Chapter 30

Feature Extraction for Classification of Proteomic Mass Spectra: A Comparative Study

Ilya Levner, Vadim Bulitko, and Guohui Lin

University of Alberta
Department of Computing Science
Edmonton, Alberta, T6G 2E8, CANADA
ilya@cs.ualberta.ca, bulitko@cs.ualberta.ca, ghlin@cs.ualberta.ca

Summary. To satisfy the ever growing need for effective screening and diagnostic tests, medical practitioners have turned their attention to high resolution, high throughput methods. One approach is to use mass spectrometry based methods for disease diagnosis. Effective diagnosis is achieved by classifying the mass spectra as belonging to healthy or diseased individuals. Unfortunately, the high resolution mass spectrometry data contains a large degree of noisy, redundant and irrelevant information, making accurate classification difficult. To overcome these obstacles, feature extraction methods are used to select or create small sets of relevant features. This paper compares existing feature selection methods to a novel wrapper-based feature selection and centroid-based classification method. A key contribution is the exposition of different feature extraction techniques, which encompass dimensionality reduction and feature selection methods. The experiments, on two cancer data sets, indicate that feature selection algorithms tend to both reduce data dimensionality and increase classification accuracy, while the dimensionality reduction techniques sacrifice performance as a result of lowering the number of features. In order to evaluate the dimensionality reduction and feature selection techniques, we use a simple classifier, thereby making the approach tractable. In relation to previous research, the proposed algorithm is very competitive in terms of (i) classification accuracy, (ii) size of feature sets, (iii) usage of computational resources during both training and classification phases.
Keywords: feature extraction, classification, mining bio-medical data, mass spectrometry, dimensionality reduction.

30.1 Introduction

Early detection of diseases, such as cancer, is critical for improving patient survival rates and medical care. To satisfy the ever growing need for effective screening and diagnostic tests, medical practitioners have turned their attention to mass spectrometry based methods. While other proteomic methods

I. Levner et al.: *Feature Extraction for Classification of Proteomic Mass Spectra: A Comparative Study*, StudFuzz **207**, 607–624 (2006)
www.springerlink.com

exist, such as PAGE*, mass spectrometry (**MS**) based approaches provide high throughput, are widely applicable, and have the potential to be highly accurate. This paper examines supervised classification in **proteomic** applications. The term proteomics will be restricted to mean the study of protein spectra, acquired by mass spectrometry techniques, to classify disease and identify potentially useful protein biomarkers. A **biomarker** is an identified protein(s) whose abundance is correlated with the state of a particular disease or condition. Currently, single biomarkers, such as PSA† used to detect prostate cancer, are relied on for disease screening and diagnosis. The identification of each biomarker, tailored for a specific disease, is a time consuming, costly and tedious process. In addition, for many diseases it is suspected that no single biomarkers exit, which are capable of producing reliable diagnoses. The following quote further motivates the use of high resolution MS techniques:

> *"The ability to distinguish sera from an unaffected individual or an individual with [for example] ovarian cancer based upon a single serum proteomic m/z feature alone is* ***not possible*** *across the entire serum study set. Accurate histological distinction is only possible when the key m/z features and their intensities are considered en masse. A limitation of individual cancer biomarkers is the lack of sensitivity and specificity when applied to large heterogeneous populations."* (Conrads et al., 2003)

While high-resolution mass spectrometry techniques are thought to have potential for accurate diagnosis due to the vast amount of information captured, they are problematic for supervised training of classifiers. Specifically, the many thousands of raw attributes forming the spectra frequently contain a large amount of redundancy, information irrelevant to a particular disease, and measurement noise. Therefore, aggressive feature extraction techniques are crucial for learning high-accuracy classifiers and realizing the full potential of mass spectrometry based disease diagnosis.

The rest of the paper is organized as follows. We first motivate the task by presenting two important disease diagnosis problems and recent studies on them. A novel combination of feature selection and classification methods is subsequently proposed and empirically evaluated on ovarian and prostate cancer data sets. The paper is concluded with discussion and future research directions.

30.1.1 Ovarian Cancer Studies

In (Petricoin et al., 2002a), genetic algorithms together with self-organizing maps were used to distinguish between healthy women and those afflicted

*The acronym PAGE stands for polyacrylamide gel electrophoresis. It is also known as 2DE for two dimensional polyacrylamide gel electrophoresis (Patterson and Aebersold, 2003).

†PSA stands for prostate specific antigen.

with ovarian cancer. Although cross-validation studies were not conducted, the approach was able to correctly classify all cancer stricken patients and 95% of healthy women, on a single test set. Motivated by the need for greater recall and precision, in (Conrads et al., 2003), a low resolution mass spectrometry technique was compared with a high resolution technique using the same ovarian cancer data set. The goal was to determine whether sensitivity and PPV[‡] (i.e., recall and precision) scores would improve by using a higher resolution spectra provided by the SELDI TOF MS hardware[§]. Keeping all other parameters fixed (including the machine learning algorithm), classification based on high resolution data achieved 100% specificity and PPV scores on the ovarian cancer data set. In contrast, none of the models based on the low resolution mass spectra could achieve perfect precision and recall scores. The researchers, therefore, concluded that the 60-fold increase in resolution improved the performance of the pattern recognition method used. Due to the low prevalence of (ovarian) cancer (Kainz, 1996), a screen test would require a 99.6% specificity to achieve a clinically acceptable positive predictive value of 10%. As a result, high resolution mass spectrometry techniques have been adopted to increase classification accuracy.

