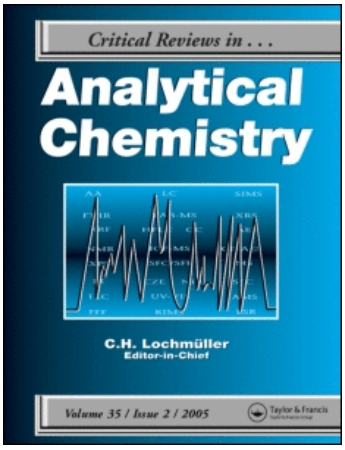

Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

Pattern Recognition

Barry K. Lavine[a]
[a] Department of Chemistry, Oklahoma State University, Stillwater, Oklahoma, USA

To cite this Article Lavine, Barry K.(2006) 'Pattern Recognition', Critical Reviews in Analytical Chemistry, 36: 3, 153 — 161
To link to this Article: DOI: 10.1080/10408340600969411
URL: http://dx.doi.org/10.1080/10408340600969411

Critical Reviews in Analytical Chemistry, 36:153–161, 2006

ISSN: 1040-8347 print / 1547-6510 online
DOI: 10.1080/10408340600969411

Pattern Recognition

Barry K. Lavine
Department of Chemistry, Oklahoma State University, Stillwater, Oklahoma, USA

Pattern recognition methods have been applied to a wide variety of chemical problems. In a typical pattern recognition study, samples are classified according to a specific property using measurements that are indirectly related to the property of interest. 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 can then be used to predict the property in samples that are not part of the original training set. In this review, the three major subdivisions of pattern recognition methodology are discussed and the analytical literature is surveyed. Much of the literature on pattern recognition focuses on novel and not so novel applications. Only the more interesting applications of pattern recognition methods are referenced in this review article.

INTRODUCTION

Many relationships in 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 that confronts an analytical chemist when investigating these types of relationships is two fold. First can a useful structure based on distinct sample groups be discerned? Second, can a sample be classified into one of these groups for the prediction of some property? To develop mathematical models that are suitable for identifying and isolating these groups or classes in multivariate data, analytical chemists have turned to pattern recognition techniques.

Pattern recognition has its origins in the field of image and signal processing where techniques were developed to categorize samples on the basis of regularities in observed data. The first applications of pattern recognition techniques to chemistry were studies involving low-resolution mass spectrometry (1–4), which appeared in the literature in 1969. Modern computers now enable these techniques to be routinely applied to a wide variety of chemical problems such as chromatographic fingerprinting (5–12), and spectral data interpretation (13–15). Numerous books and review articles have been published on this subject (16–24).

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 the property of interest. 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. Burgeoning interest in pattern recognition has prompted extension of these techniques to probe the inverse problem, i.e., given a dataset, are there regularities in the measurements that imply an underlying categorical structure and can they be used predictably.

The set of measurements that describe each sample in the data set is called a pattern, whereas the set of samples for which the property of interest and measurements are known is called the training set. The determination of the property of interest by assigning a sample to its respective class is termed recognition, hence the term "pattern recognition."

For pattern recognition analysis, each sample or individual test object is represented by a data vector $x = (x_1, x_2, x_3, x_4, \ldots . x_j, \ldots x_n)$ where xj is the value of the jth measurement variable, for example the area of a gas chromatographic peak. Each sample can be considered as a point in a high-dimensional measurement space. The dimensionality of the space corresponds to the number of measurements that are available for each sample. A basic assumption is that the Euclidean distance between a pair of points in the measurement space is inversely related to the degree of similarity between the corresponding samples. Points representing samples from one class tend to cluster in a limited region of the measurement space separate from the others. Pattern recognition is a set of numerical methods for investigating data represented in this manner in order to assess its overall structure, which is defined as the overall relation of each sample to every other sample in the data set.

Address correspondence to Barry K. Lavine, Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, USA. E-mail: bklab@chem.okstate.edu

In this review, the three major subdivisions of pattern recognition methodology are discussed: mapping and display, clustering, and classification. Specific emphasis is placed on application of pattern recognition techniques to problems in chemical and biological analyses.

Principal Component Analysis

Chemists often use graphical methods to investigate data. If the number of measurements per sample is only two or three, the data can be displayed as a graph or plot. The chemist can then search for similarities and dissimilarities among samples, find natural clusters, thereby gaining information about the structure of the data by examining the plot. When the number of measurements per sample is greater than three, a two or three-dimensional representation of the measurement space is needed that accurately reflects the distribution of the data points in the high-dimensional space. One approach to this problem of mapping multivariate data involves using a technique called principal component analysis (PCA).

PCA (25, 26) is the most widely used multivariate analysis technique in science and engineering. The overall goal of PCA is to reduce the dimensionality of a data set, while retaining the information present in it. This reduction in dimensionality can be achieved by transforming the original measurement variables into new variables, which are called principal components. Each principal component is a linear combination of the original measurement variables. Often, only two or three principal components are necessary to explain the information content of a data set in which there are a large number of interrelated measurement variables.

