio of the peak areas are applied to the input node. Each output node provides the normalized probability of each class in the training data.

influence R, then instead of establishing a new kernel in that class, the closest kernel already established is used. Thus, the number of weights can effectively be reduced by an appropriate choice of R. Optimization of the σ and R values is described below.

3. Chromatography data

Chromatograms of standard Aroclor mixtures (Supelco, Bellefonte, PA, USA) were generated on a Fisons Model 8000 gas chromatography instrument (Fisons Instruments, Danvers, MA, USA) equipped with a 20 m×0.25 mm I.D. by 0.1 μm DB-1 column. The injector and detector temperatures were 280 and 350°C, respectively. The oven temperature was 80°C for 1 min, increased 20°C/min to 220°C, increased from 220°C by 12°C/min to 320°C, followed by a constant temperature for an additional 3 min. Carrier gas flow-rate was 1.8 ml/min, with make up gas flow-rate of 25 ml/min. The instrument was controlled and data acquired with PE Nelson Turbochrom software (Perkin-Elmer Nelson Systems, Cupertino, CA, USA). After the chromatograms were acquired, they were translated into the Analytical Instrument Association (AIA) data interchange file format [17] and transferred to a Hewlett-Packard workstation. Typical chromatograms of Aroclors 1242, 1254, and 1260 are shown in Fig. 2. Several sets of standard chromatograms (0.05, 0.10, 0.20, 0.40, and 0.80 μg/ml±0.005 μg/ml for each Aroclor forming that Aroclor's calibration set) and mixture chromatograms were generated under these conditions.

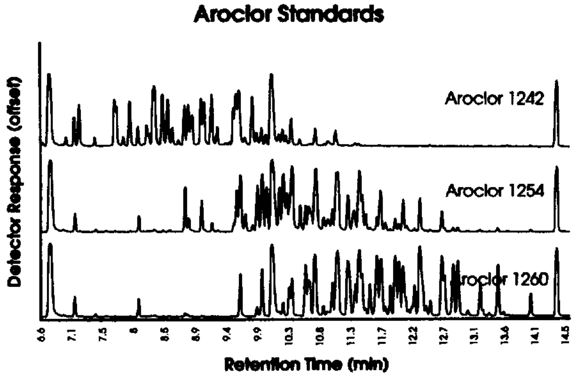

Fig. 2. Typical chromatograms of Aroclors 1242, 1254 and 1260.

Random noise was used to test the resistance of the data analysis method's determinations when processing noisy data. In order to incorporate the effects of integrator distortions, the random data was generated as chromatograms. Under the chromatographic conditions described above, 62 characteristic peak locations were identified at which PCBs in one or more of the three Aroclor standards eluted. One hundred chromatograms were created in which each of the 62 peak locations had a 50% chance of being populated with a peak. Gaussian peaks were generated at the chosen locations with an amplitude randomly distributed between zero and the maximum amplitude observed in the chromatograms of the Aroclor standards at 0.80 μg/ml. The resulting AIA data files were analyzed as unknown chromatograms.

4. Chromatogram processing

4.1. PNN software

The integrator in the Target-3 chromatography data processing software (ThruPut Systems, Orlando, FL, USA) is used to produce peak tables from the time-series chromatogram. These peak tables are processed into input data sets for the network in the form of a set of vectors, which represented the characteristic chromatographic profile of each Aroclor. A given vector consists of three values, the retention time of two peaks and their peak area ratio. Thus, a chromatogram becomes a set of vectors which contains the chromatogram signature. The process is illustrated as follows: Assume that the peak table calculated from a chromatogram contains four peaks and that these peaks are used to calculate the input vectors for the network. Each peak is characterized by its retention time (rt1, rt2, rt3, rt4) and area (area1, area2, area3, area4). Also, for the training data sets, the peaks are sorted in order of decreasing area. Thus, peak 1 has the largest area and peak 4 the lowest. In this simplified example, Table 1 is created from these four peaks and each row of the table is used as an input vector for the PNN network. As shown in Table 1, six input vectors are calculated from the four peaks. The

Table 1
Example set of inputs to the three nodes in the PNN input layer created from four peaks with retention times rt1, rt2, rt3, rt4, and areas area1, area2, area3, area4

Node 1 (retention time)	Node 2 (retention time)	Node 3 (area ratio)
rt1	rt2	area2/area1
rt1	rt3	area3/area1
rt1	rt4	area4/area1
rt2	rt3	area3/area2
rt2	rt4	area4/area2
rt3	rt4	area4/area3

output nodes of the PNN network provide the average value of the probability at the connected summation node as each vector in the vector set for a chromatogram is applied in turn to the input nodes.