Unfortunately, increasing data resolution proliferates "the curse of dimensionality", and thereby decreases the applicability of supervised classification techniques. As a result, **feature extraction** is needed to extract/select salient features in order to make classification feasible. In addition to making machine learning algorithms tractable, feature extraction can help identify the set(s) of proteins (i.e., features) that can be used as potential biomarkers. In turn, key protein identification can shed light on the nature of the disease and help develop clinical diagnostic tests and treatments.

Using the same data set, in (Lilien et al., 2003) the researchers used Principle Component Analysis (PCA) (Kirby, 2001) for dimensionality reduction and Linear Discriminant Analysis (LDA) for classification. For each of the various train/test data splits, 1000 cross-validation runs with re-sampling were conducted. When training sets were larger than 75% of the total sample size, perfect (100%) accuracy was achieved. Using only 50% of data for training, the performance dropped by 0.01%. We conclude that PCA appears to be an effective way to reduce data dimensionality.

In (Wu et al., 2003), the researchers compared two feature extraction algorithms together with several classification approaches. The T-statistic[¶] was used to rank features in terms of relevance. Then two feature subsets were greedily selected (respectively having 15 and 25 features each). Support vector machines (SVM), random forests, LDA, Quadratic Discriminant Analysis, k-nearest neighbors, and bagged/boosted decision trees were subsequently

[‡]PPV stands for Positive Predictive Value, see glossary for details.

[§]SELDI TOF MS stands for surface-enhanced laser desorption/ionization time-of-flight mass spectrometry.

[¶]The T-statistic is also known as the student-t test (Press et al., 2002).

used to classify the data. In addition, random forests were also used to select relevant features with previously mentioned algorithms used for classification. Again 15 and 25 feature sets were selected and classification algorithms applied. When the T-statistic was used as a feature extraction technique, SVM, LDA and random forests classifiers obtained the top three results (accuracy appears to be about 85%). On the other hand, classification accuracy improved to approximately 92% when random forests were used for both feature extraction and classification. Similar performance was also achieved using 1-nearest-neighbor.

The data from (Wu et al., 2003), was subsequently analyzed in (Tibshirani et al., 2004), using the nearest shrunken centroid algorithm. The ten-fold cross validated specificity was 74% with a corresponding sensitivity of 71%. Thus the balanced accuracy (BACC) of this algorithm was 72.5%. Although the accuracy of this algorithm is less than that of other methods presented, this approach used only seven features[‖] out of 91360.

30.1.2 Prostate Cancer Studies

In (Adam et al., 2002), the researchers used a decision tree algorithm to differentiate between healthy individuals and those with prostate cancer. This study also used the SELDI TOF MS to acquire the mass spectra. Receiver Operating Characteristics (ROC) curves were used to identify informative peaks which were subsequently used by the decision tree classification algorithm. The researchers did not perform cross-validation, but on a single test set the classifier achieved an 81% sensitivity and a 97% specificity, yielding a balanced accuracy (BACC) of 89%.

In (Qu et al., 2002), the performance was improved from (Adam et al., 2002) by using ROC curves to identify relevant features. For classification, the researchers used decision trees together with AdaBoost and its variant, Boosted Decision Stump Feature Selection (BDSFS) method. AdaBoost achieved perfect accuracy on the single test set for the prostate cancer data set. However, a 10-fold cross validation performance yielded average sensitivity of 98.5% and a specificity of 97.9%, for an overall BACC of 98%. For the BDSFS, the results were worse, with a sensitivity of 91.1% and a specificity of 94.3%. The researchers informally report that other classifiers had similar accuracies but were more difficult to interpret.

In (Lilien et al., 2003), the researchers again used PCA for dimensionality reduction and LDA for classification. The data set was obtained from the authors of (Adam et al., 2002). In the same fashion as with the ovarian cancer set, the researchers conducted a detailed study using various train/test set sizes. For each train/test data split, 1000 cross-validation runs (with re-sampling)

[‖]It should be noted that peak extraction and clustering were used to preprocess the data and produced 192 peaks from which 7 were used by the shrunken centroid algorithm.

Comparison of three reports for prostate cancer diagnosis based on SELDI-TOF technology.

	Adam et al. (1)	Petricoin et al. (12)	Qu et al. (29)
Diagnostic sensitivity and specificity	83%; 97%	95%; 78–83%	97–100%; 97–100%
SELDI-TOF chip type	IMAC-Cu	Hydrophobic C-16	IMAC-Cu
Distinguishing peaks, m/z[a]	4475, 5074, 5382, **7024**, 7820, 8141, 9149, 9507, **9656**	2092, 2367, 2582, 3080, 4819, 5439, 18220	Noncancer vs cancer: 3963, 4080, 6542, 6797, 6949, 6991, **7024**, 7885, 8067, 8356, **9656**, 9720 Healthy individuals vs BPH:[b] 3486, 4071, 4580, 5298, 6099, 7054, 7820, 7844, 8943
Bioinformatic analysis	**Decision tree algorithm**	**Proprietary; based on genetic algorithms and cluster analysis**	**Boosted decision tree algorithm**

Fig. 30.1. Comparison of classification techniques for prostate cancer diagnosis (reproduced from (Diamandis, 2003).) Respectively, the accuracies for (Adam et al., 2002, Petricoin et al., 2002b, Qu et al., 2002) are 89%, 83%, 98%. This comparison demonstrates the wide classification variance due to different mass spectrometry and machine learning approaches.

were conducted. When training sets were larger than 75% of the total sample size, an average accuracy of 88% was achieved. Using only 50% of data for training, the performance dropped to 86%. In comparison to ovarian cancer sets the lower accuracy suggests that this data set is much more difficult to classify correctly.

