Clinical Proteomics

ISSN 1542-6416/06/02:13–32/$30.00 (Online)

Original Article

Data Mining in Proteomic Mass Spectrometry

Asha Thomas,[1] Georgia D. Tourassi,[2] Adel S. Elmaghraby,[1] Roland Valdes Jr.,[3] and Saeed A. Jortani,[3,]*

[1]Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY; [2]Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, NC; and [3]Department of Pathology and Laboratory Medicine, University of Louisville, Louisville, KY

Abstract

Data mining application to proteomic data from mass spectrometry has gained much interest in recent years. Advances made in proteomics and mass spectrometry have resulted in considerable amount of data that cannot be easily visualized or interpreted. Mass spectral proteomic datasets are typically high dimensional but with small sample size. Consequently, advanced artificial intelligence and machine learning algorithms are increasingly being used for knowledge discovery from such datasets. Their overall goal is to extract useful information that leads to the identification of protein biomarker candidates. Such biomarkers could potentially have diagnostic value as tools for early detection, diagnosis, and prognosis of many diseases. The purpose of this review is to focus on the current trends in mining mass spectral proteomic data. Special emphasis is placed on the critical steps involved in the analysis of surface-enhanced laser desorption/ionization mass spectrometry proteomic data. Examples are drawn from previously published studies and relevant data mining terminology and techniques are explained.

*Author to whom all correspondence and reprint requests should be addressed:
Saeed A. Jortani, Department of Pathology, University of Louisville, Louisville, KY.
E-mail: sjortani@louisville.edu.

Key Words: Mass spectrometry; data mining; artificial intelligence; data analysis; SELDI; linear discriminant analysis; artificial neural networks.

Introduction

As the technology involved in biological domains like proteomics advances, large volumes of data are continuously generated. However, the data analysis required is increasingly overwhelming to the scientists in these domains. Contrarily, computational scientists are accustomed to dealing with large databases from domains such as marketing, financial institutions, and telecommunication. They typically face the challenge by the use of powerful machine learning and data mining tools. Such tools are the culmination of continuous advances made in the areas of artificial intelligence, statistics, and database management. Therefore, it is natural that the proteomics and computational intelligence domains find their paths crossed with the common goal of extracting clinically useful information from the wealth of proteomic spectral data.

The purpose of this review is to address some of the key issues involved in the application of data mining approaches to proteomic data from surface-enhanced laser desorption/ionization mass spectrometry (SELDI-MS). In addition, terminologies and popular strategies in mining of mass spectral proteomic data are described. The article is organized as follows. First, a brief introduction to data mining and proteomics is provided. Then, the need for applying data mining to proteomic data from SELDI-MS is explained. Thereafter, important steps in the mining process are outlined and explained with examples drawn from recent studies involving SELDI-MS data.

Data Mining in Proteomics

Proteomics

Proteomics is an emerging area in bioinformatics. First coined by Australian scientist Marc Wilkins in 1994 *(1)*, proteomics is a science that deals with the study of proteins and their interactions in an organism. The major focus of most proteomic studies is the discovery of the proteins of interest known as biomarkers *(2)*. The discovery of biomarkers is potentially valuable for the early detection, diagnosis, and monitoring of diseases (e.g., **refs.** *3–6*). Active and ongoing biomedical research advances coupled with the discovery of novel and powerful diagnostic proteomic tools have further strengthened the progress made.

There are several techniques for protein or peptide profiling using mass spectroscopy. Two of the major techniques intended for proteomic mass spectrometry are matrix-assisted laser desorption/ionization (MALDI-MS) *(7a)* and its extension, the SELDI-MS *(7–8)*. In both techniques, the biological matrix is mixed with the energy absorbing matrix and following shinning of the laser energy in vacuum environment, which is absorbed by the proteins causing them to get ionized. After application of an electric field, the ions accelerate through a flight tube until finally being detected by the instrument. The major difference between MALDI and SELDI is the fact that the latter uses additional chemistries on the surface of the chips to further isolate the proteins intended to be analyzed from the other molecules or proteins in the matrix *(7)*. This review will focus mainly on issues related to the data mining of SELDI proteomic data because many of the recent discoveries of potential new biomarkers have involved this technology.

Typically, protein profiles in body fluids such as serum, urine, or nipple aspirate, and in some studies tumor tissue, are analyzed using SELDI mass spectrometry (Ciphergen Biosystems, Fremont, CA). The output is a protein expression profile that is often a dataset containing information of protein peaks and their heights (intensities) in one or more

groups of patients (e.g., cancer, benign, healthy). The main goal of proteomic studies using mass spectral data is to identify protein patterns that are capable of reliably differentiating between various groups under study. The identification of these protein biomarker patterns that relate to a certain pathological state may aid in the early detection and prognosis of that disease.

However, identification of these biomarker patterns from mass spectral data is a challenging task. Issues such as few samples and large number of peaks, typical in MS data, have to be handled by advanced analytical techniques.

Data Mining

In several studies, protein profiles generated from SELDI-MS have been analyzed using statistical and data mining methods. Data mining is an approach used in scientific and business domains to extract meaningful and useful information from large and complex sets of data. The process of data mining is typically iterative *(9)* wherein previously unknown and potentially useful information is discovered by the use of powerful analytical techniques. The discovery process involves finding relationships and patterns in raw data that can be either utilized or assessed by decision makers and analysts. Steps involved in mining include data preparation, feature selection, model development or pattern recognition, and model assessment followed by application of developed model to new cases.

The early use of data mining was in financial institutions and marketing *(10–13)*. However, with an explosion in the amount of data being stored electronically, data mining has reached across telecommunications *(14)*, retail *(15)*, manufacturing *(16)*, health care *(17)*, fraud detection *(18)*, homeland security *(19)*, and biomedical domains *(20)*.

Data mining methods have evolved from advancements made in artificial intelligence, statistics, and database management. There are various data mining algorithms that are available today, based on different theoretical concepts. Some are in the form of decision trees (DTs), known for their ease of interpretability. Others, like artificial neural networks (ANNs), capture complex and nonlinear relationships in the data. ANNs are considered very powerful but are less interpretable. The advantages and disadvantages of each method when applied to SELDI-MS data are discussed in the following sections.

Application of Data Mining in SELDI-MS Proteomic Data

As previously mentioned, the use of mass spectrometry in proteomics results in a large number of high-dimensional data, where the number of features (peaks) is much greater than the number of samples. Data samples from SELDI-MS are typically comprised of hundreds to thousands of protein peaks. Such a vast number of data cannot be visually analyzed or handled by ordinary data mining tools. Tackling such a high-dimensional structure of a comparatively small-sized data set is a significantly challenging task. An example of two typical spectra for serum samples is depicted in **Fig. 1**.

In their quest for adequate tools for analyzing the data in hand and retrieving useful information, scientists in proteomics are increasingly dependent on advanced data mining techniques that can tackle issues such as the curse of dimensionality and limited data sets. These advanced techniques include artificial intelligence and machine learning.

Current practices in mining protein mass spectrometry data from SELDI include the following steps:

1. Data modeling using peaks identified by preprocessing and feature selection.
2. Careful data sampling to address the small sample size typical in SELDI-MS data.
3. Performance evaluation of the data models.