The peaks chosen to form the characteristic vector set for an Aroclor (on which the PNN is trained) must be chosen carefully. A linearity criterion and an orthogonality criterion were developed to select the peaks used to form the vector sets for each Aroclor. The linearity criterion introduces the requirement that the area of the peak selected scale linearly with the concentration of the Aroclor in the standard sample from which the chromatogram is generated. This linearity criterion requires that the area of the peak in the chromatograms of the 0.10, 0.20, 0.40, and 0.80 μg/ml standard samples are within 20% of the appropriate multiplier of the area of the same peak in the 0.05 μg/ml chromatogram. This criterion ensures that the peaks selected arises from PCB component(s) of the Aroclor mixture and do not suffer from interference from adjacent peaks. A 30% threshold in this linearity criterion is also used to investigate the sensitivity of the PNN performance to the peak selection process.

Because different Aroclors can contain the same (or different but co-eluting) chlorinated biphenyl species, a chromatogram of a sample with a mixture of Aroclors has peaks whose areas depend on components of each Aroclor. When trying to detect the signature of each Aroclor in a mixture, this interference of other Aroclors is the primary source of errors. To minimize the effect of the overlap on the PNN, the peaks selected to form each Aroclor's training vector set is also required to meet an orthogonality requirement. This orthogonality criterion requires that none of the peaks in the set chosen for one Aroclor may be present at more than 20% intensity (by area) in the chromatograms of other Aroclor standard samples at the same concentration.

Software was developed to automate the process of selecting appropriate peaks and creating the vector sets. The 20% linearity criterion and the orthogonality criterion described above, when applied to the chromatograms of pure Aroclor standards, yields 10 peaks for Aroclor 1242, 9 for Aroclor 1254, and 10 for Aroclor 1260 from our experimental calibration data set. The pairwise comparisons in each Aroclor peak set form 126 input training vectors (45 vectors from 10 peaks and 36 vectors from 9 peaks) for each chromatogram.

Unknown chromatograms are analyzed by calculating the 126-vector set from peaks that match the retention times of the 29 selected peaks. If no peak is found in a 120 s window around the expected retention time, the area of the peak is set to an arbitrary small number. Chromatogram-to-chromatogram retention time reproducibility in the data sets tested was better than 10 s.

The network is trained using the input vectors created from the 15 standard chromatograms: five chromatograms of samples spanning the concentration range for each of the three pure Aroclor standards. In standard back-propagation neural networks, the number of hidden nodes are set in advance and iteratively trained. By contrast, each input training vector becomes a node in the PNN pattern layer (shown in Fig. 1) unless it is similar (within the radius of influence) to a preexisting node. Training the PNN is simply a matter of finding the optimum value of sigma and the radius of influence [16].

The accuracy of the PNN network depends on the accuracy of the value of sigma, determined as described below. The accuracy of the value of sigma can depend on the number of input vectors, but it also depends on how well the vectors separate the different classes [14,15]. To find the optimum value of sigma and the optimum radius of influence in this application, initially we set the radius of influence to zero and chose a sigma less than one. Sigma is varied until the network recognized the Aroclors in the 15 training chromatograms at the 100% level for all three Aroclors and also accurately predicted the absence of Aroclors. When the network probability dropped below the 40% level, the Aroclor is deemed absent (and the training error is zero). This 40% threshold was identified empirically as a threshold that provided a probability of 100% when the chromatogram of a pure Aroclor was analyzed. After a sigma is identified that met these criteria, the radius of influence is optimized. Holding the sigma constant, the radius of influence is increased until the network no longer recognized the Aroclors at the 100% level and/or the probability for an absent Aroclor exceeds the 40% level.

