Development of Radial Basis Function Cascade Correlation Networks and Applications of Chemometric Techniques for Hyphenated Chromatography–Mass Spectrometry Analysis

A dissertation presented to

the faculty of

the College of Arts and Sciences of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Weiying Lu

November 2011

This dissertation titled

Development of Radial Basis Function Cascade Correlation Networks and Applications of Chemometric Techniques for Hyphenated Chromatography–Mass Spectrometry Analysis

by

WEIYING LU

has been approved for

the Department of Chemistry and Biochemistry

and the College of Arts and Sciences by

Peter de B. Harrington

Professor of Chemistry and Biochemistry

Howard Dewald

Interim Dean, College of Arts and Sciences

Abstract

LU, WEIYING, Ph.D., August 2011, Chemistry

Development of Radial Basis Function Cascade Correlation Networks and Applications of Chemometric Techniques for Hyphenated Chromatography–Mass Spectrometry Analysis (153 pp.)

Director of Dissertation: Peter de B. Harrington

A cascade correlation learning architecture has been devised for radial basis function neural networks. Cascade correlation furnishes incremental learning networks. The proposed algorithm was applied to three different datasets: a synthetic dataset and two chemical datasets. The synthetic dataset was used to test the novelty detection ability of the proposed network. In the chemical datasets, the growth regions of Italian olive oils were identified by their fatty acid profiles; mass spectra of polychlorobiphenyl compounds were classified by chlorine number. The prediction results by bootstrap Latin partition indicate the proposed neural network is useful for pattern recognition.

A discriminant based charge deconvolution analysis pipeline is proposed. The molecular weight determination (MoWeD) charge deconvolution method was applied directly to the discrimination rules obtained by the fuzzy rule-building expert system (FuRES) pattern classifier. This approach was demonstrated with synthetic electrospray ionization mass spectra. Identification of the tentative protein biomarkers by bacterial cell extracts of *Salmonella enterica* serovar typhimurium strains A1 and A19 by

liquid chromatography–electrospray ionization-mass spectrometry (LC–ESI-MS) was also demonstrated. The data analysis time was reduced by applying this approach. Furthermore, this method was less affected by noise and baseline drift.

The gasoline and kerosene collected from different locations in the United States were identified by gas chromatography/mass spectrometry (GC/MS) followed by chemometric analysis. Classifications based on two-way profile and target component ratio were compared. The projected difference resolution (PDR) mapping was applied to measure the differences among the ignitable liquid (IL) samples by their GC/MS profiles quantitatively. FuRESs were applied to classify individual ILs. The FuRES models yielded correct classification rates greater than 90% for discriminating between samples. PDR mapping, a new method for characterizing complex data sets, was consistent with the FuRES classification result.

Approved:

_____

Peter de B. Harrington

Professor of Chemistry and Biochemistry

Acknowledgments

First, I would like to thank my advisor, Prof. Peter B. Harrington. During five years of study, I learned a lot in analytical chemistry, especially in chemometrics. Prof. Harrington taught me critical thinking in science, as well as inspired me with brilliant ideas to overcome difficulties in the research projects. I would also like to thank my committee members, Profs. Glen Jackson, Hao Chen, Shiyong Wu, and Archil Gulisashvili for their helpful comments and suggestions.

The department of Chemistry and Biochemistry at Ohio University is thanked for offering me the great study opportunity to pursue my doctoral degree. I am grateful for our former and current group members. I had a great time in doing research, as well as exchanging ideas and sharing exciting moments with all of you. I also thank all the people I know during my days of study in Athens, which is too long to list here. It would have been impossible to have a wonderful experience to study at Ohio University without all of you.

Dr. John Callahan at Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration is thanked for his collaboration on the protein deconvolution research projects. Prof. J. Graham Rankin from the Forensic Science Center, Marshall University is thanked for his collaboration on the gasoline and kerosene analysis project.

Last but not least, I want to give my deepest appreciation to my parents Jianguo Lu and Yiqing Zhang, for all the love they devoted to me.

Table of Contents

List of Tables

Page

List of Figures

Page

List of Abbreviations

ACS.....American Chemical Society

ANN .....artificial neural network

ANOVA ..... analysis of variance

ASTM .....American Society for Testing and Materials

BLP .....bootstrap Latin partition

BNN ..... back-propagation network

CCN .....cascade correlation network

CDC ..... Centers for Disease Control and Prevention

CE–ESI-MS ..... capillary electrophoresis–electrospray ionization-mass spectrometry

ESI-MS ..... electrospray ionization-mass spectrometry

FTICR-MS ..... Fourier transform ion cyclotron resonance-mass spectrometry

FuRES .....fuzzy rule-building expert system

FWHM ..... full width at half maximum

GC ..... gas chromatography

GC/MS..... gas chromatography/mass spectrometry

GC–DMS .....gas chromatography–differential mobility spectrometry

HPLC .....high performance liquid chromatography

HPLC–MS ..... high performance liquid chromatography–mass spectrometry

IL..... ignitable liquid

ILR.....ignitable liquid residues

LC..... liquid chromatography

LC–ESI-MS.....
.....liquid chromatography–electrospray ionization-mass spectrometry

LC–ESI-QTOF MS.. liquid chromatography–electrospray ionization-quadrupole time-of-flight mass spectrometry

LC–MS/MS ..... liquid chromatography–tandem mass spectrometry

LDA.....linear discriminant analysis

MALDI-MS..... matrix-assisted laser desorption/ionization-mass spectrometry

MCA .....multiplicative correlation algorithm

MLF..... multi-layered feed-forward

MoWeD..... molecular weight determination

MS .....mass spectrometry

NIPALS .....non-linear iterative partial least squares

oPLS-DA .....optimal partial least squares-discriminant analysis

PCA..... principal component analysis

PCB.....polychlorobiphenyl

PCT .....principal component transformation

PDR ..... projected difference resolution

PLS-DA.....partial least squares-discriminant analysis

QTOF-MS .....quadrupole time-of-flight mass spectrometry

RBFCCN.....radial basis function cascade correlation network

RBFN..... radial basis function network

RBF.....radial basis function

RF..... random forest

RRMSEC ..... relative root mean square error of calibration

SCRBFN.....self-configuring radial basis function network

SNR .....signal-to-noise ratio

SVD ..... singular value decomposition

SVM .....support vector machine

TCCCN..... temperature constrained cascade correlation network

THRASH..... thorough high resolution analysis of spectra by Horn

TIC ..... total ion current

VisuShrink ..... visually calibrated adaptive smoothing

Chapter 1 Introduction

1.1 General Statement

This dissertation presents the development of a novel neural network for chemical data processing; biochemical and forensic applications of chemometric methods in chromatography and spectrometry methods are also introduced. Chapter 1 presents the general introduction of analytical instrumentation and chemometric methods used in this dissertation. In Chapter 2, the development of a novel neural network named radial basis function cascade correlation network is described. In Chapter 3, a discriminant based charge deconvolution analysis pipeline for protein profiling by high performance liquid chromatography–mass spectrometry (HPLC–MS) is devised. The chemometric study of ignitable liquid (IL) identification using gas chromatography/mass spectrometry (GC/MS) data is presented in Chapter 4. The summary and future works are introduced in Chapter 5. The publications and presentations associated with this dissertation are listed in the Appendices.

1.2 Chromatography and Mass Spectrometry

The analytical instrumentation used to conduct the research reported in this dissertation is briefly introduced here. Chromatography is an important separation technique in chemistry.[1] In general, an analyte is transported in a mobile phase and passes through a chromatographic column. The stationary phase materials are coated or pack inside the column. Because different components have different retention abilities in

the stationary phase, different chemical components of a mixture migrate through the column at different speed and eluted from the column at different times, which results in the chemical separation. The chromatographic measurement is referred to as a chromatogram for which the abscissa is the retention time and the ordinate is the measured quantity of material detected. Gas chromatography (GC) and liquid chromatography (LC) are commonly used chromatographic techniques, in which the mobile phases are respectively gas and liquid. High performance liquid chromatography (HPLC) was derived from LC. HPLC decreases the particle size of the stationary phase or the stationary phase support in the column. In addition, the HPLC instrument requires pumps capable of producing high pressures in the mobile phase enough to overcome the viscosity of the mobile phase as it moves through the small channels between the stationary phase particles. In addition, the pump must maintain precise flow control. The separation ability of HPLC is greatly increased compared with conventional LC.

Another widely used technique in instrumental analysis is mass spectrometry (MS).[2, 3] In a mass spectrometer, an analyte is introduced from an inlet to an ion source. Ionization takes place in the ion source that transforms compounds into ions. Afterwards, the ions are separated in a mass analyzer by their mass-to-charge ratio. By measuring the ion current at an electronic detector, the relative abundance of each ion is obtained with respect to the mass-to-charge ratio. The MS measurement is reported as a

spectrum for which the abscissa is mass-to-charge ratio and the ordinate is intensity. A mass spectrometer can be applied as a detector that is hyphenated to a GC or an HPLC instrument, respectively referred to as GC/MS and HPLC–MS. GC/MS and HPLC–MS generate two-way spectra that furnish data in a 3D matrix that can be viewed as an image. The axes of the image correspond to retention and mass-to-charge ratio and the intensity of the image encodes the ion current. Figure 1-1 gives a representation of a two-way spectrum.

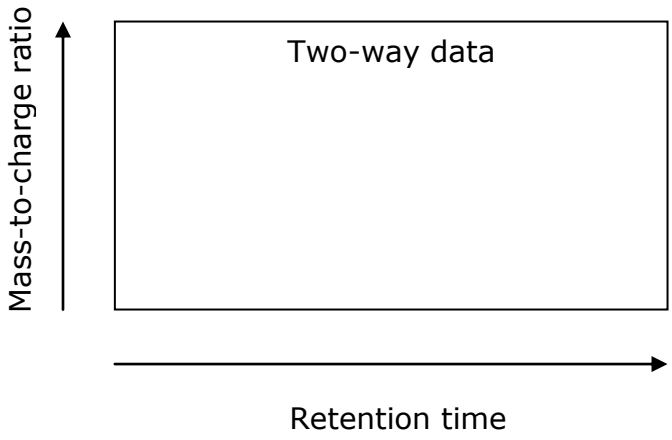

Figure 1-1. A two-way spectrum of GC/MS and HPLC–MS.

With the development of modern analytical instrumentation, the throughput and resolution of chemical measurements are increasing. The measurement can be performed rapidly using high throughput instruments. Meanwhile, the number of data points obtained from each measurement increases with the increase of resolution. The increase resolution may be

exploited to yield more analytical information obtained from a complex mixture. Therefore, the characterization of complex mixtures is possible. Nowadays, HPLC–MS is applied for a wide range of biological samples to identify and quantify compounds. The applications of HPLC–MS include food and nutritional supplements authentication[4, 5], drug analysis and metabolomics[6, 7], medical diagnostics[8, 9], etc. GC/MS also has a wide range of applications, for instance, the identification of ignitable liquids[10-12] and the detection of illicit drugs[13, 14], explosives[15, 16], and pesticides[17, 18].

The U.S. Centers for Disease Control and Prevention (CDC) estimated that each year 48 million American people are infected by foodborne diseases, of which 128 000 are hospitalized, and 3000 died.[19] As a harmful pathogenic bacterium, salmonella infections typically affect the intestines, causing symptoms such as vomiting and fever, and for some cases could be life-threatening. The CDC claimed that salmonella is the top pathogen causing hospitalization and death.[19] Because the contaminated food neither generates odor nor has visible changes in color or texture in most cases, bacteria that cause diseases are hard to detect. As a result, research on developing bacterial identification methods has attracted much attention. In the research described in Chapter 3, the protein profiling of whole cell extracts of bacteria *Salmonella enterica* are analyzed by HPLC–MS.

Arson is the leading cause of fires, and second leading cause of deaths and injuries according to the U.S. Fire Administration.[20] Therefore, arson investigation is important to the criminal justice system. The criminals may

use ignitable liquids, typically commercially available fuels or solvents, as accelerants to start a fire. Identification of the ignitable liquids is difficult because the ignitable liquids mostly are mixtures consisting of hundreds of components. In the work presented in Chapter 4, the gasoline and kerosene were analyzed by GC/MS. Chemometric analysis pipelines were proposed and evaluated to achieve an automatic and quantitative measure of identification.

1.3 Chemometrics Techniques

Chemometrics was introduced by Wold and Kowalski in the 1970's.[21-23] Chemometrics can be described as the systematic application of mathematical and statistical knowledge to chemical research. Key areas of chemometrics are experimental design, signal processing, and multivariate data analysis. Overwhelming amounts of data obtained from the chemical measurements can be used to characterize complex mixtures. As a result, it is important to develop chemometrics techniques on a computer system to process the spectra, because of the significant increase in throughput and resolution of modern analytical instruments.

Programs implementing chemometrics algorithms provide efficient, rapid, and reliable approaches to process large amounts of analytical data by computer. This dissertation focuses on research and applications of pattern classification techniques in chemometrics.

1.3.1 Notation

This dissertation follows the American Chemical Society (ACS) style. Specifically, scalars and elements of vectors and matrices are denoted as lowercase italic letters; vectors are denoted as bold lowercase letters; matrices are denoted as bold uppercase letters.

1.3.2 Pattern Classification

Pattern classification, or classification, originates from pattern recognition and machine learning in computer science.[24] Classification is a category of techniques to characterize samples into different classes. In each class, one or more samples are tested. A classifier is an algorithm used to perform classification based on the data matrix described above as input. Classification methods can be supervised or unsupervised. Unsupervised classification methods such as clustering find similarities for unlabeled data and groups the data into classes. In supervised classification, the classes are known and the classifier forces the data into one of the predetermined classes. Supervised classification on chemical data is studied in this dissertation.

The data was organized into a matrix to perform classification. In chemometrics, the sample is measured in the experiment, from which the data is obtained. In this dissertation, the data matrix for classification is organized so that the rows are observations of the samples, which are referred to as objects. When each sample is measured in several replicates, each replicate can be treated as an object. The columns are measurements,

usually referred to as variables or factors. Figure 1-2 is an example of a matrix used for classification. When dealing with two-way spectra, each intensity measurement corresponds to a mass-to-charge ratio and a retention time measurement. The two-way matrix can be reorganized into a single row. This process is referred to as unfolding. In this dissertation, different classifiers were applied to different data sets. Performances among the classifiers are compared.

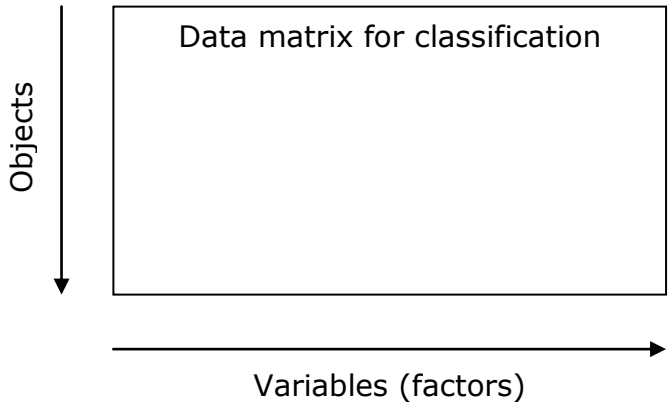

Figure 1-2. Arrangement of data matrix for classification.

1.3.3 Bootstrap Latin Partition Validation

Validation techniques evaluate the performance of the classifiers. Classification constructs mathematical models from a set of experimental data. This process of model building is referred to as training or calibration. The data used for this process is called the training data or the calibration data. Because supervised classifiers build models that force the data into the known classes, many powerful classifiers will achieve this goal by modeling

noise or spurious components in the data. This problem is especially prevalent when the data matrix is underdetermined (i.e., it has more variables than objects). The results of the model applied to the calibration data are referred to as estimates and the associated error is referred to as calibration error.

Therefore, it is always important to validate the model to assure that the model is characterizing systematic features in the data and not random components. A fundamental method of validation is accomplished by applying the model to an independent set of data. The independent set of data is referred to as the test set or the prediction set. The data are independent because they were never used for any aspect of constructing or modifying the model. The classification model is applied to the test set and the predicted classes are compared with the known classes. The classification results from the model with the independent data may be less than perfect. The term prediction is only used with the results from independent data sets. The results of the prediction set are referred to as predictions and the associated error is referred to as prediction error.

Another important term is generalization. Classification models should be general and have the ability to interpolate among different samples. Therefore, when validating models it is always useful to make sure that the data in the training and test sets are from different samples and not replicate measurements of the same sample.

Performing additional experiments could be time-consuming and costly. As a result, validation techniques without extra experiments have been developed. The most typical method is cross-validation. In cross-validation, samples are equally divided into *n* parts. Each part is extracted from the original data once and only once as the test set. The remaining part of the data after extraction is the corresponding training set. As a result, *n* pairs of training and test sets are generated. Then, *n* classification models are trained and tested by using *n* pairs of training sets and test sets.

In cross-validation, data is divided randomly so that the proportional distribution of classes in training and test sets is not maintained. As a result, bias in the classification model may occur. This may cause a problem because the models will typically minimize the largest classification errors. Therefore, the class with the largest number of objects will also have the better prediction accuracy. In extreme cases, some classes may not be present in the training set at all.

The Latin partition was developed to avoid this problem. Latin partition has an improvement over cross-validation technique by applying a constraint to maintain the proportional distribution of classes in each partitioned pair of training and test sets. Figure 1-3 demonstrates the Latin partition with three partitions in a data set that contains two classes.

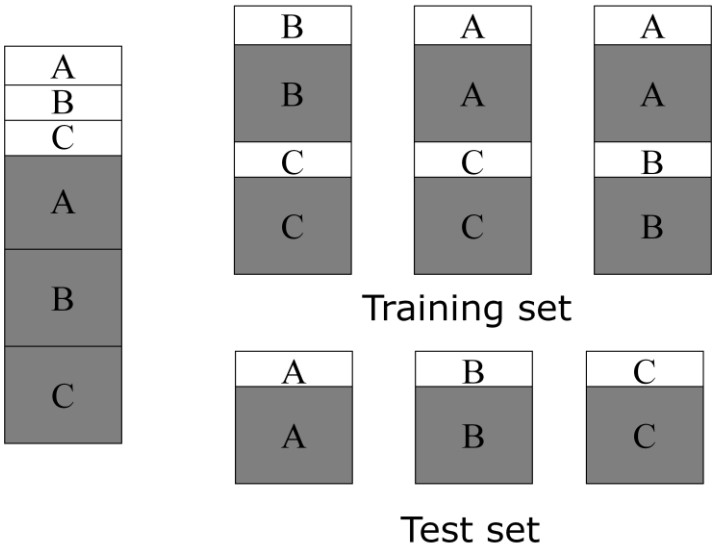

Figure 1-3. Demonstration of a Latin partition process. A, B, and C are three different partitions. Boxes in different colors are different classes in the data set.

Unlike cross-validation that repeatedly analyzes subsets of the data, bootstrapping is a method that performs multiple validations by re-sampling. The prediction results for each Latin partition can be pooled because each object is used once and only once. Because the Latin partitions are created using random sampling, the process may be repeated many times. Therefore each set of Latin partitions are independent. This process of repeating the Latin partitions is referred to as bootstrapping.

Because each bootstrap or set of Latin partition is independent the variation characterizes the reproducibility of the data and its affect on the model. The average prediction results across the bootstraps can be reported to give a general measure of prediction and the variation about the average result can be used for form confidence intervals. This approach is very important for optimizing and comparing chemometric methods and characterizing a critical source of variation. The dependency of the result on the division of the data into training and test sets.

The combined approach for model validation is referred to as Bootstrap Latin partitions (BLP).[25] In this dissertation, BLP validation is applied to compare different classifiers and different data processing methods. Through BLP many simple statistical test (e.g., t-test or ANOVA) are made available to compare prediction results and optimize classifiers.

1.3.4 Artificial Neural Networks

Artificial neural networks (ANNs) mimic the biological structure and behavior of the central nervous system.[26, 27] The study of ANN is major field

in artificial intelligence. The ANN is used as a classifier in this dissertation. The fundamental processing units of an ANN are the artificial neurons, or units. Each neuron furnishes single or multiple input data and evaluates the data by a mathematical function to a single output or multiple outputs. The function is referred to as transfer function or activation function. Typically, transfer functions are either linear functions, sigmoid functions, or radial basis functions.

The artificial neurons are organized into several layers; each layer consists of one or more artificial neurons. The network size is determined by the number of layers and number of neurons in each layer. The first layer is the input layer, and the last layer is the output layer of the network. All other layers between the input and the output layer are hidden layers.

The neural network architecture refers to the interconnection of the neurons. Many different neural network architectures exist. The most commonly applied architecture is the multi-layered feed-forward (MLF) neural network.[28] In an MLF network, a weight is multiplied to the data at each connection. The input of a neuron is the weighted sum of outputs from the previous layer, given by

$$x = \sum_{i=1}^{n} w_i x_i + b \qquad (1\text{-}1)$$

for which x is the input to the transfer function, n is the total number of neurons in the previous layer, x_i is the input signal from the i^{th} neuron in the previous layer, w_i is the weight, and b is the intercept. The intercept is implemented by adding a bias neuron to the input layer and each hidden

layer, whose output is constantly unity. The sigmoid function is typically applied as the transfer function in the hidden neuron, given by

$$f = 1/(1 + e^{-x}) \tag{1-2}$$

for which *f* is the output of the neuron.

The architecture of an MLF network is given in Figure 1-4. Neurons are given in circles; the connections are lines, and arrows give the directions of data processing. There are two input neurons, three hidden neurons, and two output neurons.

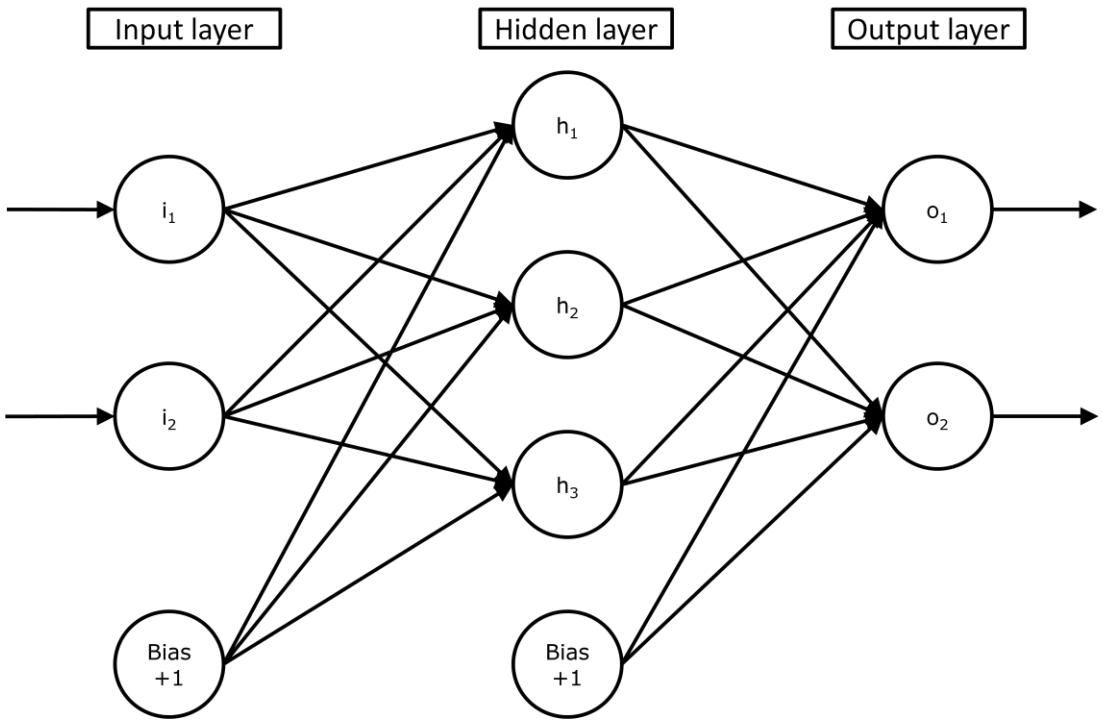

Figure 1-4. The architecture of an MLF network that contains three layers. The circles are artificial neurons; the connections are lines with arrows indicate the direction of data processing.

To determine the weights in an MLF network, a training algorithm is required. The most commonly used learning algorithm for the MLF network is back-propagation. The MLF network trained with a back-propagation learning algorithm is a back-propagation network (BNN).[28] In a BNN, the error back-propagation is used in training the neural network. The transfer functions in the input and output layers are linear functions, and the hidden layers are units with sigmoid functions. The network size and structure are fixed before training. In the training process, the weights of ANN are adjusted simultaneously based on the comparison between the error of output and the desired outputs. The adjustment routine iterates until an error threshold is achieved.

The simultaneous adjustments of all the weights in back-propagation training cause the problem of slow training and difficulties in the convergence of a BNN, especially when the size of the neural network is large.[29] In addition, a BNN cannot detect outliers because the sigmoid function applied in a BNN divides the output space by a hyperplane. In Chapter 2, a radial basis function cascade correlation network (RBFCCN) is developed to solve the problems of BNN. Comparisons with six other classifiers by synthetic and chemical data sets are reported.

1.3.5 Fuzzy Rule-Building Expert System

Besides ANNs, rule-building expert systems are also important classification techniques. The classification model of a rule-building expert system is given in a tree structure. Each rule is located in the nodes of the

classification tree. The rule contains an antecedent which determines whether a condition arises, and a consequent which directs the next rule (nodes) or indicates a classification. Univariate rule-building expert systems are the simplest form of the expert systems, where only one variable is processed in each rule. Univariate rule-building expert systems models multivariate relations by a sequential approach. Multivariate rule-building expert systems apply linear combinations of all variables to construct rules in their antecedents.

A fuzzy rule-building expert system (FuRES)[30, 31] is a classifier that builds multivariate rule with the combination of decision tree algorithm and fuzzy logic. Given an antecedent A and the consequent C, each FuRES rule is constructed to minimizes the entropy of classification $H(C|A)$, given by

$$H(C|a_j) = -\sum_{i=1}^{n} p(c_i|a_j) \ln\, p(c_i|a_j) \tag{1-3}$$

$$H(C|A) = \sum_{j=1}^{m} \left\{ H(C|a_j) \left[\sum_{i=1}^{n} p(c_i|a_j) \right] \right\} \tag{1-4}$$

for which $H(C|a_j)$ is the information entropy for consequent a_j. The number of classes *n* and *m* is the number of rule consequents. $p(c_i|a_j)$ is the conditional probabilities calculated by

$$p(c_i|a_j) = \sum_{k=1}^{n_i} \chi_A(x_k) \Big/ \sum_{k=1}^{n} \chi_A(x_k) \tag{1-5}$$

for which x_k is the projection of object *k* on the linear discriminant. χ_A is the output of fuzzy membership functions.

Unlike most classifiers, fuzzy logic are used in FuRES instead of crisp logic. The fuzzy logic is implemented by fuzzy membership functions in each branch of the classification tree. The fuzzy membership functions χ_A is given by

$$\chi_A = 1/\left(1 + e^{-x/t}\right) \quad (1\text{-}6)$$

for which t is the computational temperature that controls the degree of fuzziness; x is the input variable.