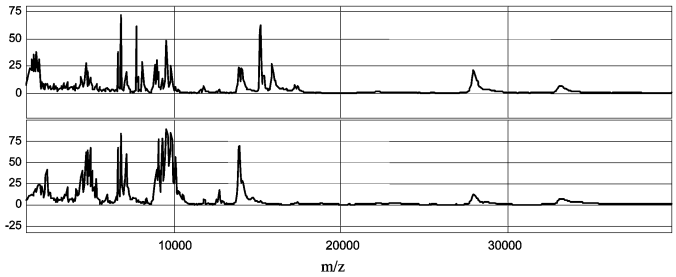

Fig. 1. Serum mass spectra. The top panel depicts the mass spectrum of a serum sample collected from a patient with inadequate heart function (test) and the bottom panel shows that of a patient with adequate function (control). Please note that the m/z for many peaks in both spectra appear to be similar, whereas others are different between the two patients. These spectra have been obtained using weak cation-exchange surfaces without prefractionation of serum samples.

The previously listed critical steps have to be carefully addressed by proteomic researchers to obtain accurate and robust decision models. The steps are repeated iteratively and modifications are made so as to probe and explore different aspects of the data. **Figure 2** outlines the mining process and the possible methods that are available under each of them. Each data-processing step is described in more detail in the following sections.

Preprocessing

The raw data obtained from SELDI-MS is noisy. The goal of preprocessing is to improve the quality of data for the subsequent steps. The results of classifier algorithms will be misleading and adversely affected when the quality of the data is poor. Thus, preprocessing and preparing the data is crucial in the analysis of raw SELDI-MS proteomic data.

Most of the published studies have used the software provided by the SELDI manufacturer for preprocessing. The SELDI software detects the locations and intensities of the proteins in each of the samples and carries out important preprocessing steps like baseline subtraction, intensity normalization, peak alignment, and peak detection. The criteria specified by the SELDI operator are used to filter the peaks.

Baseline subtraction removes the electronic and chemical noise. It is typically performed in two steps. First, the baseline is estimated using either parametric or nonparametric methods. Then, the estimated baseline is subtracted from the original mass spectrum. For example, in the SELDI software available by Ciphergen, baseline subtraction is performed by applying a filter window on the mass spectrum, estimating the average or minimum intensity within the window, and then moving the window across the spectrum to estimate the overall baseline. The size of the moving window is a critical parameter that significantly changes the outcome of the de-noising step. To date, there are no studies comparing the effectiveness of the available baseline reduction techniques and it appears that researchers optimize this step in a heuristic manner that works best with their own data sets. Although most studies are focused on reducing the low-frequency noise in the spectra, there have been many attempts to characterize and subtract the high-frequency noise components as well.

With baseline reduction completed, normalization is the next step. Because a peak in a spectrum describes only the relative amount of a protein, normalization is done to ensure meaningful comparisons across spectra. Once preprocessing is performed, the peaks obtained

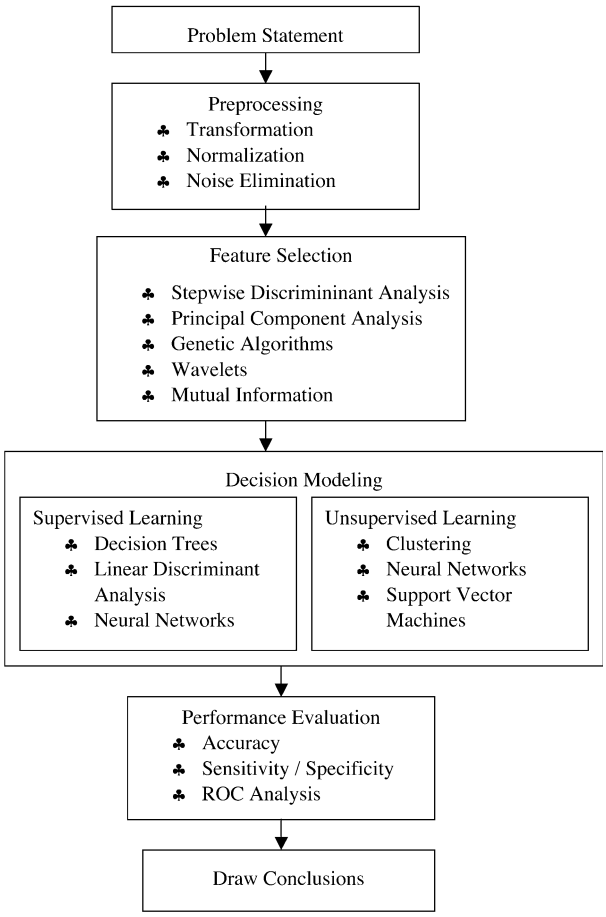

Fig. 2. Typical flowchart of the critical steps in data mining and examples of techniques available for each step. The ultimate goal is to draw conclusions based on the data set. As shown, in general, there are several options for each step. Sometimes the researchers may consider using more than one particular approach in each step.

are analyzed by further dimension reduction techniques.

Feature Selection

The Curse of Dimensionality

Most of the problems in data mining are related to the size and the complexity of the data sets. In data mining, a data set is referred to as a file or a table of rows and columns, where the rows are the records or cases and columns are the dimensions or features or attributes of the data. A data set with n dimensions is referred to as n-dimensional. Typical data mining classifiers are suited to problems where the number of features is few and the number of samples is large like the data available in retailers and manufacturers.

The straightforward approach would be to use the normalized intensity of every peak present in the spectrum as a feature. Unfortunately, in proteomic applications the number of features (dimensions) is very large (e.g., 15,000 protein peaks) and the number of data points is comparatively small (e.g., 150 patients). The data samples that have extremely high number of measurable quantities are referred to as high-dimensional data. If a one-dimensional data space has n samples, then a k-dimensional data space must have n different samples to achieve the same density. However, it is often very difficult to obtain this level of density in mass spectroscopy applications.

This problem is known in data mining as the curse of dimensionality. As the dimensionality of the input data space (i.e., the number of features) increases, it becomes exponentially more difficult to fit robust decision models (***9***). It is seen that the data often contains noisy features with very little or no information value and therefore result in the development of poor and misleading classifiers. Hence, it is simply a practical necessity to prescreen from among a large set of input variables those that are of likely utility for predicting the outputs of interest.

Prescreening the spectra to identify peaks (or features) is known as feature extraction. Feature extraction is typically done by binning. According to this process, each group of m/z points that falls within a bin is described by a value such as its average or maximum intensity. Subsequently, the characteristics of these bins (i.e., bin m/z location and estimated intensity) are used as features for data mining. The bins can be independent or

overlapping, equal-size, or adaptive. By changing the bin size and binning method, the researcher can empirically optimize the feature extraction process.

Selecting and using only the significant features for the modeling process makes the entire data mining process more accurate and effective. The data mining algorithm would be faster for a data set comprising of fewer and more relevant peaks and would result in simple and meaningful results. Thus, it is essential to remove the irrelevant and redundant features to build better models. However, it should be noted that feature selection process does not always guarantee successful selection of peaks for the classification problem. Therefore, it is necessary to validate the features that are selected by increasing the data sample size.

Feature Selection Techniques

Advances in machine learning have led to the development of automated feature selection tools. There are two types of feature selection techniques in practice today. One type analyzes each feature independently and eliminates features one at a time based on how that feature correlates to the target. Independent feature selection is a simple, straightforward, and fast process. However, it is often the case that a group of features together correlate more strongly with the desired target output. Consequently, the assumption of feature independence can be rather limiting. To address the previously mentioned limitation, feature selection techniques have been proposed where the features are analyzed in groups/subsets. The correlation between the groups of features is considered with relation to the target output. Although the process is computationally intensive, it capitalizes on the important feature interrelationships discovering critical information that is typically lost with the independent feature analysis.

