Proteomic Pattern Recognition

Ilya Levner

University of Alberta

Abstract. This report overviews the Mass Spectrometry Data Classification and Feature Extraction problem. After reviewing previous research new classification and feature extraction techniques are presented and empirically evaluated on three data sets. One of the key points made in this work, is that feature extraction techniques are composed of dimensionality reduction and feature selection methods. However, the two notions are quite different. The need for dimensionality reduction stems from the fact that classification algorithms cannot cope with the large number of input variables. On the other hand, feature selection techniques attempt to remove irrelevant and/or redundant features. Often classification algorithms cannot handle both a large number of variables and irrelevant variables that are not needed or even worse are misleading. In order to evaluate the dimensionality reduction and feature selection techniques, we use a simple classifier to evaluate performance. This makes the approach tractable. The experiments indicate that feature selection algorithms tend to both reduce data dimensionality and increase classification accuracy, while the studied dimensionality reduction technique sacrifices performance as a result of lowering the number of features a learning algorithm needs to deal with.

Keywords: BioInformatics, Mass Spectrometry, Proteomic Pattern Recognition, Classification, Feature Extraction, Multi-Resolution.

Table of Contents

1 Introduction

Early detection of diseases, such as cancer, is critical for improving patient survival rates and medical care. Modern diagnosis systems are still unreliable, slow, or nonexistent for numerous diseases. 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 exist, such as PAGE[⋆⋆], mass spectrometry (**MS**) based approaches provide very high throughput, can be widely applicable, and have the potential to be highly accurate. This study examines pattern recognition in **proteomic** applications. A good review of proteomic techniques can be found in [24]. 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, for example 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 biomarker exists, that can produce a reliable diagnoses. The need for pattern recognition is further motivated by the fact that

> *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. [4]*

The fact that there may not even be *any* biomarkers that can provide reliable screening and diagnosis has prompted research into proteomic pattern recognition(see Figure 17). This research reviews the current literature on proteomic pattern recognition with emphasis on data resolution characteristics, feature extraction and pattern analysis tools. More specifically, the high dimensionality of the MS data requires aggressive feature extraction techniques in order to make machine learning algorithms feasible. This study compares several simple yet effective techniques for feature extraction that produce results competitive with those found in the reviewed litterature on three data sets acquired for this study (described in the appendix). The rest of the paper is organized as follows. First, we motivate the need for feature extraction by a section on related research.

[⋆] See appendix for details and definitions.

[⋆⋆] PAGE- polyacrylamide gel electrophoresis. Also known as 2DE for 2 dimensional polyacrylamide gel electrophoresis. More details can be found in [24].