Compared to other classifiers such as neural networks and support vector machines, FuRES has advantages from the decision tree algorithm and fuzzy logic. FuRES model can be presented in a simple hierarchical tree structure, where general rules are located at the root of the tree and precise rules are located at the leaves. By applying the fuzzy logic, the FuRES classifier avoids ill-conditioned solutions when the data set is partially overlapped. FuRES has been successfully applied to investigate GC/MS[10], gas chromatography–differential mobility spectrometry (GC–DMS)[32, 33], and near-infrared spectroscopy[34], etc.

1.3.6 Electrospray ionization-mass spectrometry and Charge State Deconvolution Algorithms

The ionization source is an essential part of a mass spectrometer. Electrospray ionization (ESI) is a soft ionization source for MS. The electrospray ionization-mass spectrometry (ESI-MS) technique is usually hyphenated with LC, referred to as LC–ESI-MS. The ESI-MS technique was developed by Fenn et al.[35] ESI-MS is suitable to detect large biomolecules

such as proteins and peptides. Figure 1-5 is a simplified structural representation of ESI-MS. The analytes in a liquid solvent pass through a capillary tube towards a metal inlet. An electric potential, typically 1–5 kV, is applied between the capillary and the mass analyzer, causing the charge separation in the liquid, and the analyte molecules acquire charges. The charged ions migrate to the surface of the spray droplet. The electric forces pull the analyte solvent out of the capillary tube, forming a cone shaped spray of droplets known as a Taylor cone. The charged droplets break apart from the tip of the Taylor cone and move to the metal inlet. A spray known as electrospray is produced in this process. The droplets continue to break apart due to the collision and the electric field in the spray.

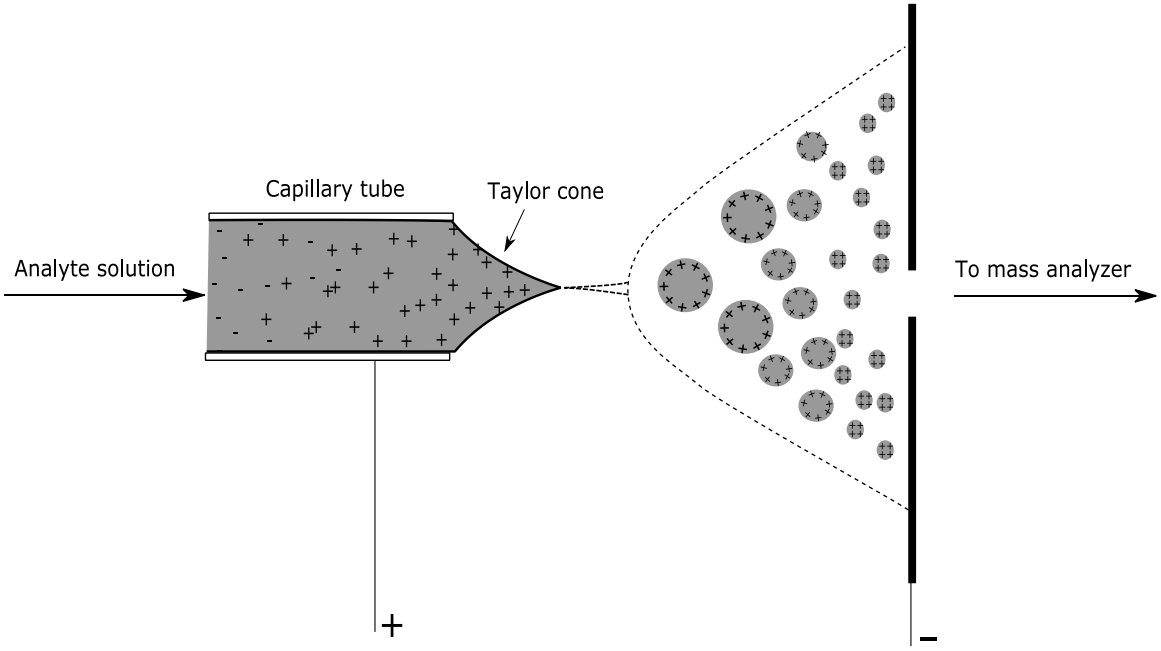

Figure 1-5. Schematic of an ESI ion source.

When analyzing proteins and peptides, the molecules in the droplets usually are multiply charged. The number of charges usually ranges from 10 to 50, depending on the size and structure of the molecule. Because different charged molecules of a same compound have different mass-to-charge ratios, ESI-MS spectra for proteins and peptides are often convolved into a series of peaks. An example of ESI-MS spectra is given in Figure 1-6.

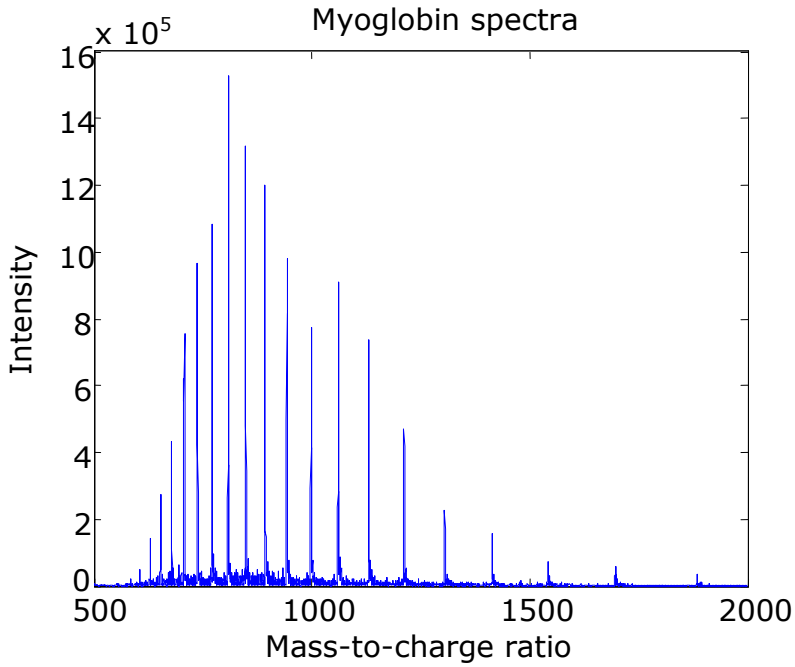

Figure 1-6. An ESI-MS spectrum of myoglobin. The spectrum was a sample spectrum of Promass software version 2.5 (Novatia, NJ).

To obtain a direct presentation about the relative intensities and molecular weights of the components, data analysis is required because of the complexity of ESI-MS spectra. Compared to manually transformation of a convolved spectrum to transform the peak clusters from the multiply charged domain to zero or singly charged domain, automated algorithms have been developed. The deconvolution algorithms were first developed by Mann et al.[36] It should be addressed that although the deconvolution technique is widely used in signal processing and image processing, the term "deconvolution" in this dissertation specifically refers to charge state deconvolution for ESI-MS. After the introduction of automated deconvolution method, many algorithms have been proposed for different types of ESI-MS instruments followed by different principles, such as maximum entropy[37-39], multiplicative correlation[40], scoring on each possible charge state pattern[41-43], the least squares method[44], minimum standard deviation[45], charge ratio analysis[46, 47], etc.

Different deconvolution techniques should be applied depending on the MS instrumentation and the size of the analyte molecules. In work described in Chapter 3, the molecular weight determination (MoWeD) method was used to process ESI-MS spectra because MoWeD performs rapidly and is suitable in processing data from quadrupole time-of-flight mass spectrometry. The deconvolved myoglobin spectrum by MoWeD is given in Figure 1-7.

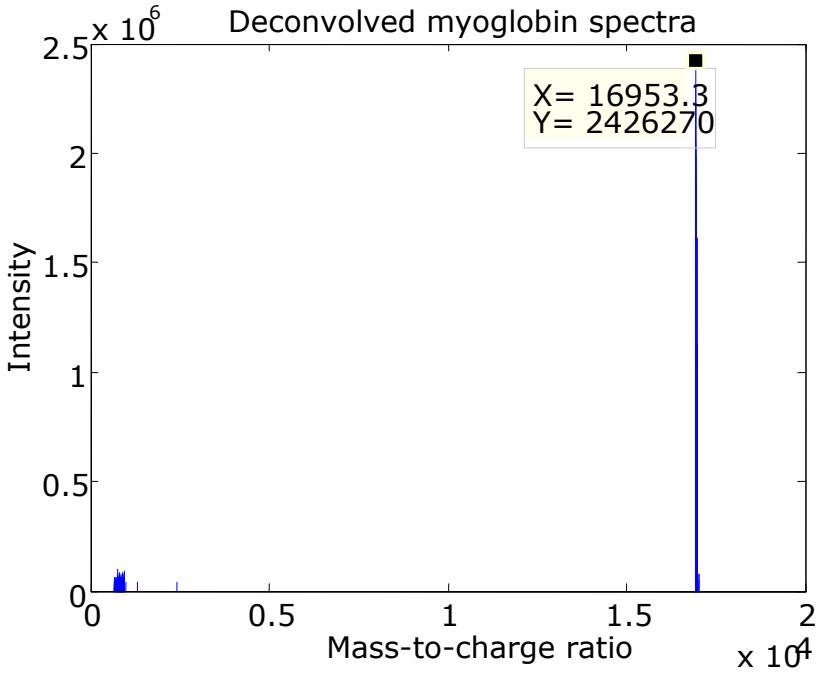

Figure 1-7. Deconvolved myoglobin spectrum by MoWeD.

The MoWeD algorithm was proposed by Pearcy and Lee[41], which performs charge state deconvolution by applying a scoring function with respect to each possible charge state. First, the peaks in the spectrum are detected by a peak-finding routine. Starting with the highest intensity peaks, this algorithm calculates and assigns charge states to each peak that is above a given noise threshold. After charge assignment of all peaks, the ESI-MS spectrum is deconvolved into the zero-charge domain. The charge assignment process contains the following steps:

1. Evaluate the range of candidate charges that is defined by the input mass-to-charge ratio range and the pre-defined maximum possible molecular mass.

2. For each candidate charge, calculate charge state patterns. Charge state patterns are a series of mass-to-charge ratio values from a contiguous charge state series.
3. A score is calculated for each candidate charge state by a scoring function. The scoring function is the number of peaks present in the charge state pattern minus the number of gaps in the charge state pattern.
4. The candidate charge with the maximum score is assigned to the peak.
5. The zero-charge spectrum is updated each time after a charge is assigned. For every point in an original peak, the molecular mass is calculated. The calculated molecular mass is the abscissa of the peak points and the peak intensity is the ordinate. According to the predefined molecular mass range and molecular mass increment, the intensities in the zero-charge spectrum are calculated by linear interpolation between adjacent peak points. The zero-charge spectrum is updated by adding the interpolated intensity.

Chapter 2 Radial Basis Function Cascade Correlation Networks

Adapted with permission from Lu, W. and Harrington, P.B.; *Algorithms* **2009**, *2*, 1045-1068.

2.1 Introduction

Artificial neural networks (ANNs) are widely used pattern recognition tools in chemometrics. The most commonly used neural network for chemists is the back-propagation neural network (BNN). The BNN is a feed forward neural network, usually trained by error back-propagation.[48, 49] BNNs have been applied to a broad range of chemical applications. Recent analytical applications of BNNs in fields such as differential mobility spectrometry[50] and near-infrared spectroscopy[51] have been reported in the literature.

Although BNN is useful, BNNs converge slowly during training especially when the network contains many hidden neurons. This slow and chaotic convergence is partially caused by the simultaneous adjustments of weights of all hidden neurons during the training of BNNs, which is referred to as the "moving target problem". To avoid this problem, a network architecture named cascade correlation network (CCN) was proposed by Fahlman and Lebiere.[29] A CCN begins training with a minimal network, which has only an input layer and an output layer. During training, the CCN determines its topology by adding and training one hidden neuron at a time, resulting in a multilayer structure. In this training strategy, the moving

target problem is avoided because only weights of single hidden neuron in the network are allowed to change at any time. CCNs have been applied to the prediction of the protein secondary structure[52] and estimation of various ion concentrations in river water for water quality monitoring[53].

A temperature constrained cascade correlation network (TCCCN)[54] combines the advantages of cascade correlation and computational temperature constraints. By adapting the sigmoid transfer function, a temperature term is added to constrain the length of the weight vector in the hidden transfer function. The temperature is adjusted so that the magnitude of the first derivative of the covariance between the output and the residual error is maximized. As a result, fast training can be achieved because of the larger gradient of the response surface. TCCCNs have been successfully applied to many areas in analytical chemistry, such as identification of toxic industrial chemicals by their ion mobility spectra[55], classification of official and unofficial rhubarb samples based on their infrared reflectance spectrometry[56], and prediction of substructure and toxicity of pesticides from low-resolution mass spectra[57], etc.

Besides BNNs and CCNs, the radial basis function network (RBFN) is another important type of neural network. An RBFN is a three-layered feed forward network, which use radial basis functions (RBFs) as transfer functions in the hidden layer. The most commonly used RBF is the Gaussian function. The number, centroids and radii of the hidden units of an RBFN can be determined by random generation, clustering, and genetic algorithms, etc.

The RBFN can also be trained by back-propagation. Wan and Harrington developed a self-configuring radial basis function network (SCRBFN).[58] In a SCRBFN, a linear averaging (LA) clustering algorithm is applied to determine the parameters of the hidden units. Class memberships of the training objects are used during clustering in the LA algorithm.

Recently, many novel supervised learning methods have gained increasing popularity, such as the support vector machine (SVM) and random forest (RF). The SVM was introduced by Vapnik.[59] The SVM first projects the training data into high-dimensional feature space by kernel functions. An optimal decision hyperplane is determined by finding the maximum margin between classes.

The RF method was developed by Breiman.[60] The RF method derives from the decision tree algorithm. During the RF training, many decision trees are trained by ensemble learning techniques, i.e. bootstrapping of the training set. Because each tree is built from different individual subset, the trees are different from each other. The classification result is then calculated by voting from all the trees built.

A radial basis function cascade correlation network (RBFCCN) that combines the advantages of CCNs and RBFNs was devised in the present work. The RBFCCN benefits from the RBF as the hidden transfer function instead of the commonly used sigmoid logistic function. The RBFCCN also has a cascade-correlation structure. The network performance was tested using both synthetic and actual chemical data sets. The partial least

squares-discriminant analysis (PLS-DA) was also tested as the standard reference method. Comparisons were made with the BNN, RBFN, SCRBFN, PLS-DA, SVM, and the RF method. Two synthetic data sets were constructed. The first data set is to evaluate classifier performance when a novel data set class is introduced. The second data set is designed to measure the performance with an imbalanced data set in that one class has much fewer objects than the other classes. Two chemical data sets were also evaluated. One is the fatty acid methyl ester measurements of Italian olive oils and the other is a collection of mass spectra from different polychlorobiphenyl (PCB) congeners. The bootstrap Latin partition (BLP)[25] validation method were used in this study.

2.2 Theory

The network architectures of an RBFN and an RBFCCN are given in Figures 2-1 and 2-2, respectively. By applying the cascade correlation algorithm, the RBFCCN has a different network topology compared with conventional RBFNs. In RBFCCNs, the transfer function applied in the hidden neuron is the Gaussian function. Unlike an RBFN that usually has only one hidden layer, the RBFCCN has a multi-layered structure. Each hidden layer contains only one neuron. In RBFCCNs, the k^{th} hidden neuron is connected with $k + l - 1$ inputs, where l denotes the number of input neurons. The output of the i^{th} object from the k^{th} hidden neuron o_{ik} is given by:

$$o_{ik} = g_k(\mathbf{x}_{ik}) = \exp\left[-\sum_{p=1}^{k+l-1}\left(x_{ikp} - \mu_{kp}\right)^2 / \left(2{\sigma_k}^2\right)\right] \qquad (2\text{-}1)$$

for which g_k is the notation of the Gaussian function; $\mathbf{x_{ik}}$ is the input vector; x_{ikp} is the corresponding p^{th} element of $\mathbf{x_{ik}}$. The μ_{kp} term denotes the p^{th} element of centroid $\boldsymbol{\mu}_\mathbf{k}$, and σ_k denotes the k^{th} radius. The o_{ik} term will depend on two factors: the Euclidean distance between the sample and the centroid and the radius. In the cascade-correlation training architecture, the hidden units are added and trained sequentially during training. The training process of an RBFCCN includes the following steps:

1. Initialize the network.
2. Add a hidden neuron to the network. Initialize this hidden neuron by setting initial values of $\boldsymbol{\mu}_\mathbf{k}$ and σ_k of the Gaussian function.
3. Train the hidden neuron. Determine the values of $\boldsymbol{\mu}_\mathbf{k}$ and σ_k.
4. Train the weights $\mathbf{W_k}$ in the output layer.
5. Repeat step 2 to step 4 until a given error threshold is achieved or a given number of hidden units were added.

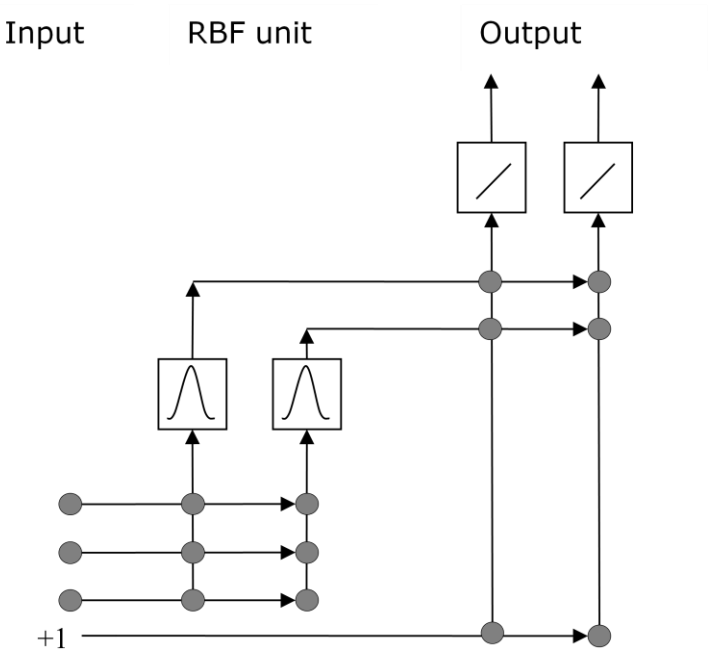

Figure 2-1. Network architecture of an RBFN. This network has three input neurons, two hidden neurons, and two output neurons.

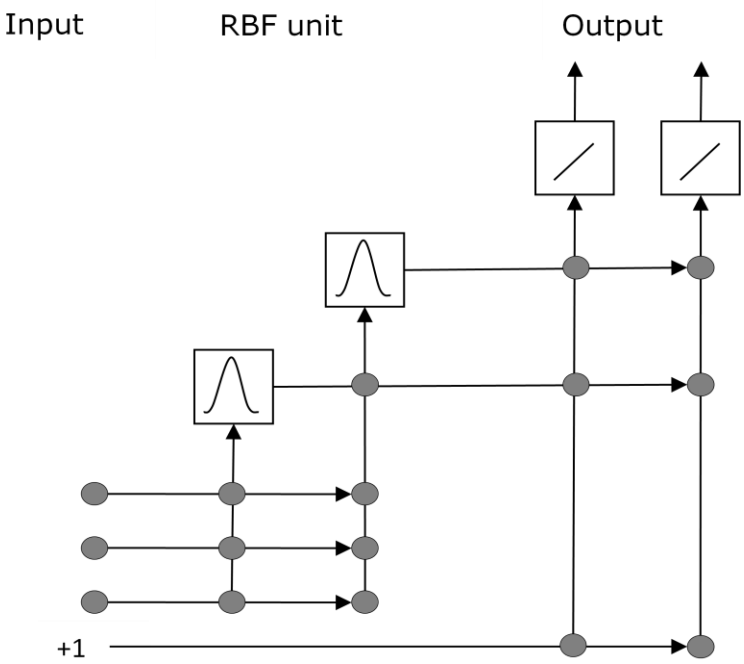

Figure 2-2. Network architecture of an RBFCCN. This network has three input neurons, two hidden neurons, and two output neurons.

2.2.1 Initialize the RBFCCN

The RBFCCN initialization is given in Figure 2-3. RBFCCN begins its training with a minimal network, which only has an input layer and an output layer. The number of input neurons *l* is equal to the number of variables of the data set. The number of output neurons *n* is equal to the number of classes in the training set. The neurons in the output layer are linear.

In this work, binary coding is used to determine the training target value. Each class has a corresponding binary sequence of unity or zero in which an element of unity indicates the identity of the object's class membership. For example, suppose an object belongs to the second class in a training set of four classes. The training target value will be encoded as (0, 1, 0, 0), i.e., the desired output vector of the trained network model is (0, 1, 0, 0).

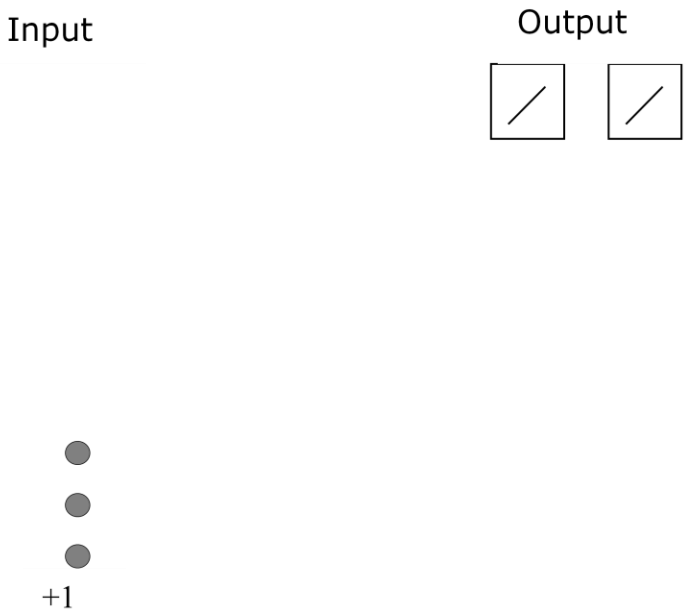

Figure 2-3. Network initialization of an RBFCCN. This network has three input neurons and two output neurons.

2.2.2 Add and Initialize a Hidden Neuron

Figures 2-4 and 2-5 demonstrate adding the first and second hidden neurons to the RBFCCN, respectively. Unlike the CCN that adds and trains a pool of candidate neurons, the RBFCCN adds and trains only one hidden neuron at a time because the initialization method applied in the RBFCCN is deterministic. The trained neuron of the RBFCCN is unique. Once the k^{th} hidden neuron is added to the RBFCCN, the centroid $\boldsymbol{\mu_k}$ is initialized with the mean vector of the target objects, and the initial radius σ_k is given by the mean of the standard deviations of the target objects. The target objects of the k^{th} hidden neuron are training objects from t_kth training class. When $k \leq n$, for which n denotes the number of training classes, $t_k = k$. When $k > n$, t_k is the class that contains the maximum total residual error among all training classes. According to the central limit theorem, all the objects from the same class tend to be normally distributed in the input space. The Gaussian function represents a class of objects. The initial hidden units represent clusters of the training data similar to LA clustering. This initialization method advantages over the random initialization method, because the value is fixed, the convergence is faster.

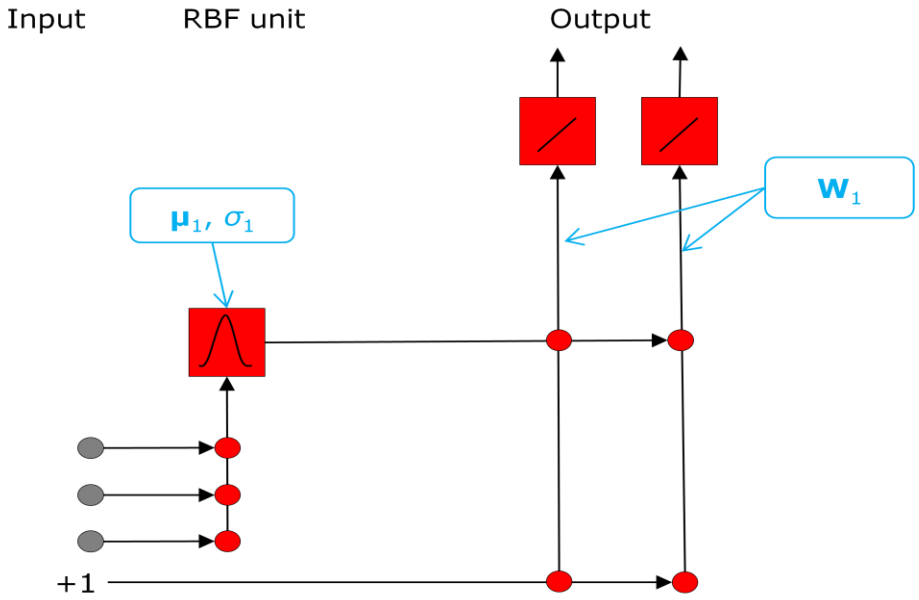

Figure 2-4. Adding first hidden neuron into an RBFCCN. This network has three input neurons, one hidden neuron and two output neurons. The neurons and connections being trained are red. $\boldsymbol{\mu_1}$, σ_1 and $\mathbf{W_1}$ are parameters to be trained.

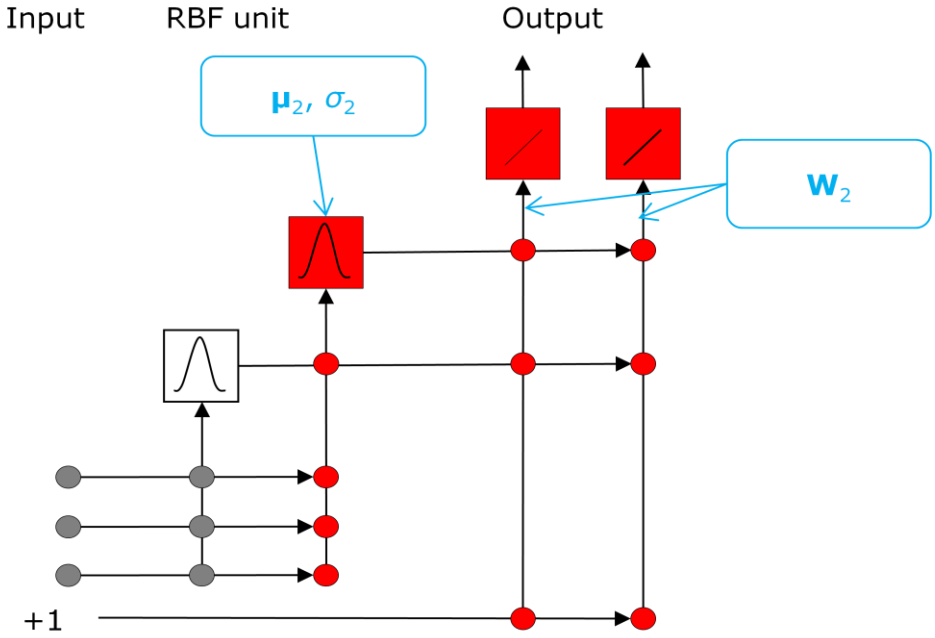

Figure 2-5. Adding second hidden neuron into an RBFCCN. This network has three input neurons, two hidden neurons and two output neurons. The neurons and connections being trained are red. $\boldsymbol{\mu_2}$, σ_2 and $\mathbf{W_2}$ are parameters to be trained.