Examples of independent feature selection techniques (also called filter methods) are receiver operating characteristics (ROC) analysis, statistical tests (i.e., *t*-test, Wilcoxon test, χ^2 test), wavelet transforms, information gain, and so on. The grouped feature selection techniques include stepwise feature selection, genetic algorithms (GAs), correlation-based feature selection, and principal component analysis (PCA), to name a few. Feature selection is either pursued as a separate step before decision modeling or as a step conducted in parallel with decision modeling (called wrapper method) **(9)**. Wrappers use the classification algorithm as an integral part of the feature selection process. In other words, the feature selection process is optimized with respect to the particular classifier at hand. The following sections provide a brief description of some feature selection techniques that have been proposed for SELDI-MS proteomic analysis. Independent and grouped feature analysis techniques are presented separately.

Independent Feature Selection

Statistical techniques based on *t*-, *F*, and χ^2 test statistics have been used in feature selection of SELDI-MS data. Liu et al. **(21)** used criteria based on the previously listed statistics for identifying discriminatory peaks from the ovarian cancer data set **(22)**. Normalization of data during preprocessing was required prior to performing *t*-test, whereas the other two techniques were independent of this step. Contrary, χ^2 test statistics were used for selecting the features to be used in the neural network model for the detection of renal cancer **(23)**. Wilcoxon test was used by Sorace et al. **(24)** for the ovarian cancer data set to rank features and to understand the underlying properties of the data. Kozak et al. **(25)** performed *t*-test and Wilcoxon test on the data set comprising ovarian and benign ovarian tumors. ANOVA was used by Wagner et al. **(26)** on a prostate cancer data set. The peaks prevalent

in less than 30 samples were removed before applying feature selection to the remaining peaks. Mann-Whitney test has also been used to rank features ***(27)***.

Similarly, ROC analysis has been exploited by Adam et al. ***(28)*** and Qu et al. ***(29)*** as the criteria to rank the peaks for healthy and benign cases in prostate cancer detection before classification using a DT. In both studies, a predefined threshold of the area under the ROC curve (AUC) was selected as the cutoff to determine significant peaks for further classification. Briefly, AUC is a performance index that summarizes the ROC curve. The area under the ROC curve is always bounded between 0.5 and 1. An area of 0.5 for a peak suggests that it had no discriminatory power and 1 suggests its ability to perfectly differentiate between two groups. In the above studies ***(28,29)***, there were no peaks with perfect AUC, which implied that no peak was independently capable of discriminating perfectly the groups. ROC analysis will be discussed further under performance evaluation.

Mutual information is a feature selection technique that is based on the information content present in the features ***(30)***. Thus far, it has been used in analyzing proteomic mass spectral data for the prediction of lung cancer ***(31)***. If a particular feature is strongly correlated with the target output, it should have high mutual information. The features with low information content are eliminated in this method. Depending on the underlying statistical distribution of the data, mutual information captures general statistical dependencies between each feature and the target output. Therefore, mutual information is better suited to nonlinear decision analysis.

Wavelet transforms, popularly used in medical imaging and general image processing, have been successfully applied by Zhu et al. ***(32)*** to proteomic data before being classified with a recursive tree-based algorithm. Wavelet analysis attempts to decompose the proteomic spectra using multiscale filter banks. After decomposition, large wavelet coefficients were condensed to more significant and informative coefficients that can distinguish between the two groups based on their within-group and between-group sum of squares ***(32)***. Wavelet transform has also been used by Qu et al. ***(33)*** for feature selection of SELDI-MS data prior to constructing a linear discriminant analysis (LDA) function. The objective of Qu's study was to determine if SELDI-MS could be used for early prostate cancer detection.

Grouped Feature Selection

Stepwise analysis is a very popular linear feature selection algorithm based on concepts in linear statistics. As its name suggests, features are reviewed and selected in a step-by-step manner and at each step the variables are evaluated to determine which contributes most in differentiating between the groups. Stepwise analysis compares the means of the different peaks and selects the peaks with the maximum difference in the mean values in the groups. The next peak is added such that the power of the model increases in combination with the first one and this process continues and the result of the "forward stepwise analysis" is a list of peaks in their order of significance. This process can also be done in reverse (backward stepwise elimination). Initially, all variables are included in the model. The variable that least contributes to the classification is eliminated. Thus, only highly contributing variables are retained in the model. Li et al. ***(34)*** used forward and backward stepwise feature selection available in the ProPeak software to rank the peaks for the breast cancer data. Stepwise analysis and Wilcoxon tests were repeatedly applied on the ovarian cancer data in **ref.** ***24*** to select the features and the classification rules were developed by visual inspection and binning based on the *p*-values.

The simplicity and ease in understanding the results of stepwise analysis has made it a

very popular tool in feature selection. However, the stepwise method takes into consideration only the linear relationships between the peaks and does not account for any nonlinear dependencies between them. Thus, stepwise analysis cannot be considered an optimal method if the features selected will be used to develop a nonlinear model in ANN and similar nonlinear decision modeling algorithms. If nonlinear components are present in the data and they are important in the classification problem, then linear feature techniques would be less suited.

In such situations, computationally intensive algorithms like those employed by GAs would be better feature selection tools. GAs, developed by researchers in the field of artificial intelligence (*35*) are based on certain aspects of natural evolution, genetic inheritance, and survival of the fittest. GAs explore all possible subsets to obtain a set of features that will discriminate between the groups of training data. By operating on feature sets in parallel, GAs obtain a global optimal result and are known to be superior in generating robust results for feature selection when presented with high-dimensional data. However, the use of GAs is a computationally complex process, especially when the number of features is very large.

It should be noted that a GA does not always guarantee success. Being a stochastic system, there are situations when a fast convergence occurs, which may halt the process of evolution. There are practical limitations on the number of iterations (generations) of a GA and unlimited number of population size, which may cause the GA to converge to a local optimal solution. These algorithms are nevertheless robust search algorithms that are very powerful in discovering solutions for complex high-dimensional problems where the search space is complex and poorly understood and where traditional methods fail. In general, GAs maintain and evaluate potential solutions, generate better and robust solutions, and thereby improve the quality of decision making. Conrad et al. (*36*) and Petricoin et al. (*37*) used GAs for peak selection and self-organizing maps for classification of SELDI-MS proteomic data in cancer diagnosis.

PCA is an unsupervised feature selection technique used to summarize data in high-dimensional space into a few dimensions. Each dimension or principal component represents a linear combination of the original variables. The first principal component accounts for most of the variability of the data. The next principal component accounts for the variability not accounted for by the first component and so on. Typically, the first few principal components are identified and then the data set is projected onto these components for dimension reduction. There are several applications of PCA for feature selection in MS data (*38–41*). For example, Lilien et al. used a specific algorithm called Q5 in which PCA was used for feature selection and LDA for model development (*38*).