4.2. Standard chromatography data analysis

For the purposes of performance comparison, several other data processing methods are used to process the same Aroclor data. First, the data is processed using the standard commercial chromatographic data processing software in which calibration curves relating peak area to concentration are created using peaks in the individual Aroclor chromatograms [1]. Samples are analyzed with this method by calculating the unknown concentration from the appropriate peak area/concentration calibration line for each peak found that corresponds to a peak in the calibration set for each Aroclor. The reported concentration for each Aroclor is determined to be the average concentration calculated using each peak in that Aroclor's calibration set. Outliers are identified using Chauvenet's criterion and not used to calculate the average or standard deviation that form the reported result for each Aroclor [18]. The standard deviation of the concentrations calculated using each peak is used as a confidence level in the reported answer. When the average concentration is larger than three times the standard deviation from zero, the Aroclor is identified in the sample.

4.3. Linear regression analysis

Linear regression is also used to process the test data sets of Aroclors and Aroclor mixtures. The linear regression analysis software was developed using the Matlab (Mathworks, Nantik, MA, USA) numerical computation system. The peaks selected as representative of each Aroclor are used as the basis vector that was regressed against a vector of matching peaks in the unknown. Since the regression fitting procedure does not require peaks that are unique to each Aroclor, all the peaks that could be attributed to any Aroclor are used in the basis vector. In the three Aroclor calibration sets, 62 peaks are attributable to one or more Aroclors. The three basis vectors (one for each Aroclor) are found by singular value decomposition of the normalized 62 peak vectors in the calibration sets of five chromatograms spanning the concentration range for pure samples of each of the three Aroclors. An Aroclor is deemed to be detected by linear regression analysis when the calculated concentration is both greater than three times the variance in the fit of that basis vector and greater than the 0.05 μg/ml detection limit. The high variance resulting in fits to the random noise inputs are the primary discriminator against false positive detections.

4.4. Iterative PNN

Even using peak sets selected with the orthogonality criterion, the PNN has difficulty identifying Aroclors that are minor components (a factor of five or less in concentration compared to the Aroclor in highest concentration) of Aroclor mixtures. This difficulty arises from the inability to find completely orthogonal peak sets for any Aroclor. Even with the 20% orthogonality criterion that ensures that no other Aroclor has a major overlapping peak, when the concentration difference is large, the Aroclor in the largest concentration provides a large total contributions to some of the peaks in the sets of the other Aroclors in the mixture. In addition, even when there is not a direct overlapping interference, a large area peak adjacent in retention time to a small area peak interferes with the area determination of the small peak [19]. In some severe cases, the integrator does not resolve the small peak and so the area of the large peak, with the sum of the two areas, is used in the input vector calculations. To address this problem, an iterative PNN method was developed. This method subtracts the chromatograms of the Aroclors positively identified by the PNN from the unknown chromatogram and the resulting difference chromatogram is reanalyzed by the PNN in order to identify Aroclors present at lower concentration. Concentrations of the identified Aroclors are determined using a simple method developed by Lea et al. [20], although any method to calculate the concentration can be used.

The first step in the iterative method is to analyze the vector set calculated from an unknown chromatogram with the trained PNN. The result is a probability for each Aroclor on which the network was trained. If one of the output probabilities is greater than 80%, the iterative method is executed. If more than one output probability exceeds 80%, the Aroclor with the highest probability is used. In the iterative method, the concentration of the Aroclor with the highest probability is calculated with the Lea method [20] using the same peak set chosen to calculate the PNN input vectors. The calculated concentration is used to scale the areas of the Aroclor peaks in the peak table of the chromatogram of the 0.20 μg/ml standard, and then this peak set is subtracted from the table of unknown peaks. Following subtraction, the new peak set is used to calculate new input vectors for the Aroclors that were not identified in the first pass and these vectors are applied to the trained PNN. If the new probability of the remaining Aroclor or Aroclors (in the case of a network trained for three Aroclors) increases to 70% or more in the second PNN analysis, the quantization and subtraction process is performed again to remove the effect of that positively-identified Aroclor. Finally (for a network trained to recognize three Aroclors) the presence of the third Aroclor can be investigated by the identification and subtraction process. The final result is the highest probability found for each Aroclor component, at any stage of the iteration.