Because of the correlations among measurement variables, dimensionality reduction is possible with PCA. Consider a set of samples characterized by two measurements, x_1 and x_2. Figure 1 shows a plot of these data in a two-dimensional measurement space. The coordinate axes or basis vectors of this measurement space are the variables x_1 and x_2. There appears to be a relationship between these two measurement variables. This relationship suggests that x_1 and x_2 are correlated, since fixing the value of x_1 limits the range of values possible for x_2. If x_1 and x_2 were uncorrelated, the enclosed rectangle (shown in Figure 1) would be fully populated by data points. Because information is defined as the scatter of points in a measurement

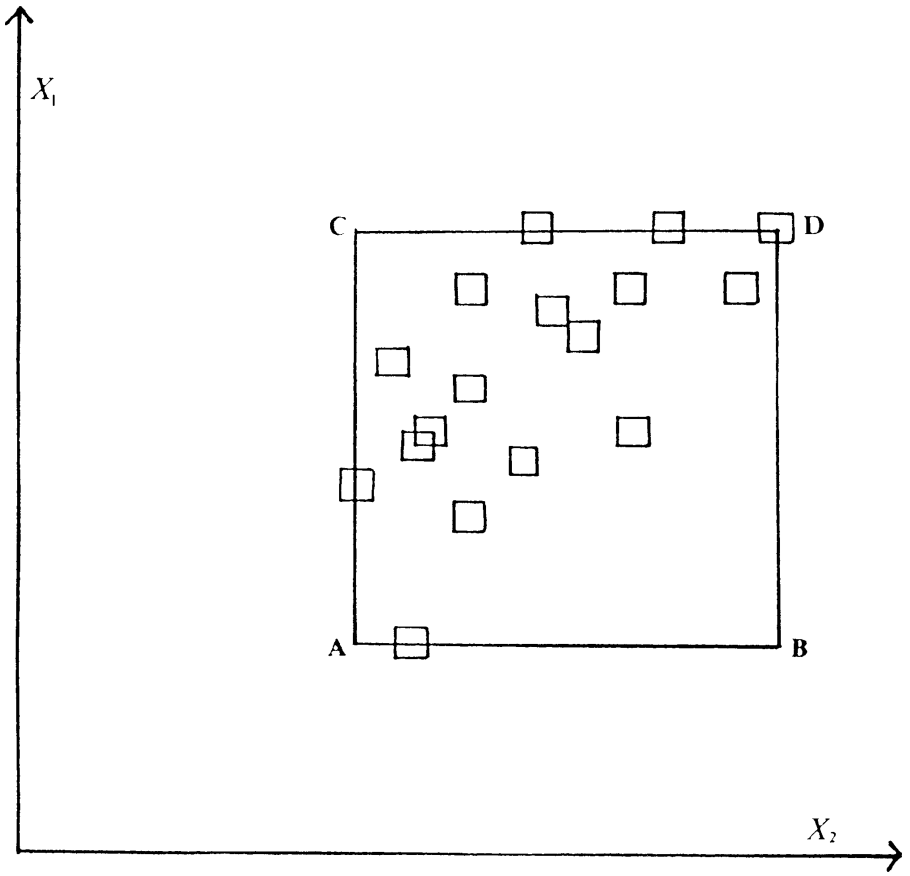

FIG. 1. Fifteen samples plotted in a two-dimensional data space. The data points are restricted to a small region of the measurement space defined by the vertices A-D of the rectangle because x_1 and x_2 are correlated. (Adapted from *NBS J. Res.*, 1985, 190(6), 465–476.)

$$X = \begin{bmatrix} 0.5 & 0.5 & 1.0 \\ 1.9 & 0.7 & 2.6 \\ 2.0 & 2.0 & 4.0 \\ 0.3 & 1.8 & 2.1 \\ 1.9 & 1.7 & 3.6 \\ 1.2 & 0.2 & 1.4 \\ 1.9 & 0.9 & 2.8 \end{bmatrix}$$

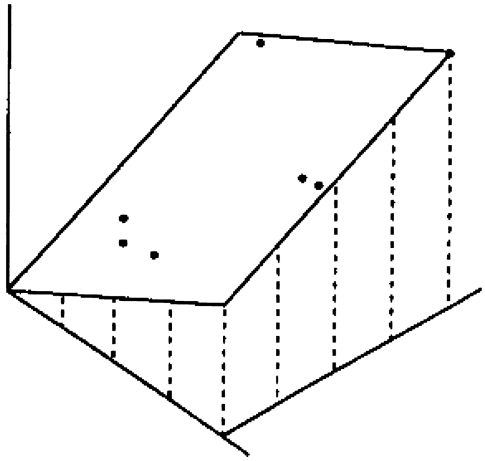

FIG. 2. The sample (data) points may reside in a subspace of the original measurement space in the case of strongly correlated measurement variables. (Adapted from *Multivariate Pattern Recognition in Chemometrics*, Elsevier Science Publishers, Amsterdam, 1992.)

space, it is evident that correlations among the measurement variables decrease the information content of the measurement space. The data points, which are restricted to a small region of the measurement space due to correlations among the variables, may reside in a subspace when the measurement variables are highly correlated (see Figure 2). Variables that are highly correlated or have a great deal of redundancy are said to be collinear.

High collinearity between variables—as measured by their correlation—is a strong indication that a new set of basis vectors can be found that will be better at conveying the information content present in the data than axes defined by the original measurement variables. The new basis set which is linked to variation in the data can be used to develop a new coordinate system for displaying the data. The principal components of the data define the variance-based axes of this new coordinate system. The largest principal component is formed by determining the direction of largest variation in the original measurement space and modeling it via a line fitted by linear least squares (see Figure 3). 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. The number of principal components that can be extracted from the data is smaller of either the number of samples or number of measurements in the data set, as this number defines the largest number of independent variables in our data. A measure of the amount of information conveyed by each principal component is its variance. For this reason, the principal components are usually arranged in order of decreasing variance: the most informative principal component is first, and the least informative is the last.