Other than an exhaustive search of every possible feature combination, there is no particular feature selection strategy that guarantees optimal results. Consequently, it is advised to explore a variety of strategies given a particular proteomic data set. For example, Liu et al. (*21*) investigated several feature selection and classification techniques using a publicly available ovarian cancer data set. Specifically, three feature selection techniques (χ^2 method, *t*-test, and correlation analysis) were used in their study. The features selected by these techniques were then fed into diverse classification methods such as support vector machines (SVM), DTs, k-nearest neighbor (KNN), and naïve-Bayesian algorithms. Results with and without the feature selection phase were acquired to determine the importance of the feature selection process. The study clearly demonstrated that feature selection is a critical step but it should always be

considered in relation to the classification step. For this particular study, SVM classification algorithm and KNN reported the best classification accuracy (100%) when used with the peaks selected by correlation-based feature analysis. In contrast, models built by randomly selecting features from the data resulted in poor accuracy.

In another study, Li et al. *(42)* compared the performance of a SVM classifier using the filter approach (a statistical test using a distance measure) and GAs in ovarian cancer detection and diagnosis. The main goal of the study was to demonstrate the feasibility of applying artificial intelligence techniques on serum proteomic data. In this study, GAs emerged as the superior feature selection strategy when used in combination with a SVM classifier.

To summarize, there is no general guideline on which feature selection strategy should be used. It is generally understood that grouped feature selection techniques that are capable of capturing nonlinear relationships among the available features should be used in combination with nonlinear models. Typically, researchers should investigate various techniques to empirically optimize the feature selection process. Once the features are selected, data modeling is the next step.

Classifiers and Data Modeling

Humans and animals acquire the ability to learn and recognize through interaction with the environment. Learning from data has been an area of interest for researchers in statistics and computer science. Machine learning algorithms are able to make inferences about a sample of data through familiarization and repeated interaction with the data. These algorithms vary in their training techniques, final goal, and representation of data.

A learning process would typically comprise of the task of learning and developing rules or functions from the given data set of samples. The development of mathematically accurate rules and functions to describe data is called data modeling. The model developed identifies the properties of the different classes and what separates them to make a correct classification. In the next phase, called testing, the developed model is validated with new data to verify that the model produces accurate results. The learning phase and estimation of the model is implemented and described using different learning methods or algorithms.

A function that describes or approximates the data is of the form f (X, w) where X is the input and w is a parameter of the function. The function f can be linear or nonlinear. Machine learning classifiers that are based on nonlinear decision function include ANN, self-organizing maps (SOMs), and GAs. LDA is an example of a linear classifier.

Machine learning algorithms are of two types—supervised and unsupervised. In supervised learning (also called "learning with a teacher"), prior knowledge is available about the class to which each case (sample) belongs. The training data set comprises of input values and their corresponding output classes (provided by teacher). During the training phase, the training data is used to determine how the features are to be selected, weighed, and combined so as to discriminate between the classes. The testing phase involves application of the weighted features to classify a new test data whose class is not known and which the decision model has not seen before. Thus, the goal of classification methods is to build models of the data set in hand and use that model to classify new samples. The process of learning would involve creating a model so that the predictions of the model are close to the desired target. If the model is able to classify new data correctly we have reason to believe that it is a good model. A wide range of algorithms has been developed for

supervised learning (DTs, SVM, logistic regression, et al.).

In unsupervised learning ("learning without a teacher"), the group to which each sample belongs is unknown or ignored and data is grouped together based on similarity measures. The learning process involves no teacher and the algorithm must identify patterns in the data. Often, unsupervised learning may lead to more than one possible solution. Clustering and Kohonen's SOMs are typical examples of unsupervised learning that have been used in studies analyzing mass spectroscopy data. ANNs, on the other hand, come as supervised and unsupervised learning algorithms.

In both learning techniques, the goal is to predict (classify) or describe the data by developing models of the data, which are then used to classify or describe new cases. If the data has just two or three features, then classifying data would be easy. However, developing models can be a daunting task if there are many features to analyze. High-dimensional data is not only hard to visualize, but all possible combinations should be considered by exhaustive searching techniques during the training phase when the model is developed. A large number of dimensions with very few samples leads to what is often referred to as over-fit or over-trained models. Over-fit models cannot generalize and fail to classify new cases with the desired accuracy.

Supervised Classifiers

LDA is a supervised linear modeling approach and was one of the first to be applied to proteomic mass spectral data. LDA is a noniterative and deterministic classification approach. LDA analysis computes exact solutions, it is simple to implement, and easy to understand. LDA is typically inefficient in classifying high-dimensional data and thus feature selection techniques should be applied before any modeling can be done. LDA captures potential linear relationships in the data and represents it in the form of a linear function. However, nonlinear and complex relationships cannot be fully captured when a linear classifier is used.

LDA has been used in SELDI-MS proteomic studies conducted by Lilien et al. where the LDA was a part of their Q5 algorithm ***(38)***. Wagner et al. ***(26)*** analyzed prostate cancer data using both LDA and QDA (quadratic discriminant analysis). The feature reduction step helped in reducing the number of peaks by 97%. In another study, Qu et al. ***(34)*** analyzed prostate cancer data comprising 45,538 peaks. They developed an LDA model using 12 peaks selected by wavelet transformation and a procedure based on Mahanobis distance.

DT analysis is another popular supervised classification technique ***(9)***. The resulting model is in the form of a hierarchical tree structure. The main advantage of DTs is that they are expressed as a set of rules. As they are visually presented in an easy to interpret form, DTs are very popular in several domains including proteomics. A DT's goal is to identify a set of variables (peaks) that can be used to classify cases or samples into specific groups. A DT performs classification through a systematic process referred to as recursive partitioning ***(9)***. At each node of the tree, a test (presence/absence/intensity of peaks) is applied to one or more variables (peaks) that will have one of two outcomes. The outcome will lead to a split into a leaf node or to another decision node where another test is applied. Peaks are selected for a split based on a cost function, which is the measure of heterogeneity of the descendant nodes. The cost function is often an entropy measure or gain in information. The splitting is made and the tree model is built until there is no gain in information associated with a split or if the there are no more nodes left. Finally, classification of new test data is done by following

the tree rules (i.e., by moving through the tree until a leaf is encountered).

Popular programs for constructing DTs are C4.5 ***(43)*** and classification and regression tree (CART) ***(44)***. The C4.5 algorithm was used by Won et al. ***(37)*** to classify the data set of 36 samples and 119 peaks into renal cell carcinoma (RCC) and healthy patients. The 119 peaks were identified by preprocessing using the Ciphergen Protein Chip software. The DT identified five discriminatory peaks for model development. A CART model was developed by Markey et al. ***(46)*** to classify 41 serum samples with 21 peaks into lung cancer and controls. The peaks used in this model were identified by mass spectrometry. Adam et al. ***(28)*** used CART to classify serum proteomic patterns identified by Ciphergen System software (i.e., the Biomarker Wizard), into prostate cancer, benign prostate hyperplasia (BPH), and controls. Three hundred and twenty-six serum samples from 167 PCA patients, 77 BPH patients, and 82 healthy men were used in this particular study. More than 60,000 peaks in the range of 2000–40,000 Da were selected by the SELDI software and was reduced to 772 by preprocessing and then to 124 by peak selection using ROC analysis. Finally, the DT selected nine peaks to develop the model. Biomarker pattern software (BPS) based on the CART decision model was used by Zhang et al. ***(47)*** to classify 156 urine samples into transitional cell carcinoma, benign urogenital diseases, and controls. The process started with peak detection phase using the Biomarker Wizard software. A signal-to-noise ratio of 5 was used to filter the peaks that were in the range 2000 to 50,000 Da. Four hundred peaks produced by peak detection were used in DT construction. The training algorithm identified five discriminatory peaks and developed the tree model. BPS was also used by Kang ***(48)*** to identify serum biomarkers that distinguish between severe acute respiratory syndrome (SARS) and non-SARS samples.