2.2.3 Train the Hidden Neurons

The training strategy of the hidden neurons is adopted from that of the CCN. After initialization, the centroids $\boldsymbol{\mu_k}$ and radii σ_k are trained by maximizing the absolute value of the covariance between the output and the target value of a hidden unit by appropriate optimization algorithms. The covariance C_k from the k^{th} hidden unit is given by

$$C_k = \sum_{i=1}^{m} (o_{ik} - \bar{o}_k) y_{ik} \tag{2-2}$$

for which o_{ik} is the output of the i^{th} observation and the k^{th} hidden neuron; y_{ik} is the corresponding target value; m is the total number of training objects. Once a hidden neuron is trained, the centroid and radius of it will remain unchanged for the rest of the network training process.

Instead of using all training objects as target values, only objects in t_kth training class are selected as the target value for the training of the hidden neurons, for which t_k is the target class membership used in initializing k^{th} hidden neuron. As a result, the target value y_{ik} of the i^{th} object and the k^{th} hidden neuron is given by

$$y_{ik} = \begin{cases} 1 \ (c_i = t_k) \\ 0 \ (c_i \neq t_k) \end{cases} \tag{2-3}$$

for which c_i is the class membership of the i^{th} training object.

2.2.4 Train the Weights in the Output Layer

The weights in the output layer are recalculated and stored after each hidden neuron is trained. As the case in TCCCNs, the input units do not connect to the output units directly. The predicted value $\hat{\mathbf{Y}}_\mathbf{k}$ of the network

with k^{th} hidden neurons is calculated by the product of the output matrix $\mathbf{O_k}$ and the weight matrix $\mathbf{W_k}$, which is given by

$$\widehat{\mathbf{Y}}_\mathbf{k} = \mathbf{O_k W_k} \tag{2-4}$$

for which $\mathbf{O_k}$ is the output matrix for the hidden neurons. The matrix is augmented with a column of unity, which allows a bias or intercept value to be calculated. Therefore, the output matrix $\mathbf{O_k}$ has m rows and $k + 1$ columns, for which m denotes the total number of training objects. The weight matrix $\mathbf{W_k}$ stores the weight vector of the output layer. The $\mathbf{W_k}$ matrix has $k + 1$ rows and n columns, for which n denotes the number of classes of the training object. Singular value decomposition (SVD) is applied to determine the values of the weight vectors, given by

$$\mathbf{O_k} = \mathbf{U_k S_k V_k}^{\mathrm{T}} \tag{2-5}$$

in which $\mathbf{U_k}$ and $\mathbf{V_k}$ are eigenvectors that respectively span the column and row spaces for the $\mathbf{O_j}$ matrix, and $\mathbf{S_k}$ is the singular value matrix. By using SVD, the pseudoinverse of $\mathbf{O_k}$ can be computed with $\mathbf{V_k S_k}^{-1}\mathbf{U_k}^{\mathrm{T}}$. According to Eq. 2-4, $\mathbf{W_k}$ is the least squares fit of the targets onto the outputs of the hidden layer units and is given by

$$\mathbf{W_k} = \mathbf{Y V_k S_k}^{-1}\mathbf{U_k}^{\mathrm{T}} \tag{2-6}$$

for which $\mathbf{Y}$ is the target value matrix of the whole training set.

2.2.5 Evaluate the Stopping Condition of RBFCCN

The RBFCCN can be trained until a given number of hidden units were added and trained, or a given error threshold is achieved. The relative root

mean square error of calibration (RRMSEC) was used in this work. The RRMSEC is given by

$$\text{RRMSEC} = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{y}_{ij} - y_{ij}\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_j\right)^2}} \tag{2-7}$$

for which m is the total number of training objects, n is the number of classes, y_{ij} is the target value for the i^{th} object and class j, $\hat{y}_{ij}$ is the network model output for object i and class j, and $\bar{y}_j$ is the average target value for class j. To have a relative metric, the standard error of calibration is corrected by the standard deviation. By applying the RRMSEC thresholds, the experimental results only depend on different network topologies. Different training algorithms such as QuickProp, Rprop, and Bayesian approach affect the convergence time and achieve equivalent classification accuracies for the training sets.

Figure 2-6 gives the RRMSEC with respect to hidden unit number trained by the RBFCCN. The RRMSEC thresholds were determined by training an RBFCCN model using one training data set from the bootstrap Latin partition until the RRMSEC is not significantly improved. Once the RRMSEC threshold is determined, it is applied to train all the other neural networks. Of course, this method is biased in favor of the RBFCCN but it is required so that all the other reference classifiers have the same performance. However, the primary goal of this research is to compare the prediction accuracies and the ability of the different classifiers to generalize when trained to similar target values of classification accuracies. Because

the classifiers for comparison are inherently different, the RRMSEC threshold is only applied to train the network models.

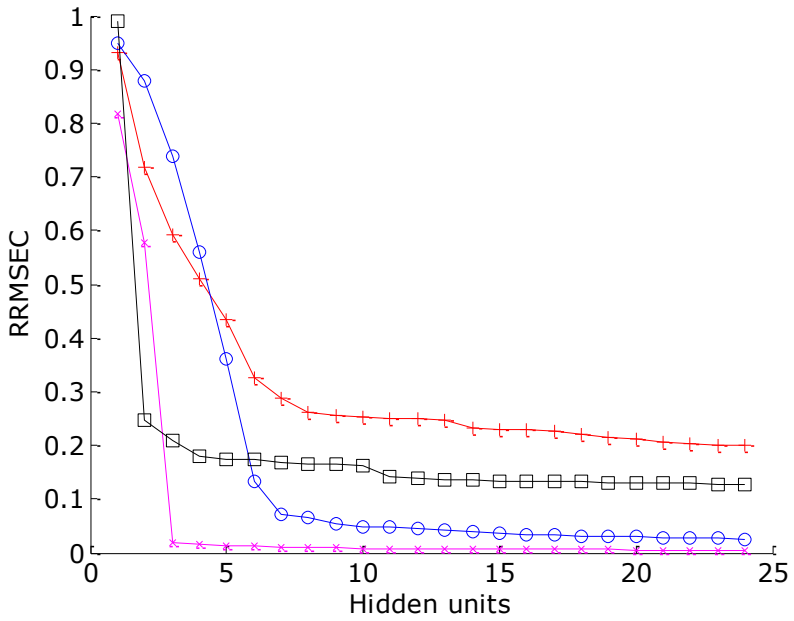

Figure 2-6. The RRMSEC with respect to hidden unit number trained by the RBFCCN using one training data set from the bootstrapped Latin partition. Magenta line with x: novel class data set; black line with □: imbalanced data set; red line with +: Italian olive oil data set; blue line with o: PCB data set.

2.2.6 Identify the Class Membership

The class membership of an object is determined by its corresponding output vector using the following strategies. When all the outputs are below 0.5, the object is labeled as unknown. Otherwise, the class is determined by the winner-take-all method, in which the unknown is classified by the index of the maximum element in the output vector. The SVM and RF have their own novel class evaluation procedure, which is not performed simultaneously with classification. The RF evaluates outliers by a metric which is referred to as proximity. The proximity measure is the fraction of trees in which a pair of objects is in the same terminal node among the entire bootstrapped trees. The proximity is based on the rule that similar objects should be in the same terminal nodes more often than dissimilar ones. The degree of the outlyingness is then calculated by the sum of squared proximities between that objects and all other objects in the same class. The novel class evaluation procedure of SVM is referred to as one-class SVM, which models the inlier into a hyper-sphere in the data space. The SVM and RF method were excluded from the novel class evaluation.

2.2.7 Advantages of RBFCCN

The RBFCCN offers several advantages. The cascade-correlation architecture has the ability of incremental learning. Incremental learning refers that the network build and expand itself during training by adding and training one hidden unit at a time. First, the incremental learning ability avoids the moving target problem in the BNN and the network converges

rapidly. Second, the cascade-correlation architecture does not require determining the size of the network prior to training. Cascade-correlation networks can be trained to a threshold of residual error instead. Third, multiple networks can be obtained by training only once. These trained networks are networks with hidden units ranging from one to the total number of hidden units added to the cascade-correlation network. Fourth, by using RBF transfer functions, RBFCCNs are suitable for performing novel class evaluation, i.e., the ability to identify unknown data or outliers in a data set.

2.3 Experimental Section

2.3.1 General Information

All calculations were performed on an AMD Athlon XP 3000+ personal computer running Microsoft Windows XP SP3 operating system. The programs were in-house scripts in MATLAB version 7.5, except for the analysis of variance (ANOVA), SVM, and RF. ANOVA was performed in Microsoft Excel version 12.0. The SVM calculations were performed by the LIBSVM software version 2.89 with MATLAB interface.[61] The RF program were obtained from reference.[62] The training of RBFCCN was implemented through fminbnd and fminunc functions by their default parameters from the MATLAB optimization toolbox version 3.1.2. The fminunc function uses the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton method with a cubic line search procedure. The fminbnd function is based on golden section search and parabolic interpolation algorithm. In the RBFCCN, RBFN, and SCRBFN,

the weights of the output neurons were updated by the SVD algorithm. The SVD algorithm is implemented by the MATLAB function pinv. The PLS-DA and all ANNs apply binary coding to determine the classes of the outputs.

Instead of training neural networks to achieve the minimum error of an external validation set, all the neural networks (the BNNs, SCRBFNs, RBFNs, and RBFCCNs) compared in this work were trained to a given RRMSEC in each data set. All the neural networks and PLS-DA applied the binary coding method to set the training target value, and the method to identify the class membership stated above. The BNNs used in this work consists of three layers: one input layer, one hidden layer and one output layer. The sigmoid neuron was used in the hidden layer, and the output layer was linear. The two-stage training method of RBFN is applied. The centroids and radii of RBFN are initialized by the K-means clustering, and optimized by back-propagation. The centroid of the k^{th} hidden neuron $\boldsymbol{\mu_k}$ is initialized by the mean of the objects in the k^{th} cluster, and the radius of the k^{th} hidden neuron σ_k is initialized by

$$\sigma_k = \left(\sum_{q=1}^{3} \left\| \boldsymbol{\mu_k} - \boldsymbol{\mu_q} \right\|^2 \right) \Big/ \sqrt{3} \qquad (2\text{-}8)$$

for which $\boldsymbol{\mu_q}$ is the three nearest neighbors of $\boldsymbol{\mu_k}$. The details of this method is described in the reference[63]. In the SCRBFNs, the parameter λ in the linear averaging clustering algorithm was adjusted gradually to achieve the RRMSEC.

For the RBFN model, the number of hidden neurons *h* equals to the number of training classes. For the BNN model, *h* is empirically proposed by

$$h = \begin{cases} n & (l \leq n) \\ \text{round}((l+n)/2) & (l > n) \end{cases} \quad \text{(2-9)}$$

for which *l* denotes the number of variables of the data set, *n* denotes the number of classes of the training object, round denotes round to the closest integer. Because two synthetic data sets were relatively simple in data size that have less variables and classes, *h* were fixed without further evaluations. To demonstrate the numbers of hidden neurons was appropriate, independent tests were performed by evaluating BNNs on two chemical data sets with half *h* and double *h* hidden neurons so that the network performances can be observed by significantly decreasing or increasing the hidden layer size.

Table 2-1 gives the average prediction accuracies of the BNN models of Italian olive oil and the training set of the PCB data set. For the Italian olive oil data set, the BNN models with 7 and 14 hidden neurons did not significantly differ with respect to prediction accuracy. The BNN models with four hidden neuron had too few hidden units to model the data sufficiently. For the PCB dataset, the effect of the three different numbers of hidden neurons on the prediction results was not significant. The BNNs with extra hidden neurons will not overfit the data if trained to the same RRMSEC. As a result, the heuristic equation of *h* was appropriate.

Table 2-1. Average prediction accuracies of the BNN models with 95% confidence intervals of Italian olive oil and the training set of the PCB data set. The BNN was trained by different number of hidden neurons with 30 BLPs.

Data set	Number of hidden neurons	Prediction accuracy
Olive oil	7	95.5 ± 0.3
	14	95.9 ± 0.4
	4	87.7 ± 0.2
PCB	13	99.9 ± 0.1
	26	99.9 ± 0.1
	7	99.9 ± 0.1

To determine the learning rates and momenta of the BNNs and RBFNs, these networks were trained by three different sets of learning rates and momenta with BLPs. Thirty bootstraps and two partitions were applied. Table 2-2 gives the prediction results by the Italian olive oil data set and the training set of the PCB data set. The training parameters of the back-propagation networks did not significantly affect the comparison of the modeling methods. These sets of learning parameters were also trained by the two synthetic data sets and same results were obtained. For each data set, there was no statistical difference of the BNN and RBFN prediction results at a 95% confidence interval by two-way ANOVA with interaction. Therefore, the learning rates and momenta were fixed respectively at 0.001 and 0.5 for all further evaluations.

Table 2-2. Average prediction accuracies of the BNN and RBFN models with 95% confidence intervals of Italian olive oil and the PCB data sets. The BNN and RBFN were trained by three different sets of learning rates and momenta with 30 BLPs.

Data set	Learning rate	Momentum	BNN	RBFN
Olive oil	0.001	0.5	95.5 ± 0.3	92.0 ± 0.7
	1×10^{-4}	0.5	95.4 ± 0.3	91.9 ± 0.7
	0.001	0	95.6 ± 0.3	91.5 ± 0.6
PCB	0.001	0.5	99.9 ± 0.1	92 ± 6
	1×10^{-4}	0.5	99.5 ± 0.2	91 ± 6
	0.001	0	99.9 ± 0.1	94 ± 5

The PLS-DA was implemented by the non-linear iterative partial least squares (NIPALS) algorithm[64]. The number of latent variables was determined by minimizing the root mean squared prediction error in each test. As a result, the PLS-DA was a biased reference method. The numbers of latent variables in the PLS-DA models may vary between runs.

All the SVMs used the Gaussian RBF as their kernel functions. Two SVM parameters: the cost c and the RBF kernel parameter γ must be adjusted before each prediction. The grid search of parameter pairs (c, γ), in which $c = 2^i$, i = -2, -1, 0, …, 20; $\gamma = 2^j$, j = -10, -9, -8, …, 10, was performed to determine their value by achieving the best training accuracies. The defaults of the remaining parameters were used. Because the result of the RF algorithm is not sensitive to the parameter selected, 1000 trees with the default setting of the number of variables to split on at each node is used in all evaluations.

Thirty bootstraps and two Latin partitions were applied to evaluate all the data sets in this study. The results are reported as prediction accuracy, which is the percentage of correctly predicted objects. Four data sets were tested, including the novel class data set, imbalanced data set, Italian olive oil data set, and the PCB data set. The numbers of variables, objects, and classes of data sets are given in Table 2-3. The modeling parameters of the ANNs, PLS-DA, SVM, and the RF method are given in Table 2-4. Similar to the latent variables used in the PLS-DA models, the numbers of hidden neurons used to train SCRBFN and RBFCCN models may vary between different runs. Therefore, only typical latent variables and numbers of hidden neurons are reported.

Table 2-3. The numbers of variables, objects, and classes of the data sets evaluated.

	Novel class		**Imbalanced**		**Olive oil**	**PCB**	
	Training	Test	Training	Test	BLP validation	BLP validation	External validation
Variables	2	2	2	2	8	18[a]	18[a]
Objects	400	100	610	10	478	131	154
Classes	4	1	3	1	6	7	8[b]

[a]This number is the number of variables after the modulo preprocessing method.

[b]The PCB non-PCB compounds and PCB congeners that contain 1, 9 and 10 chlorine atoms were considered as one class.

Table 2-4. The modeling parameters of the ANNs, PLS-DA, SVM, and the RF method. Hidden units are the number of hidden units in the trained network model. Latent variables are the number of latent variables used in the PLS-DA models. The RBF kernel parameter is denoted by γ in the SVM method. Mtry is the number of variables to split on at each node in the RF method.

	Modeling parameters	**Novel class**	**Imbalanced**	**Olive oil**	**PCB**
	RRMSEC threshold	0.02	0.2	0.4	0.1
BNN	Learning rate	0.001	0.001	0.001	0.001
	Momentum	0.5	0.5	0.5	0.5
	Hidden units	4	3	7	13
RBFN	Learning rate	0.001	0.001	0.001	0.001
	Momentum	0.5	0.5	0.5	0.5
	Hidden units	4	3	6	7
SCRBFN	Hidden units	-	17	~6	~20-30
RBFCCN	Hidden units	4	3	~6	~8
PLS-DA	Latent variables	-	2	~8	~16-18
SVM	Cost	-	2^{10}	2^{10}	2^{13}
	γ	-	2^{-1}	1	2^{-5}
RF	Number of trees	-	1000	1000	1000
	Mtry	-	1	2	4

2.3.2 Detection of a Novel Class Using a Synthetic Data Set

This synthetic data set was designed to test the BNN, RBFN, and RBFCCN abilities to respond to a novel class during prediction. The training set comprised two variables and four classes. Each training class and the test objects had 100 objects. Each class was normally distributed with means of (0.0, 0.0), (40.0, 0.0), (0.0, 40.0), and (40.0, 40.0), respectively, with a standard deviation of 1.5. The test objects were distributed about a mean of (20.0, 20.0) with a standard deviation of 1.5. Both networks were trained repeatedly 30 times on this data set to obtain statistically reliable results.

2.3.3 Synthetic Imbalanced Data Set

An imbalanced data set is a data set that the numbers of objects are not equal in each class. This data set was designed to compare the performances when the data set is highly imbalanced. This data set has two variables. The training set comprised three normally distributed classes. Two classes are majority classes, which have 300 objects respectively distributed with means of (3.0, 0.0), (-3.0, 0.0) and with standard deviations of unity. The other training class is the minority class that has only 10 objects distributed about a mean of (0.0, 0.0) with a standard deviation of 0.1. The test class has the same distribution with the minority training class. The ANNs were trained to the RRMSEC thresholds of 0.2. The network performances were evaluated by predicting the minority class in the training

set. All modeling methods were reconstructed 30 times to obtain statistically reliable results.

2.3.4 Italian Olive Oil Data Set

Italian olive oil data were obtained from references[65, 66]. This data set is a well-studied standard reference data set. Different source regions of Italian olive oil were classified by the profile of eight different fatty acids. To minimize the effect of class imbalance and obtain fair comparison results, objects from smaller classes that have less than 50 objects were removed from the evaluation data. The number of classes was six. Each variable in the training sets was scaled between 0 and 1. The variables of the test sets in each Latin partition were scaled using the range acquired from the training set to obtain unbiased results. The RRMSEC threshold for training was 0.4.

2.3.5 PCB Data Set

In the PCB data set, PCB congeners with different numbers of chlorine atoms were classified by their electron ionization mass spectra.[54, 58] The mass spectra were obtained from reference[67]. These spectra were split into the training set and the external validation set. The PCB congeners in the training set contained 2 to 8 chlorine atoms. Most of the PCB congeners have duplicate spectra. Among these duplicate spectra, the spectra with the lowest record numbers were selected for training, because these spectra are of the highest quality in the reference library. The PCB congeners in the external validation set contained 0 to 10 chlorine atoms. The external validation set was built from the remaining replicate spectra. This data set

comprised of PCB congeners that have less than 10 objects, and non-PCB compounds. The non-PCB compounds and PCB congeners that contain 1, 9 and 10 chlorine atoms were uniquely different from any of the training classes. The external validation set contains 45 unique spectra.

Each spectrum was mean-centered and normalized to unit vector length. The spectra were transformed to a unit mass-to-charge ratio scale that ranged from 50 to 550 Th and any peaks outside this range were excluded. Because the raw data were underdetermined, i.e., there were more variables than objects, the dimensions of PCB data set were further reduced by using the modulo preprocessing method[68, 69]. This compression method is especially effective for mass spectral data. Based on the previous study[54] by the principal component analysis (PCA), the divisor value of 18 was chosen. The compressed spectra were centered about their mean and normalized to unit vector length. The training RRMSEC thresholds were 0.1.

2.4 Results and Discussion

2.4.1 Detection of a Novel Class Using a Synthetic Data Set

The bivariate plot of the synthetic data set is given in Figure 2-7. The response surface of the BNN is in Figure 2-8. RBFN and RBFCCN networks have similar response surfaces, given in Figure 2-9. For each sampling point, the maximum of the output neurons is plotted. Because of the different shapes and properties of the sigmoid function and the Gaussian function, these networks have unique response surfaces. The BNN model gave an open, sigmoidal shaped response surface that divides the output space into

regions that correspond to the four classes. When the BNN model extrapolates outside the region defined by the data objects, the response can be larger than unity, which is occurs when the output units are linear. Alternatively, the RBFCCN and RBFN had a Gaussian shaped response surface that has a finite span of the output space, which is closed and compact. The maximum response of RBFCCN is unity.

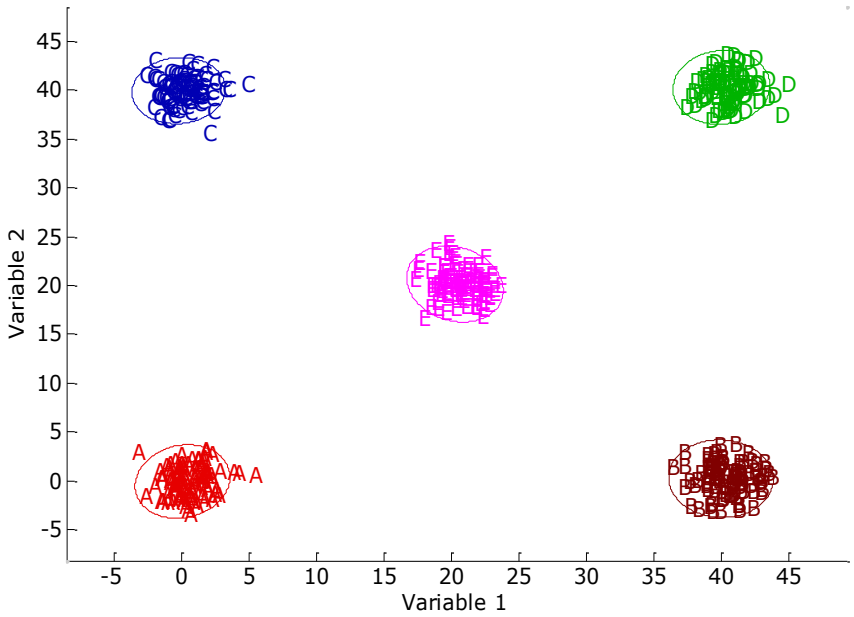

Figure 2-7. Two-variable plot of the synthetic novel class data set. A, B, C, and D denote the training sets, and E denotes the test set. The 95% confidence intervals were calculated around each training class.

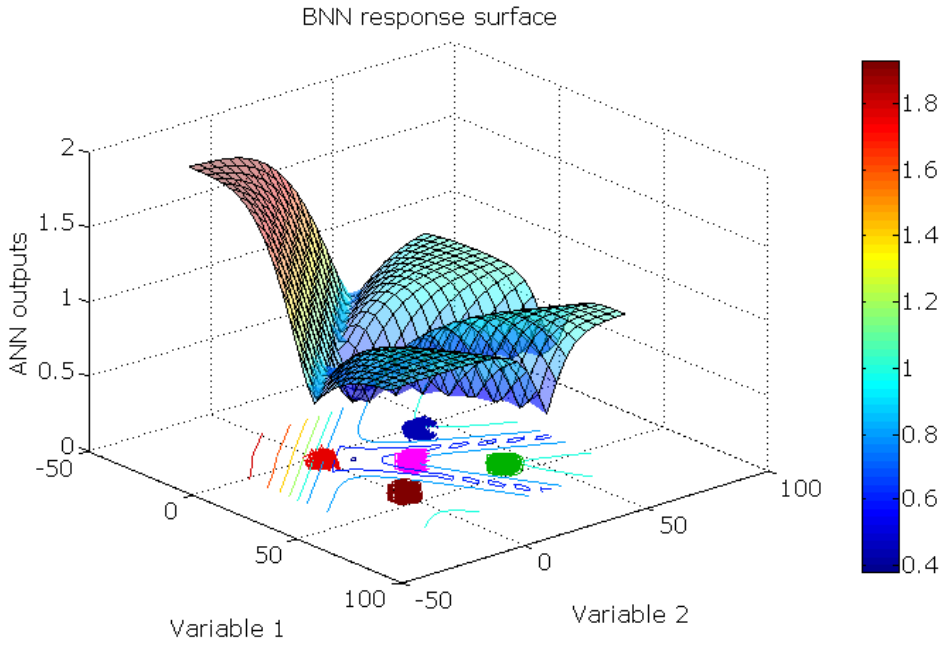

Figure 2-8. The BNN response surface of the synthetic novel class data set. For each sampling point, the maximum of the output neurons is plotted.

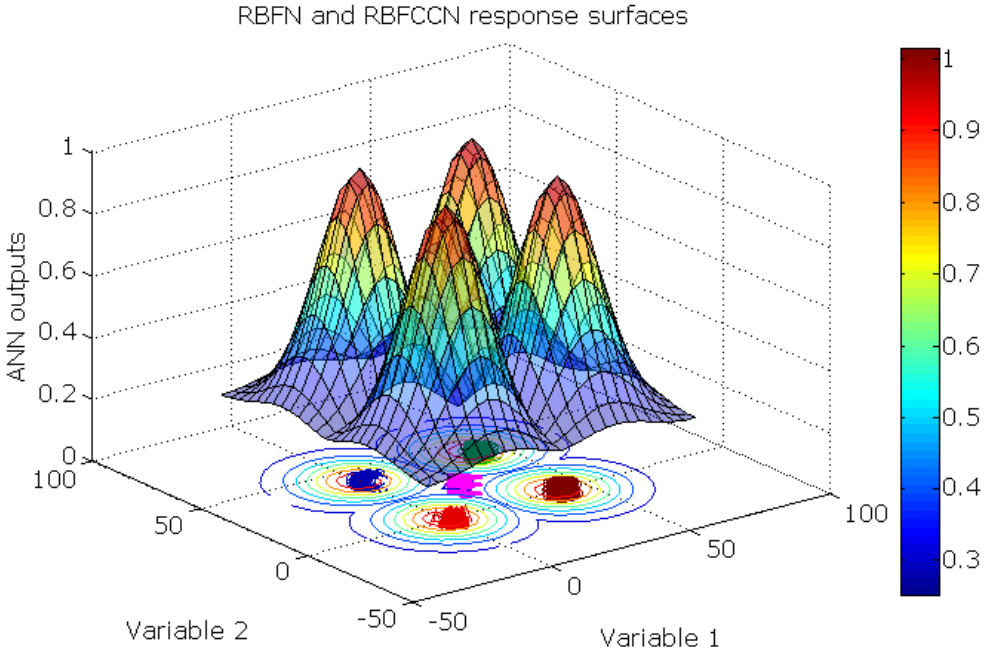

Figure 2-9. The RBFN and RBFCCN response surface of the synthetic novel class data set. For each sampling point, the maximum of the output neurons is plotted.

The test set was designed to be uniquely different from the data in the training set. The ideal prediction results of these test objects should be no excitation from any of the output neurons, i.e., the outputs are (0, 0, 0, 0). Figure 2-10 gives different outputs of the test set with respect to the different models. The outputs of RBFCCN and RBFN were the same for all repeats. The trained BNN models have different weights, hence different response functions each time they are trained. Because in most cases one output element was larger than 0.5, the BNN models misclassified most of the test objects as one of the training classes. The RBFCCN and RBFN models correctly identified all test objects as unknown. Compared to the RBFN models, the prediction of RBFCCN models were closer to the ideal solution, because the outputs from the RBFN models spread more widely than the RBFCCN models.