In (Petricoin et al., 2002b, Wulfkuhle et al., 2003), researchers used Genetic Algorithms (GA's) for feature extraction and Self Organizing Maps (SOM's) for classification of prostate cancer. This approach achieved a 95% specificity and a 71% sensitivity, for a balanced accuracy of 83%. Although cross validation was carried out, the results were not presented.

In (Diamandis, 2003), the aforementioned studies on prostate cancer raised the following question: Why do the features and classification performance vary so drastically across studies? Indeed, results reproduced in Figure 30.1, indicate that different SELDI-TOF approaches combined with different machine learning techniques for pattern recognition produce highly variable results. This observation further motivates the need for comparative studies done on a regular basis using several mass spectrometry techniques in conjunction with a number of machine learning approaches. We attempt to carry out such a study in this paper.

30.2 Existing Feature Extraction and Classification Methods

Feature extraction is central to the fields of machine learning, pattern recognition and data mining. This section introduces algorithms used in this study. More details on the algorithms used within this study can be found in Part 1, Chapters 3 and 4.

30.2.1 Centroid Classification Method

A fast and simple algorithm for classification is the centroid method (Hastie et al., 2001, Park et al., 2003). This algorithm assumes that the target classes correspond to individual (single) clusters and uses the cluster means (or centroids) to determine the class of a new sample point. A prototype pattern for class C_j is defined as the arithmetic mean:

$$\boldsymbol{\mu}_{C_j} = \frac{1}{|C_j|} \sum_{\boldsymbol{x}_i \in C_j} \boldsymbol{x}_i$$

where $\boldsymbol{x}_i$'s are the training samples labeled as class C_j. Recall that the training sample is a MS spectra represented as a multi-dimensional vector (denoted in bold). In a similar fashion, we can obtain a prototypical vector for all the other classes. During classification, the class label of an unknown sample $\boldsymbol{x}$ is determined as:

$$C(\boldsymbol{x}) = \arg\min_{C_j} \; d(\boldsymbol{\mu}_{C_j}, \boldsymbol{x})$$

where $d(\boldsymbol{x}, \boldsymbol{y})$ is a distance function or:

$$C(\boldsymbol{x}) = \arg\max_{C_j} \; s(\boldsymbol{\mu}_{C_j}, \boldsymbol{x})$$

where $s(\boldsymbol{x}, \boldsymbol{y})$ is a similarity metric. This simple classifier will form the basis of our studies. It works with any number of features and its run-time complexity is proportional to the number of features and the complexity of the distance or similarity metric used. Preliminary experiments were conducted to establish which similarity/distance metric is most appropriate for the centroid classification algorithm**, and the L_1 distance metric was selected. Defined by:

$$L_1(\boldsymbol{x}, \boldsymbol{\mu}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|_1 \tag{30.1}$$

with $\|\boldsymbol{y}\|_1 = \sum_i^N |y(i)|$, and $y(i)$ being the value of the i^{th} feature. The value $L_1(\boldsymbol{x}, \boldsymbol{\mu})$ has a linear cost in the number of features. In this study, data sets contain two classes and hence the number of calls to a metric is also two. Therefore, the centroid classifier, at run-time, is linear in the number of features. During training, two prototypes are computed and the cost of computing each prototype is $O(mN)$, where N is the number of features and m is the number of training samples which belong to a given class. Note that m only varies between data sets and not during training or feature selection processes. Thus, we can view m as a constant and conclude that the centroid classifier has $O(N)$ cost in the training phase.

**Due to space restrictions, the results are not shown. A companion technical report (Levner, 2004) provides experimental details and supplementary material.

30.2.2 Nearest Shrunken Centroid

A special purpose feature selection algorithm for the nearest centroid algorithm was developed by Tibshirani et al. and presented in (Hastie et al., 2001, Tibshirani et al., 2003, 2004). The algorithm, related to the lasso method described in Part 1, Chapter 1, Section 4, tries to shrink the class prototypes ($\boldsymbol{\mu}_{C_j}$) towards the overall mean:

$$\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \tag{30.2}$$

Briefly, the algorithm calculates:

$$\boldsymbol{d_j} = \frac{\boldsymbol{\mu}_{C_j} - \boldsymbol{\mu}}{m_j(\boldsymbol{s})} \tag{30.3}$$

where $m_j = \sqrt{\frac{1}{|C_j|} - \frac{1}{m}}$, $\boldsymbol{s}$ is a vector of pooled within class variances for each feature and division is done component wise. We can now view the class centroid as:

$$\boldsymbol{\mu}_{C_j} = \boldsymbol{\mu} + m_j(\boldsymbol{s} \cdot \boldsymbol{d_j}) \tag{30.4}$$

where $\cdot$ denotes component wise multiplication. By decreasing $\boldsymbol{d_j}$ we can move the class centroid towards the overall centroid. When a component of the class centroid is equal to the corresponding component of the overall mean for all classes, the feature no longer plays a part in classification and is effectively removed. Hence as $\boldsymbol{d_j}$ shrinks progressively more features are removed.