Overall, DTs are easy to interpret and can be represented in the form of if–then rules. The training and testing times are reasonable when compared with neural networks and they can handle large number of features. Despite these advantages, they are not recommended when there are large amounts of missing data and in problems involving complex data distributions. The instability associated with DTs causes small changes in data to result in entirely different decision rules. Therefore, interpreting DT results should be exercised with caution.

Thus far, the discussion was on supervised linear classifiers. Another popular classifier that has been used in SELDI-MS protein data analysis is the ANN ***(49)***. ANNs are powerful nonlinear classifiers, based on artificial intelligence principles. Inspired by the human brain, they comprise of highly connected and nonlinear units referred to as neurons that are connected by weights. The interunit connection weights stores the processing ability of the neural network. This is obtained by the process of learning from the training data set. There are two different types of neural networks—supervised neural networks and unsupervised neural networks.

By far, the most popular ANN architecture in medical decision making has been the traditional backpropagation neural network, introduced by Rumelhart et al. ***(50)***. A simple representation of a neural network is as a set of interconnected layer of neurons with weighted interconnections reflecting various influences of each neuron on the others (**Fig. 3** illustrates a simplified neural network model). Training is performed by using a data set with known inputs and outputs. In the backpropagation training algorithm, the neural network output is compared with its expected output for each training example. Computation starts at the output layer and propagates backward to the hidden layer(s). The difference between the desired output and actual output (error) is calculated and propagated back through the

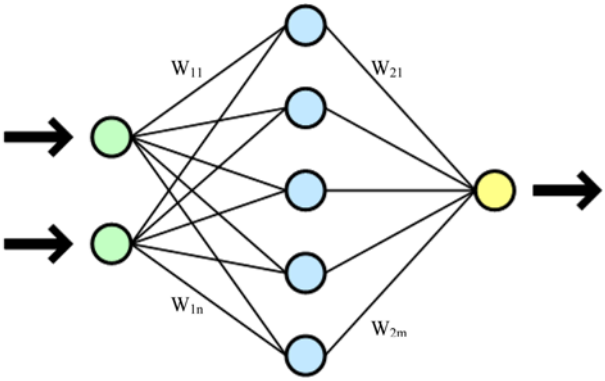

Fig. 3. A schematic representation of an artificial neural network.

network. The neural network weights are adjusted in an iterative manner until the network achieves a predefined threshold of minimum squared error for the whole training set.

As with all supervised classifiers, development of an ANN involves both the training and validation phases of the network. During the learning phase, the data is presented repeatedly to the ANN and it learns from that by storing the decision rules as weights. Following the training phase is the validation phase where new cases are presented for classification. Rogers et al. *(23)* used 48 data samples from patients with RCC, 38 controls, and 20 benign cases to train a feed-forward neural network model. They reported sensitivity and specificity values in the range 98.3 to 100% during training and 81.8 to 83.3% during testing. However, further testing was performed later to assess the robustness of the model with a group of 80 samples to get sensitivity and specificity in the 41 to 76.6% range, which called for further evaluation of contributing factors. SELDI-MS coupled with ANN was used by Mian et al. *(51)* to identify protein patterns associated with either control/drug treatment for chemotherapy-sensitive cells or response/nonresponse for chemotherapy-resistant breast cancer cells, respectively for two popular chemotherapeutic drugs. The model developed was able to classify the test data as resistant or as sensitive to a particular drug and discriminate between chemotherapy-sensitive and -resistant breast cancer cells with high accuracy. Ball et al. *(52)* used multilayered perception ANN with a backpropagation algorithm to mine the SELDI-analyzed data of tumors and controls. Poon et al. *(53)* used neural networks to discriminate hepatocellular carcinoma (HCC) from chronic liver disease (CLD). Two hundred and fifty significant peaks identified by significance analysis of microarrays software were used to develop a feed-forward backpropagation ANN model. ROC curve analysis showed that the ANN was useful in differentiating HCC and CLD cases regardless of serum concentrations. They reported a sensitivity and specificity for HCC and CLD cases as approx 95% and approx 90%, respectively.

However, ANNs are not devoid of limitations. The main criticism regarding ANNs is their "black box" nature. It is very difficult to determine how the neural network makes the decisions in classifying data. An additional disadvantage is their low computational efficiency during the training phase. The optimal theoretical size of data required for training and ANN is almost never met in medical research. Small sample size often leads to overfitting. Furthermore, the architecture of the ANN for getting the best results is subjective. Changes in the ANN architecture, training parameters, and starting point of training will result in different models and solutions. The training algorithms often get trapped in the local minima resulting in suboptimal results. Despite all the issues surrounding the architecture and training of ANNs, this group of classifiers is known to produce generalized results and often report better accuracy than traditional statistical techniques in medical diagnosis. In addition, neural networks can handle problems with large amounts of diverse features and are known to perform well with complex data distributions.

Unsupervised Classifiers

Clustering is an unsupervised learning technique (*9*), wherein data samples are grouped into clusters, based on a measure of association so that similar groups are in one cluster. Input to cluster analysis is a set of samples and a measure of association (similarity or dissimilarity between two samples). The output is a group of clusters and a generalized description of each cluster. Hierarchical clustering is a popular clustering algorithm (*9*) that has been used in several SELDI studies. The data is grouped into a sequence of clusters in a bottom-up manner and the result is displayed in the form of a dendrogram or a tree structure. The number of clusters in the data is not prespecified in the problem. Agglomerative hierarchical clustering is a type of hierarchical clustering algorithm where each sample is regarded as a cluster. In the agglomerative clustering technique, pairs of clusters closer to each other are merged based on their similarity measure and this is repeated until the entire data is in one cluster.

Poon et al. (*53*) used two-way hierarchical clustering to differentiate HCC from CLD. The result of the study reported was that the algorithm separated HCC and CLD cases into two major clusters. However, there was no mention of the model evaluation. Purohit et al. (*38*) performed hierarchical cluster analysis for classifying data into diseased and healthy. The PCA feature selection method was first applied on the data. This combination of PCA and cluster analysis could correctly identify 68% of the patients and 100% of the healthy group. They indicated the possibility of the diseased people to be in a different disease stage.

SOM is a nonlinear clustering technique and was first introduced by Kohonen (*54*). Similar in design to the human brain, an SOM consists of interconnected neurons and has two layers; the input layer, which contains the raw data and the output layer, which comprises of the clusters. The process of learning involves successive passes through the input layer until no new information is obtained on subsequent passes. The result of this process is a set of clusters that best represent the differences in the data. SOMs follow unsupervised learning and depend on the self-organizing nature of the data for the formation of best set of clusters.

SOMs have been used in the Petricoin et al. (*22*) study in combination with GAs to identify ovarian cancer biomarkers and to classify the samples into healthy, benign, and cancer cases. The clustering algorithm was trained on 50 healthy women and 50 women in different stages of ovarian cancer and tested on 116 new cases, yielding a sensitivity of 100% and specificity of 95%.