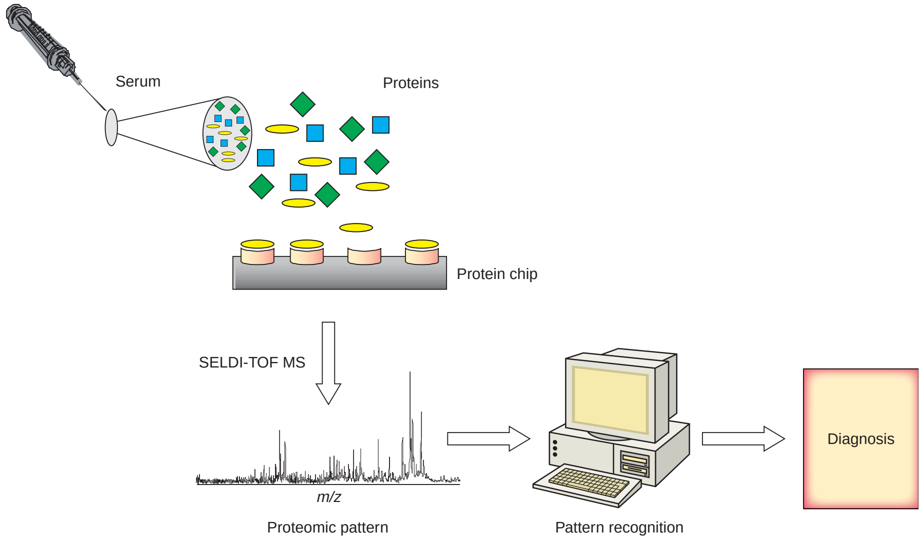


Fig. 1. Conceptual view of proteomic pattern recognition for disease diagnosis. A sample drawn from the patient is applied to a protein chip which is made up of a specific chromatographic surface. After several washing steps and the application of an energy-absorbing molecule, the species that are retained on the surface of the chip are analyzed via mass spectrometry. The surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) is used to acquire the proteomic patterns. The species bound to the array surface can be ionized by matrix-assisted laser desorption/ionization (MALDI) and their mass-to-charge (m/z) ratios measured by TOF MS. The result is a mass spectrum of the species that bound to and subsequently desorbed from the array surface. The pattern of peaks within the spectrum is analyzed using pattern recognition software to diagnose the biological sample. [4]

Next, we outline the feature extraction algorithms which are compared in the subsequent section on experimental results. The paper is concluded with a discussion of results and future research.

2 Related Research

We start the review by briefly examining a study evaluating two different mass spectrometry techniques. Motivated by the need for greater recall and precision[⋆⋆⋆], in [4] the standard TOF MS technique was compared to the SELDI TOF MS approach. The goal was to determine whether sensitivity and PPV

[⋆⋆⋆] see appendix for definitions of precision and recall.

(i.e., recall and precision) scores would improve by using a higher resolution spectra provided by the SELDI TOF MS technique (Figure 2 depicts a spectrum at two different resolutions). Keeping all other parameters constant, including the machine learning algorithm, classification based on high resolution data achieved 100% specificity and PPV scores on the ovarian cancer data set using the SELDI based MS technique[†]. In contrast, none of the models[‡] based on the low resolution mass spectra could achieve perfect precision and recall scores. The researchers concluded that the high resolution technique, which increased resolution 60 fold, improves the performance of the pattern recognition technique used. In a field where the outcome of a test can mean the difference between life and death and whose goal is to improve the quality of life, sending healthy patients for unnecessary surgery or letting misdiagnosed patients die is extremely undesirable. Therefore, pattern recognition techniques need highly **relevant** information in order to make the most accurate and reliable predictions. According to [13], due to the low prevalence of (ovarian) cancer a screen test would require a 99.6% specificity to achieve a clinical acceptable positive predictive value of 10%. As a result, high resolution MS techniques have been adopted to increase the recall and precision of the provided diagnosis. Unfortunately, increasing the data resolution proliferates ”the curse of dimensionality”, making a large number of modern day ML (Machine Learning) techniques intractable. As a result, **feature extraction** is needed to extract/select salient features in order to make pattern recognition techniques feasible. In addition to making ML algorithms feasible, 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 could shed light on the nature of the disease and help develop clinical diagnostic tests and treatments.

The rest of this section surveys previous research on proteomic pattern recognition with the emphasis on feature extraction and classification techniques.

2.1 Lung Cancer Studies

In September 2002, the First Annual Conference on Proteomics Data mining presented two challenges. The first, was to cluster a set of MS samples into two groups with the labels (diseased or healthy) of the patients unknown. The second challenge was essentially the same but included the labels. The two challenges exemplify the conceptual division of machine learning into unsupervised and supervised techniques. *The project concentrates on supervised techniques for classification but addresses unsupervised methods for the task of feature ex-*

[†] Unfortunately, these results are not cross validated. I.e., only a single test set was used to evaluate performance

[‡] For each technique a total of 108 models were created using the approach in [10], which used SOM’s and genetic algorithms to learn the models.