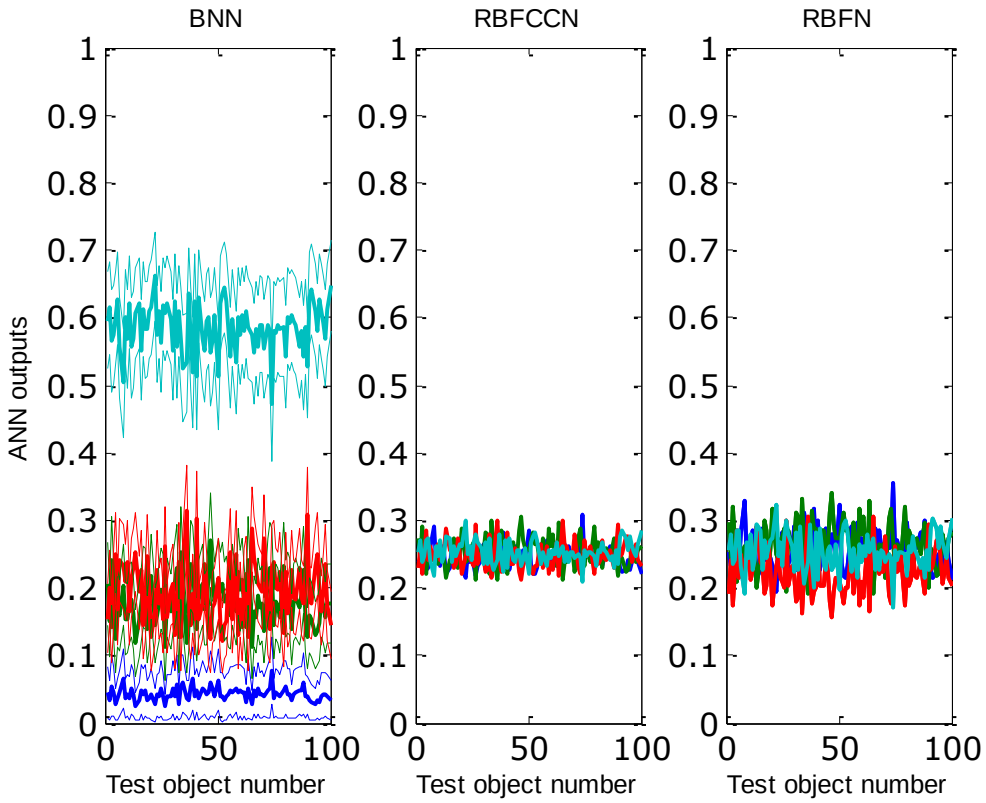

Figure 2-10. Average prediction outputs from the test set. BNN, RBFCCN and RBFN models were obtained by training each network 30 times. The 95% confidence intervals are indicated as the thin lines around the BNN outputs. Different colors represent excitations from different output neurons.

2.4.2 Synthetic Imbalanced Data Set

The bivariate plot of the synthetic imbalanced data set is given in Figure 2-11. It can be observed that objects in two majority classes have larger spans than the minority classes in the input space. The predictions of small classes by different ANNs are given in Table 2-5. The prediction of the SCRBFN and RBFCCN models are better than the prediction results of the BNN and PLS-DA models. The RBFCCN, SVM, and RF methods have better predictions among all seven methods. The RBFN models have slightly worse prediction results than the three methods above. The trained models of the ANNs will have a relatively loose fits to the training set by setting the training error threshold to 0.2.

The BNN and PLS-DA models trend to first model the majority classes in the prediction class. As a result, predictions of minority classes are poor.

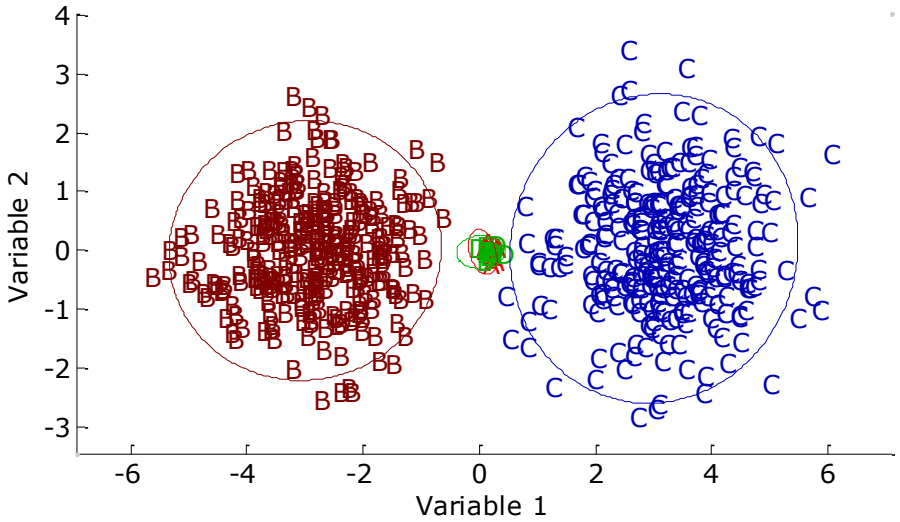

Figure 2-11. Two-variable plot of the synthetic imbalanced data set. A (red), B, and C denote the training classes. D (green) denotes test class. The 95% confidence intervals were calculated around each training class.

Table 2-5. Average numbers of correctly predicted objects with 95% confidence intervals from class D in an imbalanced data set by different models. All modeling methods were reconstructed 30 times.

	Total	**BNN**	**SCRBFN**	**RBFN**	**RBFCCN**	**PLS-DA**	**SVM**	**RF**
Correctly predicted	10	0	7	9.1 ± 0.1	10	0	10	10

2.4.3 Italian Olive Oil Data Set

The principal component scores of the Italian olive oil data set are given in Figure 2-12. From this plot, it can be seen that objects in the same classes form clusters, but the confidence intervals are overlapped with each other. The prediction accuracies of different ANN models are given in Table 2-6. The SVM and RF models have the highest prediction accuracy of 97.9%. The results calculated by the BNN and RBFCCN models are better than the

results calculated by the SCRBFN and RBFN models. The PLS-DA models yield a lower average prediction accuracy of 89.8%. A two-way ANOVA with interaction at a significance level of 0.05 was performed to analyze the sources of variation and prediction accuracies. The results of ANOVA are given in Table 2-7. Different modeling methods, different source regions and the interaction between the classifiers show significant differences in prediction. The ANOVA results indicate that the methods evaluated have different performances. The SVM and RF perform best among the methods evaluated. The RBFCCN and BNN have statistically better performance in predicting this data set compared to PLS-DA, RBFN, and SCRBFN.

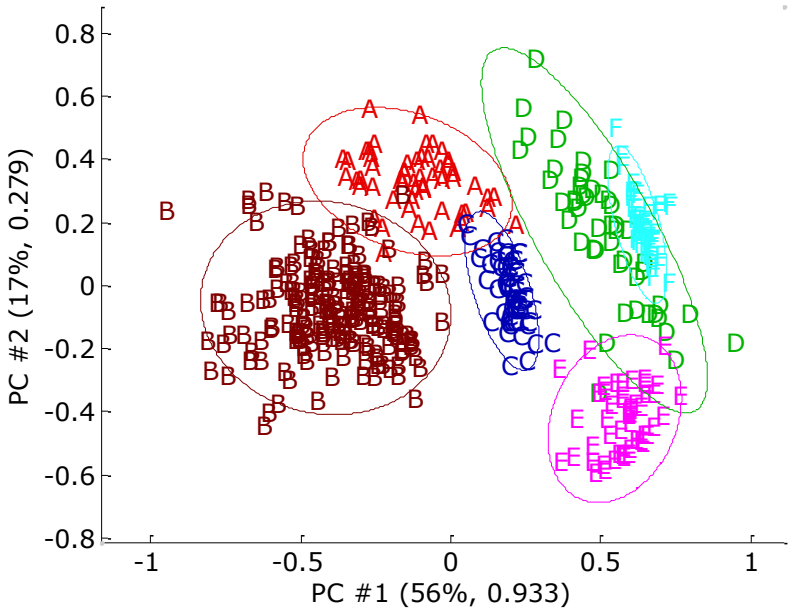

Figure 2-12. A principal component score plot for the olive oil data set. Each axis is labeled with the percent total variance and the absolute eigenvalue. Each observation of the data set was scaled to [0, 1]. The 95% confidence intervals appear as an ellipse around each class. The sources regions are: (A) Calabria; (B) South Apulia; (C) Inland Sardinia; (D) East Liguria; (E) West Liguria; (F) Umbria.

Table 2-6. Average numbers of correctly predicted objects with 95% confidence intervals of Italian olive oil data set by different modeling methods with 30 BLPs.

Source regions	Total	BNN	SCRBFN	RBFN	RBFCCN
Calabria	56	50.9 ± 0.6	50.3 ± 0.5	52.5 ± 0.6	52.7 ± 0.2
South Apulia	206	203.8 ± 0.5	199.5 ± 0.6	203.5 ± 0.4	200.1 ± 0.3
Inland Sardinia	65	65	58.3 ± 0.3	63.4 ± 0.4	64.2 ± 0.2
East Liguria	50	38 ± 1	37.4 ± 0.7	24.4 ± 3.6	35.4 ± 0.7
West Liguria	50	48.2 ± 0.3	43.3 ± 0.5	47.6 ± 0.8	48.5 ± 0.3
Umbria	51	50.4 ± 0.4	40.4 ± 0.4	48.7 ± 0.8	50.9 ± 0.1
Prediction accuracy (%)		95.5 ± 0.3	89.8 ± 0.3	92.0 ± 0.7	94.5 ± 0.2
Source regions	**Total**	**PLS-DA**	**SVM**	**RF**	
Calabria	56	40.9 ± 0.5	53.6 ± 0.4	53.2 ± 0.4	
South Apulia	206	202.4 ± 0.4	202.5 ± 0.6	203.5 ± 0.7	
Inland Sardinia	65	63.7 ± 0.4	65	65	
East Liguria	50	27 ± 1	47.3 ± 0.4	46.4 ± 0.4	
West Liguria	50	48.4 ± 0.3	48.9 ± 0.4	49.0 ± 0.2	
Umbria	51	47.2 ± 0.6	50.9 ± 0.1	50.9 ± 0.1	
Prediction accuracy (%)		89.8 ± 0.3	97.9 ± 0.2	97.9 ± 0.2	

Table 2-7. ANOVA table of the Italian olive oil data set by different source regions and modeling methods. F_{crit} is the critical value.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	F_{crit}
Source regions	4.44	5	0.89	1919.9	2.22
Modeling methods	0.53	6	8.79×10^{-2}	189.8	2.11
Interaction	2.03	30	6.78×10^{-2}	146.5	1.47
Within	0.56	1218	4.63×10^{-4}		
Total	1.49	1259			

2.4.4 PCB Data Set

The principal component scores of the PCB data are given in Figure 2-13. The principal components and mean were calculated only from the training set. The training set was labeled with upper case letters. The external validation set was projected onto the first two principal components, labeled with underlined lower case letters. The external validation scores were more dispersed than the training set. A part of the external validation scores are outside of the 95% confidence interval of their respective class because of their low quality. The principal component scores of non-PCB compounds and PCB congeners that contain 1, 9 and 10 chlorine atoms are uniquely different from the training set. The BLP internal validation for the training set alone was first performed. The prediction accuracies of internal validations are given in Table 2-8. The average prediction accuracies of the SVM, RF, RBFCCN, and BNN models are better than the average prediction accuracy of the SCRBFN, RBFN, and PLS-DA model.

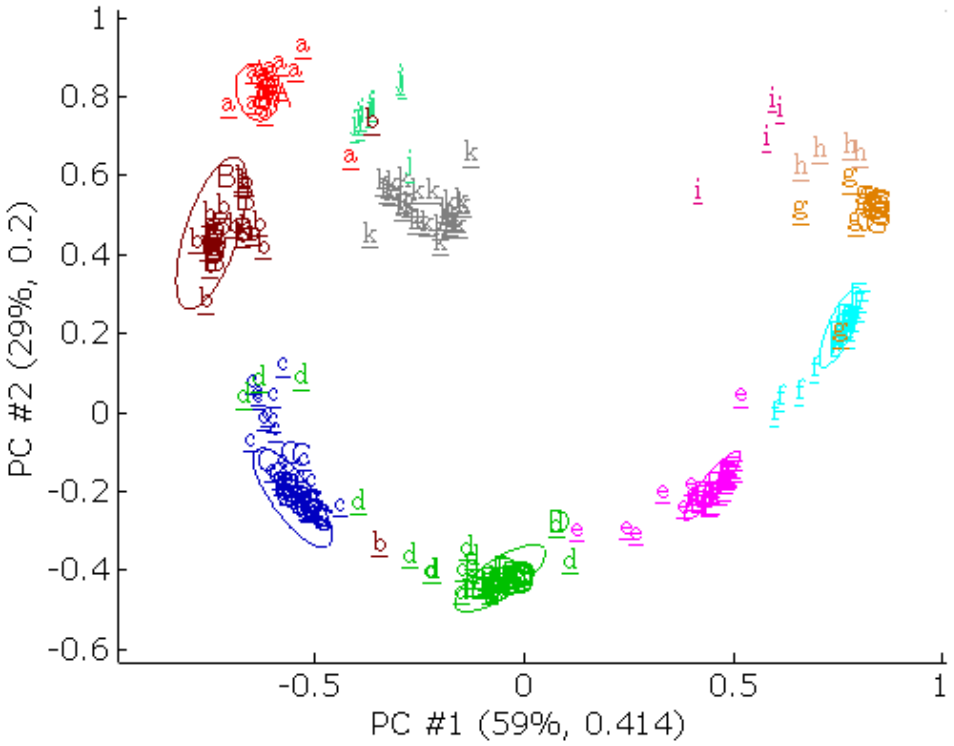

Figure 2-13. A principal component score plot for the PCB data set. The letters with upper case represents the training set. The underlined letters with lower case represents the external validation set. The external validation set was projected onto the first two principal components from the training set. Each axis is labeled with the percent total variance and the absolute eigenvalue from the training set. The 95% confidence intervals were calculated and given as an ellipse around each class from the training set. The PCB congeners are: (A) 2; (B) 3; (C) 4; (D) 5; (E) 6; (F) 7; (G) 8; (H) 9; (i) 10; (j) 1; (k) 0, the numbers denotes the number of chlorine atoms in the PCB congeners.

Table 2-8. Average numbers of correctly predicted spectra with 95% confidence intervals of PCB data set by different modeling methods with 30 BLPs.

Cl number	Total	BNN	SCRBFN	RBFN	RBFCCN
2	10	10	8.3 ± 0.5	8.0 ± 0.8	10
3	12	12	11.1 ± 0.8	11.0 ± 0.7	12
4	28	28	26 ± 2	26 ± 2	28
5	29	28.9 ± 0.1	27 ± 2	27 ± 2	28
6	24	24	23.3 ± 0.9	23 ± 1	24
7	18	18	17 ± 1	17 ± 1	18
8	10	10	9.0 ± 0.7	9.4 ± 0.6	10
Prediction accuracy (%)		99.9 ± 0.1	93 ± 5	92 ± 6	99.2
Cl number	**Total**	**PLS-DA**	**SVM**	**RF**	
2	10	9.9 ± 0.1	10	10	
3	12	11.9 ± 0.1	12	11.9 ± 0.1	
4	28	27.5 ± 0.3	28	28	
5	29	26.2 ± 0.5	29	29	
6	24	22.0 ± 0.4	24	24	
7	18	14.8 ± 0.6	18	18	
8	10	10.0 ± 0.1	10	10	
Prediction accuracy (%)		93.3 ± 0.6	100	99.9 ± 0.1	

After internal validation, the entire training set was trained and the external validation set was predicted repeatedly 30 times. The prediction accuracies of external validation set are given in Table 2-9. The prediction accuracy without novel classes (i.e., classes that were not used during training) is the prediction accuracy calculated by the external validation set excluding the non-PCB congeners that contain 1, 9 and 10 chlorine atoms. The total prediction accuracy is the prediction accuracy calculated by the complete external validation set. Because the prediction set contained low quality spectra that make the data set more difficult to classify, the result is generally worse than BLP validation of the training set. The SVM, BNN, and RF method obtained better results than other methods. The RBFCCN models yielded average prediction accuracy of 81.7% without the novel classes, which was ranked fifth among seven methods. The RBFCCN and SCRBFN models respectively indentified 95.6% and 100% of the unknown objects. The BNN and RBFN models were capable of classifying the test objects, but they can hardly identify the unknown objects. This result is consistent with the result from the synthetic novel class data set. As a result, the BNN and PLS models yielded total prediction accuracies less than 65%.

Table 2-9. Average numbers of correctly predicted spectra with 95% confidence intervals of PCB external validation data set. All modeling methods were reconstructed 30 times. The prediction accuracy without unknown is the prediction accuracy calculated by the external validation set excluding the non-PCB compounds and PCB congeners that contain 0, 1, 9 and 10 chlorine atoms. The total prediction accuracy is the prediction accuracy calculated by the complete external validation set.

Cl number	Total	BNN	SCRBFN	RBFN	RBFCCN
2	13	13	11	12.1 ± 0.1	12
3	20	17.5 ± 0.4	17	16.6 ± 0.2	16
4	28	26.8 ± 0.4	19	24.9 ± 0.2	27
5	21	17.2 ± 0.4	20	15.5 ± 0.4	16
6	13	11.3 ± 0.3	10	11.0 ± 0.5	7
7	7	6.7 ± 0.2	5	5.0 ± 0.1	6
8	7	6	3	4.8 ± 0.4	5
0,1,9,10	45	0.3 ± 0.4	45	23.5 ± 8.3	43
Prediction accuracy without unknown (%)		90.4 ± 0.6	78.0	82.4 ± 0.5	81.7
Total prediction accuracy (%)		64.0 ± 0.5	84.4	74 ± 6	85.7

Cl number	Total	PLS-DA	SVM	RF
2	13	10	13	13
3	20	14	19	18.6 ± 0.2
4	28	28	28	26.3 ± 0.2
5	21	11	17	15.0 ± 0.1
6	13	8	11	11.1 ± 0.2
7	7	7	7	5
8	7	6	6	6
0,1,9,10	45	0	-	-
Prediction accuracy without unknown (%)		77.1	92.7	87.2 ± 0.4
Total prediction accuracy (%)		54.5	-	-

2.5 Conclusions

The proposed RBFCCN network combines the concepts of RBFN and CCN. During the training of RBF hidden units, an RBFCCN applies both the initialization technique similar to that of the SCRBFN and the optimization technique of CCNs. The cascade correlation algorithm furnishes the incremental learning ability of the RBFCCN. The incremental learning ability ensures the RBFCCN automatically builds its network topology during training. Before training RBFCCNs, no prior information about network topology is required. As a result, training RBFCCNs are more convenient than training BNNs. Another advantage of cascade-correlated structure is that it avoids the moving target problem and converges more rapidly than the BNNs.

RBFCCNs, BNNs, RBFNs, SCRBFNs, PLS-DAs, SVMs and RFs were tested with four data sets. The test results were obtained with statistical measurements of confidence intervals. The SVM and RF methods proved their excellence over the neural network approaches on these classification problems. All three neural networks were generally yielded better performance than PLS-DA in prediction. Compared with the RBFN and SCRBFN models in four test data sets, the RBFCCN models generally yielded better prediction accuracies. The RBF transfer function applied in RBFCCNs makes RBFCCNs a reliable approach for novel class evaluation. RBFCCNs generally yielded better novel class evaluation ability compared with RBFNs, BNNs and PLS-DA by setting an output threshold 0.5. The RBFCCN is also

capable of modeling imbalanced data. The RBFCCN was statistically shown to be a robust and effective classification algorithm for chemometrics, especially in novel class evaluation and outlier detection.

Future work will involve in developing novel training methods to train the networks more rapidly. Investigations of different optimization algorithms such as the genetic algorithms and particle swarm optimizations to train RBFCCNs are necessary. In addition, it is important to compare RBFCCNs with other methods for outlier or novel class evaluation, such as one-class SVM in chemical data sets.

Chapter 3 A Discriminant Based Charge Deconvolution Analysis Pipeline for Protein Profiling of Whole Cell Extracts Using Liquid Chromatography–Electrospray Ionization-Quadrupole Time-of-Flight Mass Spectrometry

Adapted with permission from Lu, W., Callahan, J.H., Fry, F.S., Andrzejewski, D., Musser, S.M., and Harrington, P.B.; *Talanta* **2011**, *84(4)*, 1180-1187.

3.1 Introduction

Electrospray ionization-mass spectrometry (ESI-MS) methods such as liquid chromatography–electrospray ionization-mass spectrometry (LC–ESI-MS) and capillary electrophoresis–electrospray ionization-mass spectrometry (CE–ESI-MS)[70] have been applied to the analysis of intact proteins in recent years. CE–ESI-MS has been applied to the diagnosis of cancer[71], to detect glycoproteins[72] and ribosomal proteins of *Escherichia coli*[73], etc. Protein expression profiling analyses on bacteria whole cell extracts without proteolytic digestion by liquid chromatography–electrospray ionization-quadrupole time-of-flight mass spectrometry (LC–ESI-QTOF MS) have been reported.[74-77] A new analysis pipeline to process the LC–ESI-MS data for whole cell extracts is proposed. LC–ESI-MS spectra of biological samples often have hundreds of component peaks, and some peaks may have a low signal-to-noise ratio (SNR). The proposed analysis pipeline applies data processing methods including wavelet denoising, baseline removal, peak binning, peak centroiding, and multivariate classification to extract proteomic

information. The analysis can be automated and the results are easier to reproduce than manual spectra inspection.[74, 78-80]

For biomolecules such as proteins, ESI-MS often produces multiply charged spectra. This multiplicity of charge states gives an envelope of peaks in a spectrum for each component. The charge state deconvolution method, also known as deconvolution, is a method that determines the molecular mass of a biomolecule from multiply charged ESI-MS peaks.[35] Deconvolution transforms a multiply charged ESI-MS spectrum into a zero-charge or singly charged spectrum. The zero-charge spectrum is a spectrum that each component data point has an abscissa of its molecular mass. Similarly, in the singly charged spectrum, each component data point has an abscissa that is the mass-to-charge ratio of its singly charged ion.

All multiply charged spectra were deconvolved into zero-charge spectra in this work. Many deconvolution methods have been proposed, such as thorough high resolution analysis of spectra by Horn (THRASH)[44], molecular weight determination (MoWeD)[41], maximum entropy deconvolution[37], multiplicative correlation algorithm (MCA)[40], Zscore[42], etc. Generally, different deconvolution algorithms are designed for low-resolution and high-resolution ESI-MS data, based on whether the isotopic peaks are resolved.[42] MoWeD, MCA, and maximum entropy deconvolution were suitable for low-resolution mass spectrometry such as quadrupole time-of-flight mass spectrometry (QTOF-MS), which has unresolved isotopic peaks, particularly for high-charge states and high molecular weight proteins.

THRASH was designed for high-resolution mass spectrometry (resolved isotopic peaks) such as Fourier transform ion cyclotron resonance–mass spectrometry (FTICR-MS). Zscore has different deconvolution routines for each resolution. The deconvolution method was originally designed to deconvolve ESI-MS spectra. To deconvolve a two-way LC–ESI-MS spectrum, the deconvolution should be performed on the binned ESI-MS scan in a given retention time window.

The deconvolved spectra can be used as input data to multivariate pattern classifiers, and the discriminant rules with candidate biomarker information can be obtained. In this work, this general approach was performed by the MoWeD deconvolution algorithm and the fuzzy rule-building expert system (FuRES)[31] as the pattern classifier. This general analysis approach is named MoWeD–FuRES. The MoWeD deconvolution algorithm was chosen because this algorithm is relatively efficient, simple, and is suitable for the deconvolution of low-resolution mass spectra. FuRES combines the advantage of the fuzzy logic data analysis with the decision tree algorithm. FuRES has been successfully applied in matrix-assisted laser desorption/ionization-mass spectrometry (MALDI-MS) data analysis such as mouse age identification[81], premature or at-term deliveries classifications of amniotic fluids[82]. FuRES is also capable of two-way data analysis in applications such as classification of jet fuels[33] and ignitable liquids[53].

In general analysis approaches such as MoWeD–FuRES, applying deconvolution and the data processing methods on an individual mass

spectrum is time-consuming, especially when there are many samples and the noise threshold in the deconvolution method is low. A novel hyphenated approach named FuRES–MoWeD is proposed in the analysis pipeline, which applies charge deconvolution algorithm to multivariate pattern recognition rules. Because FuRES was performed prior to the deconvolution, only peaks correlated to each class are deconvolved. As a result, the FuRES–MoWeD approach is more robust in terms of the classification ability than MoWeD–FuRES with respect to baseline drift and noise. In addition, this approach is efficient because the deconvolution is performed only once for a set LC–ESI-MS of spectra. FuRES–MoWeD is compared to the MoWeD–FuRES approach on a synthetic data set and a *Salmonella enterica* strain identification data set.

3.2 Theory

The FuRES–MoWeD analysis approach focuses on effectively obtaining molecular component differences between different classes of samples. Figure 3-1 gives the flowcharts of FuRES–MoWeD and MoWeD–FuRES approaches for comparison. Both approaches combine the pattern recognition methods with the charge deconvolution. In MoWeD–FuRES approach, charge deconvolution is applied to every ESI-MS scan. When the chromatographic separation time is long, the deconvolution will be computationally demanding. As a result, the deconvolution should be performed after binning the spectra in a given retention time interval.

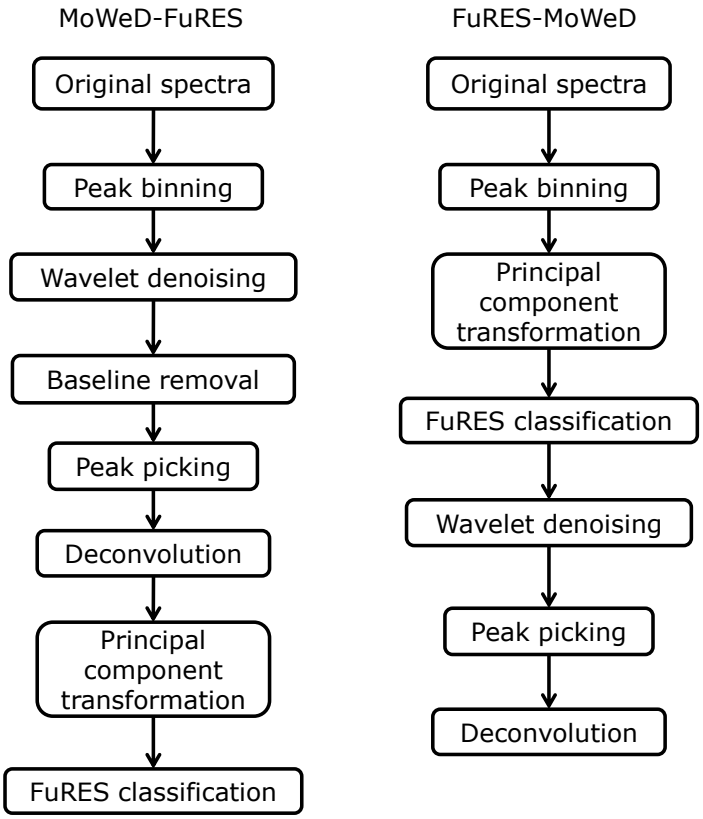

Figure 3-1. Diagram of key steps in a general LC–ESI-MS analysis pipeline (MoWeD–FuRES) and the analysis pipeline proposed in this work (FuRES–MoWeD).