In samples with categorical data, KNN may be used to compute the similarity between samples and clusters. This technique is based on the distances from immediate neighbors. The class of an input pattern is chosen as the class of the majority of its KNN. Euclidean and Mahanobis distances are often used as the distance metrics (*9*).

There are several clustering algorithms available in the literature and could easily confound a new researcher trying to get a suitable algorithm. All clustering algorithms will produce clusters in a given data irrespective of whether or not they exist. Therefore, cluster algorithms should only be applied on a data set that is expected to contain clusters. Once clusters are formed out of the data, they should be evaluated using existing validation techniques (external and internal criterion analysis) for cluster validation. However, existing validation techniques are subjective in nature and data analysts should keep in mind that there is no best clustering algorithm. Therefore, empirical study is advised by exploring several algorithms on a given data set.

Combining Classifiers

Although most of the recent proteomic studies have been performed using individual classifiers, the approach of combining classifiers is a promising strategy to reduce the classification prediction error and to improve decision-making performance. The final classifier is a set of base classifiers, each of which makes its own classification and they are combined to constitute the single classification result of the entire classifier.

Breiman (*55*) refers to classifiers as unstable if a small change in data causes large changes in classification and they often have high bias and low variance. The opposite is true for stable classifiers. DTs and neural nets are considered to be unstable classifiers, whereas LDA and KNN are considered to be stable. If classifiers are combined, the variance and error are usually lowered. Bagging and boosting are two such techniques that have been used for combining classifiers in data mining and they have also been applied in proteomic data analysis.

Boosting of weak classifiers is performed by taking a weighted vote of each classifier. Boosting applies the training algorithm sequentially to the training sample and the sample is reweighed to give importance to cases that are not correctly predicted. A weighted majority vote of the classifiers is taken for the decision making. This method often results in a stronger classifier and is very effective in DTs, which are considered to be unstable classifiers, as it boosts the performance of existing weak learners, reduces the test error and variance, avoids over-fitting, and thus leads to an improvement of performance. For example, a boosted DT was used in the early detection of prostate cancer in the studies conducted by Qu et al. (*29*). The study aimed at developing a classifier to discriminate men with prostate cancer from those with BPH and controls.

Bagging is another technique that has been used to combine predictors and was first used by Breiman in data mining applications. Many replicates of the available data set having the same size as the original data set are drawn with replacement from the training sample (called bootstrap samples). The learning algorithm is then applied to each bootstrap sample and validation is then performed on the samples where training is not done. The final classification is completed by taking a majority vote where each classifier has equal weight in voting. When the individual results from each classifier vary considerably from one another, taking an average of the results will result in a more accurate prediction. However, if the results are similar, averaging the results may not result in a better performance. Thus, the advantage of bagging is clearly when the performance of individual classifiers is uncorrelated. Izmirlan (*56*) used bagging in a proteomic random forest classification algorithm where DTs were bagged using 632 cross-validation and randomly selected feature sets.

The result of combining classifiers often improves performance, especially when combining classifiers that do not think alike and have low correlation. Using a combined approach, individual classifiers complement each other by capturing information that individually they miss.

Sampling

One of the major challenges faced in the application of machine learning algorithms to medical data is the validation of a trained model with new test data. The process of decision modeling requires that the model be developed by training on a certain set of the data (training set), which is followed by the validation of the model on another set of data not previously seen during training (testing set). The obvious way to handle this issue is by dividing the data into train and test sets before model building by stratified random sampling.

Stratified random sampling was used in **refs.** *28* and *45*. These studies report the performance of the decision models based on a single train-test data split.

However, medical data is quite often very difficult and expensive to acquire. Thus, there are not enough available cases to divide into train-test subsets. Furthermore, the noise inherent in most medical data and the complex relationships between features requires a sample size large enough to efficiently model the data with accuracy. In addition, the size of the test set controls the statistical power and confidence in the developed decision model. Consequently, sophisticated sampling strategies are required to capitalize on the available data (*57*). The following is a brief description of data sampling strategies recommended in medical decision making.

Cross-Validation

Cross-validation is by far the most popular data sampling strategy. The data is randomly divided into two sets. The decision model is trained on the first and tested on the second. This random splitting process is repeated several times to reduce the selection bias. The average of all the test estimates gives the average error of the model. If the data set used for training is too small, the model may not be able to predict test cases well. A small test set may not result in an accurately validated classifier and can have a large error rate. Consequently, different train–test ratios (e.g., 50–50%, 75–25%, etc.) are explored with cross-validation.

A common implementation is the k-fold cross-validation. The data is partitioned into k-disjoint sets. Training of the classifier is done on the k-1 sets and testing on remaining one data set. This is done for all the k-subsets resulting in k-models and the estimated error will be the average of the k-error rates. For example, 10-fold cross-validation divides the data into 10 groups. Nine groups are used for training and testing is done on the left out-group. This is repeated 10 times until each of the 10 groups has served as the test group. The average test error of the 10 groups is the final test error estimate and gives an approximate idea of the quality of the model for the classification of the data. K-fold cross-validation requires the careful random stratification of the 10 groups. Ten-fold cross-validation was used in **ref.** *57* to discriminate between early-stage melanoma patients with and without melanoma recurrence. The sensitivity and specificity were 72 and 75%, respectively. Wagner et al. (*26*) stressed the significance of cross-validation with various decision models. Their study showed results based on a 90- to 10%-data split, randomly repeated 100 times. Zhang et al. (*47*) employed the same data sampling strategy.

A noteworthy special case is the popular leave-one-out cross-validation approach, a special case of the k-fold cross-validation implementation. If the data comprises of k-samples, the data model is trained using the k-1 samples and tested on the remaining one sample. This is repeated k times until every sample has served as a test case. The average error on the k samples is the estimated test error. The main advantage of the technique is that the decision model uses almost all available cases for training without compromising the statistical significance of the testing phase. However, the technique can be time consuming with large sample sizes and elaborate decision models. Therefore, it is used predominantly in resampling of small data sets.

There are several published examples of studies using variations of the cross-validation sampling scheme *(see* **refs.** *32*,*38*,*39*,*42*,*46*,*59*). In a comparative study, Zhu et al. (*32*) used both leave-one-out and k-fold cross-validation to assess the validity of their DT model and achieved an error rate of 14.5 and 20% for leave-one-out and k-fold cross-validation, respectively showing that leave-one-out has a lower bias than the latter.

Bootstrap

The bootstrap method involves the generation of a training set by sampling with replacement. This is done *n* times for a data set of n cases. The data is trained on the bootstrap set and tested on the original set. It is recommended that the process is repeated many times (>200). The final performance estimate is the average of all bootstrap estimates. This is a computationally intensive process even for small size data sets. Mian et al. *(51)* used bootstrap validation with 86 independent ANN models. The variation in the model's performance, accuracy of classifying test data, and identification of potential outliers were assessed through the bootstrap resampling.

Performance Assessment of Models

The last phase of the data mining process is the assessment of the models developed by the previously described machine learning algorithms. The different methods for evaluating the models' ability to classify new test cases are discussed below.

Accuracy

Accuracy of classification is calculated by taking the ratio of the number of correctly classified samples to the total number of samples in the test data. However, when the prevalence of a particular class is higher than the other class, the majority class will cause a bias in the result. In such a scenario, the accuracy measure can be misleading. SELDI-MS-based proteomic studies that have used accuracy to report results include those in **refs.** *25*,*31*,*39*,*42*,*51*, and *52*.