Principal component analysis takes advantage of the fact that a large amount of data is generated in a calibration or a pattern recognition study. The data have a great deal of redundancy and therefore a great deal of collinearity. Because the measurement variables are correlated, 100-point spectra do not necessarily require 100 independent axes to define the position of the sample points. Employing principal component analysis, the original measurement variables that constitute a correlated axis system can be converted into a system which removes correlation by forcing the new axes to be independent and orthogonal, a requirement that greatly simplifies the data because the correlations present in the spectral data usually allows us to use far

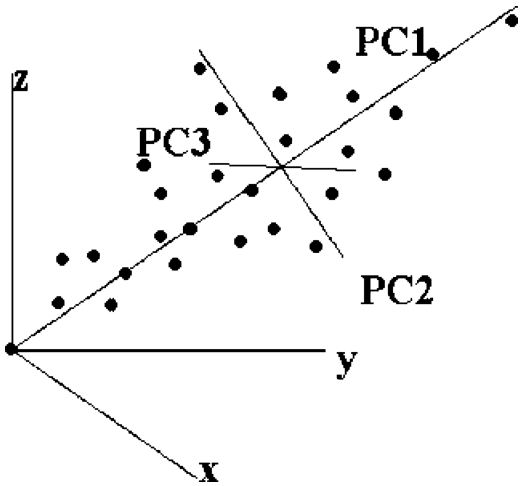

FIG. 3. Principal component axes defining a new set of basis vectors for the measurement space whose coordinate axes are X, Y, and Z. The third principal component describes only noise in the data. (Courtesy of Applied Spectroscopy, 1995, 49(12), 14A–30A with permission.)

fewer axes to represent the sample points. In other words, the spectra for a set of gasoline samples may reside in a subspace of the original 100-dimensional measurement space, and a plot of the two or three largest principal components of the data can help us to visualize the relative position of the samples in this subspace.

PCA has allowed chemist to successfully tackle problems in infrared library searching. Isenhour (27) has shown that data compression by PCA can decrease the time required to perform an infrared library search. A method for decreasing the time required to perform a standard library search based on PCA of infrared spectra has also been developed by Small and coworkers (28). PCA is used to calculate a new set of basis vectors, which dramatically reduce the dimensionality of the original search space. Spectra are projected onto a principal component map of the data where they are represented as a single point in a plane. Library spectra that lie close to the unknown are retained for standard library searching. The use of pattern recognition techniques such as PCA to compress and interpret infrared, NMR and mass spectra has been reviewed (29, 30).

One of the most interesting applications of PCA is chemical imaging, which is a combination of molecular spectroscopy and digital imaging. Data sets generated by chemical imaging are large, multivariate in nature, and require significant processing. Esbensen (31, 32) has shown that pattern recognition techniques can be used to evaluate spectral images generated by the so-called angle measure technique. Using PCA, he was able to differentiate powders by type. Bright (33) utilized PCA to analyze compositional electron microprobe X-ray maps. PCA evaluation of surface enhanced Raman images proved to be a powerful tool for the separation of spectral information of various intracellular components (34). Cosmic spectra spike artifacts were successfully removed from hyperspectral imaging data by taking advantage of the ability of PCA to identify outliers (35).

Booksh (36) has developed an implemented a Raman imaging system that uses principal component mapping to investigate the spatial variations of chemical and physical properties at interfaces in glass-reinforced composites on a colloidal scale. He has demonstrated that PCA analyzed Raman is more interpretable than traditional univariate methods. Sasic (37) and Sekulic (38) reported similar conclusions in more recent studies. Multivariate analysis of visible and near infrared images of works of art using PCA show that it is possible to obtain segmentation and a classification of painted zones by pigments with different chemical composition. This preliminary result has strong implications with regards to the detection of art forgeries (39). A near IR imaging system using a focal plane array detector was developed to differentiate common household plastics by type for recycling purposes. Sorting of plastics by type is crucial to ensure the economic viability of recycling because the most valuable reprocessed plastics are prepared from pure polymer streams. Using PCA to analyze the images, Wienke (40) demonstrated the feasibility of the proposed method.

Another interesting application of PCA is microarrays, which allow for the expression level of thousands of genes to be measured simultaneously. In these studies, mRNA is extracted from a collection of cells and is hybridized on the array, which consists of a probe set for the targeted genes of interest. Each probe has its own well-defined place on the microarray. Identification of an individual hybridization is possible due to the fluorescent tagging of the mRNA. Datasets generated by these arrays consist of a small number of observations (e.g., 20–100 samples) on a large number of measurement variables (e.g., 20,000 genes). Each variable indicates whether a particular gene or protein is under or over expressed. The samples in these data sets have other attributes associated with them such as a class label denoting the pathology of the subject for which the sample was taken.