In the proposed FuRES–MoWeD approach, FuRES classification is applied to the LC–ESI-MS spectra prior to charge deconvolution. The resultant FuRES discriminant comprises a two-way image of retention time and mass-to-charge ratio. The spectrum of each retention time in the discriminant is then deconvolved from the FuRES discriminant image. Because the discriminant image contains fewer data points than a complete set of sample data, FuRES–MoWeD approach reduces the analysis time compared to MoWeD–FuRES approach. In addition, because the discriminants only retain relevant information for the classification, there is less noise in the discriminants than the sample spectra. Therefore, the analysis result is less affected by noise in the spectra than MoWeD–FuRES approach.

Compared to a LC–ESI-MS spectrum, a FuRES discrimination rule is different because the discrimination rule contains both positive and negative peaks that respectively represent the relative importance for two classes. To extract the zero-charge spectrum from the FuRES discrimination rule, first the deconvolution was performed on the absolute value of the discrimination rule spectrum. Afterwards, the sign of each peak of the zero-charge spectrum from the deconvolved rule was determined by counting and comparing the number of positive and negative peaks in the charge state pattern calculated from the MoWeD algorithm (step 3). If the number of positive rule peaks in the pattern is larger than the negative peaks, the deconvolved rule peak is positive, and vice versa. The signs of the peaks in

the zero-charge spectrum indicate the corresponding class. Positive peaks are selective for spectra that are partitioned to the right branch of the classification tree by the discrimination rule, and contrarily, the negative peaks are selective to the left branch of the classification tree.

3.3 Experimental Section

3.3.1 Synthetic Data Set

A synthetic data set with two classes was generated by methods adapted from the reference[83]. Class A comprises 30 synthetic ESI-MS spectra of horse heart myoglobin with baseline noise. The isotopic distribution of this protein was calculated by the polynomial algorithms[84] with a permutation threshold of 0.01. The charge distributions were calculated by the binomial distribution, assuming each basic amino acid in the protein has a probability of 0.5 to receive a proton. Each peak was a Gaussian peak with a full width at half maximum (FWHM) resolution of 10 000. The pure signal was normalized to a maximum intensity of 10. It is assumed that in LC–ESI-MS spectra of cell extracts, many chemical impurities irrelevant to identification will cause a bell-shaped noise in the mass spectrum. The noise model is further validated by Section 4.2. The chemical noise was simulated by a Gaussian function with amplitude of 30, a center of 0.4 and a standard deviation of 0.1 at the 0–1 abscissa. According to the literature[83], Poisson distributed shot noise was added. Each spectrum was sampled in a mass-to-charge ratio range of 550–2000 Th with an increment of 0.1 Th. Class B comprises 30 spectra containing baseline noise only. All spectra were

normalized to unit length. The data set was stored as a 60 × 14 501 matrix, for which the spectra were rows.

3.3.2 Bacterium Identification Data Set

Two strains, designated A1 and A19, of *Salmonella enterica* reference set A (SAR A)[85] were analyzed. Strains A1 and A19 are classified as part of the typhimurium serovar. All bacteria were grown 24 h on tryptic soy agar plates (Difco Laboratories, Sparks, MD). The cells were first vortexed to form a slurry of cells in 70% ethanol. Two hundred microliters of the slurry was collected and centrifuged to a pellet, and the 70% ethanol was removed. Proteins were extracted from bacterial cells with a 50:45:5 solution of acetonitrile (J.T. Baker, Phillipsburg, NJ), HPLC-grade water (J.T. Baker), and formic acid (Sigma–Aldrich, St. Louis, MO). The cells were mixed with 1 mL of the extraction solution and placed in a Barocycler extraction tube (Pressure Biosciences Inc., Boston, MA). In the Barocycler, the sample is exposed to cycles of high (35 kpsi) and low (0 kpsi) pressure. Each pressure is maintained for 25 s. A series of 10 cycles was performed to extract the proteins. The cellular debris was then centrifuged to a pellet, and the clear solution extract was removed.

Separations of protein extracts were performed on an Agilent 1100 HPLC system (Agilent Technologies, Palo Alto, CA) installed with two 150 mm × 2.1 mm Prosphere P-HR (Alltech Associates, Deerfield, IL) columns. The columns were sequentially connected to improve the chromatographic resolution. Mobile phase A and B was respectively 5% acetic acid in water,

and 5% acetic acid in acetonitrile. After sample injection, the solvent composition was held constant at 10% B for 5 min, linearly increased to 50% B between 5 and 70 min, followed by a linear increase to 90% B at between 70 and 80 min. The solvent composition was then linearly changed back to 10% B by 110 min. The flow was split after the column with approximately 25% of the flow going to the mass spectrometer while the remaining eluent was diverted to an HP 1100 fraction collector.

Mass spectra were acquired on a Waters Q-Tof Premier mass spectrometer (Waters, Milford, MA) over a mass-to-charge ratio range of 550–2000 Th using electrospray ionization in the positive ion mode. The scan time was 2.0 s with a 0.1 s interscan delay. The spectra were collected in continuum mode. Five replicates were obtained for each bacterial strain.

The LC–ESI-MS spectra of *Salmonella enterica* were converted into ASCII text files using the Databridge program with MassLynx version 4.0 (Waters, Milford, MA). The text files were then imported into MATLAB. The original spectra were binned by a mass-to-charge ratio increment of 0.1 Th and a retention time increment of 0.1 min. The mass-to-charge ratio cutoff range was 550–2000 Th and the run time cutoff was 85 min. The MS scans were stored as rows in the matrix. Each binned spectrum was stored as an 851 × 14 501 matrix.

3.3.3 Data Processing

All calculations were performed on a personal computer equipped with a Core i7 940 CPU and 12 GB memory running Microsoft Windows XP x64

SP2 operating system. All programs were in-house scripts written in MATLAB version 7.11 (The MathWorks Inc., Natick, MA). The MATLAB code used in this study are available upon request from the corresponding author.

In MoWeD–FuRES approach, each ESI-MS scan was denoised by a wavelet denoising method, in which the nonlinear discrete wavelet transform is applied. In FuRES–MoWeD approach, the same denoising method was applied to the FuRES discrimination rule. Wavelet denoising was performed by using the ThreshWave function in the WaveLab toolbox for MATLAB version 8.5.[86] The wavelet filter used was the Symmlet level 4. The threshold was determined by the visually calibrated adaptive smoothing (VisuShrink) method[87]. Soft thresholding technique was applied on the wavelet coefficients. After applying wavelet denoising, the signal and noise components were separated. The wavelet based noise estimation spectra were then used to calculate the SNR for each peak.

The component peaks were identified by finding local maxima in the spectrum with SNRs larger than 3. The range of the component peak is defined by a starting point and an ending point. The starting point and ending point are two local minima of the spectrum, which are nearest to the peak maxima on both sides. When some peaks with low intensities appeared near a large peak and none of these neighboring peaks were more than three times more intensive than the large peak, these small peaks were considered to be the post-translational modification of this protein. Therefore, the same charge states were assigned to these small peaks.

In MoWeD–FuRES approach, the spectra were baseline corrected by using an iterative 10th order polynomial fitting after wavelet denoising. The iteration stopped when the number of the fitted points was less than 10% of the number of points in the spectrum or the median of the absolute value of the residuals converged.

The MoWeD charge deconvolution method was applied on the processed spectra and the FuRES discrimination rule for MoWeD–FuRES and FuRES–MoWeD approaches, respectively. The maximum possible molecular mass was 50 kDa. After deconvolution, the mass spectra were transformed to 550–50 000 Da with a molecular mass increment of 1 Da. For the bacterium identification data set, each deconvolved two-way spectrum was stored as an 851 × 49 451 matrix.

Each two-way object in the bacterium identification data set was unfolded to a vector that respectively had 42 082 801 and 12 340 351 points for the MoWeD–FuRES and FuRES–MoWeD approach. All vectors were normalized to unit length. All unfolded vectors were stored as rows in a 10 × 42 082 801 sparse matrix and a 10 × 12 340 351 sparse matrix for MoWeD–FuRES and FuRES–MoWeD approach, respectively. The principal component transformation (PCT) was applied as a lossless compression method before FuRES modeling. After PCT compression, the number of variables equals to the number of objects in the training set. Therefore, the computation time and memory requirement of the FuRES classification were greatly reduced.

3.4 Results and Discussion

3.4.1 Synthetic Data Set

The examples of generated synthetic ESI-MS spectra were demonstrated in Figure 3-3. The peak of horse heart myoglobin could not be observed in class A because of relatively low concentrations compared with the baseline noise. Figure 3-3 demonstrates the input vectors for MoWeD deconvolution and the FuRES discrimination rules from the synthetic data set by two processing methods. Because the pure signal is three times weaker than the baseline, the MoWeD–FuRES could not identify the protein because each protein feature was not extracted from noise when deconvolution was applied on the individual unprocessed spectra. The protein signals were identified as high frequency noise and deconvolved to a high molecular mass domain. However, the FuRES–MoWeD could generate the discrimination rules correctly because the protein features were correctly extracted from the whole set of spectra by FuRES discrimination rules, which made it possible to distinguish between the protein peaks and noise.

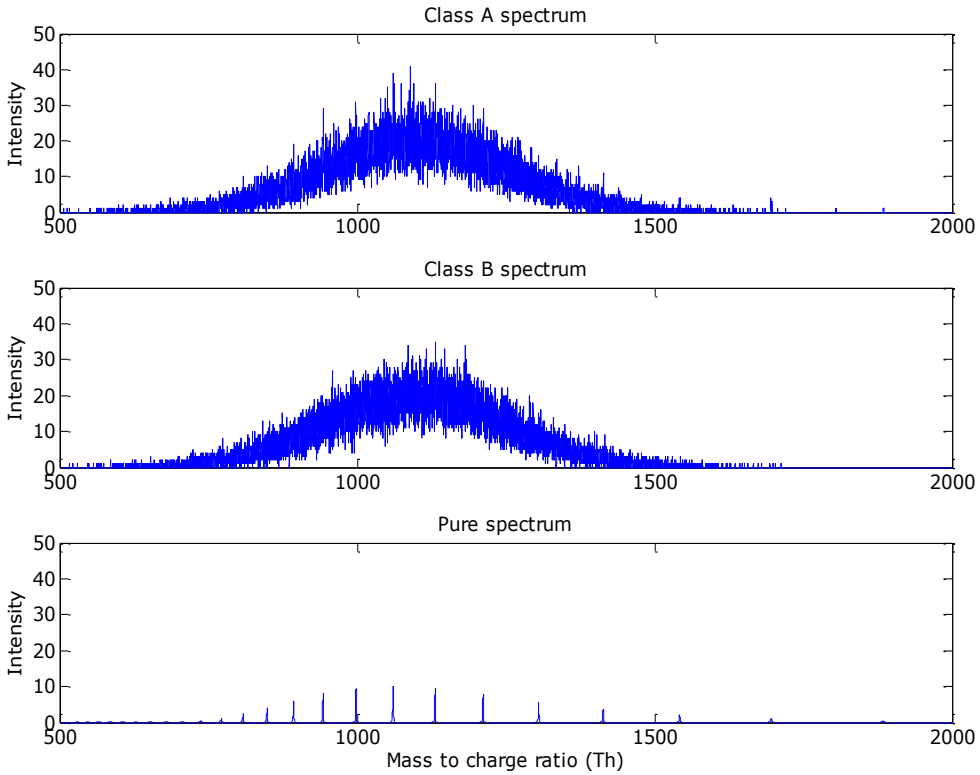

Figure 3-2. Examples of synthetic horse heart myoglobin spectra (Class A and Class B) and the pure spectrum.

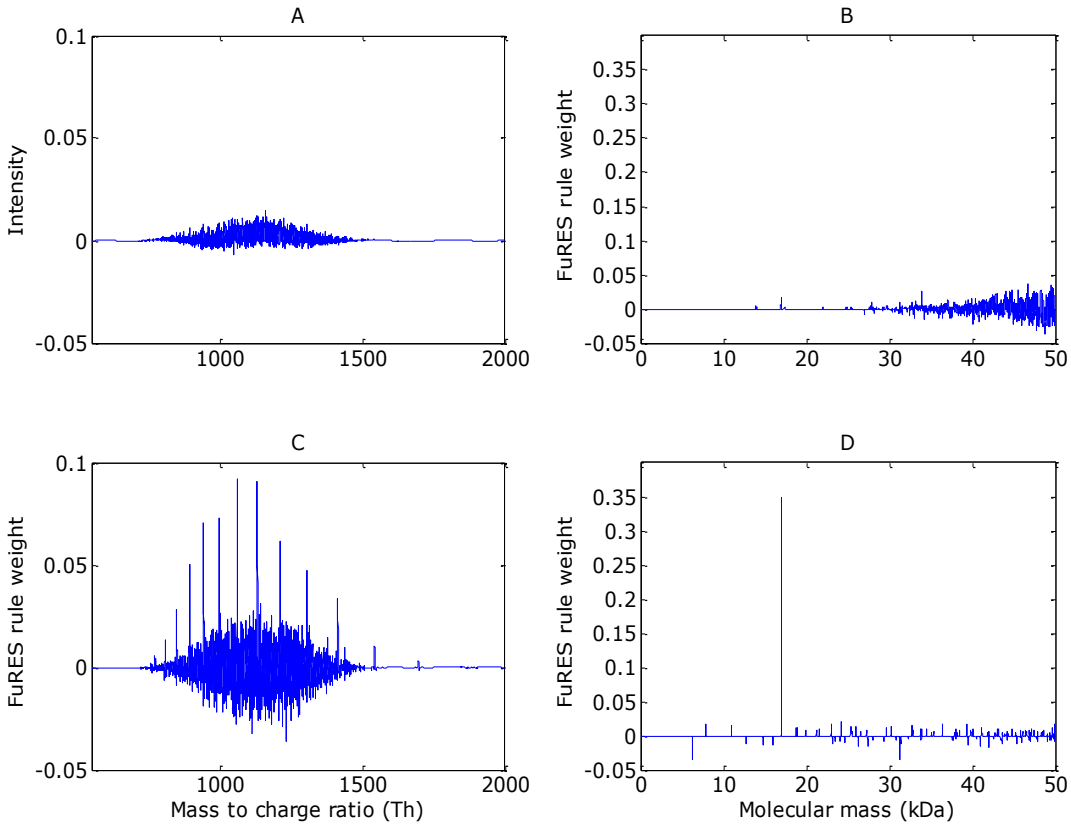

Figure 3-3. The mass spectra and the discrimination rules of the synthetic data set processed by MoWeD–FuRES (top panels) and FuRES–MoWeD (bottom panels) approach. (A) An example of the denoised and baseline corrected spectrum as the input data for deconvolution, (B) the final MoWeD–FuRES discriminant rule, (C) the denoised and baseline corrected FuRES discriminant rule as the input data for deconvolution, (D) the final FuRES–MoWeD discriminant rule.

3.4.2 Bacterium Identification Data Set

The example LC–ESI-MS spectra of two strains of *Salmonella enterica* are given in Figure 3-5. The total ion chromatogram and total mass profile are also shown. The two-way spectra are transformed and plotted in logarithmic scale. From this figure, it is concluded that the two-way spectra of the two bacteria strains are similar. In addition, the noise components form a bell shape in the middle mass-to-charge ratio range and the retention time from 20 to 80 min. This phenomenon is also observed from the total ion chromatogram and the total mass profile. The characteristics of noise in the experimental data were consistent with the noise simulated in the previous data set. The deconvolution results by MoWeD for a single spectrum of *Salmonella enterica* strains A1 and A19 are given in Figure 3-5. The profiles generated by MoWeD were consistent with the spectra calculated by a commercial software package, ProTrawler 6 (BioAnalyte, Cambridge, MA). The comparison indicates the MoWeD deconvolution is a suitable method in this analysis pipeline for the bacterium identification data set. Because the deconvolution method used in ProTrawler is not publicly available, further comparisons were not performed.

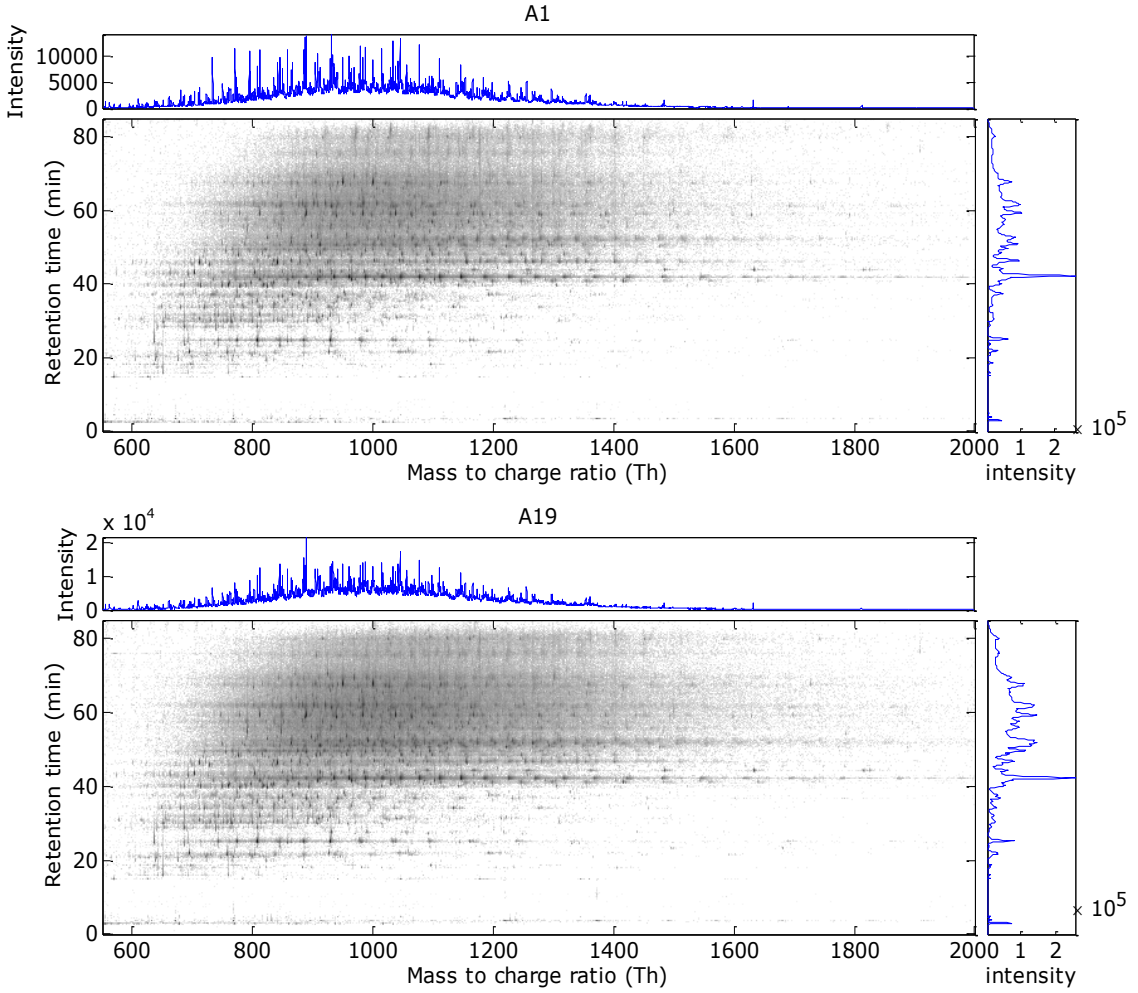

Figure 3-4. Two-way LC–ESI-MS data objects of *Salmonella enterica* strains A1 and A19. The intensity in the two-way image is plotted in logarithmic scale to compare the amount of noise. The total ion chromatogram and total mass spectra are besides the two-way image.

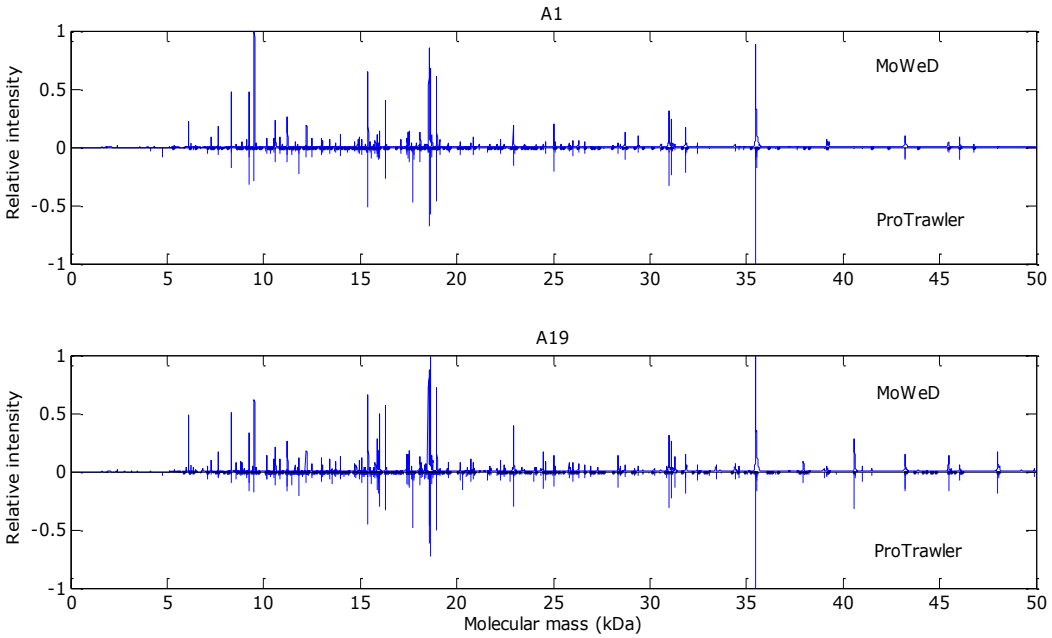

Figure 3-5. Comparison of the deconvolved protein profiles by MoWeD and the ProTrawler on a representative spectrum of Salmonella enterica strains A1 (top panel) and A19 (bottom panel).

When dealing with a two-way LC–ESI-MS data set, the data processing speed will be a more important factor compared to an ESI-MS data set, because the size of the data set is usually hundreds to thousands times larger. The effectiveness of the proposed FuRES–MoWeD approach is apparent in Table 3-1 by comparing run times and the total number of deconvolution routine evaluations. The actual computation time will vary by many factors such as the software and hardware configuration of computer system, the choice of programming language, etc. However, a general comparison of algorithmic efficiency between two approaches can be made by program profiling. The run time was measured by the cputime function in MATLAB. Because the deconvolution calculation was performed only on FuRES discriminant, fewer deconvolution routine evaluations are needed. Moreover, the baseline correction routine is not performed. The FuRES–MoWeD approach required less time than the MoWeD–FuRES to perform.

Table 3-1. Comparisons on run time and total number of deconvolution routine evaluations of bacterium identification data set.

	MoWeD–FuRES	**FuRES–MoWeD**
Total deconvolution routine evaluations	8510	851
Run time for deconvolution routine (min:s)	12:49	1:46
Total run time (min:s)	25:18	3:08

The FuRES discrimination rules calculated by FuRES–MoWeD and MoWeD–FuRES approaches are given in Figures 3-6 and 3-7. In Figure 3-6, the discrimination rules were summed along the retention time dimension. The positive value indicates the corresponding component has characteristically high concentration in class A1 over A19, and vice versa. Similar summed spectra were obtained, which indicates the results of the two approaches are consistent in terms of detecting major proteomic features. The protein signals in both low and high molecular weight range can be observed sufficiently, which means the proposed method improves the detectability of high MW proteins when incorporated with FuRES. LC–ESI-MS provides the molecular weight and retention time of the proteins, which is insufficient to identify proteins due to various effects such as post-translational modification, mass errors and isotopic distribution. The five most abundant peaks and the corresponding tentative protein search results by SwissProt/TrEMBL database[88, 89] are listed in Table 3-2. After searching the database with organism keyword "Salmonella typhimurium", a list that contains 15 744 protein entries were exported. The theoretical average molecular weights were then calculated by Compute pI/Mw, which is a part of ExPASy proteomics tools[90]. The search result was a protein entry that has the least mass difference between the observed mass and the calculated mass. Four out of five proteins were matched for both approaches.

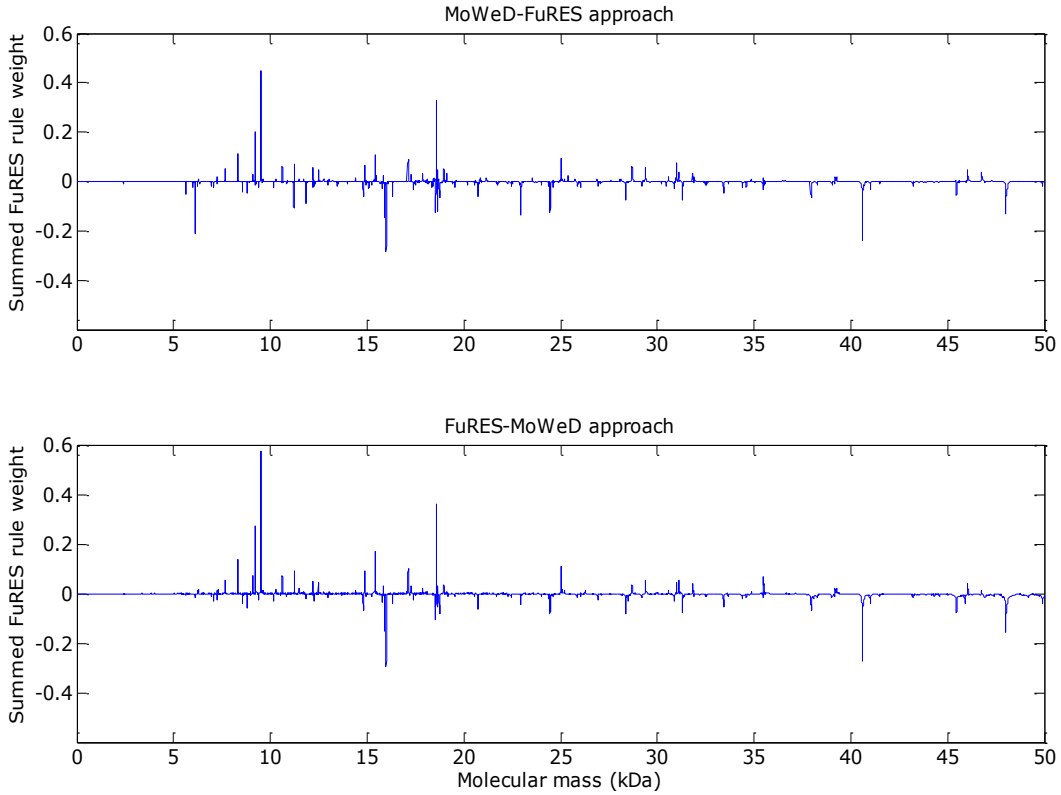

Figure 3-6. Comparison of the summed discrimination rules calculated by MoWeD–FuRES (top panel) and FuRES–MoWeD (bottom panel) approaches. The discrimination rules are summed along the retention time dimension. The positive value indicates the corresponding component has characteristically high concentration in class A1 over A19, and vice versa. The protein signals in both low and high molecular weight range were sufficiently extracted.