Sensitivity/Specificity

In two-class samples, there are four possible outcomes when the decision model is tested. They are true-positive, true-negative, false-positive, and false-negative results. Sensitivity (true-positive rate) is the ratio of the number of correctly classified positive samples over the total number of positive samples. High sensitivity is much desired in medical diagnosis where the impact of wrongly predicting a diseased person as healthy is high. False-positive rate is the probability that a healthy subject is wrongly classified as diseased (referred to as specificity). High specificity is desirable where a false alarm would result in unwanted elaborate tests and treatments. Ideally, for perfect classification, both sensitivity and specificity should be 1 (100%). The clinically acceptable sensitivity and specificity depends on the application.

Several studies have reported their results using sensitivity and specificity as the performance indices (*see* **refs.** *23*,*24*,*28*,*33*,*45–47*). The main limitation of using sensitivity and specificity as the only evaluation indices is their dependence on class prevalence and the decision threshold. Therefore, it is difficult to directly compare the results of studies that are reported using only sensitivity and specificity measures.

ROC Analysis

Based on the classical signal detection theory, ROC analysis reports the sensitivity and specificity of a decision model for all possible decision thresholds. Conventionally, a ROC curve plots the true-positive fraction (or sensitivity) vs the false-positive fraction (or [1-specificity]) for a wide and continuous range of decision thresholds and provides a more meaningful and valid measure of classification performance. Furthermore, ROC curves can be used to determine optimum decision levels that maximize accuracy, average benefit, or other measures of clinical efficacy. The AUC gives a complete picture of the performance of a model and can be used to compare the performance of multiple models. The more the curve is shifted to the upper left corner of the graph (specificity = sensitivity = 1), the better the diagnostic performance. The area index varies between 0.5 (representing chance behavior) and

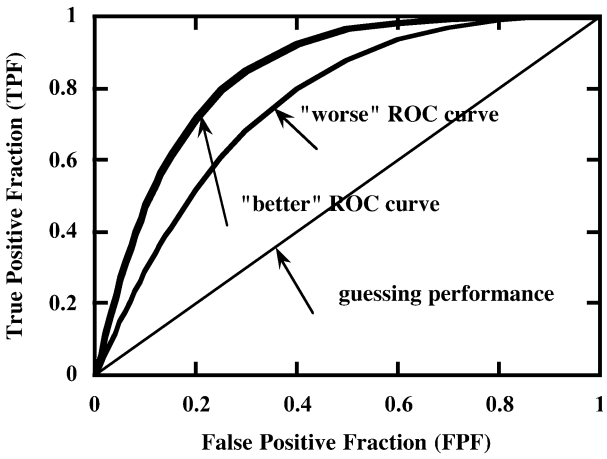

Fig. 4. A typical ROC curve. The closer the line is to the upper left had corner of the graph, the performance of the test is considered to be best with the least false-positives or -negatives.

1.0 (representing perfect performance). Generally, a higher value of the area index indicates better performance. More importantly, ROC analysis is threshold and prevalence independent. **Figure 4** shows typical ROC curves.

The usage of ROC in data mining is still below its full potential. It is highly recommended that researchers embrace this technique, since comparison of results across different studies is straightforward. ROC analysis has been utilized in several SELDI proteomic studies ***(25,28,34,42,46)***.

Conclusions

Data mining is a data-driven process where the results obtained largely depend on the data being analyzed. The methods employed for feature selection, classification, data sampling, and performance evaluation drive the process and alter final results. Thus, it is recommended to explore more than one technique to make comparisons and better understand the problem in hand. The promise and opportunities of combination of data mining approaches and the generated proteomic mass spectrometry data for discovery of novel biomarkers with diagnostic value is obvious. However, caution should be exercised in application applying various data mining techniques in this regard. This review made an effort to reinstate the critical issues and limitation of data mining applications to be addressed by researchers analyzing SELDI-MS proteomic data for extracting clinically useful information.

Acknowledgments

Funding support for AT was in part by a grant from NIH-NCRR 5 P20 16480 (principal investigator: Nigel Cooper).

References

1 Wilkins, M. R., Sanchez, J. C., Gooley, A. A., et al. (1996) Progress with genome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* **13,** 19–50.

2 Coombes, K. (2005) Analysis of mass spectrometry profiles of the serum proteome. *Clin. Chem.* **51,** 1–2.

3 Rodland, K. D. (2004) Proteomics and cancer diagnosis. *Clin. Bioch.* **37,** 579–583.

4 Liotta, L. A., Ardekani, A. M., Hitt, B. H., et al. (2003) General keynote: proteomic patterns in sera serve as biomarkers of ovarian cancer. *Gynecol. Oncol.* **88,** S25–S28.

5 Conrads, T. P., Fusaro, V. A., Ross, S., et al. (2004) High-resolution serum proteomic features for ovarian cancer detection. *Endocr. Relat. Cancer* **11,** 163–178.

6 Yip, T. C., Chan, J. W., Cho, W. C., et al. (2005) Protein chip array profiling analysis in patients with severe acute respiratory syndrome identified serum amyloid: a protein as a biomarker potentially useful in monitoring the extent of pneumonia. *Clin. Chem.* **51,** 47–55.

7 Coombes, K. R., Koomen, J. M., Baggerly, K. A., Morris, J. S., and Kobayashi, R. (2005) Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics* **1,** 41–52.

8 Hong, H., Dragan, Y., Epstein, J., et al. (2005) Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS). *BMC Bioinformatics* **15,** S5.

9 Kantardzic, M. (2002) *Data Mining: Concepts, Methods, Models, and Algorithms*. Wiley and IEEE Press, New York.

10 Wilson, R. L. and Sharda, R. (1994) Bankruptcy prediction using neural networks. *Decision Support Systems* **11,** 545–557.

11 Barr, D. S. and Mani, G. (1994) Using Neural Nets to manage investments. *AI Expert* 1994; 16–21.

12 Sung, T. K., Chang, N., and Lee, G. (1999) Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. *J. Manag. Info. Sys.* **1,** 63–85.

13 Shaw, M. J., Subramaniam, G., Tan, G. W., and Welge, M. E. (2001) Knowledge management and data mining for marketing. *Dec. Supp. Sys.* **31,** 127–137.

14 Daskalaki, S., Kopanas, I., Goudara, M., and Avouris, N. (2003) Data mining for decision support on customer insolvency in telecommunications business. *Eur. J. Oper. Res.* **145,** 239–255.

15 Haa, S. H., Baeb, S. M., and Parkb, S. C. (2002) Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. *Comp. and Indus. Eng.* **43,** 801–820.

16 Caskey, K. R. (2001) A manufacturing problem solving environment combining evaluation, search, and generalisation methods. *Computers in Industry* **44,** 175–187.

17 Kusiak, A., Dixon, B., and Shaha, S. (2005) Predicting survival time for kidney dialysis patients: a data mining approach. *Comp. Biol. Med.* **35,** 311–327.

18 Chen, W. H., Hsu, S. H., and Shen, H. P. (2005) Application of SVM and ANN for intrusion detection. *Comp. and Oper. Res.* **32,** 2617–2634.

19 Seifert, J. W. (2004) Data mining and the search for security: challenges connecting the dots and databases. *Government Information Quarterly* **21,** 461–480.