PCA is ideally suited for tackling classification problems involving microarray data because the principal component plot can provide the user with a visual graphic denoting interclass versus intraclass differences. Because most of the measurement variables in a microarray experiment do not contain information about the pathology of the sample, these uninformative variables can obscure information present in the data about the class membership of the sample. Recently, a genetic algorithm (GA) has been developed that selects features which optimize the separation of the classes in a plot of the two or three largest principal components of the data. Because the largest principal components capture the bulk of the variance in the data, the features chosen by the GA contain information primarily about differences between the classes in a data set. The principal component analysis routine embedded in the fitness function of the GA acts as an information filter, significantly reducing the size of the search space since it restricts the search to feature sets whose principal component plots show clustering on the basis of class. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a smart one pass procedure for feature selection, classification, and prediction. The efficacy of this procedure has been demonstrated in several studies [41–48] including cancer classification by gene microarrays (41) and differentiation of bacteria at the subspecies and strain level by multiarray sensors (42).

Cluster Analysis

The basic objective of cluster analysis to determine the structural characteristics of a data set by organizing the data into subgroups or clusters. These methods are based on the principle that the distances between pairs of points in the high dimensional measurement space are inversely related to their degree of similarity. Although several different types of clustering algorithms exist, e.g., K-means, FCV, and Kohonen neural networks, hierarchical methods are by far the most popular because of the intuitive nature associated with the interpretation of the results. The starting point for this algorithm is a distance or similarity matrix. Distances between all pairs of points in the data set are computed, and the resulting distance matrix is scanned to

identify the nearest pair. These two points are combined to yield a new point which is located midway between the two original points. The distance matrix for the data set is recomputed. This involves measuring the distances between this new point and every other point in the data set. Again, the distance matrix is scanned for the smallest value, and the nearest point pair is combined to form a single point. This procedure is repeated until every point has been linked. The result of this procedure is a diagram called a dendogram, which is a visual representation of the relationships between the samples in a data set. The dendogram will have information about the number of clusters present in the data. However, the interpretation of the data structure will very much depend upon the criteria used to assess similarity. There are also a variety of ways to cluster the data depending on how the distance between a data point and a cluster of points is measured, which gives this algorithm considerable flexibility when probing the structure of a data set.

Using hierarchical clustering, Ramos (49) was able to differentiate various species of Mycobacterium on the basis of HPLC profiles of their mycolic acid profiles. Goodacre (50) has demonstrated that FT-IR fingerprinting when coupled with cluster analysis can detect changes in the global metabolic profiles of plants in response to biotic interactions. He has also shown that hierarchical clustering of surface enhanced Raman spectra of bacteria isolates associated with urinary tract infections can be discriminated to strain level for a sample grouping of *Escherichia coli* (51). One of the more interesting applications of hierarchical clustering is the sensor array, which allow for the identification, classification, and in some cases quantification of organic vapors. Unlike traditional chemical sensing, an individual sensor is not highly selective towards the analyte of interest but the pattern of the array's response can be used to differentiate various analytes. When coupled with hierarchical clustering, a colorimetric sensor array for detection of organics in water, which consisted of hydrophobic dyes on a hydrophobic membrane, could readily distinguish subtle structural features, e.g., primary versus secondary versus cyclic amines (52). A quartz crystal microbalance sensor array was developed to evaluate volatile degradation compounds found in used engine oil (53). The headspace of new and used petroleum products was sampled by the sensor array. Hierarchical clustering of the data revealed patterns that were characteristic of new and used oil. Furthermore, the new oils clustered into groups separated by mileage. Using conducting polymer sensor arrays, Sadik (54) has demonstrated that chlorinated organic phenols can be readily differentiated from other volatile and semivolatile chlorinated compounds when clustering techniques based on differences in Euclidean distances between pairs of points are used.

Classification

The overall goal of any pattern recognition study is the development of a classification rule that can accurately predict the class membership of an unknown sample. For this reason, classification methods are consistently used to analyze multivariate data. Unlike clustering or mapping and display techniques, these methods use *a priori* information about the class assignment of the samples. Although clustering and mapping and display techniques are powerful methods for analyzing the structure of a data set, they are often not sufficient for developing a classifier.

Classification methods can be divided into two distinct groups: partition based methods, and similarity based methods. Partition based methods, e.g., multilayered feed forward neural networks and support vector machines (55, 56), divide a data space into different regions. In the simplest case, that of a binary classifier, the high dimensional measurement space is divided into two regions. Samples that share a common property are found on one side of the decision surface whereas those samples comprising the other category are found on the other side. The decision surface is developed using a training procedure in which the internal structure of the network or machine is adjusted empirically to obtain a best match between the output and the anticipated result for a set of input data that serves as the training set. Neural networks and support vector machines can discern subtle patterns in noisy and nonlinear data. However, spurious or chance classification can be a serious problem. To successfully exploit the advantages of neural networks, it is necessary to properly train the network (57, 58) using both a validation set and a training set with the appropriate point distribution in the data space. This may not always be feasible due to the limited size of the data sets routinely generated by analytical chemists. Although only a few publications have appeared in the analytical literature on applications of support vector machines (59, 60), it is anticipated that interest in these methods will be burgeoning in the near future because of their superior predictive power due to transverse learning (61). Recently, Sadik and Land (62) have demonstrated a significant increase in classification accuracy of a sensor array towards organophosphate nerve agent simulants when support vector machines were used compared to artificial neural networks.