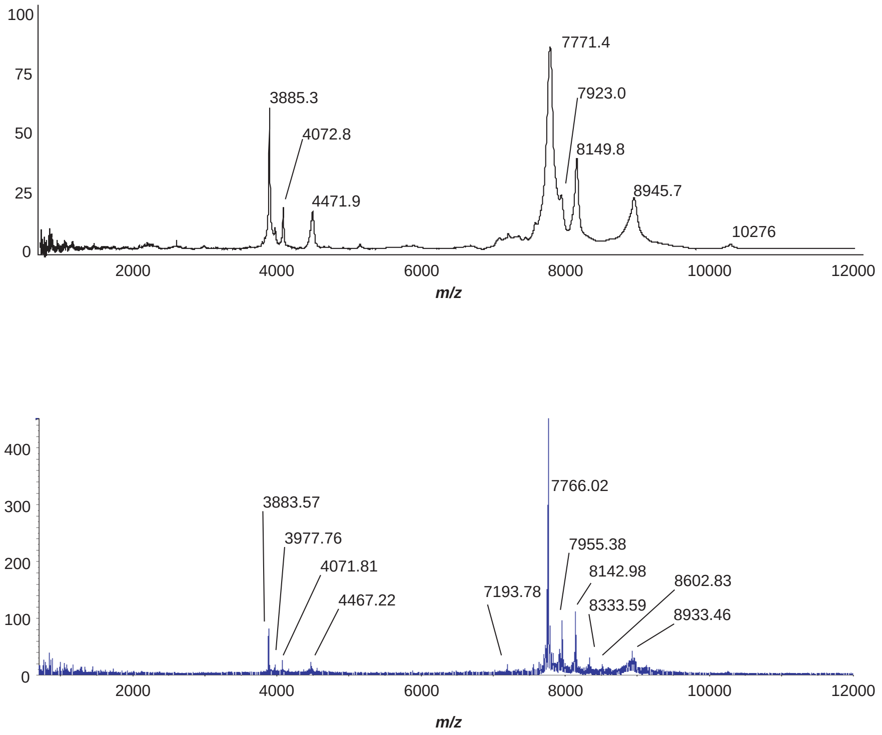


Fig. 2. TOP Standard TOF MS spectra . **BOTTOM** Sixty fold increase in resolution as a result of using SELDI TOF MS. The spectra represents an abundance histogram (y-axis in parts-per-million) of ionized proteins with specific mass-to-charge (m/z) ratios (x-axis). An increase in resolution basically increases the number of histogram bins or buckets as depicted in this figure. [4]

traction. In both challenges, presented by the Departments of Radiology and Biostatistics at Duke University, the spectra of 41 patients (24 healthy and 17 individuals with lung cancer) was given to the participants. The extremely small sample size is yet another constraint within the proteomics field that makes the pattern analysis that much harder.

In [31] the researches used local linear regression to adaptively smooth the spectra and reduce noise. Peaks were then hand-extracted and normalized using the $L1$-norm. Using top 13 intensity peaks as input into an SVM module, the procedure achieved an error rate of 2% in a leave-one-out cross-validation (**LOOCV**) manner. While the researchers were able to achieve excellent performance, the high level of human intervention detracts from the goal of automated pattern recognition.

Good performance on the supervised lung cancer set was achieved in [2]. Through visual inspection, the researchers removed sinusoidal noise, performed baseline subtraction, and normalized the spectra using $L1$ metric. Interestingly, the fractionation process[§] was reversed much like in [31] and the spectra smoothed. The process effectively created a blurred version of the original high resolution spectra acquired by the MALDI TOF MS technique. Next, max filtering was used to select peaks and further reduce the dimensionality of the data. After several more processing steps that reduced the size of each data vector from 60,831 to 506, a genetic algorithm (GA) was used to find informative peaks that discriminate between healthy and cancer stricken individuals. Finally, Fisher's linear discriminant analysis was performed on the selected peaks. The procedure was repeated in a leave-one-out cross-validation (**LOOCV**) manner. The best model misclassified 3 out of 41 samples giving an error rate of 7.3%.