20 Barrera, J., Cesar, R. M., Ferreira, J. E., and Gubitoso, M. D. (2004) An environment for knowledge discovery in biology. *Comp. Biol. Med.* **34,** 427–447.

21 Liu, H., Li, J., and Wong, L. (2002) A comparative study on feature selection and classification methods using gene expression profiles. *Genome Informatics* **13,** 51–60.

22 Petricoin, M. F., Ardekani, A. M., Hitt, B. A., et al. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577.

23 Rogers, M. A., Clarke, P., Noble, J., et al. (2003) Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. *Cancer Res.* **63,** 6971–6983.

24 Sorace, J. M. and Zhan, M. (2003) A data review and reassessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **4,** 24.

25 Kozak, K. R., Amneus, M. W., Pusey, S. M., et al. (2003) Identification of biomarkers for ovarian cancer using strong anion-exchange proteinchips: potential use in diagnosis and prognosis. *PNAS* **100,** 14,666–14,671.

26 Wagner, M., Naik, D. N., Pothn, A., et al. (2004) Computational protein biomarker prediction: a case study for prostate cancer. *BMC Bioinformatics* **5,** 26.

27 Zhukov, T. A., Johnson, R. A., Cantor, A. B., Clark, R. A., and Tockman, M. S. (2003) Discovery of distinct protein profiles specific for lung tumors and pre-malignat lung lesions by SELDI mass spectrometry. *Lung Cancer* **40,** 267–279.

28 Adam, B. L., Qu, Y., Davis, J. W., et al. (2002) Serum protein finger printing coupled with a pattern-matching algorithm distinguishes prostate cancer from benign hyperplasia and healthy men. *Cancer Research* **62,** 3609–3614.

29 Qu, Y., Adam, B. L., Yasui, Y., et al. (2002) Boosted decision tree analysis of surface enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.* **48,** 1835–1843.

30 Tourassi, G. D., Frederick, E. D., Markey, M. M., and Floyd, C. E. (2001) Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Med. Phys.* **28,** 2394–2402.

31 Hilario, M., Kalousis, A., Müller, M., and Pellegrini, C. (2003) Machine learning approaches to lung cancer: prediction from mass spectra. *Proteomics* **3,** 1716–1719.

32 Zhu, H., Yu, C. Y., and Zhang, H. (2003) Tree based disease classification using protein data. *Proteomics* **3,** 1673–1677.

33 Qu, Y., Adam, B. L., Thornquist, M., Potter, J. D., Thompson, M. L., and Yasui, Y. (2003) Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics* **59,** 143–151.

34 Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. A., and Chan, D. W. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin. Chem.* **47,** 1296–1304.

35 Holland, J. H. (1994) *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications of Biology, Control and Artificial Intelligence, 3rd ed.* MIT Press, Cambridge, MA.

36 Conrads, T. P., Zhou, M., Petricoin, E. F., Liotta, L., and Veenstra, T. D. (2003) Cancer diagnosis using proteomic patterns. *Expert Rev. Mol. Diagn.* **3,** 411–420.

37 Petricoin, E. F. and Liotta, L. A. (2004) SELDI-TOF based proteomic pattern diagnostics for early detection of cancer. *Curr. Opin. Biotech.* **15,** 24–30.

38 Lilien, R. H., Farid, H., and Donald, B. R. (2003) Probabilisitic disease classification of expression—dependent proteomic data from mass spectrometry of human serum. *J. Comp. Biol.* **10,** 925–946.

39 Purohit, P. V. and Rocke, D. M. (2003) Discriminant models for high-throughput proteomics mass spectrometer data. *Proteomics* **3,** 1699–1703.

40 Slotta, D. J., Heath, L. S., Ramakrishnan, N., Helm, R., and Potts, M. (2003) Clustering mass spectrometry data using order statistics. *Proteomics* **3,** 1687–1691.

41 Coombes, K. R., Fritsche, H. A., Clarke, C., et al. (2003) Quality control and peak finding from nipple aspirate fluid by surface enhanced laser desorption and ionization. *Clin. Chem.* **49,** 1615–1623.

42 Li, L., Tang, H., Wu, Z., et al. (2004) Data mining techniques for cancer detection using serum proteomic profiling. *Artif. Intel. Med.* **32,** 71–83.

43 Quinlan, J. R. (1986) Introduction of decision trees. *Machine Learning* **1,** 81–106.

44 Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) Classification *and Regression Trees.* Wadsworth International Group, Belmont, CA.

45 Won, Y., Song, H. J., Kang, T. W., Kim, J. J., Han, B. D., and Lee, S. W. (2003) Pattern analysis of serum proteome distinguished renal cell carcinoma from other urologic diseases and healthy persons. *Proteomics* **3,** 2310–2316.

46 Markey, M. K., Tourassi, G. D., and Floyd, C. E., Jr. Decision Tree classification of proteins identified by mass spectrometry of blood samples from people with and without lung cancer. *Proteomics* **3,** 1678–1679.

47 Zhang, Y. F., Wu, D. L., Liu, W. W., et al. (2004) Tree analysis of mass spectral urine profiles discriminates transitional cell carcinoma of the bladder from non cancer patient. *Clin. Biochem.* **37,** 772–779.

48 Kang, X., Xu, Y., Wu, X., et al. (2005) Proteomic fingerprints for potential application to early diagnosis of severe acute respiratory syndrome. *Clin. Chem.* **51,** 56–64.

49 Bishop, C. M. (1995) *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, UK .

50 Rumelhart, D., Hinton, G., and Williams, R. (1988) Learning internal representations by error propagation. In: *Neurocomputing,* (Anderson, J. and Rosenfeld, E.), MIT Press, Cambridge, MA, pp. 675–695.

51 Mian, S., Ball, G., Hornbuckle, J., et al. (2003) A prototype methodology combining surface enhanced laser desorption/ionization protein chip technology and artificial neural network algorithms to predict the chemoresponsiveness of breast cancer cell lines exposed to Paclitaxel and Doxorubicin under in vitro condition. *Proteomics* **3,** 1725–1737.

52 Ball, G., Mian, S., Allibone, R. O., et al. (2002) An integrated approach using artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers. *Bioinformatics* **18,** 395–404.

53 Poon, T. C. W., Yip, T., Chan, A. T. C., Yip, C., Yip, V., and Mok, T. S. K. (2003) Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. *Clin. Chem.* **49,** 752–760.

54 Kohonen, T. (1995) *Self Organizing Maps.* Springer Publishers, Berlin, Germany.

55 Breiman, L. (1996) Bagging predictors. *Machine Learning* **24,** 123–140.

56 Izmirilan, G. (2004) Application of random forest classification algorithm to a SELDI-TOF Proteomics study in the setting of a cancer prevention trial. *Ann. NY Acad. Sci.* **1020,** 154–174.

57 Tourassi, G. D. and Floyd, C. E. (1997) The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. *Med. Dec. Mak.* **17,** 186–192.

58 Wilson, L. L., Tran, L., Morton, D. L., and Hoon, D. S. B. (2004) Detection of Differentially expressed proteins in early-stage melanoma patients using SELDI-TOF mass spectrometry. *Ann. NY Acad. Sci.* **1022,** 317–322.

59 Tatay, J. W., Feng, X., Sobczak, N., et al. (2003) Multiple approaches to data mining of proteomic data based on statistical and pattern classification methods. *Proteomics* **3,** 1704–1709.