30.2.3 Ordered and Sequential Feature Selection

Using the aforementioned centroid method as the base classifier, we can select features with SFS (Sequential Forward Selection) technique or via an ordered feature selection approach. Both of these wrapper-based techniques incrementally build a feature set by adding one feature at a time to the active (i.e., previously selected) set of features and invoking the nearest centroid classifier using the active feature set. Sequential Forward (respectively Backward) selection (SFS and SBS) methods start from an empty (respectively full) set of features and at each step add (respectively remove) a single feature that produces the greatest increase in performance. In contrast, the ordered feature selection approach first evaluates each of the N features independently of all others. The features are then ranked according to the performance of the base classifier (i.e., the nearest centroid classifier in our case). Once ranked and sorted, the ordered feature selection approach incrementally adds the topmost ranked feature to the active set. In total, N feature subsets are tried, where s_1 contains a single top ranked feature, s_2 contains the two top ranked features, and so on until s_N is tried. In contrast, to the SFS procedure, ordered feature

selection is linear in the number of calls to the base classifier since at each stage the top ranked feature is added to the active set and the newly created active set is evaluated by the base classifier. Since there are only N features the total number of calls to the base classifier is $2N$, N initial calls to rank individual features, and N times to evaluate the ever larger subsets $s_1, ..., s_N$. Unlike the SFS algorithm, the greedy approach will not stop until all N sets have been tried. The final stage of the algorithm merely selects the feature set producing the best classification accuracy on the particular data set.

30.2.4 Univariate Statistical Tests

Instead of ranking features by invoking a classifier, one can use filter ranking based on statistical tests. In general, univariate statistical tests analyze each feature independently of others. The student-t (T-test) and the Kolmogorov-Smirnov (KS-test) (Press et al., 2002) algorithms are common examples. Both tests compare feature values from samples belonging to class i to feature values from samples belonging to class j. The goal is to determine if the feature values for class i come from a different distribution than those for class j. The key difference between the two tests are the assumptions they make. The T-test assumes that both distributions have identical variance, and makes no assumptions as to whether the two distributions are discrete or continuous. On the other hand, the KS-test assumes that the two distributions are continuous, but makes no other assumptions.

In the case of the T-test, the null hypothesis is $\mu_A = \mu_B$, representing that the mean of feature value for class A is the same as the mean of the feature values for class B. In the case of the KS-test, the null hypothesis is $cdf(A) = cdf(B)$, meaning that feature values from both classes have an identical cumulative distribution. Both tests determine if the observed differences are statistically significant and return a score representing the probability that the null hypothesis is true. Thus, features can be ranked using either of these statistics according to the significance score of each feature. In addition, the two tests can be combined together into a composite statistic. While many possible composition strategies exist, we limit our experiments to a simple multiplicative composition, whereby the T-test significance score is multiplied together with the KS-test significance score (referred to as the T*KS-test henceforth).

Both the benefits and drawbacks of these statistical tests stem from the assumption that features are independent. On one hand, the independence assumption makes these approaches very fast. On the other hand, the independence assumption may not hold for all data sets. Technical details on these and other statistical tests can be found in (Hastie et al., 2001, Press et al., 2002).

Recall that in (Wu et al., 2003), the T-test and random forests were used for feature extraction teamed with a number of classifiers. The researchers used the T-test to rank each feature but chose to test classification algorithms

with 15 and 25 top-ranked features. Their line of research appears more focused on comparing classifiers rather than the two feature extractors (T-test and random forests). In contrast, we show that feature ranking coupled with ordered feature selection can automatically find a feature subset of arbitrary size that improves performance (with respect to using either a single best feature or using all features).

30.2.5 Dimensionality Reduction

Recall that feature selection algorithms attempt to select relevant features with respect to the performance task, or conversely remove redundant or irrelevant ones. In contrast, dimensionality reduction algorithms attempt to extract features capable of reconstructing the original high dimensional data. For example, PCA (Kirby, 2001) attempts to find a linear combination of principal components that preserves the data variance. In proteomic pattern recognition the most common technique is down sampling. This technique filters the spectra and sub-samples it to reduce the dimensionality. A common approach is to convolve the spectrum with a uniform filter at regular intervals (windows). This technique, essentially removes high frequency components. In order to test the conjecture made in (Conrads et al., 2003), that higher resolution data tends to improve classification performance, we will use this approach to test the merit of dimensionality reduction via down sampling.

30.3 Experimental Results

We conducted experiments on the ovarian and prostate data sets, previously used in (Petricoin et al., 2002a) and (Petricoin et al., 2002b). The ovarian cancer set includes sera from 91 controls and 162 ovarian cancers patients. Acquired from (Johann, 2003), each data sample contains 15,156 features. The prostate cancer data set is composed of 322 samples in total, and was also acquired from (Johann, 2003). There are 190 serum samples from patients with benign prostate whose PSA levels are greater than four, 63 samples with no evidence of disease and PSA level less than one, 26 samples with prostate cancer with PSA levels four through ten, and 43 samples with prostate cancer and PSA levels greater than ten. Again, each sample is a histogram with 15,156 bins, with each bin corresponding to a single feature.

For all experiments, each data set was split into three subsets of equal size. Each test fold used one of the three subsets with the remaining two subsets used for training. We ran two sets of experiments. The first optimized performance directly on the test set. For a given feature selection technique, this approach produces a single feature set and hence makes feature analysis possible. The drawback of this approach is that performance estimates are overly optimistic. To get a better performance estimate, a second set of experiments was carried out. It optimized performance on the training set.

Specifically, we used a leave-one-out cross-validation (LOOCV) internal loop based solely on the training set to select features using a subset of the most promising algorithms. The reported accuracy for all experiments is the average classification accuracy over the three test folds and the error bars represent one standard deviation. Accuracy is taken as the arithmetic mean of sensitivity and specificity. This measure is related to BER (Balanced Error Rate) and can be analogously thought of as balanced accuracy ($BACC$), where $BER = 1 - BACC$.