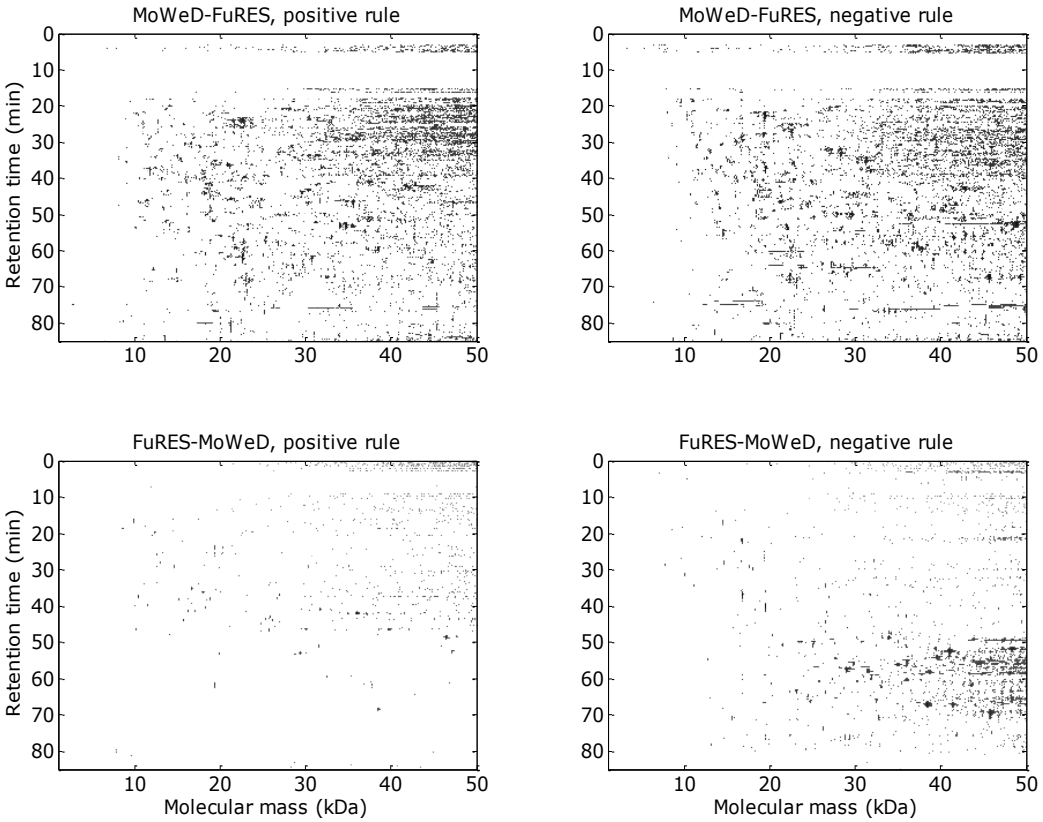

Figure 3-7. The discrimination rules calculated by MoWeD–FuRES (top panels) and FuRES–MoWeD (bottom panels) approaches. The rule weights are plotted in logarithmic scale to compare the amount of noise.

Table 3-2. Five largest peaks observed in the summed discrimination rules and the tentative SwissProt/TrEMBL database search results for FuRES–MoWeD and MoWeD–FuRES approaches

	Observed molecular mass (Da)	Calculated average molecular mass[a] (Da)	Relative intensity	Class[b]	Tentative accession ID
MoWeD–FuRES	9520	9520.96	1	A1	P0A1R6
	18,585	18586.08	0.7316	A1	Q7CQV9
	15,990	15988.81	0.6395	A19	P0A1J7
	40,595	40595.34	0.5315	A19	P19576
	6092	6094.94	0.4738	A19	D0ZSD2
FuRES–MoWeD	9520	9520.96	1	A1	P0A1R6
	18,585	18586.08	0.6270	A1	Q7CQV9
	15,990	15988.81	0.5171	A19	P0A1J7
	40,595	40595.34	0.4764	A19	P19576
	9239	9239.61	0.4704	A1	P0A1R8

[a]Results were obtained by Compute pI/Mw, which is a part of ExPASy proteomics tools.

[b]Class means the characteristic protein in the specified strain has relatively high concentration against the other strain.

The noise in LC–ESI-MS spectra possibly comes from the polymer contaminants from column degradation and other baseline noise components. These noise components can affect the deconvolution routine because noise peaks can be mistaken as signal peaks, which is a typical case that creates high mass artifacts in the deconvolved spectra. Figure 3-7 demonstrates the two-way LC–ESI-MS discrimination rules. The intensities were plotted in logarithmic scale to compare the amount of noise. It can be observed that there is a greater amount of noise in the rules obtained by the MoWeD–FuRES approach in the high mass range than that from the FuRES–MoWeD approach. The noise in the rules image is consistent with the results from the synthetic data set and previous study[74]. These artifacts were retained by using MoWeD–FuRES. In the FuRES–MoWeD approach however, noise levels are reduced in the FuRES discrimination rules before deconvolution, because the noise did not correlate with the systematic differences between the two bacterial strains. As a result, there are fewer artifact peaks in the FuRES–MoWeD approach than MoWeD–FuRES.

Theoretically, the FuRES classification rule is generated to minimize the entropy of classification between different classes, so that relevant protein signals are extracted for deconvolution without the cost of sensitivity and specificity. When and only when a protein signal is not characteristic for differentiating between the two rule consequents, that signal will not be present in the rule and thus be omitted from deconvolution. In practice, low-lying signals were correctly identified in the simulated data set. Additionally,

the locations of low-intensity peaks were consistent between the two comparing approaches in Figure 3-6. The reason is that the FuRES rule looks for correlation among features of the spectra. There is a signal averaging benefit among the multiple charge state peaks and among objects common to each rule consequent.

Another concern is about the treatment of multiple charge variants of a protein, because the multiple charge envelope of a protein is different between different runs. Because the FuRES classification rule only picks signals with systematic differences between samples, the random variation of the multiple charge distribution between different samples will be omitted. On the other hand, the FuRES rules treat systematic multiple charge variants as separate signals. However, a systematic change of variations between classes could possibly be caused by conformational changes[91], when the conformation and envelopes are on opposing consequents of the rule. Although the molecular weight remains unchanged, the proposed approach retains the specific proteins as tentative biomarkers in this regard. Again, further structure elucidation work is required for confirmation.

The FuRES–MoWeD uses the unprocessed spectra as training data, while the MoWeD–FuRES uses the deconvolved spectra. Therefore, the FuRES models obtained from these two methods have different predictive powers. Because both approaches apply pattern recognition methods during processing, internal validation methods can be applied. BLP validation[25] was applied to compare the FuRES classification ability on the deconvolved

spectra and the unprocessed spectra. The number of bootstraps and the number of partitions were respectively 10 and 2. The prediction results are in Table 3-3. The prediction accuracy was 89 ± 5% and 100% for MoWeD–FuRES and FuRES, respectively. This lower prediction accuracy indicates the loss of information in wavelet denoising, baseline removal and peak picking. The difference of the prediction results is statistically significant by two-way analysis of variance (ANOVA) with interaction at a significance level of 0.05.

Table 3-3. The confusion matrix of average correctly predicted objects with 95% confidence intervals between the two approaches of FuRES models. Each class contains five data objects.

	MoWeD–FuRES		**FuRES**	
	A1	A19	A1	A19
A1	4.6 ± 0.4	0.4 ± 0.4	5	0
A19	0.7 ± 0.5	4.3 ± 0.5	0	5

Besides the validation of the two approaches, the BLP can be used to validate the biomarker candidate in the deconvolved FuRES rule. Figure 3-8 is the average FuRES–MoWeD discrimination rule with upper and lower 95% confidence intervals. The plot is zoomed out into 25–30 kDa for demonstration purposes. The average rule is plotted as a thick blue line and the confidence intervals are plotted as thin gray lines. The rules are summed along the retention time dimension. The average is obtained from 20

MoWeD deconvolved FuRES discriminants by BLP. Peaks in the discriminant that extend beyond the confidence interval are significant at a 0.05 probability. Most peaks in this plot are greater than their respective confidence intervals, indicating the biomarker candidates are significantly characteristic features in the FuRES-MoWeD model.

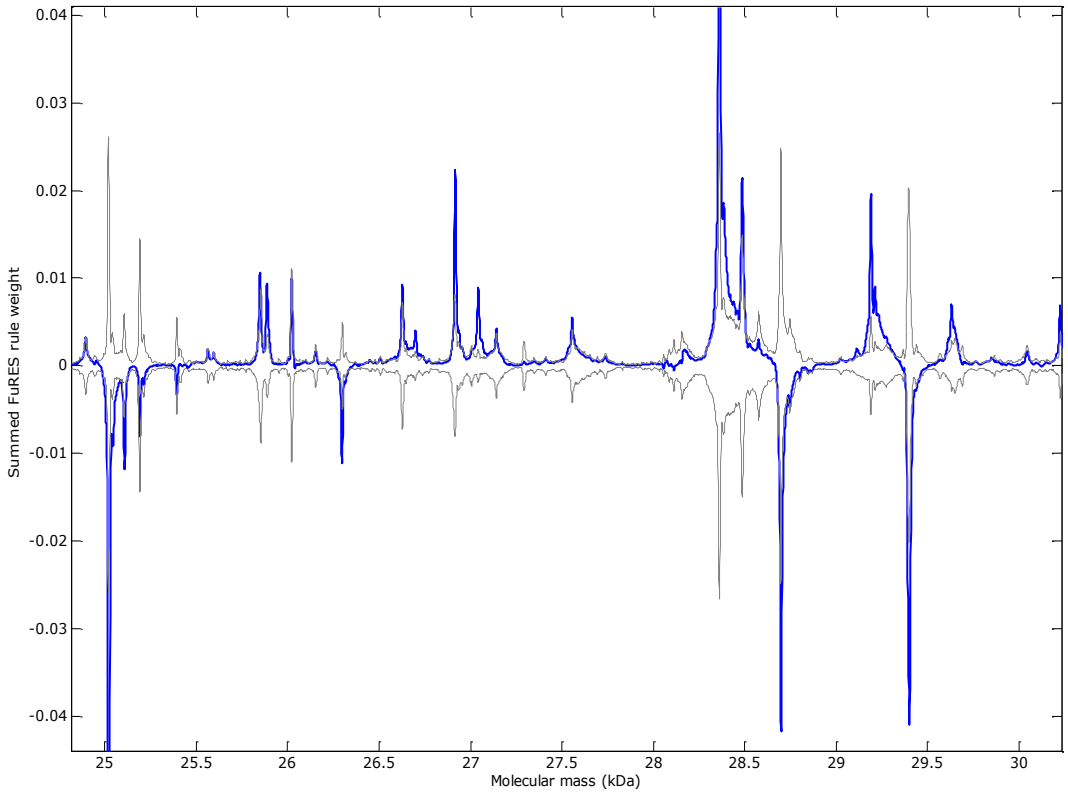

Figure 3-8. The average FuRES–MoWeD discrimination rule with upper and lower 95% confidence intervals. The plot is zoomed out into 25–30 kDa. The average rule is plotted as a thick blue line and the confidence intervals are plotted as thin gray lines. The rules are summed along the retention time dimension. The average is obtained from 20 FuRES discriminants by 10 bootstraps of two Latin partitions. The magnitudes of most peaks are greater than their respective confidence intervals.

3.5 Conclusions

The proposed FuRES–MoWeD approach could rapidly find the features in complex sets of ESI-MS data. This approach was demonstrated by a synthetic ESI-MS data set and a LC–ESI-MS data set for bacteria strain identification, with the comparison of the MoWeD–FuRES approach. The resultant discrimination rule indicates that biomarker candidates can be found when signal-to-noise ratios in the spectra are low by FuRES–MoWeD approach. Performing the charge deconvolution on the FuRES discriminant rules as opposed to each individual across replicates yielded models less affected by baseline noise and was an order of magnitude more computationally efficient. In addition, the models obtained by the two approaches were evaluated by using bootstrapped Latin partitions to furnish statistical relevance and confidence intervals.

This proposed approach was not limited to FuRES models and MoWeD methods demonstrated in the current study. Future work will involve the applications of deconvolution on rules of other classifiers such as partial least squares-discriminant analysis, and using other deconvolution methods such as Zscore and MCA. This analysis pipeline is also feasible in the applications of high-resolution ESI-MS and CE–ESI-MS. The pipeline can be advantageous because the increased data size produced by high-resolution ESI-MS usually demands rapid analysis methods. As a pre-screening method, the proposed pipeline can be also applied in differential protein expression by liquid chromatography–tandem mass spectrometry (LC–

MS/MS) that provides helpful fragmentation information to identify protein biomarkers specifically. Following the basic concept of the FuRES–MoWeD approach, the potential application can be extended further to other areas of proteomics. The processing method applied to the discriminant is not limited to charge deconvolution for LC–MS data. When processing LC–MS or LC–MS/MS profiles, a database search routine for peptide fragments can be applied to the discriminant instead. Database searches on discriminants could potentially be useful when dealing with protein digest samples, where small peptide peaks are present.

Chapter 4 Ignitable Liquid Identification Using Gas Chromatography/Mass Spectrometry Data by Projected Difference Resolution Mapping and Fuzzy Rule-Building Expert System Classification

4.1 Introduction

Ignitable liquid (IL) identification is an important topic in arson crime investigation, because liquids such as gasoline and kerosene are commonly used as accelerants in arson crimes. The American Society for Testing and Materials (ASTM) E1618 standard method has been devised to identify different types of ILs by gas chromatography/mass spectrometry (GC/MS).[92] More specific classification is required in some actual casework. For example, when comparing fire debris samples with an IL sample found in a suspect's possession or on his clothing, classification of the type of ILs is not specific because of the wide availability of gasoline and kerosene and their variability due to crude oil source, production processes, and blending at the refinery or in storage tanks at retail outlets. Therefore, it is important to be able to compare two or more samples in a case to determine if the ignitable liquid residues share a common source. The recent National Academy of Science report recommends that pattern recognition techniques (of which ignitable liquid residue analysis is one) have established "error rates" to meet "Daubert" rules of evidence in court.[93, 94] This research was undertaken to establish such rates for comparing patterns in GC/MS analyses of ignitable liquid residues (ILR).

In recent years, GC/MS data analysis of ignitable liquids by chemometric techniques has been reported. A variety of methods such as covariance mapping[95, 96], trace organic compound analysis by principal component analysis (PCA) and linear discriminant analysis (LDA)[97-101], and artificial neural networks[102, 103] have been applied, which demonstrates the utility of chemometrics to forensic investigations. Analysis by chemometric techniques can be performed more efficiently and objectively than manual operations by an expert, because the predictions are performed automatically once the model is generated. The analytical result is quantitative, unbiased, and can be validated by cross-validation and bootstrap Latin partition (BLP)[25] techniques that estimate prediction abilities of models with given confidence intervals, as opposed to the subjective opinions furnished by an expert.

Gasoline and kerosene samples collected from the refinery, distribution terminals, and retail outlets on different dates were analyzed by GC/MS.[104] Both two-way GC/MS profiles and target component ratio data were processed. Classification by partial least squares-discriminant analysis (PLS-DA) and fuzzy rule-building expert systems (FuRESs) were performed to identify each sample. The projected difference resolution (PDR) metric is calculated to demonstrate the separations between each pair of IL samples as well.

Two-way GC/MS profiles and target component ratios are applied as two comparative inputs for classification in this study. A two-way GC/MS

profile retains both retention properties and mass spectrum information as an image of GC/MS measurement.[10, 32] In addition, no component peaks were discarded. Figure 4-1 is a two-way GC/MS image of a gasoline sample.

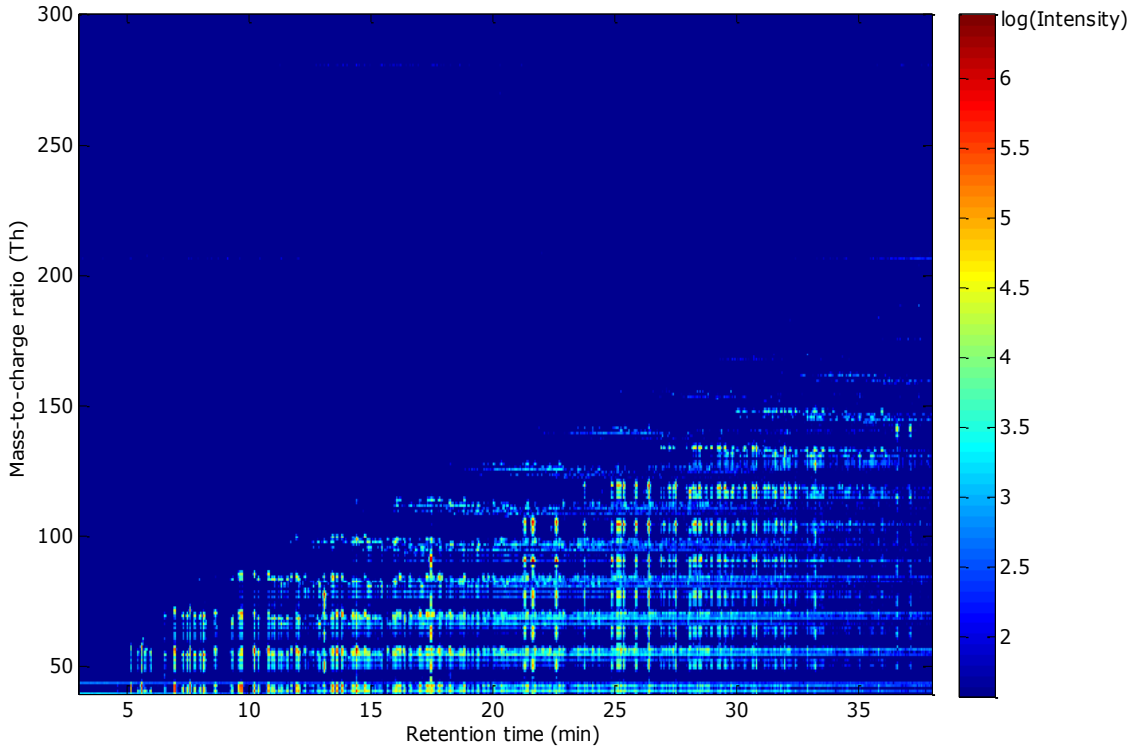

Figure 4-1. Two-way GC/MS data of a gasoline sample. The peak intensities were plotted in logarithmic scale to show more detail from the smaller peaks.

Target component ratio is an analysis method developed for comparing neat samples to fire debris submitted from casework.[105] The target component ratio is calculated by searching specified retention time windows for the presence of peaks of specific compounds. The ratios of peak areas from the sequential peaks are calculated to establish a unique profile for the identification of ignitable liquids. The target component ratio method was reported to be a possible method to compare gasoline residues in fire debris to a gasoline source. Different IL samples from different locations have unique production processes (crude oil sources and blending for time of year and altitude) and different storage conditions, which results in characteristic chemical compositions. The theoretical assumption of the target component ratio method is the speed of the evaporation stays constant for components that have similar chemical structure in ignitable liquids and consequently similar gas chromatographic retention times (especially on a non-polar polydimethylsilicone capillary column). The ratio of an adjacent peak pair of an evaporated or burnt sample should have an insignificant change compared to the neat sample. Figure 4-2 gives a comparison of the total ion current and the target component ratio profile. The component ratios are given in bar plots. Each bar was connected to the corresponding peak pair from which the component ratio is calculated.

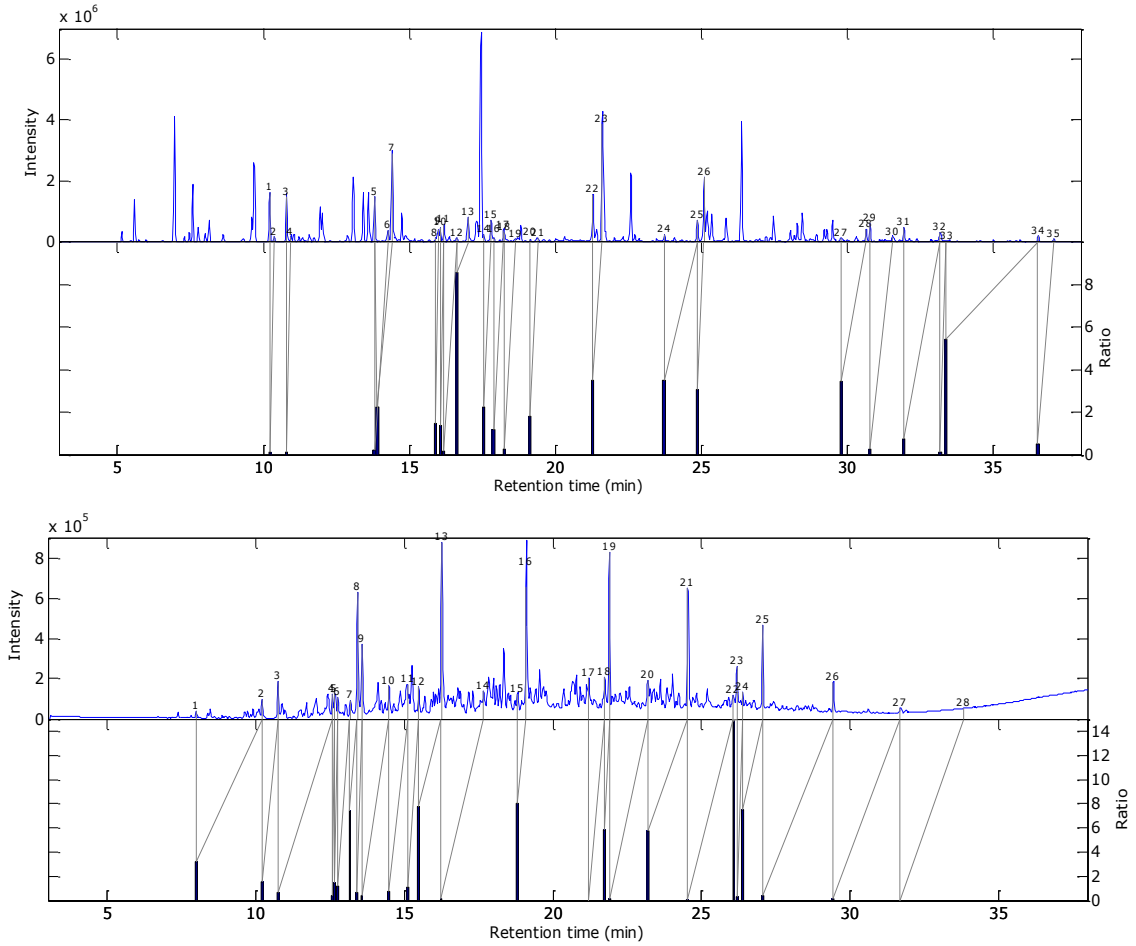

Figure 4-2. TIC chromatograms and component ratios of a gasoline sample (Top panel) and a kerosene sample (Bottom panel). The component peaks are labeled with numbers. Each component ratio bar was located under the peak that is denominator in the corresponding ratio. Each bar was connected to a pair of corresponding component peaks by lines.

The PDR metric is an analog to chromatographic resolution.[106] PDR is applied to measure the separation of pairs of samples quantitatively in a multivariate data space. Given a GC/MS data set that comprises two classes, the number of variables (i.e., number of data points, which is calculated by the number of retention time measurements times the number of mass-to-charge ratio measurements) is *n*, the numbers of objects (i.e., amount of GC/MS spectra) in classes a and b are respectively m_1 and m_2. The data matrices $\mathbf{X_a}$ and $\mathbf{X_b}$ respectively have sizes of $m_1 \times n$ and $m_2 \times n$, in which each row is a two-way GC/MS data. The PDR measure of class separation $R_s(a, b)$ is a scalar calculated by

$$\mathrm{R_s(a,b)} = \frac{|\bar{\mathbf{t}}_\mathbf{a} - \bar{\mathbf{t}}_\mathbf{b}|}{2(s_\mathrm{a} + s_\mathrm{b})} \tag{4-1}$$

for which $\mathbf{t_a}$ and $\mathbf{t_b}$ are the scores for the two classes obtained by projecting the objects onto the difference vector $\overline{\mathbf{X_a}} - \overline{\mathbf{X_b}}$ of the class averages, given by

$$\mathbf{t_a} = \mathbf{X_a}(\overline{\mathbf{X_a}} - \overline{\mathbf{X_b}})^\mathrm{T} \tag{4-2}$$

$$\mathbf{t_b} = \mathbf{X_b}(\overline{\mathbf{X_a}} - \overline{\mathbf{X_b}})^\mathrm{T} \tag{4-3}$$

for which $\overline{\mathbf{X_a}}$ and $\overline{\mathbf{X_b}}$ are the average class vectors that have a length of *n*. The column vectors $\mathbf{t_a}$ and $\mathbf{t_b}$ have lengths of m_1 and m_2, respectively. From the projections the averages $\bar{t}_a$ and $\bar{t}_b$ and their corresponding standard deviations s_a and s_b are calculated.

PDR is proposed as a straightforward multivariate measure for rapidly quantifying the separation of multivariate data objects for a pair of classes.

The smaller the PDR, the harder to predict two classes by multivariate pattern recognition methods. Generally, a well-resolved separation of two classes has a PDR value greater than 1.5, which is comparable to the minimum resolution for baseline resolution between a pair of chromatographic peaks. When the data set contains more than two classes, the PDR metric for each pair of classes is systematically calculated for all combinations of pairs. The PDR matrix can be as a triangle that measures the separation of each pair of classes.

Column bleeding was observed in the measurement of kerosene samples because of the higher column temperatures that are required for elution, which may bias the pattern classification of the two-way profiles. As a result, the baseline was corrected by a procedure based on the PCA result. For a two-way GC/MS data matrix **X,** mass spectrometry scans are stored by rows and extracted ion chromatograms are stored by columns. First, the spectrum segment of the final 1 minute retention time window was selected to acquire background mass spectral scans. The PCA is performed on this background matrix of mass spectra for each sample. By performing the classification based on background subtraction using 1–10 largest principal components, it is concluded that the loading of the first principal component $\mathbf{v_1}$ characterize the mass spectrum of column bleeding impurities. Therefore, background correction is obtained by

$$\mathbf{X_c} = \mathbf{X} - \mathbf{X} \cdot \mathbf{v_1} \cdot \mathbf{v_1}' \qquad (4\text{-}4)$$

for which $\mathbf{X_c}$ is the baseline corrected spectrum. Figure 4-3 is a comparison of total ion current (TIC) chromatograms showing the effect of the background correction.

Two approaches were applied to determine the number of latent variables in the partial least squares-discriminant analysis (PLS-DA)[64] method. In the optimal partial least squares-discriminant analysis (oPLS-DA), the number of latent variables is determined by achieving highest prediction accuracy for the prediction set. As a result, oPLS-DA is positively biased, which is applied as a reference method.

The other PLS-DA method is unbiased because the prediction set is not used to determine the number of latent variables in the PLS-DA model. The procedure for unbiased PLS-DA training is similar to the BLP method[107]. First, the training set is split into two subsets using Latin partitions. Then, each partitioned subset is used once for prediction and once for model-building. The procedure is bootstrapped 10 times. Lastly, the prediction accuracies are averaged across the 10 bootstraps, the number of latent variables is determined by achieving highest average prediction accuracy.