Consider an example analysis of a sample created with 0.05 μg/ml of Aroclor 1242, 0.20 μg/ml of Aroclor 1254, and 0.80 μg/ml of Aroclor 1260 (all nominal concentrations plus or minus 5%). The chromatogram of this sample is processed into a peak table and the 29 peaks in the peak sets for the three Aroclor targets are found. Applying the 126 vector set calculated from the 29 peaks to the trained PNN results in a output probability of 69% for Aroclor 1242, 29% for Aroclor 1254, and 100% for Aroclor 1260. The concentration of Aroclor 1260 was calculated to be 784±11 μg/ml using the 10 characteristic peaks in the peak set. Subtracting 3.92 times the area of each peak in the Aroclor 1260 chromatogram at 0.20 μg/ml that overlaps the 10 peaks in the Aroclor 1242 set and the nine peaks in the Aroclor 1254 set results in a new 126 vector set for the unknown chromatogram (reusing the 10 Aroclor 1260 peaks with the new values for the Aroclor 1242 and 1250 peaks). When the PNN operates on this input vector, the probability with which Aroclor 1242 is identified increases to 100% and the probability with which Aroclor 1254 is identified increases to 44% (the probability for Aroclor 1260 does not change). Subtracting the estimated contribution of Aroclor 1242 (at 70±20 μg/ml) from the nine peaks in the Aroclor 1254 peak set yields a third 126 vector set. Processing this vector set with the PNN results in the final reported identification probability for Aroclor 1254 of 84%. The final result of the iterative analysis is that Aroclor 1242 and 1260 are identified with a probability of 100% and Aroclor 1254 is identified with a probability of 84%.

5. Results

In Table 2 we give the Aroclor concentrations of the samples from which chromatographic data sets were generated to test this method. The data sets are numbered 1 through 33. There are 12 data sets of a single Aroclor sample, five data sets with a mixture of two Aroclors and 16 data sets with a mixture of three Aroclors. The results of the PNN analysis of the mixture data described in Table 2 are given in the first three columns of Table 3. A neural net probability above 40% indicates that the Aroclor is present. The (non-iterative) PNN predicts 88 correctly with no false positives and 11 false negatives. A false positive occurs when the Aroclor is predicted to be present when it is absent. A false negative is the opposite, the Aroclor is predicted to be absent when

Table 2
Concentration values for the Aroclor mixtures processed by the gas chromatograph

DataSet	Aroclor 1242	Aroclor 1254	Aroclor 1260
1	0.05[1]	0.05[1]	0.05[1]
2	0.10[1]	0.10[1]	0.10[1]
3	0.20[1]	0.20[1]	0.20[1]
4	0.40[1]	0.40	0.40[1]
5	0.80[1]	0.80	0.80[1]
6	0	0[3]	0.80
7	0	0.80	0
8	0.80	0[3]	0
9	0.80[1]	0.80	0.05[1]
10	0	0.05[1]	0.10[1]
11	0.05	0	0.10
12	0.20[1]	0.80	0.80[1]
13	0.20[1]	0.05[1]	0.80
14	0.10[1]	0.80	0.40[1]
15	0.40[1]	0.80	0[3]
16	0.80	0.10[1]	0.20[1]
17	0.80	0.05[1,2]	0.80[1]
18	0.05[1]	0.80	0[3]
19	0.05[1,2]	0.20[1]	0.80
20	0	0.40	0.05[1]
21	0.80[1]	0.20[1]	0.20[1]
22	0.20	0	0
23	0	0.20	0
24	0	0	0.20
25	0.20	0	0
26	0	0.20	0
27	0	0	0.20
28	0.20	0	0
29	0	0.20	0
30	0	0	0.20
31	0.20	0.05[1]	0.40
32	0.20	0.05[1]	0.40
33	0.20	0.05[1]	0.40

Concentrations are given in μg/ml.

[1] False negative determinations by standard analysis.

[2] False negative determinations by linear regression analysis.

[3] False positive determinations by linear regression analysis.

it is present. The false negative results are indicated in Table 3.

As discussed, false negative PNN results occurred when the sample contains mixtures of Aroclors with large differences in concentration, such as in data set 17. Using this iterative approach to identifying minor mixture components results in the successful identification of all components in the data set described in Table 2, eliminating the false negative determinations. The results of the iterative PNN method appear in the last three columns of Table 3. The success of the iterative PNN relies on the insensitivity of the PNN to minor variations in the input ratios. This insensitivity is critical since additional distortions of the area ratios are introduced as a result of subtracting the interfering chromatograms based on the rough concentration estimations provided by the Lea method.