Dimensionality Reduction

We progressively down-sampled the spectra by averaging each sample spectra using a uniform filter. In other words, given a window of size w we averaged w adjacent features (i.e., m/z values) into a single new feature. The window was then shifted by w features and the process repeated. For each trial we increased size of the window w. This effectively produces data with progressively lower resolution and reduced dimensionality. For each down sampled data set we used the centroid classifier. The results, presented in Figure 30.2, show that classification performance decreases as the size of the filter increases. However, the decrease is clearly non-monotonic and, in essence, very noisy. This noise can be attributed to either the filtering or the sub-sampling stages of the down-sampling process. To determine which of the two components produced the oscillations in classification accuracy, another experiment was carried out.

In the second experiment we performed frequency based data filtering. The procedure first transformed each spectra into the frequency domain via the Fast Fourier Transform (FFT). Then a low pass filter was applied to the frequency coefficients in order to remove the high frequency components. The final stage transformed the filtered data back to the spatial domain. By varying the size of the low pass filter, the number of frequency coefficients used in reconstructing the MS spectra was varied and, in essence, considered feature selection in the frequency domain. Clearly the loss in accuracy, shown on the right side of Figure 30.2, is much more monotonic in comparison with the down-sampling method (the left-hand side). This suggests that the majority of oscillations result from the sub-sampling step rather than the frequency filtering step. This led us to the conjecture that down-sampling is in general detrimental to classification performance. To further investigate this hypothesis, we ran the centroid classifier on each individual feature for the down-sampled spectra and found the classification performance inferior to the performance of a single best feature from the non down-sampled spectra. This further supported the claim that down-sampling appears detrimental to classification accuracy. The conclusions drawn are in line with those in (Conrads et al., 2003) where changes in resolution created by different MS techniques produced similar results. Because the MS spectra are histograms describing the ion concentrations based on the mass-to-charge ratios, the low resolution techniques effectively aggregate distinct ion concentrations into a

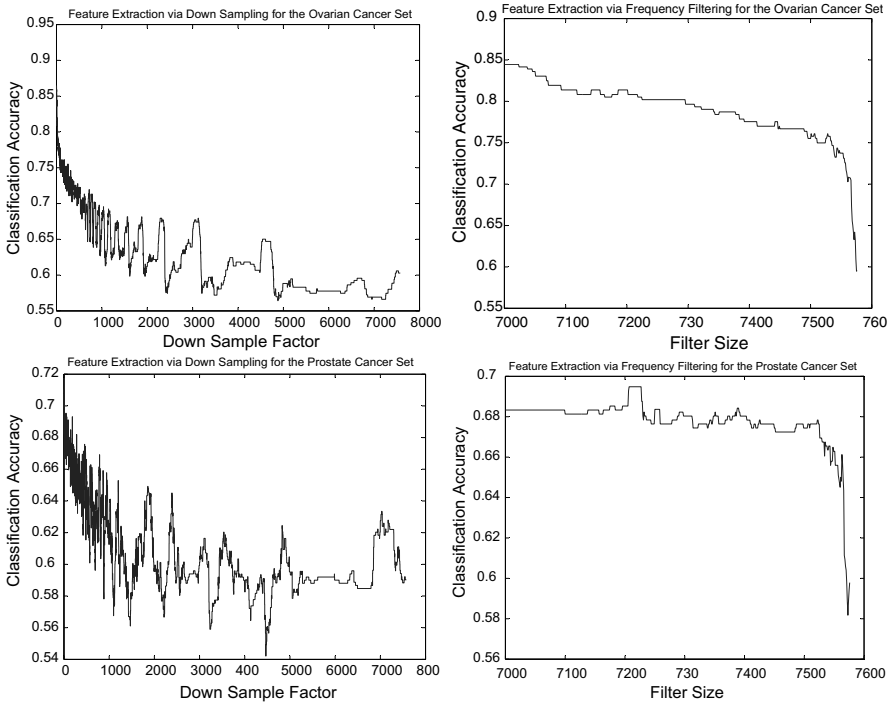

Fig. 30.2. Classification accuracy on progressively down-sampled data. **Top:** ovarian cancer data set. **Bottom:** prostate cancer data set. **Left:** Down-sampling Performance. The down-sample factor indicates the ratio of original number of features to the number of features after down-sampling. As the down-sample factor increases, the number of features decreases. **Right:** Frequency filtering. While all data sets exhibit oscillations, the performance nevertheless gradually declines as the dimensionality of the data is reduced as indicated by the increasing down-sampling factor on the x-axis.

single bin. Hence, down-sampling, whether due to low-resolution MS hardware or done deliberately in software to reduce data dimensionality, appears to lower diagnosis performance.

30.3.1 Ordered and Sequential Feature Selection

To compute the exact relevance of individual features, the centroid classifier was ran on individual features. Histogram plots for each data set are shown in Figure 30.3. Each plot represents the distribution of features with respect to classification accuracy and shows that a very large number of features are essentially irrelevant and/or redundant with respect to diagnosis. This provides further unfavorable evidence for the down-sampling approach, which in essence, aggregates individual features together. Such an approach would