Classification of data by neural networks remains an important subject as evidenced by the fact that artificial neural networks had the second largest number of citations among pattern recognition techniques in the Chemical Abstract database. (PCA had the largest number of citations.) Although it is not practical to discuss all of these applications, it is worthwhile to present a few representative studies that will give the reader a flavor for the scope and breadth of this technique. Artificial neural networks have been used to interpret two-dimensional gel electrophoresis spot patterns (63) obtained from Streptomyces coelicolor and controls and for classifying modified starches from their IR spectra (64). Harrington (65) has trained a neural network to predict the toxicity of organic phosphate pesticides by identifying the active substructure and then using the neural network to screen GC/MS data for environmentally hazardous compounds. Neural networks have been used to classify clusters in a scatter diagram formed from the images in a multispectral data set (66). Authentic and adulterated orange juice samples have been correctly identified using a neural network developed

from the concentration of flavanones and trace element components of the juice samples (67). Potatoes have been classified by type using isoelectrophoretic focusing patterns and feed forward neural networks (68). Wines can be classified by geographic region based on their trace metal content using an artificial neural network (69).

Similarity based classifiers use the Euclidean or Mahalonobis distance to classify samples in a data set. In K-NN, a sample is classified according to the majority vote of its k-nearest neighbors, where k is an odd number, e.g., 1, 3, or 5. For a given sample, the Euclidean distance is computed between the sample and every other point in the data set. These distances are arranged from smallest to largest to define the sample's k-nearest neighbors. Based on the class label of the majority of the sample's k-nearest neighbors, the sample is assigned to a class in the data set. If the assigned class and the actual class label of the sample match, the sample is considered correctly classified. The overall classification success rate is calculated over the entire set of points. Although K-NN cannot furnish a statement about the reliability of a particular classification, the 1-nearest neighbor rule has an error rate that is twice as large as the Bayes classifier, which is the optimum classifier for any set of data. For this reason, K-NN is often used as a benchmark against which to judge other methods.

In linear discriminant analysis (LDA), an observation is assigned to the class with the smallest discriminant score, which is the square of the Mahalonobis distance between the sample and the class centroid. In LDA, each class is assumed to have the same correlation structure. The inverse of the covariance matrix is computed by first estimating the covariance matrix for each class in the data set and then averaging these matrices to yield a pooled estimate. When the covariance structure for each class is different, quadratic discriminant analysis (QDA) is used. The sample's discriminant score is determined for each class by first computing the square of the Mahalonobis distance between the sample and the class centroid. The log of the determinant of the covariance matrix, which conveys information about the size, shape and orientation of the data cloud for each class is then computed and added to the square of the Mahalonobis distance to obtain the discriminant score.

A sample is assigned to the class with the smallest discriminant score. For correlated data, the square of the Mahalonobis distance often contains as much information as the discriminant score. Both LDA and QDA can provide the user with a probabilistic statement about the reliability of a classification. For classes that are well separated, LDA will perform well, even when the classes do not have the same covariance structure. However, when there are more features than samples in the data, maximum likelihood methods fail to provide a reliable estimate of the inverse of the covariance matrix. In these situations, covariance stabilization is crucial. Wold (70) was the first to address the issue of covariance stabilization in discriminant analysis by developing a biased estimator for the covariance matrix. His method is called SIMCA. Friedman and Frank (71) later revestigated the issue of covariance stabilization and formulated a different approach to the development of biased estimators of the inverse of the class covariance matrix. Their method is called regularized discriminant analysis (RDA). From a practical standpoint, RDA will generally outperform SIMCA, LDA, or QDA since the biased estimates of the inverse of the covariance matrix are more stable than those obtained in SIMCA and LDA and QDA are actually special cases of RDA. Recently, Lewis (72) has evaluated QDA, LDA, K-NN, SIMCA and RDA using data from an array of 20 compositionally different carbon black polymer composite chemiresitor vapor detectors, and observed that RDA was the best performing discriminant. Stetter (73) has reported that neural networks can outperform K-NN when instrumental parameters associated with the sensor array are taken into account during the training. Lavine (74) has demonstrated that underground fuel spills can be traced to their source using discriminant analysis and that SIMCA and RDA outperform feed forward neural networks.

Practical Considerations

The choice of the training set is important in any pattern recognition study. Each class must be well represented in the training set. Experimental variables must be controlled or otherwise accounted for by selection of suitable samples that take into account all sources of variability in the data, e.g., lot-to-lot variability. Experiment artifacts such as instrumental drift or sloping baseline must be minimized. For this reason, there is strong interest in preprocessing techniques such as wavelets (75) since they can leverage information from data through denoising and compression as well as contribute to the development of new algorithms to mine information from multivariate data. Pattern recognition methods generally perform better when wavelet based descriptors are used instead of the original data.

Features that contain information about differences in the source profile of each class must be present in the data. Otherwise, the classifier is likely to discover rules, which do not work well on samples that are not part of the original data set. Even if these features are present in the data, feature selection is often crucial to ensure a successful pattern recognition study, since a large number of irrelevant features can introduce so much noise that a good classification of the data may not be obtained. When these irrelevant features are removed, a clear and well-separated class structure is often found. The deletion of irrelevant variable is therefore an important goal of feature selection.