Standard chromatography data processing of the data described in Table 2 (using Target-3) yields 61 correct results and 38 false negative results. The data sets where the standard analysis yielded false negative results are indicated in Table 2. The high rate of false negative determinations arises from the inability to find unique peaks in the Aroclor chromatograms with which to build the area/concentration relationships. When more than one Aroclor contributes a PCB component to a peak used to identify and quantify any Aroclor, the peak area/concentration relationship determined for that peak from the single-Aroclor calibration sets is not valid. The resulting variance in the concentration calculated using each peak increases the standard deviation of the average answer to the point that a negative identification is made. In this standard manual analysis procedure, peaks are chosen that exhibited baseline resolution and the linear behavior over the concentration range in the calibration sets. When the standard analysis is performed with the peaks chosen for their orthogonality between the set of three Aroclors, the analysis of the data described in Table 2 results in 19 false negative results, however, the quantization performance, not reported in this work, degrades substantially.

Processing the data described in Table 2 with the linear regression technique yields 93 correct results, four false positive results, and two false negative results. The data sets on which the erroneous results occur are indicated in Table 2. The false positive results occur when the basis vectors describing one Aroclor has a statistically significant fit to the peak pattern when that Aroclor is not present. These false positive results occur when at least one Aroclor component of the mixture is present at the 0.80 μg/ml highest concentrations. At these high concentrations, even PCBs present in minor amounts generate peaks in the chromatogram that exceed the integrator area cutoff and so appear in the peak table. Since these PCBs do not produce peaks that are

Table 3
Results of PNN neural net analysis

Data set	Single PNN analysis			Iterative PNN analysis		
	Aroclor 1242	Aroclor 1254	Aroclor 1260	Aroclor 1242	Aroclor 1254	Aroclor 1260
1	95	45	83	98	82	96
2	98	45	86	100	80	100
3	95	45	88	100	71	96
4	95	48	86	100	80	93
5	93	48	86	98	78	96
6	12	21	100	24	16	100
7	5	100	29	31	100	2
8	100	33	7	100	11	20
9	93	100	31[1]	100	100	91
10	7	38[1]	95	24	73	100
11	100	26	100	100	26	100
12	57	43	86	96	76	89
13	100	26[1]	100	100	56	100
14	33[1]	55	74	96	89	91
15	76	100	29	100	100	2
16	100	55	91	100	80	96
17	100	33[1]	100	100	69	100
18	24[1]	100	29	100	100	2
19	60	29[1]	100	100	84	100
20	5	79	33[1]	31	91	82
21	100	57	81	100	76	98
22	100	29	19	100	11	20
23	7	100	24	24	100	13
24	7	21	100	13	16	100
25	100	31	12	100	6	16
26	7	100	26	13	100	20
27	5	21	100	11	13	100
28	100	29	12	100	9	16
29	7	100	21	13	100	6
30	7	21	100	13	16	100
31	100	26[1]	100	100	67	100
32	100	29[1]	100	100	62	100
33	100	29[1]	100	100	87	100

Neural net results are given as percent probabilities.

[1] False negative determinations by PNN analysis.

major features of the chromatograms of the Aroclor at all concentrations, they are not included in the basis vector calculated for that Aroclor and so are fit using other basis vectors of other Aroclors in the linear regression fitting.

In order to test the sensitivity of the PNN method to the requirement that peaks be linear in the concentration range, a second set of vectors was created using the same orthogonality criterion, but relaxing the linearity criterion to 30%. With this relaxed criterion, sets of 17 peaks for Aroclor 1242, 9 for Aroclor 1254, and 13 for Aroclor 1260 were selected. From these peak sets, 250 vectors were calculated for each chromatogram. The PNN trained with 250 vectors calculated for each training chromatogram gave 89 correct predictions with two false positives and eight false negatives. The false negatives occurred for Aroclor 1254 in data sets 13, 17, 19, 31, 32, and 33 in Table 2, and occurred for Aroclor 1260 in data sets 9 and 20, the false positive determinations occurred on Aroclor 1242 in data set 6 and Aroclor 1254 in data set 8 of Table 2. The iterative application of this PNN did not result in any false negative or false positive results, demonstrating that peak linearity is not critical to the PNN performance.

To be useful for environmental screening analyses and automated analyses, data processing techniques must also recognize the absence of Aroclors in the sample. In order to test the resistance of these data processing methods to false positive determinations, the 100 randomly-generated artificial chromatograms described above were analyzed. Standard data processing with Target-3 yielded 21 false positive determinations out of the 300 possible determinations. Analyzing the chromatograms with the linear regression technique described above yielded 187 false positive determinations. The PNN technique and the iterative PNN method both gave no false positive results. These results are consistent with our experience in the analyzing of non-Aroclor-contaminated soil and oil samples in our laboratory.