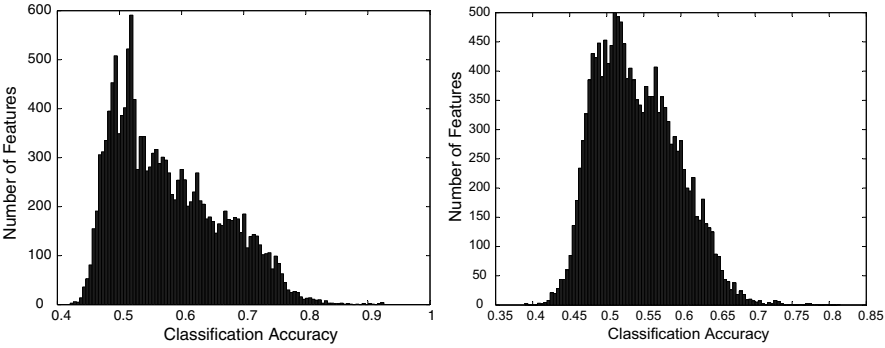

Fig. 30.3. Performance using individual features **Left:** ovarian cancer data set. **Right:** prostate cancer data set. The histograms show the number of features with a specific classification accuracy on a single test fold when individual features are used.

inevitably merge relevant and irrelevant (or redundant) features together and decrease the overall performance as evidenced by the experimental results of the previous section. Interestingly, there are a number of features within each data set that produce classification accuracies below 50%. These features can mislead and confuse the classifier.

Once each feature was ranked and the feature set sorted, ordered feature selection was used. In addition, the SFS procedure was also employed to select *relevant* feature sets. The results are presented in Figure 30.4 and are discussed in the next section.

30.3.2 Performance Comparison

Figure 30.4 presents the best performance for each feature extraction technique on each data set. Clearly, SFS coupled with the centroid algorithm produced superior results in comparison to the other algorithms tested in terms of feature set size and classification accuracy.

On the ovarian cancer data set, classification based on four features selected via SFS had the same accuracy of 98.0%, tieing with a set composed of 48 features created by the ordered feature selection. Previously, PCA coupled with LDA produced the only perfect cross-validated classification accuracy (Lilien et al., 2003). On the prostate cancer data set, the SFS classifier increased the base classification accuracy from 69.7% to 94% using only 11 of 15,154 features. In contrast, PCA coupled with LDA produced an accuracy of 88% (Lilien et al., 2003) . In (Qu et al., 2002), the boosted decision stumps produced an impressive 98% accuracy on the same data set. However, we were unable to get this set and used the data set from (Petricoin et al., 2002b), where the accuracy using GA's combined with SOM's was only 83%. Overall the SFS/centroid system appeared competitive with the previous approaches

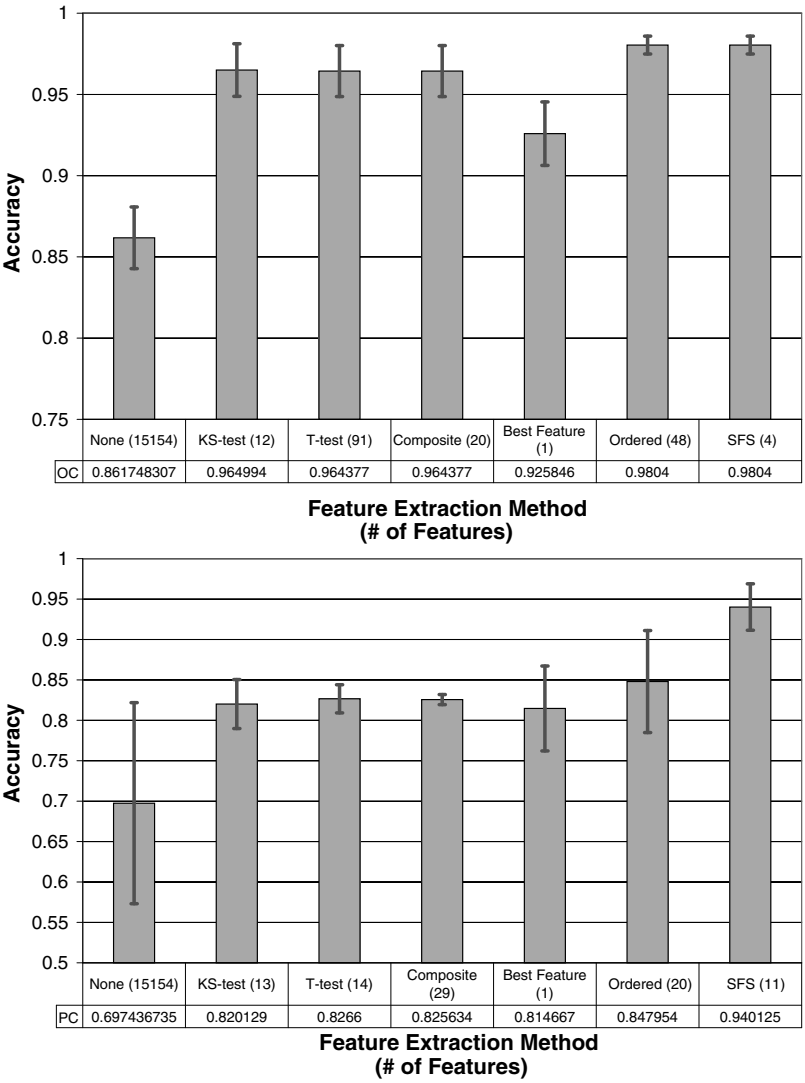

Fig. 30.4. Performance of Feature Extraction Algorithms optimized on the test sets. **Top:** ovarian cancer data set. **Bottom:** prostate cancer data set.

in terms of classification accuracy but produced considerably smaller feature sets. Note that the PCA/LDA approach always uses n features corresponding to n eigenvectors. Since the rank of the covariance matrix is bounded by the number of samples, n is necessarily upper bounded by the number of training samples, and was set to this upper bound in (Lilien et al., 2003). Furthermore, boosted decision stumps used to classify the prostate cancer data set needed

Table 30.1. Active feature set extracted by the SFS procedure. **Top:** Ovarian cancer data set. **Bottom:** Prostate cancer data set. Column 1 shows the order each feature was added to the active set. Column 2 contains the feature index. Column 3 shows classification accuracy using just the one feature. Column 4-6 present the rank of each feature using the T-test, KS-test, and T*KS-test with respect to topmost ranked feature. The SFS procedure does not appear to select the same features as any of the ordered FS methods.