For averaging techniques such as K-NN or discriminant analysis, feature selection is vital since signal is averaged with noise over a large number of variables with a loss of discernible signal amplitude when noisy features are not removed from the data. With neural networks, the presence of too many irrelevant features may cause the network to focus its attention on idiosyncrasies of individual samples due to the nets ability to approximate a variety of complex functions in higher dimensional space. For complex chromatographic and spectroscopic data sets, it is important to identify and delete features that contain information

about experimental artifacts or other systematic variations in the data not related to legitimate differences between the classes represented in the study. In many studies, it is inevitable that relationships exist among sets of conditions used to generate the data and the patterns that result. The existence of these complicating relationships is an inherent part of fingerprint type data (76, 77). If the basis of classification for samples in the training set is other than legitimate group differences, unfavorable classification results can be obtained for the prediction set despite a linearly separable training set. Pattern recognition is about reasoning, using the available information about the problem to uncover information contained within the data. Data preprocessing, feature selection, and classification are an integral part of this reasoning process. Each plays a role in uncovering the information contained within the data.

REFERENCES

1. P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, Computerized learning machines applied to chemical problems: convergence rate and predictive ability of adaptive binary pattern classifiers. *Analytical Chemistry* 41 (1969):690–695.
2. P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, Computerized learning machines applied to chemical problems: multicategory pattern classification by least squares. *Analytical Chemistry* 41 (1969):695–700.
3. P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, Computerized learning machines applied to chemical problems: molecular structure parameters from low resolution mass spectrometry. *Analytical Chemistry* 42 (1970):1387–1394.
4. P. C. Jurs, Machine intelligence applied to chemical systems: prediction and reliability improvement in classification of low resolution mass spectrometry data *Analytical Chemistry* 43 (1971):22–26
5. D. Brodnjak-Voncina, Z. C. Kodba, and M. Novic, Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. *Chemometrics and Intelligent Laboratory Systems* 75 (2005):31–43.
6. C. A. Johnson, H. Topoff, R. K. Vander Meer, and B. K. Lavine, "Do these eggs smell funny to you? Egg discrimination by formica hosts of the slave-making ant, *Polyergus breviceps* (Hymenoptera: Formicidae). *Behavior Ecology and Sociobiology* 57(2005):245–255.
7. C. Simo, P. Martin-Alvarez, C. Barbas, and A. Cifuentes, Application of stepwise discriminant analysis to classify commercial orange juices using chiral micellar electrokinetic chromatography-laser induced fluorescence data of amino acids. *Electrophoresis* 25 (2004), 2885–2891.
8. C. R. Johnson, H. Topoff, R. K. Vander Meer, and B. K. Lavine, Queens ripe for the killing: When a host queen becomes the target of aggression by the slave-maker ant queen, Polyergus breviceps. *Animal Behavior* 64 (2002):807–815.
9. C. Johnson, R. K. Vander Meer, and B. K. Lavine, Changes in the cuticular hydrocarbon profile of the slave-maker ant queen, polyergus breviceps, after killing a formica queen. *Journal of Chemical Ecology* 27 (2001):1787–1804.
10. B. K. Lavine, D. Brzozowski, J. Ritter, A. J. Moores, and H. T. Mayfield, Fuel Spill identification by selective fractionation prior to gas chromatography I. Water soluble components. *Journal of Chromatographic Science* 39 (2001):501–506.
11. R. L. White, P. D. Wentzell, M. A. Beasy, D. S. Clark, and D. W. Grund, Taxonomy of Amanita mushrooms by pattern recognition of amino acid chromatographic data. *Analytica Chimica Acta* 277 (1993):333–346.
12. H. A. Clark and P. C. Jurs, Classification of crude oil gas chromatograms by pattern recognition techniques. *Analytical Chemistry* 51(1979):616–623).
13. P. J. Gemperline and N. R. Boyer, Classification of near-infrared spectra using wavelength distances: comparison to the Mahalanobis distance and residual variance methods. *Analytical Chemistry* 67(1995):160–166.
14. N. K. Shah and P. J. Gemperline, Combination of the Mahalanobis distance and residual variance pattern recognition techniques for classification of near-infrared spectra. *Analytical Chemistry* 62 (1990):465–470.
15. P. J. Gemperline, L. D. Webber, and F. O. Cox, Raw materials testing using soft independent modeling of class analogy analysis of near-infrared reflectance spectra. *Analytical Chemistry* 61(2) 1989:138–144.
16. P. C. Jurs, Pattern recognition used to investigate multivariate data in analytical chemistry. *Science* 232 (1986):1219–1224.
17. P. C. Jurs**,** G. A. Bakken, and H. E. McClelland, Computational methods for the analysis of chemical sensor array data from volatile analytes. *Chemical Reviews* 100 (7) (2000):2649–2678.
18. D. L. Massart and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis* (John Wiley & Sons, New York, 1983).
19. R. G. Brereton (Ed.), *Multivariate Pattern Recognition in Chemometrics* (Elsevier, Amsterdam, 1992).
20. G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition* (John Wiley & Sons, New York, 1992).
21. B. D. Ripley, *Pattern Recognition and Neural Networks* (University Press, Cambridge, 1996).
22. K. R. Beebe, R. J. Pell, and M. B. Seasholtz, *Chemometrics: A Practical Guide* (John Wiley & Sons, New York, 1998).
23. K. J. Siddiqui and D. Eastwood, *Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring*, SPIE Proceedings, Volume 3854, September 1999.
24. R. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*, 2nd Ed. (Wiley-Interscience, New York, 2000).
25. J. E. Jackson, *A User's Guide to Principal Component Analysis* (John Wiley & Sons, New York, 1991).
26. I. T. Jolliffe, *Principal Component Analysis*, (Springer-Verlag, New York, 1986).
27. C. P. Wang and T. L. Isenhour, Infrared library search on principal component analyzed Fourier transform absorption spectra. *Analytical Chemistry* 41 (1987):185–194.
28. J. M. Bjerga and G. W. Small, Automated selection of library subsets for infrared searching. *Analytical Chemistry* 62 (1990):226–233.
29. G. W. Small, Automated spectral interpretation. *Analytical Chemistry* 59 (1987):535A–545A.
30. H. J. Luinge, Automated interpretation of vibrational spectra. *Vibrational Spectroscopy* 1 (1990):3–18.
31. J. Huang, and K. H. Esbensen, Applications of angle measure technique (AMT) in image analysis Part I. A new methodology for in situ powder characterization. *Chemolab* 54 (2000):1–19.