FuRES is based on the decision tree algorithm and fuzzy logic theory, where each branch of the decision tree model is a multivariate fuzzy rule.[31] FuRES has been successfully applied in forensic researches to analyze two-way GC/MS and gas chromatography–differential mobility spectrometry (GC–DMS) data.[10, 32]

OPLS-DA, PLS-DA, and FuRES were validated by the BLP method[25]. The PDRs of each training set is calculated for comparison as well. Bootstrapping is a re-sampling method. Latin partition is a block cross-validation, in which the class distributions are maintained between the training set and the prediction set. BLP provides validation results with confidence intervals by running the evaluation repeatedly with different training and prediction set partitions.

4.2 Experimental

4.2.1 Sample Collection

The details of the sample collections and GC/MS experiments were described in the literature[104]. A number of gasoline and most kerosene samples come from the Quality Assurance Laboratory of Marathon Ashland Refinery, Catlettsburg, KY. The quality assurance laboratory collects samples from Marathon refineries and distribution terminals. The gasoline samples were regular grade (87 octane) gasoline. Most of the samples were from the Midwestern states where Marathon distribution terminals are located. Other gasoline samples were collected at a number of retail outlets across the US and over several years. A few kerosene samples were purchased locally (Huntington, WV) from service stations or home improvement stores. The locations of samples are listed in Table 4-1.

Table 4-1. Sources of the collected gasoline and kerosene samples.

Gasoline		**Kerosene**	
States	Number of samples	States	Number of samples
AL	1	IL	2
CA	1	IN	2
CT	1	KY	2
FL	2	MI	2
GA	3	MN	1
IL	1	OH	1
IN	5	TN	1
KY	2	WI	2
LA	1	Total	13
MI	7		
NC	1		
OH	8		
TN	5		
TX	1		
WI	1		
WV	2		
Total	42		

4.2.1 GC/MS Measurement

Spectra from neat samples were used in this work. There were 126 objects and 42 classes in the gasoline data set, and 39 objects and 13 classes in the kerosene data set. Each sample was analyzed in triplicate.

All spectra were collected on an Agilent 6890N gas chromatograph coupled to a 5973 mass selective detector. The column was a 60 m long Varian DB-1 (100% dimethylpolysiloxane) column with an internal diameter of 250 μm, and film thickness of 1 μm. The carrier gas was ultrahigh purity helium. The injection volume was 0.1 μL. For gasoline samples, the split flow was 200:1. The temperature programming for gasoline was initial 35 °C with a 2.00 min hold time, and a linear temperature ramp of 5 °C/min to 250 °C with 1.33 min hold time. For kerosene samples, the split flow was 50:1. The temperature programming for kerosene was initial 100 °C with 1.00 min hold time, and a linear temperature ramp of 5 °C/min to a final temperature of 275 °C with 5.00 min hold time. The solvent delay was 3.00 min.

4.2.2 Data Processing

All calculations were performed on a personal computer equipped with a Core i7 940 processor and 12 GB memory running Microsoft Windows XP x64 SP2 operating system. All programs were in-house scripts written in MATLAB and C++ programming language using MATLAB version 7.11 (The MathWorks, Inc. Natick, MA) and Microsoft Visual C++ version 10.0, except that the component peaks were identified by AMDIS version 2.69 (National Institute of Standards and Technology, Gaithersburg, MD).

The first preprocessing step was binning of the spectra along mass-to-charge ratio dimension to an increment of 1 Th, and an increment of 0.02 min along the retention time dimension. The mass-to-charge ratio range of 40—300 Th and the retention time range of 3.00—38.00 min were used for further analysis. Next, the baseline correction was performed on each GC/MS spectrum. The PCA was performed by the singular value decomposition (SVD) method. Each two-way spectrum was then reshaped into a vector that comprised 457 011 data points. Each vector was normalized to unit vector length. The sizes of the data matrices were 126 × 457 011 and 39 × 457 011 for gasoline and kerosene, respectively.

Because the size of the two-way GC/MS spectra and the number of classes is large, the principal component transformation (PCT) is used as a lossless compression method. The PDR metric, oPLS-DA, PLS-DA, and FuRES classification were performed after PCT. PCT compression was implemented by the SVD method. All principal component scores were used for further calculation. When applying BLP validation to the PCT compressed PLS-DA and FuRES classification models, the principal component scores of each prediction set were calculated by projecting the prediction set onto the principal component loadings calculated from the training set to achieve unbiased results. After compression, the number of variables was greatly reduced to the number of objects, i.e., the number was respectively reduced from 457 011 to 126 and 39, and the compression ratio was respectively 1:3627 and 1:11 718 for gasoline and kerosene data sets.

The component ratio was calculated by dividing peak areas of designated components. Twenty-one and twenty-five target component ratios were respectively calculated for gasoline and kerosene spectra, which are listed in Tables 4-2 and 4-3. The chromatographic components were identified by using "Use Retention Time" analysis in AMDIS software by a pre-built mass spectra library that contains the target components. If one or both of the component peaks were not detected, the corresponding ratio is set to zero. The component ratio was preprocessed by autoscaling, by which each variable is subtracted by its mean and then divided by the standard deviation.

Table 4-2. Target compound list, estimated retention time, and corresponding ratios identified in each gasoline sample

Peak	Compound	Est. RT (min)	Ratio	Peak Ratio
1	3-methylpentane	10.21		
2	2-methyl-1-pentene	10.37	1	2/1
3	*n*-hexane	10.79		
4	2-hexene	10.95	2	4/3
5	3-methylhexane	13.80		
6	1,3-dimethylcyclopentane	14.26	3	6/5
7	2,2,4-trimethylpentane	14.40	4	7/5
8	dimethylcyclopentane	15.90		
9	methylcyclohexane	15.98	5	9/8
10	2,5-dimethylhexane	16.06		
11	2,4-dimethylhexane	16.18	6	11/10
12	1,2,3-trimethylcyclopentane	16.63	7	12/11
13	2,3,4-trimethylpentane	17.00	8	13/12
14	dimethylhexane	17.54		
15	3-methylheptane	17.80	9	15/14
16	2,3,4-trimethylhexane	17.88		
17	1,3-dimethylcyclohexane	18.22	10	17/16
18	2,2,5-trimethylhexane	18.26		
19	1-ethyl-3-methylcyclopentane	18.62	11	19/18
20	1,2-dimethylcyclohexane	19.15		
21	1,4-dimethylcyclohexane	19.40	12	21/20
22	ethylbenzene	21.28		
23	*m,p*-xylene	21.60	13	23/22
24	isopropylbenzene	23.73		
25	*n*-propylbenzene	24.86	14	25/24
26	3-ethyltoluene	25.09	15	26/25
27	a methyl indane	29.78		
28	tetramethylbenzene	30.64	16	28/27
29	tetramethylbenzene	30.78		
30	a methyl indane	31.54	17	30/29
31	a methyl indane	31.94		
32	naphthalene	33.18	18	32/31
33	a dimethyl indane	33.39	19	33/32
34	2-methylnaphthalene	36.53	20	34/33
35	1-methylnaphthalene	37.08	21	35/34

Table 4-3. Target compound list, estimated retention time, and corresponding ratios identified in each kerosene sample

Peak	Compound	Est. RT (min)	Ratio	Peak Ratio
1	toluene	7.98		
2	*p*-xylene	10.20	1	2/1
3	nonane	10.74	2	3/2
4	1-ethyl-2-methyl-Benzene	12.57	3	4/3
5	ethyl-methyl-benzene	12.65	4	5/4
6	a trimethylbenzene	12.76	5	6/5
7	a ethylmethylbenzene	13.17	6	7/6
8	decane	13.41	7	8/7
9	1,2,4-trimethyl-benzene	13.57	8	9/8
10	a trimethylbenzene	14.48	9	10/9
11	a diethylbenzene	15.10	10	11/10
12	3-methyl-decane	15.47	11	12/11
13	undecane	16.25	12	13/12
14	a methyl-trans-decalin	17.66	13	14/13
15	1,2,3,4-tetrahydronapthalene	18.81		
16	dodecane	19.11	14	16/15
17	a dihydro-dimethyl-1H-indene	21.20		
18	1,2,3,4-tetrahydro-6-	21.74	15	18/17
19	tridecane	21.89	16	19/18
20	a tetrahydrodimethylnapthalene	23.19	17	20/19
21	tetradecane	24.54	18	21/20
22	1,7-dimethyl-napthalene	26.08	19	22/21
23	2,6,10,14-tetramethyl-	26.19	20	23/22
24	3-methyl-tetradecane	26.39	21	24/23
25	pentadecane	27.06	22	25/24
26	hexadecane	29.44	23	26/25
27	heptadecane	31.70	24	27/26
28	octadecane	33.84	25	28/27

Both oPLS-DA and PLS-DA was implemented by the nonlinear iterative partial least squares (NIPALS) algorithm[108]. The classification method was PLS2. In PLS-DA, oPLS-DA, and FuRES, binary encoding was applied to build the target (response) matrix **Y**. For instance, when a data set has three classes, the first class will be encoded as (1, 0, 0). The number of bootstraps was 10 and the number of partitions was 3 in BLP evaluations. Each object will be predicted 10 times, and each sample that ran in triplicate will be predicted 30 times in the BLP validation process.

4.3 Results and Discussion

4.3.1 Baseline Correction

The TIC chromatograms of a gasoline and a kerosene sample are given in Figure 4-4. The effect of baseline correction is demonstrated. Because there is no significant column bleeding from the original spectra of gasoline, no significant change in TIC chromatograms is observed after baseline correction. In kerosene samples, the baseline goes upwards in the uncorrected TIC profile. Compared to the uncorrected spectra, the baseline of the corrected spectra is improved. The pattern classification and PDR metric of the spectra before baseline correction is performed. Comparisons were made between baseline corrected spectra and original spectra. The result is given in section 3 below.

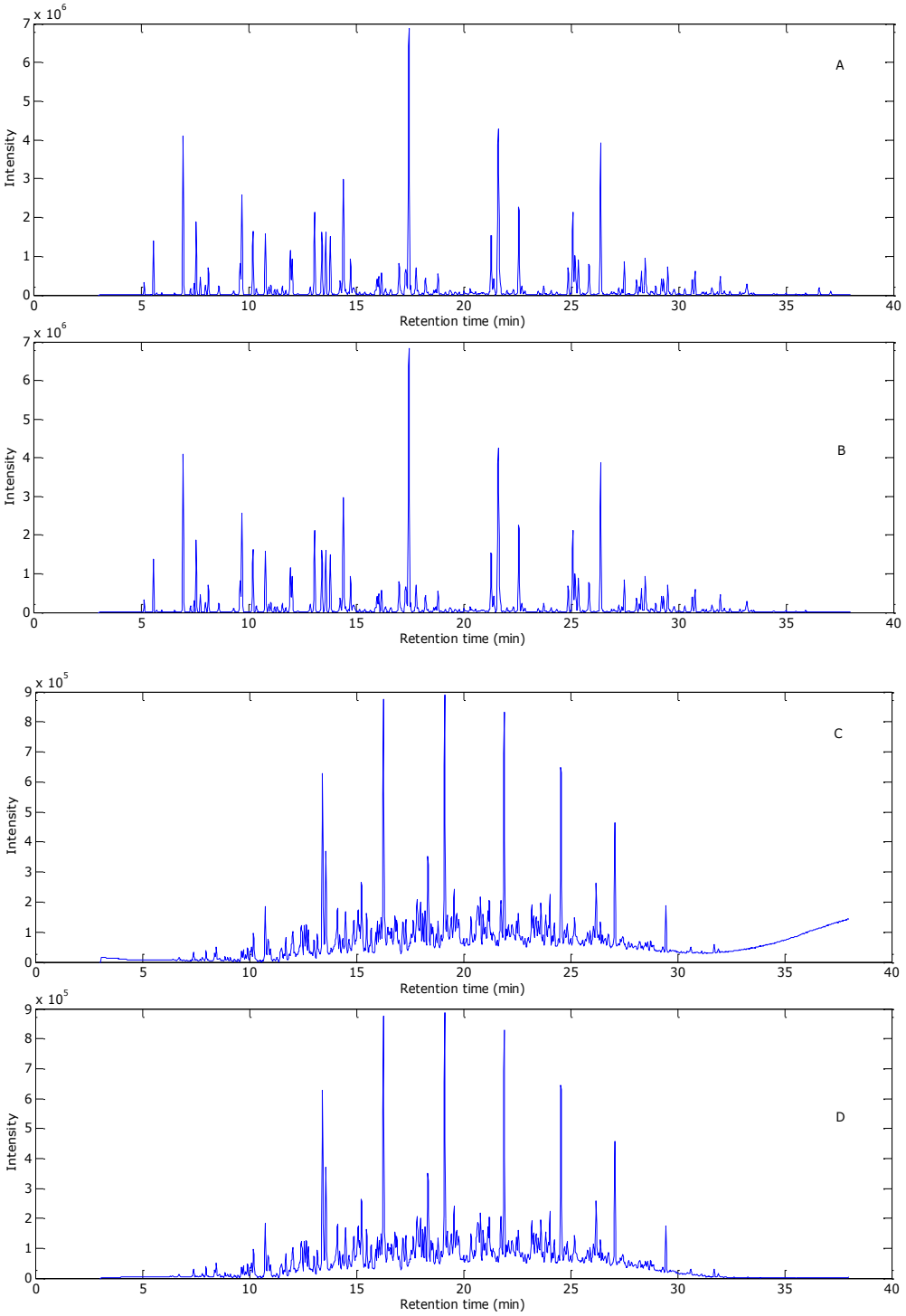

Figure 4-3. TIC chromatograms of a gasoline sample and a kerosene sample. (A) A gasoline sample before baseline correction, (B) the gasoline sample after baseline correction, (C) a kerosene sample before baseline correction, (D) the kerosene sample after baseline correction.

4.3.2 PDR Mapping

The PDR mapping of gasoline and kerosene samples by the two-way profile method is given in Figures 4-4 and 4-5, respectively. The geometric mean of PDRs are plotted in grayscale, which are measured repeatedly by removing one replicate from each class, a total of nine combinations of subsets for a pair of classes. The darkness of the box indicates the PDR value. All PDR values that are greater than or equal to 5 are plotted in white. The numbers printed in the box are the number of times out of a total of 60 times that an object was misclassified between the pair of classes during the BLP validation by FuRES. For example, sample 42 and sample 40 was misclassified as each other 20 times in Figure 4-4. Most of the misclassifications of the classes are located in gray boxes, indicating that the PDR metric effectively measures the predictive ability of the classifiers. It can be concluded the lower the PDR between two classes, the more likely misclassification will occur.

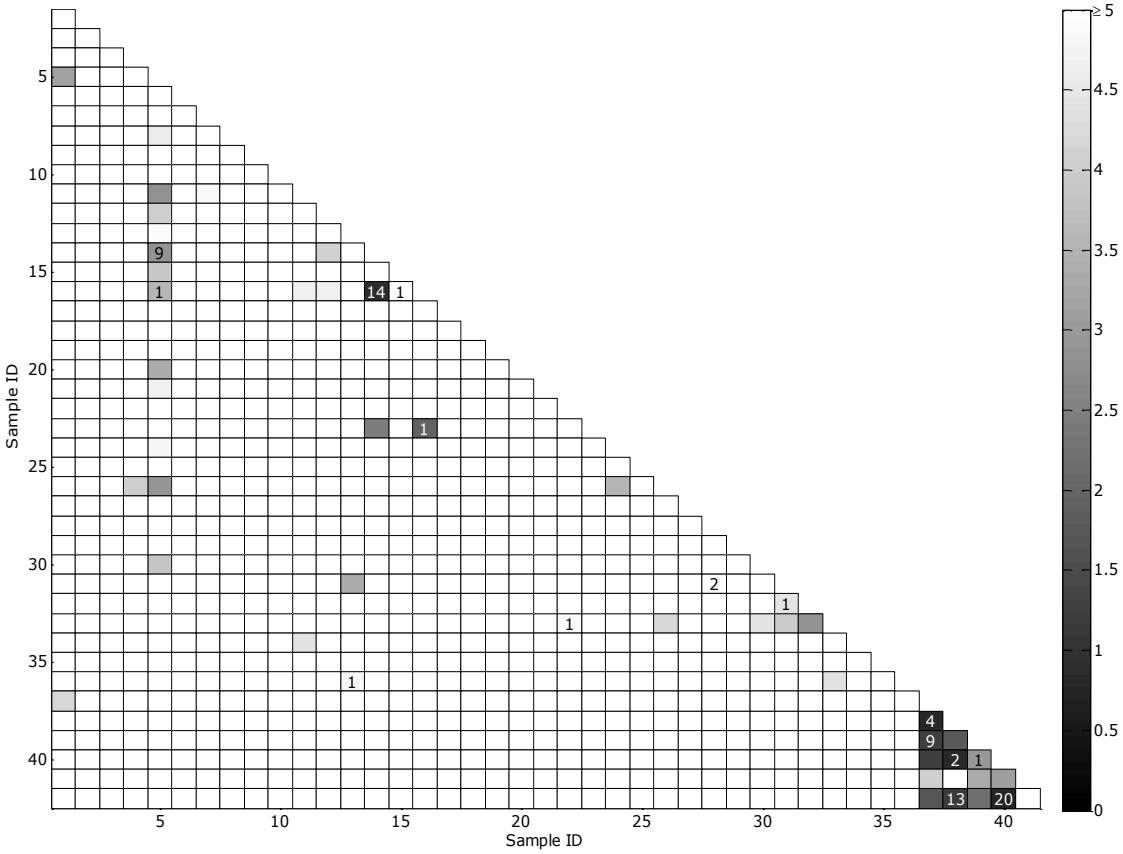

Figure 4-4. The PDR mapping of gasoline samples by the two-way profile method. The PDR values and the FuRES prediction use different bootstrap approaches. The PDR values are encoded by color intensity, which is the geometric mean of all possible subsets of Latin partitions. All PDR values that are greater than or equal to 5 are plotted in white. In a pair of classes that comprised of six objects, the subsets that comprised of four objects were obtained by removing one out of three objects in each class, which results nine possible combinations of subsets. The numbers in the box are the numbers of misclassifications between the corresponding pair of samples out of a total of 60 times by the BLP validation of the FuRES model.

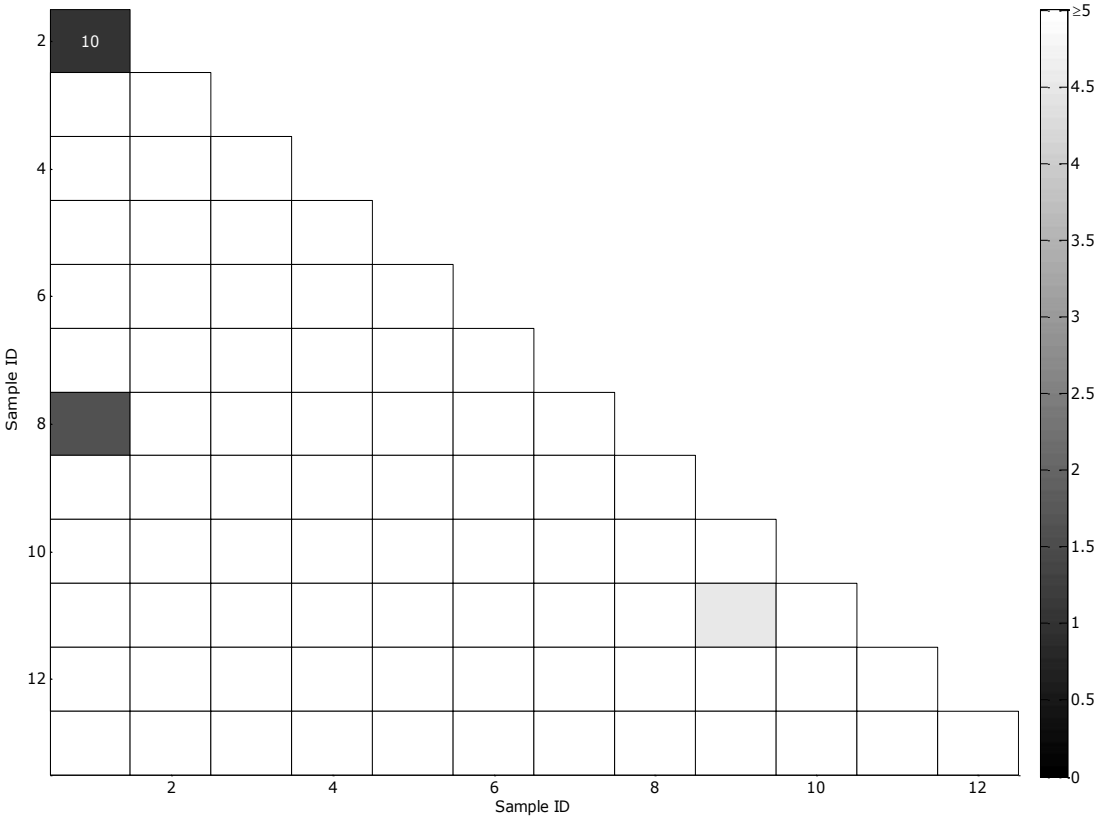

Figure 4-5. The PDR mapping of kerosene samples by the two-way profile method. The figure is plotted following the same method as described in Figure 4-4.

4.3.3 Pattern Classification

The BLP validation of PDR metric, oPLS-DA, PLS-DA, and FuRES are given in Table 4-4. The effect of baseline correction is evaluated. Although the prediction accuracy is not improved for two-way profiles after baseline correction, the PDRs were improved significantly in the kerosene spectra, indicating that the separation between each pair of classes was generally improved.

Both the two-way profile and component ratio methods achieved prediction accuracies greater than 90% using the FuRES classifier. For the gasoline data set, the two-way profile method and the component ratio method performed equally well. The two-way profile method achieved higher prediction accuracies than the component ratio method for the kerosene data set because the two-way profiles retain more chemical information. The loss of peak information manifests itself in lower PDR values. The PDRs of the component ratio method is lower than the two-way profile method for both gasoline and kerosene data. It is essentially a differential transformation so there is a loss in signal-to-noise ratio, which can be expected with any differential transformation.

Table 4-4. PDRs and prediction accuracies of oPLS-DA, PLS-DA and FuRES with 95% confidence intervals by BLP validation. Both full two-way profile and component ratio methods are reported.

	Gasoline	**Kerosene**
Total number of objects	126	39
Two-way profile, original spectra		
Geometric mean PDR	16 ± 3	17 ± 8
oPLS-DA (%)	99 ± 0	100 ± 0
PLS-DA (%)	95 ± 2	83 ± 6
FuRES (%)	93 ± 3	97 ± 0
Two-way profile with baseline correction		
Geometric mean PDR	16 ± 2	41 ± 15
oPLS-DA (%)	99 ± 1	100 ± 0
PLS-DA (%)	94 ± 2	92 ± 5
FuRES (%)	94 ± 2	97 ± 0
Component ratio		
Geometric mean PDR	8 ± 1	9 ± 2
oPLS-DA (%)	81 ± 5	81 ± 7
PLS-DA (%)	48 ± 7	62 ± 5
FuRES (%)	94 ± 3	91 ± 6

The two-way spectra contain noise. As a positively biased method, oPLS-DA achieved higher prediction accuracies for the gasoline data because of overfitting the data. FuRES is a soft classifier and is inherently resistant to overfitting. However, for the component ratio method overfitting is mostly avoided because the training data set is overdetermined (i.e., fewer variables than objects). As a result, the FuRES method achieved better predictions for the component ratio data than the biased oPLS-DA method. The unbiased PLS-DA method achieved marginally better prediction accuracies for the two-way gasoline data that are statistically insignificant. Unbiased PLS-DA performs worse than the FuRES method for the rest of the data sets, especially in the classification of component ratio data. The performance demonstrated that FuRES is a powerful classifier for samples measured by GC/MS.

4.4 Conclusions

In this study, different IL samples were identified using several chemometric techniques. PDR measured the separation between the different samples and the results were presented as heat maps. PLS-DA and FuRES was used to build classification models. The models were validated by BLP validation. FuRES for both the gasoline and kerosene data sets predicted the classes with greater than 90% accuracy. Furthermore, the results of PDRs and pattern classifications were consistent in both data sets. The results indicated the usefulness of various chemometric methods including baseline correction by PCA, PCT, PDR, PLS-DA, and FuRES to the forensic

analysis workflow of the IL identification task by GC/MS. A novel method, PDR mapping, is presented for the first time for characterizing complex data sets.

The work also demonstrates the usefulness of both the two-way profile and component ratio methods. The PCT compressed two-way profile keeps both the gas chromatographic and mass spectrometric information, which is useful in comparing unevaporated samples. Although less accurate in kerosene sample prediction, the component ratio method is an effective method that provides an approach to compare unevaporated gasoline samples and fire debris. Future work will involve the identification of ILR from fire debris by chemometrics, as well as to perform a feature selection study on the choice of component peaks.

Chapter 5 Summary

The development of classification techniques is a large and active field that involves knowledge in chemistry, mathematics, statistics, computer science, etc. It is important that the classifier combines some features in data analysis so as to process complex data set such as data sets that contain outliers and the imbalanced data. In Chapter 2, the RBFCCN is proposed as a new classification method with the ability to detect outliers in the data set. In addition, the RBFCCN has the advantage of incremental learning. The RBFCCN has been compared to other classification methods. The result proves RBFCCN provides attractive abilities, including novelty detection, classification on imbalanced data, and incremental learning without sacrificing the classification performance compared to other RBFNs.

In future studies, RBFCCN may be applied to other research projects. Other than the comparison of prediction performances against other classifiers, it is useful to conduct other studies on the comparisons of RBFCCNs against other outlier diagnostic techniques, such as one-class SVM.

Chapters 3 and 4 demonstrate applications of chemometrics, especially multivariate classification techniques, in the fields of chromatography and spectroscopy. In Chapter 3, the biomarker candidate of ESI-MS bacteria is obtained through the FuRES–MoWeD pipeline. This technique is demonstrated with the biomarker candidate discovery study on the strain identification of *Salmonella enterica*. The data processing speed is an order of magnitude faster than other comparing methods. The project is of great

importance in food safety monitoring because some specific bacteria strains could be life-threatening pathogens. The experiment investigates the possibility to combine pattern classification to other processing methods, in this case charge state deconvolution.

The concept of the hyphenated processing approach has great potential in a wide area of proteomics. Specifically, database searches on the discriminants would potentially be useful when dealing with protein digest samples, for which small peptide peaks are present. Future studies will be to apply the discriminant to a proteomic database search routine when processing LC–MS and LC–MS/MS data. Additionally, the proposed FuRES–MoWeD pipeline will be compared to clustering methods, which is a widely used unsupervised classifier in proteomics and bioinformatics to find biomarkers.

In Chapter 4, individual gasoline and kerosene samples are identified, which is more specific than the current identification techniques that characterize different types of ignitable liquids. The proposed classification method generates the results based on an entire set of experiment that consists of many samples. This project is important in the forensic investigation of arson cases because the identification result was reported in an average classification rate with statistical measure of confidence, which is more specific than the commonly used pair-wise comparison in standard forensic procedures. The result indicates the potential use of chemometric modeling techniques towards the evidence investigation of ignitable liquids.

In addition, PDR mapping is proposed as a convenient data visualization technique for characterizing complex data sets. The PDR mapping technique presented the separation between classes, which is consistent with the classification results.