Order Added	Feature Index	Individual Feature Accuracy	T-Test	KS-Test	T*KS -Test
1	1679	0.9258	5003	9289	14129
2	541	0.8303	8185	7502	9272
3	1046	0.62	9012	13276	5997
4	2236	0.9104	4855	5501	7953

Order Added	Feature Index	Individual Feature Accuracy	T-Test	KS-Test	T*KS -Test
1	2400	0.8147	2106	1880	1499
2	6842	0.6393	7823	14543	11650
3	2667	0.6246	1756	7601	13111
4	6371	0.5776	5600	609	4297
5	2005	0.5262	7128	11984	8482
6	1182	0.5147	12400	6180	890
7	7604	0.6328	7694	12788	5943
8	462	0.4531	11165	14343	11810
9	659	0.5868	13282	11766	11307
10	187	0.4994	14893	1807	5032
11	467	0.6036	12602	8744	2272

500 stumps to achieve the aforementioned accuracy. In contrast, the SFS/-centroid method selected only 5 and 11 features for the ovarian and prostate cancer data sets respectively, while producing comparable classification accuracy.

Active Feature Sets

The relationship between the features selected by the SFS procedure and the corresponding rankings based on statistical tests is illustrated in Table 30.1. Each table examines the features selected by the SFS procedure for the ovarian and prostate cancer data sets. In both cases, the features added to the active set are ranked far from first by the statistical tests. In addition, individual feature performance does not appear to be an effective indicator of classification performance within a set of features. In fact, the eighth and tenth features have individual classification accuracies of less than 50% on the prostate data. Furthermore, not a single feature selected by any of the ordered feature selection approaches appears in the active set produced by

Table 30.2. Active feature set extracted from the prostate cancer data set by the SFS procedure. Column 1 shows the order each feature was added into the active set. Column 2 contains the feature index. Column 3 provides the actual mass-to-charge ratio of each feature. The last three columns present nearby (±500 Da) features found previously in (Adam et al., 2002, Qu et al., 2002, Petricoin et al., 2002b). Clearly the SFS procedure found a set of features very different than the other algorithms.

Order Added	Feature Index	M/Z	Adam et al.	Qu et al.	Petricoin et al.
1	2400	**500.8**			
2	6842	**4074.8**	4475	3963; 4080; 4071	
3	2667	**618.6**			
4	6371	**3533.0**		3486	3080
5	2005	**349.4**			
6	1182	**121.3**			
7	7604	**5033.3**	5074	5289	4819; 5439
8	462	**18.4**			
9	659	**37.6**			
10	187	**3.0**			
11	467	**18.8**			

the SFS procedure. However, the ordered approaches do improve performance in comparison to the classification accuracy based on the full feature set. This indicates that there are a number of relevant features related to the presence/absence of cancer.

To further examine the features extracted by the SFS, we compared the active sets extracted by this procedure for the prostate cancer set to the features selected using other approaches surveyed in the previous research literature (refer to Figure 30.1). The results are summarized in Table 30.2. Clearly, very few common features are observed. As hypothesized in (Diamandis, 2003), it appears that different algorithms extract different relevant features based on their internal machinery and bias. A crucial goal for future research is therefore, to determine which, if any, features can serve as potential biomarkers, and shed light on the nature of cancer, and possibly even its cure.

30.3.3 LOOCV Performance

The previous section presented results with classification performance optimized directly on the test set. While this approach produces feature sets that can be analyzed easily, algorithm performance may be grossly optimistic. To produce a more realistic performance estimate we re-ran our experiments with feature selection done using leave-one-out cross-validation (LOOCV) within the training set. This procedure was also repeated 3 times for each external test set. Due to the increased cost of the LOOCV procedure, we selected SFS, KS-test, T-test and also the nearest shrunken centroid algorithm for comparison. Results are presented in Figure 30.5. The LOOCV performance estimates are similar to performance optimized on test sets for the ovarian cancer. However, for the prostate cancer LOOCV performance is substantially lower.

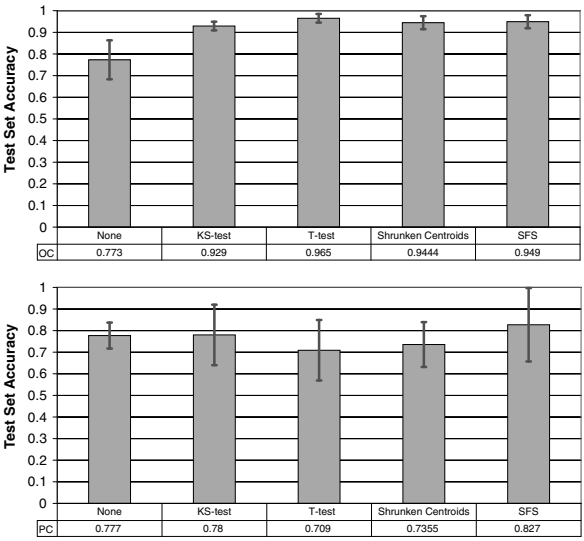

Fig. 30.5. Performance of Feature Extraction Algorithms optimized using LOOCV within the training set. The balanced accuracy is averaged over 3 test folds unseen during training. **Top:** ovarian cancer data set. **Bottom:** prostate cancer data set.