32. J. Huang, and K. H. Esbensen, Applications of AMT (angle measure technique) in image analysis Part II: Prediction of powder functional properties and mixing components. *Chemometrics and Intelligent Laboratory Systems* 57 (2001):37–56.
33. D. Bright, Measurement of chemical components using scatter diagrams with principal component analysis in *Microbeam Anal. Proc. Annu. Conf. Microbeam Anal Soc.* 29th, ed. E. Etz, (VCH, New York, 1995), 403–404.
34. C. Eliasson, A. Loren, J. Engelbrektsson, M. Josefson, J. Abrahamsson, and K. Abrahamsson, Surface-enhanced Raman scattering of single living lymphocytes with multivariate evaluation. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy* 61A (2005):755–760.
35. D. Ben-Amotz and D. Zhang, Removal of cosmic spikes from hyper-spectral images using a hybrid upper bound spectrum method. *Applied Spectroscopy* 56 (2002):91–98.
36. C. M. Stellman, K. S. Booksh, A. R. Muroski, M. P. Nelson, and M. L. Myrick, Principal component mapping applied to Raman microspectroscopy of fiber-reinforced polymer composites. *Science and Engineering Composite Materials* 7(1998):51–80.
37. S. Sasic, D. Clark, J. C. Mitchell, and M. Snowden, A comparison of Raman chemical images produced by univariate and multivariate data processing—A simulation with an example from pharmaceutical practice. *Analyst* 129 (2004):1001–1007.
38. L. Zhang, M. J. Henson, and S. Sekulic, Multivariate data analysis for Raman imaging of a model pharmaceutical tablet. *Analytica Chimica Acta* 545 (2005):262–278.
39. S. Baronti, A. Casini, F. Lott, and S. Porciani, Principal component analysis of visible and near infrared multispectral images of works of art. *Chemometrics and Intelligent Laboratory Systems* 39 (1997):103–114.
40. W. H. A. M. van den Broek, D. Wienke, W. J. Melssen, R. Feldhoff, T. Huth-Fehre, T. Kantimm, and L. M. C. Buydens, Application of a spectroscopic infrared focal plane array sensor for on-line identification of plastic waste. *Applied Spectroscopy* 51(1997):856–865.
41. J. Karasinski, S. Andreescu, O. A Sadik, B. Lavine, and M. N. Vora, Multiarray sensors with pattern recognition for the detection, classification, and differentiation of bacteria at subspecies and strain levels. *Analytical Chemistry* 77(2005):7941–7949.
42. B. K. Lavine, C. E. Davidson, and W. T. Rayens, Machine learning based pattern recognition applied to microarray data, in *Combinatorial Chemistry and High Throughput Screening* 7(2004):115–131.
43. B. K. Lavine and M. Vora, Identification of Africanized honeybees. *Journal of Chromatography A* 1096(2005):69–75.
44. B. K. Lavine, C. E. Davidson, and A. J. Moores, Genetic algorithms for spectral pattern recognition. *Vibrational Spectroscopy* 28 (2002):83–95.
45. B. K. Lavine, A. Vesanen, D. M. Brzozowski, and H. T. Mayfield Authentication of fuel standards using gas chromatography/pattern recognition techniques. *Analytical Letters* 34 (2001):281– 294.
46. B. K. Lavine, C. E. Davidson, A. J. Moores, and P. R. Griffiths, Raman spectroscopy and genetic algorithms for the classification of wood types. *Applied Spectroscopy* 55 (2001):960–966.
47. B. K. Lavine, D. Brzozowski, A .J. Moores, C. E. Davidson, and H. T. Mayfield, Genetic algorithm for fuel spill identification. *Analytica Chimica Acta* 437 (2001):233–246.
48. B. K. Lavine, J. Ritter, A. J. Moores, M. Wilson, A. Faruque, and H. T. Mayfield, Source identification of underground fuel spills by solid phase micro-extraction/high-resolution gas chromatography/genetic algorithms. *Analytical Chemistry* 72 (2000):423–431
49. L. S. Ramos, Characterization of mycobacteria species by HPLC and pattern recognition. *Journal of Chromatographic Science* 32 (1994):219–227.
50. E. Gidman, R. Goodacre, B. Emmet, R. A. Smith, and D. Gwynn-Jones, Investigating plant-plant interference by metabolic fingerprinting. *Phytochemistry* 63 (2003):705–710.
51. R. M. Jarvis and R. Goodacre, Discrimination of bacteria using surface-enhanced Raman spectroscopy. *Analytical Chemistry* 76 (2004):40–47.
52. C. Zheng and K. S. Suslick, A colorometric sensor array for organics in water. *Journal of the American Chemical Society* 127 (2005):11548–11549.
53. K. Sepcic, M. Josowicz, J. Janata, and T. Selby, Diagnosis of used engine oil based on gas phase analysis. *Analyst* 129 (2004):1070–1075.
54. M. Masila, A. Sargent, and O. A. Sadik, Pattern recognition studies of halogenated organic compounds using conducting polymer sensor arrays. *Electroanalysis* 10 (1998):312–320.
55. Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Method*, (Cambridge University Press, UK, 2000).
56. Vojislav Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models (Complex Adaptive Systems)*, (MIT Press, Cambridge, MA, 2001).
57. T. B. Blank and S. D. Brown, Nonlinear multivariate mapping of chemical data using feed-forward neural networks. *Analytical Chemistry* 65 (1993):3081–3089.
58. T. B. Blank and S. D. Brown, Data processing using neural networks. *Analytica Chimica Acta* 277 (1993):273–287.
59. K. Brudzewski, A. Kesik, K. Kolodziejczyk, U. Zborowska, and J. Ulaczyk, Gasoline quality prediction using gas chromatography and FTIR spectroscopy: An artificial intelligence approach. *Fuel* 85 (2006):553–558.
60. S. Zomer, R. G. Brereton, J. F. Carter, and C. Eckers, Support vector machines for the discrimination of analytical chemical data: Application to the determination of tablet production by pyrolysis gas chromatography-mass spectrometry. *Analyst* 129 (2004):175–181.
61. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer-Verlag, NY, 1995).
62. O. Sadik, W. H. Land, A. Wanekaya, M. Uematsu, M. Embrechts, L. Wong, D. Leibensperger, and A. Volykin, Detection of classification of organophosphate nerve agent simulants using support vector machines with multiarray sensors. *Journal of Chemical Information and Computer Sciences* 44 (2002):499–507.
63. J. Vohradsky, Adaptive classification of two-dimensional gel electrophoresis spot patterns by neural networks and cluster analysis. *Electrophoresis* 18 (1997):2749–2754.
64. L. Dolmatova, C. Ruckerbusch, N. Dupuy, J. P. Huvenne, and P. Legrand, Identification of modified starches using infrared spectroscopy and artificial neural network processing. *Applied Spetroscopy* 52 (1998):329–338.
65. C. Cai and P. deHarrington, Prediction of substructure and toxicity of pesticides with temperature constrained cascade correlation