The future studies of the chemometric characterization of ignitable liquids will be focused on the development of instrumentations, because intelligent instrumentations can be greatly benefitted from the chemometric classification techniques developed in Chapter 4. In field applications such as industrial quality assurance of refinery product or forensic analysis to study the identification of fuel residues from fire debris, it is desirable to apply easy-to-use portable analytical instruments. With the recent developments of GC/MS miniaturization, some commercialized portable GC/MS models are available nowadays. The chemometric model developed in Chapter 4 can be programmed into microchips installed in a portable GC/MS so as to obtain an intelligent instrument for rapid on-site characterization of ignitable liquids.

References

1. Skoog, D. A.; Holler, F. J.; Nieman, T. A., In *Principles of Instrumental Analysis 5th Edition*, Thomson Brooks/Cole: Pacific Grove, CA, 1998; pp 674-794.

2. Skoog, D. A.; Holler, F. J.; Nieman, T. A., In *Principles of Instrumental Analysis 5th Edition*, Thomson Brooks/Cole: Pacific Grove, CA, 1998; pp 498-534.

3. Gross, J. H., In *Mass Spectrometry: A Textbook*, Springer: New York, 2004; pp 1-12.

4. You, Q.; Wang, B. W.; Chen, F.; Huang, Z. L.; Wang, X.; Luo, P. G. Comparison of Anthocyanins and Phenolics in Organically and Conventionally Grown Blueberries in Selected Cultivars. *Food Chemistry* **2011,** *125(1)*, 201-208.

5. Lee, M. J.; Choi, J. S.; Cha, S. W.; Lee, K. S.; Lee, Z. W.; Hwang, G. S.; Lee, S. H.; Kamal, A. H. M.; Jung, Y. A.; Seung, N. S.; Woo, S. H. Variation in the Ginsenoside Profiles of Cultivated Ginseng (Panax Ginseng CA Meyer) Landraces in Korea. *Process Biochemistry* **2011,** *46(1)*, 258-264.

6. Kertesz, V.; Van Berkel, G. J. Liquid Microjunction Surface Sampling Coupled with High-Pressure Liquid Chromatography-Electrospray Ionization-Mass Spectrometry for Analysis of Drugs and Metabolites in Whole-Body Thin Tissue Sections. *Analytical Chemistry* **2010,** *82(14)*, 5917-5921.

7. Barfield, M.; Wheller, R. Use of Dried Plasma Spots in the Determination of Pharmacokinetics in Clinical Studies: Validation of a Quantitative Bioanalytical Method. *Analytical Chemistry* **2011,** *83(1)*, 118-124.

8. Mimmi, M. C.; Picotti, P.; Corazza, A.; Betto, E.; Pucillo, C. E.; Cesaratto, L.; Cedolini, C.; Londero, V.; Zuiani, C.; Bazzocchi, M.; Esposito, G. High-Performance Metabolic Marker Assessment in Breast Cancer Tissue by Mass Spectrometry. *Clinical Chemistry and Laboratory Medicine* **2011,** *49(2)*, 317-324.

9. Silva, E. G.; Lopez, P. R.; Atkinson, E. N.; Fente, C. A. A New Approach for Identifying Patients with Ovarian Epithelial Neoplasms Based on High-Resolution Mass Spectrometry. *American Journal of Clinical Pathology* **2010,** *134(6)*, 903-909.

10. Lu, Y.; Chen, P.; Harrington, P. B. Comparison of Differential Mobility Spectrometry and Mass Spectrometry for Gas Chromatographic Detection of Ignitable Liquids from Fire Debris Using Projected Difference Resolution. *Analytical and Bioanalytical Chemistry* **2009,** *394(8)*, 2061-2067.

11. Sun, X. B.; Zimmermann, C. M.; Jackson, G. P.; Bunker, C. E.; Harrington, P. B. Classification of Jet Fuels by Fuzzy Rule-Building Expert Systems Applied to Three-Way Data by Fast Gas Chromatography-Fast Scanning Quadrupole Ion Trap Mass Spectrometry. *Talanta* **2011,** *83(4)*, 1260-1268.

12. Gregg, S. D.; Campbell, J. L.; Fisher, J. W.; Bartlett, M. G. Methods for the Characterization of Jet Propellent-8: Vapor and Aerosol. *Biomedical Chromatography* **2007,** *21(5)*, 463-472.

13. Langel, K.; Gunnar, T.; Ariniemi, K.; Rajamaki, O.; Lillsunde, P. A Validated Method for the Detection and Quantitation of 50 Drugs of Abuse and Medicinal Drugs in Oral Fluid by Gas Chromatography-Mass Spectrometry. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **2011,** *879(13-14)*, 859-870.

14. Pieri, M.; Castiglia, L.; Miraglia, N.; Guadagni, R.; Malorni, L.; Sannolo, N.; Acampora, A.; Della Casa, E. Study of the Fragmentation Pattern of Ketamine-Heptafluorobutyramide by Gas Chromatography/Electron Ionization Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2010,** *24(1)*, 49-56.

15. Routon, B. J.; Kocher, B. B.; Goodpaster, J. V. Discriminating Hodgdon Pyrodex (R) and Triple Seven (R) Using Gas Chromatography-Mass Spectrometry. *Journal of Forensic Sciences* **2011,** *56(1)*, 194-199.

16. Yamaguchi, S.; Uchimura, T.; Imasaka, T.; Imasaka, T. Gas Chromatography/Time-of-Flight Mass Spectrometry of Triacetone Triperoxide Based on Femtosecond Laser Ionization. *Rapid Communications in Mass Spectrometry* **2009,** *23(19)*, 3101-3106.

17. Shen, C. Y.; Cao, X. W.; Shen, W. J.; Jiang, Y. A.; Zhao, Z. Y.; Wu, B.; Yu, K. Y.; Liu, H.; Lian, H. Z. Determination of 17 Pyrethroid Residues in Troublesome Matrices by Gas Chromatography/Mass Spectrometry with Negative Chemical Ionization. *Talanta* **2011,** *84(1)*, 141-147.

18. Lavagnini, I.; Urbani, A.; Magno, F. Overall Calibration Procedure Via a Statistically Based Matrix-Comprehensive Approach in the Stir Bar Sorptive Extraction-Thermal Desorption-Gas Chromatography-Mass Spectrometry Analysis of Pesticide Residues in Fruit-Based Soft Drinks. *Talanta* **2011,** *83(5)*, 1754-1762.

19. Centers for Disease Control and Prevention. CDC Estimates of Foodborne Illness in the United States. http://www.cdc.gov/foodborneburden/2011-foodborne-estimates.html (accessed May 2011)

20. U.S. Fire Administration, *Topical Fire Research Series, Vol. 1*. 2001; pp 1-3.

21. Kiralj, R.; Ferreira, M. M. C. The Past, Present, and Future of Chemometrics Worldwide: Some Etymological, Linguistic, and Bibliometric Investigations. *Journal of Chemometrics* **2006,** *20(6-7)*, 247-272.

22. Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; Jong, S. D.; Lewi, P. J.; Smeyers-verbeke, J., Handbook of Chemometrics and Qualimetrics: Part A. In *Data Handling in Science and Technology ; V. 20a.*, Elsevier: New York, 1997; pp 1-18.

23. Brereton, R. G., In *Applied Chemometrics for Scientists*, John Wiley & Sons: Chichester, West Sussex, UK, 2007; pp 1-8.

24. Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; Jong, S. D.; Lewi, P. J.; Smeyers-verbeke, J., Handbook of Chemometrics and Qualimetrics: Part B. In *Data Handling in Science and Technology ; V. 20b.*, Elsevier: New York, 1997; pp 207-242.

25. Harrington, P. B. Statistical Validation of Classification and Calibration Models Using Bootstrapped Latin Partitions. *Trac-Trends in Analytical Chemistry* **2006,** *25(11)*, 1112-1124.

26. Masters, T., Practical Neural Network Recipes in C++. In Academic Press: Boston, 1993; pp 78-116.

27. Basheer, I. A.; Hajmeer, M. Artificial Neural Networks: Fundamentals, Computing, Design, and Application. *Journal of Microbiological Methods* **2000,** *43(1)*, 3-31.

28. Bishop, C., In *Neural Networks for Pattern Recognition*, Oxford University Press: Oxford, UK, 1995; pp 116-163.

29. Fahlman, S. E.; Lebiere, C. *The Cascade-Correlation Learning Architecture*; Report CMU-CS-90-100; Carnegie Mellon University: Pittsburgh, PA, 1991; pp 1-13.

30. Jurs, P. C., Computer-Enhanced Analytical Spectroscopy, Volume 3. In *Modern Analytical Chemistry.*, Plenum Press: New York, 1992; Vol. 3, pp 239-260.

31. Harrington, P. B. Fuzzy Multivariate Rule-Building Expert Systems - Minimal Neural Networks. *Journal of Chemometrics* **1991,** *5(5)*, 467-486.

32. Lu, Y.; Harrington, P. B. Forensic Application of Gas Chromatography–Differential Mobility Spectrometry with Two-Way Classification of Ignitable Liquids from Fire Debris. *Analytical Chemistry* **2007,** *79(17)*, 6752-6759.

33. Rearden, P.; Harrington, P. B.; Karnes, J. J.; Bunker, C. E. Fuzzy Rule-Building Expert System Classification of Fuel Using Solid-Phase Microextraction Two-Way Gas Chromatography Differential Mobility Spectrometric Data. *Analytical Chemistry* **2007,** *79(4)*, 1485-1491.

34. Xu, Z. F.; Bunker, C. E.; Harrington, P. D. Classification of Jet Fuel Properties by near-Infrared Spectroscopy Using Fuzzy Rule-Building Expert Systems and Support Vector Machines. *Applied Spectroscopy* **2010,** *64(11)*, 1251-1258.

35. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass-Spectrometry of Large Biomolecules. *Science* **1989,** *246(4926)*, 64-71.

36. Mann, M.; Meng, C. K.; Fenn, J. B. Interpreting Mass-Spectra of Multiply Charged Ions. *Analytical Chemistry* **1989,** *61(15)*, 1702-1708.

37. Ferrige, A. G.; Seddon, M. J.; Green, B. N.; Jarvis, S. A.; Skilling, J. Disentangling Electrospray Spectra with Maximum-Entropy. *Rapid Communications in Mass Spectrometry* **1992,** *6(11)*, 707-711.

38. Reinhold, B. B.; Reinhold, V. N. Electrospray Ionization Mass-Spectrometry - Deconvolution by an Entropy-Based Algorithm. *Journal of the American Society for Mass Spectrometry* **1992,** *3(3)*, 207-215.

39. Ferrige, A. G.; Seddon, M. J.; Jarvis, S. Maximum-Entropy Deconvolution in Electrospray Mass-Spectrometry. *Rapid Communications in Mass Spectrometry* **1991,** *5(8)*, 374-377.

40. Hagen, J. J.; Monnig, C. A. Method for Estimating Molecular-Mass from Electrospray Spectra. *Analytical Chemistry* **1994,** *66(11)*, 1877-1883.

41. Pearcy, J. O.; Lee, T. D. MoWeD, a Computer Program to Rapidly Deconvolute Low Resolution Electrospray Liquid Chromatography/Mass Spectrometry Runs to Determine Component Molecular Weights. *Journal of the American Society for Mass Spectrometry* **2001,** *12(5)*, 599-606.

42. Zhang, Z. Q.; Marshall, A. G. A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra. *Journal of the American Society for Mass Spectrometry* **1998,** *9(3)*, 225-233.

43. Zheng, H. R.; Ojha, P. C.; McClean, S.; Black, N. D.; Hughes, J. G.; Shaw, C. Heuristic Charge Assignment for Deconvolution of Electrospray Ionization Mass Spectra. *Rapid Communications in Mass Spectrometry* **2003,** *17(5)*, 429-436.

44. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *Journal of the American Society for Mass Spectrometry* **2000,** *11(4)*, 320-332.

45. Tseng, Y. H.; Uetrecht, C.; Heck, A. J. R.; Peng, W. P. Interpreting the Charge State Assignment in Electrospray Mass Spectra of Bioparticles. *Analytical Chemistry* **2011,** *83(6)*, 1960-1968.

46. Maleknia, S. D.; Downard, K. M. Charge Ratio Analysis Method: Approach for the Deconvolution of Electrospray Mass Spectra. *Analytical Chemistry* **2005,** *77(1)*, 111-119.

47. Maleknia, S. D.; Green, D. C. eCRAM Computer Algorithm for Implementation of the Charge Ratio Analysis Method to Deconvolute Electrospray Ionization Mass Spectra. *International Journal of Mass Spectrometry* **2010,** *290(1)*, 1-8.

48. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986,** *323(6088)*, 533-536.

49. Rumelhart, D. E.; Macclelland, J. L., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: 1: Foundations*. 2nd print. ed.; MIT Press: Cambridge, MA., 1986; pp 318-362.

50. Eiceman, G. A.; Wang, M.; Prasad, S.; Schmidt, H.; Tadjimukhamedov, F. K.; Lavine, B. K.; Mirjankar, N. Pattern Recognition Analysis of Differential Mobility Spectra with Classification by Chemical Family. *Analytica Chimica Acta* **2006,** *579(1)*, 1-10.

51. Marengo, E.; Bobba, M.; Robotti, E.; Lenti, M. Hydroxyl and Acid Number Prediction in Polyester Resins by near Infrared Spectroscopy and Artificial Neural Networks. *Analytica Chimica Acta* **2004,** *511(2)*, 313-322.

52. Wood, M. J.; Hirst, J. D. Predicting Protein Secondary Structure by Cascade-Correlation Neural Networks. *Bioinformatics* **2004,** *20(3)*, 419-420.

53. Diamantopoulou, M. J.; Antonopoulos, V. Z.; Papamichail, D. M. Cascade Correlation Artificial Neural Networks for Estimating Missing Monthly Values of Water Quality Parameters in Rivers. *Water Resources Management* **2007,** *21(3)*, 649-662.

54. Harrington, P. B. Temperature-Constrained Cascade Correlation Networks. *Analytical Chemistry* **1998,** *70(7)*, 1297-1306.

55. Chen, P.; Harrington, P. B. Discriminant Analysis of Fused Positive and Negative Ion Mobility Spectra Using Multivariate Self-Modeling Mixture Analysis and Neural Networks. *Applied Spectroscopy* **2008,** *62(2)*, 133-141.

56. Wang, F.; Zhang, Z.; Cui, X.; Harrington, P. B. Identification of Rhubarbs by Using Nir Spectrometry and Temperature-Constrained Cascade Correlation Networks. *Talanta* **2006,** *70(5)*, 1170-1176.

57. Wan, C.; Harrington, P. B. Screening GC-MS Data for Carbamate Pesticides with Temperature-Constrained Cascade Correlation Neural Networks. *Analytica Chimica Acta* **2000,** *408(1-2)*, 1-12.

58. Wan, C.; Harrington, P. B. Self-Configuring Radial Basis Function Neural Networks for Chemical Pattern Recognition. *Journal of Chemical Information and Computer Science* **1999,** *39(6)*, 1049-1056.

59. Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995,** *20(3)*, 273-297.

60. Breiman, L. Random Forests. *Machine Learning* **2001,** *45(1)*, 5-32.

61. Chang, C.; Lin, C. Libsvm: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed June 2009)

62. Jaiantilal, A. Randomforest-Matlab http://code.google.com/p/randomforest-matlab/ (accessed June 2009)

63. Walczak, B.; Massart, D. L. Local Modelling with Radial Basis Function Networks. *Chemometrics and Intelligent Laboratory Systems* **2000,** *50(2)*, 179-198.

64. Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression - a Tutorial. *Analytica Chimica Acta* **1986,** *185*, 1-17.

65. Clarkson University's FTP Archive. ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/Original/oliveoil.dat (accessed July 2008)

66. Hopke, P. K.; Massart, D. L. Reference Data Sets for Chemometrical Methods Testing. *Chemometrics and Intelligent Laborary Systems* **1993,** *19(1)*, 35-41.

67. McLafferty, F. W., *Registry of Mass Spectral Data, 5th Edition*. John Wiley & Sons: New York, 1989.

68. Crawford, L. R.; Morrison, J. D. Computer Methods in Analytical Mass Spectrometry. Identification of an Unknown Compound in a Catalog. *Analytical Chemistry* **1968,** *40(10)*, 1464-1469.

69. Tandler, P. J.; Butcher, J. A.; Hu, T.; Harrington, P. B. Analysis of Plastic Recycling Products by Expert Systems. *Analytica Chimica Acta* **1995,** *312(3)*, 231-244.

70. Haselberg, R.; de Jong, G. J.; Somsen, G. W. Capillary Electrophoresis-Mass Spectrometry for the Analysis of Intact Proteins. *Journal of Chromatography A* **2007,** *1159(1-2)*, 81-109.

71. Theodorescu, D.; Schiffer, E.; Bauer, H. W.; Douwes, F.; Eichhorn, F.; Polley, R.; Schmidt, T.; Schofer, W.; Zurbig, P.; Good, D. M.; Coon, J. J.; Mischak, H. Discovery and Validation of Urinary Biomarkers for Propstate Cancer. *Proteomics Clinical Applications* **2008,** *2(4)*, 556-570.

72. Puerta, A.; Bergquist, J. Development of a CE-MS Method to Analyze Components of the Potential Biomarker Vascular Endothelial Growth Factor 165. *Electrophoresis* **2009,** *30(13)*, 2355-2365.

73. Moini, M.; Huang, H. Application of Capillary Electrophoresis/ Electrospray Ionization-Mass Spectrometry to Subcellular Proteomics of Escherichia Coli Ribosomal Proteins. *ELECTROPHORESIS 25(13)*, 1981-1987.

74. Williams, T. L.; Leopold, P.; Musser, S. Automated Postprocessing of Electrospray LC/MS Data for Profiling Protein Expression in Bacteria. *Analytical Chemistry* **2002,** *74(22)*, 5807-5813.

75. Everley, R. A.; Mott, T. M.; Wyatt, S. A.; Toney, D. M.; Croley, T. R. Liquid Chromatography/Mass Spectrometry Characterization of Escherichia Coli and Shigella Species. *Journal of the American Society for Mass Spectrometry* **2008,** *19(11)*, 1621-1628.

76. Mott, T. M.; Everley, R. A.; Wyatt, S. A.; Toney, D. M.; Croley, T. R. Comparison of MALDI-TOF/MS and LC-QTOF/MS Methods for the Identification of Enteric Bacteria. *International Journal of Mass Spectrometry* **2010,** *291(1-2)*, 24-32.

77. Williams, T. L.; Monday, S. R.; Edelson-Mammel, S.; Buchanan, R.; Musser, S. M. A Top-Down Proteomics Approach for Differentiating Thermal Resistant Strains of Enterobacter Sakazakii. *PROTEOMICS* **2005,** *5(16)*, 4161-4169.

78. Trygg, J.; Holmes, E.; Lundstedt, T. Chemometrics in Metabonomics. *Journal of Proteome Research* **2007,** *6(2)*, 469-479.

79. Bijlsma, S.; Bobeldijk, L.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K. Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Analytical Chemistry* **2006,** *78(2)*, 567-574.

80. Michaud, F. T.; Garnier, A.; Lemieux, L.; Duchesne, C. Multivariate Analysis of Single Quadrupole LC-MS Spectra for Routine Characterization and Quantification of Intact Proteins. *Proteomics* **2009,** *9(3)*, 512-520.

81. Harrington, P. B.; Laurent, C.; Levinson, D. F.; Levitt, P.; Markey, S. P. Bootstrap Classification and Point-Based Feature Selection from Age-Staged Mouse Cerebellum Tissues of Matrix Assisted Laser Desorption/Ionization Mass Spectra Using a Fuzzy Rule-Building Expert System. *Analytica Chimica Acta* **2007,** *599(2)*, 219-231.

82. Harrington, P. d. B.; Vieira, N. E.; Chen, P.; Espinoza, J.; Nien, J. K.; Romero, R.; Yergey, A. L. Proteomic Analysis of Amniotic Fluids Using Analysis of Variance-Principal Component Analysis and Fuzzy Rule-Building Expert Systems Applied to Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *Chemometrics and Intelligent Laboratory Systems* **2006,** *82(1-2)*, 283-293.

83. Schulz-Trieglaff, O.; Pfeifer, N.; Gropl, C.; Kohlbacher, O.; Reinert, K. LC-MSsim - a Simulation Software for Liquid Chromatography Mass Spectrometry Data. *BMC Bioinformatics* **2008,** *9*.

84. Hibbert, D. B. A Prolog Program for the Calculation of Isotope Distributions in Mass-Spectrometry. *Chemometrics and Intelligent Laborary Systems* **1989,** *6(3)*, 203-212.

85. Beltran, P.; Plock, S. A.; Smith, N. H.; Whittam, T. S.; Old, D. C.; Selander, R. K. Reference Collection of Strains of the Salmonella-Typhimurium Complex from Natural-Populations. *J. Gen. Microbiol.* **1991,** *137*, 601-606.

86. Donoho, D.; Duncan, M.; Huo, X.; Levi-Tsabari, O. Wavelab 850. http://www-stat.stanford.edu/~wavelab/ (accessed December 2009)

87. Donoho, D. L.; Johnstone, J. M. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika* **1994,** *81(3)*, 425-455.

88. The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* **2009,** *38(suppl 1)*, D142-D148.

89. Jain, E.; Bairoch, A.; Duvaud, S.; Phan, I.; Redaschi, N.; Suzek, B.; Martin, M.; McGarvey, P.; Gasteiger, E. Infrastructure for the Life Sciences: Design and Implementation of the UniProt Website. *BMC Bioinformatics* **2009,** *10(1)*, 136.

90. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A., Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, Walker, J. M., Ed. Humana Press: Totowa, New Jersey, USA, 2005; pp 571-607.

91. Konermann, L.; Collings, B. A.; Douglas, D. J. Cytochrome C Folding Kinetics Studied by Time-Resolved Electrospray Ionization Mass Spectrometry. *Biochemistry* **1997,** *36(18)*, 5554-5559.

92. *American Society for Testing Materials ASTM E1618 - 10, "Standard Test Method for Ignitable Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography-Mass Spectrometry"*. 2010.

93. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC, 2009.

94. DeHaan, J. D.; Icove, D. J., In *Kirk's Fire Investigation*, 7th ed.; Pearson: Upper Saddle River, NJ, 2011; p 679.

95. Sigman, M. E.; Williams, M. R. Covariance Mapping in the Analysis of Ignitable Liquids by Gas Chromatography/Mass Spectrometry. *Analytical Chemistry* **2006,** *78(5)*, 1713-1718.

96. Sigman, M. E.; Williams, M. R.; Ivy, R. G. Individualization of Gasoline Samples by Covariance Mapping and Gas Chromatography/Mass Spectrometry. *Analytical Chemistry* **2007,** *79(9)*, 3462-3468.

97. Sandercock, P. M. L.; Du Pasquier, E. Chemical Fingerprinting of Unevaporated Automotive Gasoline Samples. *Forensic Science International* **2003,** *134(1)*, 1-10.

98. Sandercock, P. M. L.; Du Pasquier, E. Chemical Fingerprinting of Gasoline - 2. Comparison of Unevaporated and Evaporated Automotive Gasoline Samples. *Forensic Science International* **2004,** *140(1)*, 43-59.

99. Sandercock, P. M. L.; Du Pasquier, E. Chemical Fingerprinting of Gasoline - Part 3. Comparison of Unevaporated Automotive Gasoline Samples from Australia and New Zealand. *Forensic Science International* **2004,** *140(1)*, 71-77.

100. Hupp, A. M.; Marshall, L. J.; Campbell, D. I.; Smith, R. W.; McGuffin, V. L. Chemometric Analysis of Diesel Fuel for Forensic and Environmental Applications. *Analytica Chimica Acta* **2008,** *606(2)*, 159-171.

101. Monfreda, M.; Gregori, A. Differentiation of Unevaporated Gasoline Samples According to Their Brands, by SPME-GC-MS and Multivariate Statistical Analysis. *Journal of Forensic Sciences* **2011,** *56(2)*, 372-380.

102. Desa, W. N. S. M.; Daeid, N. N.; Ismail, D.; Savage, K. Application of Unsupervised Chemometric Analysis and Self-Organizing Feature Map (SOFM) for the Classification of Lighter Fuels. *Analytical Chemistry 82(15)*, 6395-6400.

103. Doble, P.; Sandercock, M.; Du Pasquier, E.; Petocz, P.; Roux, C.; Dawson, M. Classification of Premium and Regular Gasoline by Gas Chromatography/Mass Spectrometry, Principal Component Analysis and Artificial Neural Networks. *Forensic Science International* **2003,** *132(1)*, 26-39.

104. Rankin, J. G.; Bondra, A.; Trader, C.; Lu, W.; Harrington, P. In *Target Compound Ratios and Chemometric Analyses for the Individualization of Neat Ignitable Liquids and Residues from Fire Debris*, Conference Proceedings of the 12th International Interflam Conference, London, July 5-7, 2010; Interscience Communications: London, 2010; pp 1305-1320.

105. Barnes, A. T.; Dolan, J. A.; Kuk, R. J.; Siegel, J. A. Comparison of Gasolines Using Gas Chromatography-Mass Spectrometry and Target Ion Response. *Journal of Forensic Sciences* **2004,** *49(5)*, 1018-1023.

106. Cao, L. Nonlinear Wavelet Compression Methods for Ion Analyses and Dynamic Modeling of Complex Systems. OHIO University, Athens, OH, 2004. p 180.

107. Harrington, P. D.; Kister, J.; Artaud, J.; Dupuy, N. Automated Principal Component-Based Orthogonal Signal Correction Applied to Fused near Infrared-Mid-Infrared Spectra of French Olive Oils. *Analytical Chemistry* **2009,** *81(17)*, 7160-7169.

108. Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta* **1986,** *185*, 1-17.

Appendix A

Publications

Lu, W.; Rankin, J. G.; Bondra, A.; Trader, C.; Heeren, A.; Harrington, P.B. Ignitable liquid identification using gas chromatography/mass spectrometry data by projected difference resolution mapping and fuzzy rule-building expert system classification. Submitted to *Forensic Science International*.

Lu,W.; Callahan, J.H.; Fry, F.S.; Andrzejewski, D.; Musser, S.M.; Harrington, P.B. A discriminant based charge deconvolution analysis pipeline for protein profiling of whole cell extracts using liquid chromatography–electrospray ionization-quadrupole time-of-flight mass spectrometry. *Talanta* **2011**, 84(4), 1180-1187.

Rankin, J. G.; Bondra, A.; Trader, C.; Lu, W.; Harrington, P. In *Target Compound Ratios and Chemometric Analyses for the Individualization of Neat Ignitable Liquids and Residues from Fire Debris*, Conference Proceedings of the 12th International Interflam Conference, London, July 5-7, 2010; Interscience Communications: London, 2010; pp 1305-1320.

Lu, W.; Harrington, P.B. Radial basis function cascade correlation networks. *Algorithms* **2009**, 2, 1045-1068.

Appendix B

Presentations

Harrington, P.B.; Lu, W.; Lu Y.; Harnly, J. M., "Novel algorithm for registration of hyperspectral images and retention time alignment of multiway data", the 36th Federation of Analytical Chemistry and Spectroscopy Societies (FACSS 2009).

Lu, W. and Harrington, P.B., "Theory and analytical applications of the temperature constrained radial basis function neural networks", the 59th Annual Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy (Pittcon 2008).

Thesis and Dissertation Services