5.1. Summary and conclusions

The method described in this paper is one of many possible approaches to the Aroclor classification problem. The PNN network was developed because it satisfied the constraints for typical use in environmental testing laboratories. Most important, it can be trained with a small data set (for example, the standard calibration data set normally collected when building a conventional analysis method) and it is resistant to false positive classifications. The method can learn nonlinearities in the data as is evidenced by the results of Section 5. It can be used on environmental samples containing one or multiple numbers of Aroclors. Finally, because the PNN result is a probability of identification, the result is naturally interpreted as a target identification. The format of the output as an identification probability allows a new result category, the tentative positive. The PNN technique is being incorporated into a system for automated analysis. In this system, the third tentative positive result category is defined to be a probability between 40% and 70%. Tentative positive results will indicate that the result needs further validation before a final determination can be made.

We have shown that probabilistic neural networks are an effective classifier of unknown samples containing no Aroclors, a pure Aroclor and Aroclor mixtures. However, in samples containing a mixture of Aroclors at widely different concentration levels, all three methods described in this paper (the PNN network, the Linear regression method, and the standard method) have difficulty classifying the sample correctly. Furthermore, the linear regression and standard analysis techniques exhibit a high rate of false positive results on random noise. The iterative application of the PNN described overcomes this difficulty and is able to classify all unknown experimental and artificial data sets presented in this study without error.

Further studies will investigate techniques for identifying Aroclors in the presence of weathering (volatilization loss) and degradation. The approach to weathered data will be somewhat different from that presented here. Algorithms will be developed to create data sets that have the characteristics of weathered data and will be tested on neural networks trained on weathered data. Weathered data is currently being produced and studied at the University of Southern Indiana for this study.

References

[1] US EPA Office of Solid Waste and Emergency Response, Washington, DC, SW 846 method 8000 Gas Chromatography; method 8080 Organochlorine pesticides and polychlorinated biphenyls by gas chromatography, 3rd ed., November 1986.
[2] R.P. Eganhouse, Anal. Chem. 63 (1991) 2130.
[3] J.W. Elling, S.M. Mniszewski, J.D. Zahrt, L.N. Klatt, J. Chromatogr. Sci. 32 (1994) 213.
[4] B.K. Lavine, A. Stine, H.T. Mayfield, Anal. Chim. Acta 277 (1993) 357.
[5] J.R. Long, H.T. Mayfield, M.V. Henly, P.R. Kromann, Anal. Chem. 63 (1991) 1256.
[6] B.K. Lavine, H. Mayfield, P.R. Kromann, A. Faruque, Anal. Chem. 67 (1995) 3846.
[7] S. deJong, Mikrochim. Acta, II (1991) 93.
[8] N. Dimov, A. Tsoutsoulova, Perfumer Flavorist 12 (1988) 45.
[9] M.K. Park, J. Cho, N.Y. Kim, J.H. Park, Anal. Chim. Acta 284 (1993) 73.
[10] P.E. Shaw, B.S. Buslig, M.G. Moshonas, J. Agric. Food Chem. 41 (1993) 809.
[11] L.A. Wilson, J.H. Ding, A.E. Woods, J. Assoc. Off. Anal. Chem. 2 (1991) 248.
[12] D.S. Burdick, W.S. Rayens, J. Chemometrics 1 (1987) 157.
[13] W.S. Rayens, J. Chem. 4 (1990) 159.
[14] D.F. Specht, IEEE Trans. Neural Networks 1 (1990) 111.

[15] D.F. Specht, Neural Networks 3 (1990) 109.

[16] Neural Computing, Neuralware Inc., Penn Center West Building IV, Pittsburgh PA, 1994.

[17] Analytical Instrument Association Chromatography data standard specification, Version 1.0 (1992), 225 Reinekers Lane, Suite 625, Alexandria, VA 22314-2875, USA.

[18] J.R. Taylor, An Introduction to Error Analysis, University Science Books, Mill Valley, CA, 1982.

[19] V.R. Meyer, LC·GC 13 (1995) 252.

[20] R.E. Lea, R. Bramston-Cook, P. Tschida, Anal. Chem. 55 (1983) 626.