network from low-resolution mass spectra. *Analytical Chemistry* 71 (1999):4134–4141.
66. C. G. H. Walker, Analysis of multispectral microscope images using neural networks. *Surface and Interface Analysis* 24 (1996):173–180.
67. H. P. Dettmar, G. S. Barbour, K. T. Blackwell, T. P. Vogl, D. L. Alkon, F. S. Fry, J. E. Totayh, and T. L. Chambers, Orange juice classification with a biologically based neural network. *Journal of Combinatorial Chemistry* 20 (1996):261–266.
68. K. Jensen, T. Tygsen, C. Kesmir, I. M. Skovgaard, and I. Sondergaard, Classification of potato varieties using isoelectrophoretic focusing patterns, neural nets and statistical methods. *Journal of Agriculture and Food Chemistry* 45 (1997):158–161.
69. L. X. Sun, K. Danzer, and G. Thiel, Classification of wine samples by means of artificial neural networks and discriminant analytical methods. *Fresenius Journal of Analytical Chemistry* 359 (1997):143–149.
70. S. Wold, Pattern recognition by means of disjoint principal component models. *Pattern Recognition* 8 (1976):127–139.
71. I. Frank, and J. Friedman, Classification: Oldtimers and newcomers. *Journal of Chemometrics* 3 (1989):463–475.
72. M. C. Burl, B. C. Sisk, T. P. Vaid and N. S. Lewis, Classification performance of carbon black polymer composite vapor detector arrays as a function of array size and detector composition. *Sensors and Actuators B* B87 (2002):130–149.
73. J. R. Stetter, M. W. Findlay, K. M. Schroeder, C. Yue, and W. R. Penrose, Quality classification of grain using a sensor array and pattern recognition. *Analytica Chimica Acta* 248 (1993):1–11.
74. B. K. Lavine, A. Faruque, P. Kroman, and H. T. Mayfield, Source identification of fuel spills by pattern recognition analysis of high speed gas chromatograms. *Analytical Chemistry* 67 (1995):3846–3852
75. F. Chau, Y. Liang, J. Gao, and X. Shao, *Chemometrics—From Basics to Wavelet Transform* (Wiley Interscience, New York, 2004).
76. J. A. Pino, J. E. McMurry, P. C. Jurs, B. K. Lavine, and A. M. Harper, Applications of pyrolysis/gas chromatography/pattern recognition to the detection of CF heterozygotes. *Analytical Chemistry* 57 (1985):295–302
77. B. K. Lavine, H. Mayfield, and A. Stine, Application of pattern recognition techniques to problems in advanced pollution monitoring. *Process and Quality Control* 2 (1992):347–355.