Table 30.3. Computational times, in CPU seconds, taken by each algorithm for the LOOCV feature selection.

CPU Time (sec)	None	KS-test	T-test	SFS	Shrunken Centroids
Ovarian	0.87	24.55	623.37	2175.75	33115.00
Prostate	1.31	25.4	639.14	3269.37	33115.00

The running times of each algorithm are presented in Table 30.3. Although the nearest centroid takes the greatest amount of time, the running time is dictated by the number of shrunken centroids examined during the LOOCV stage. Recall that decreasing $\boldsymbol{d_j}$ shrinks the class centroid (for each class). Hence the number of times we decrease $\boldsymbol{d_j}$ directly impacts performance. In our case we used 200 progressively shrunken centroid sets and picked the best one using LOOCV.

30.4 Conclusion

Mass spectrometry disease diagnosis is an emerging field poised to improve the quality of medical diagnosis. However, the large dimensionality of the data requires the use of feature extraction techniques prior to data mining and

classification. This paper analyzed statistical and wrapper-based approaches to feature selection as well as dimensionality reduction via down-sampling. Experimental results indicate that down-sampling appears detrimental to classification performance, while feature selection techniques, in particular sequential forward selection coupled with a fast but simple nearest centroid classifier, can greatly reduce the dimensionality of the data and improve classification accuracy. Future research will investigate how the selected features impact classification accuracy when used in conjunction with more sophisticated classifiers, such as Artificial Neural Networks and Support Vector Machines. From a biological perspective, it is of interest to investigate the nature of the selected features. As potential biomarkers, these features may shed light on the cause or even the cure to cancer and other disease.

30.5 Acknowledgements

We would like to thank Lihong Li and Greg Lee for their various contributions. Ovarian and prostate cancer data sets provided by the National Cancer Institute, Clinical Proteomics Program Databank (Johann, 2003). Funding for this research was provided by University of Alberta, National Science and Engineering Research Council, and Alberta Ingenuity Center for Machine Learning.

Glossary

In this section we define the various measures used. Respectively, TP, TN, FP, FN, stand for the number of true positive, true negative, false positive, false negative samples at classification time.

Sensitivity $\frac{TP}{TP+FN}$ is also known as Recall.

Specificity $\frac{TN}{TN+FP}$

PPV (Positive Predictive Value) $\frac{TP}{TP+FP}$. is also known as Precision.

NPV (Negative Predictive Value) $\frac{TN}{TP+FP}$

Accuracy defined as $\frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$ in this paper.

References

B. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, and Jr. G. L. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62(13):3609–3614, 2002.

T. P. Conrads, M. Zhou, E. F. Petricoin III, L. Liotta, and T. D. Veenstra. Cancer diagnosis using proteomic patterns. *Expert Reviews in Molecular Diagnostics*, 3 (4):411–420, 2003.

E. Diamandis. Proteomic patterns in biological fluinds: Do they represent the future of cancer diagnostics. *Clinical Chemistry (Point/CounterPoint)*, 48(8):1272–1278, 2003.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer Verlag, New York, 2001.

D. Johann. Clinical proteomics program databank. Technical report, National Cancer Institute, Center for Cancer Research, NCI-FDA Clinical Proteomics Program, 2003. http://ncifdaproteomics.com/ppatterns.php.

C. Kainz. Early detection and preoperative diagnosis of ovarian carcinoma (article in german). *Wien Med Wochenschr*, 146(1–2):2–7, 1996.

Michael Kirby. *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. John Wiley & Sons, New York, 2001.

I. Levner. Proteomic pattern recognition. Technical report, University of Alberta, April 2004. No: TR04-10.

R.H. Lilien, H. Farid, and B. R. Donald. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *Computational Biology*, 10(6), 2003.

H. Park, M. Jeon, and J. B. Rosen. Lower dimensional representation of text data based on centroids and least squares. *BIT*, 43(2):1–22, 2003.

S. D. Patterson and R. H. Aebersold. Proteomics: The first decade and beyond. *Nature, Genetics Supplement*, 33:311–323, 2003.

E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359 (9306):572–577, 2002a.

E. F. Petricoin, D.K. Ornstein, C. P. Paweletz, A. Ardekani, P.S. Hackett, B. A. Hitt, A. Velassco, C.Trucco, L. Wiegand, K. Wood, C. Simone, P. J. Levine, W. M. Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. A. Liotta. Serum preteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute*, 94(20):1576–1578, 2002b.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientifi Computing, Second Edition*. Cambridge University Press, 2002.

Y. Qu, B. Adam, Y. Yasui, M. D. Ward, L. H. Cazares, P. F. Schellhammer, Z. Feng, O. J. Semmes, and Jr. G. L. Wright. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48(10):1835–1843, 2002.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 18 (1):104–117, 2003.

R. Tibshirani, T. Hastiey, B. Narasimhanz, S. Soltys, G. Shi, A. Koong, and Q. Le. Sample classifcation from protein mass spectrometry by 'peak probability contrasts'. *BioInformatics*, 2004.

B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classifcation of ovarian cancer using mass spectrometry data. *BioInformatics*, 19(13), 2003.

J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin. Proteomic applications for the early detection of cancer. *Nature Reviews*, 3:267–275, 2003.