In [34] wavelets were used to reduce the dimensionality of the data and R-trees (similar to decision trees) were used to classify the data. In a LOOCV test 6 out of 41 samples were misclassified, producing an error of 14.5%. (Ten fold CV produced an error of 20%.) A similar approach was used in [18] where wavelets were also used for dimensionality reduction. Then partial least squares coupled with discriminant analysis achieved a classification error of 15%. The researchers indicate that neural networks and SVM's achieved a higher accuracy but did not report cross validated results using these techniques. A similar approach using partial least squares together with logistic regression was used in [27] and achieved perfect accuracy using a LOOCV setup. However, even the authors state that there was "considerable effort involved in the preprocessing of the data" [27].

In [12], an information theoretic approach was taken in order to extract relevant features. After, once again merging fractions into reliable peaks, the researchers used entropy to select relevant peaks from the blurred spectra. Once relevant peaks were identified a boolean mask was created to indicate whether a feature was relevant. The researchers hypothesized that the presence of a relevant peak was more important than its intensity, therefore all subsequent work used the binary mask as the input vector. The training phase used unsupervised clustering and fuzzy logic applied sequentially in a feedback loop. The precess recursed until stable clusters were found. Prototypical vectors (centroids) were then created, corresponding to individual clusters. The nearest-neighbor method is then used to classify a new sample based on the distance to class prototypes. The distance metric seems to be also leaned from the training data using a k-means like neural network. Results, however, are considerably poorer than those of previous works.

In [7] the data set was normalized by combining adjacent mass spectra together and there by reducing the input dimensionality from the original 60,831 features to just 1,676. Next, baseline subtraction was performed and absolute peak heights were converted into relative peak heights. The data was then buck-

[§] The fractionation process is discussed in the appendix eted using an entropy measure. The number of bins was derived by means of an MDL technique. Feature selection was done via two techniques, Relief-F and InfoGain. InfoGain, like other statistical techniques considers the relevance of each feature independently and is based on the entropy measure mentioned previously. In contrast Relief-F [17] uses a nearest-neighbor like approach and is more context sensitive. In conjunction with the two feature extraction (FE) techniques, the researchers used six machine learning (ML) algorithms, namely: a multi-layer perceptron, an artificial neural network, a Naive Bays algorithm, a nearest-neighbor algorithm, two decision tree algorithms and a set cover rule learner; in order to determine the best (FE,ML) pair. Over a 10-fold cross validation the best result of 2.4% error was achieved using the Relief-F together with the multi-layer perceptron algorithm.

The rest of the entries used techniques that yielded ever decreasing performance results. In [21], the CART algorithm was used after the data was bucketed into 13 bins. No statistically significant results were given, however, on the test set the approach achieved a 10% error. In [22], the researchers manually selected peaks and ran a number of statistical algorithms such as logistic regression, linear discriminant analysis and decision trees. Since no cross-validation was not performed, deciphering a clear winner is difficult is this particular setting. Finally, in [29], the researchers concentrated solely on the unsupervised clustering problem and used order statistics as a distance metric on pre-ranked peaks. The distance metric was used by PCA algorithm to cluster the unlabelled data. This technique seems to be closely related to multi-dimensional scaling (MDS) algorithm commonly used to factor distance matrices.

It is interesting to note that a number of researchers (for example in [2, 18, 31, 12, 7]) chose to reduce the dimensionality of the data by merging neighboring features together. This effectively reduces the resolution of the data and contradicts the claim made in [4] that higher resolution spectra leads to better classification accuracy. Therefore an interesting line of research is use a rigorous empirical study in order to validate the conjecture that higher resolution data indeed produces better classification performance.

2.2 Ovarian and Breast Cancer Studies

In [19] the researchers aimed at identifying breast cancer from MS spectra. In total 169 samples were used in this study, with 103 patients having been diagnosed with breast cancer. The initial data was log-normalized to reduce sample variance. Using a specialized version of structural risk minimization algorithm[¶], the features were ranked according to the amount each contributed towards maximal sample separability. Multi-variate logistic regression was then used to build a classifier that discriminated between healthy and unhealthy individuals. In a 20-fold cross-validation study, using a 70/30% train/test set ratio, the approach

[¶] The exact algorithm was Unified Maximum Separability Analysis (UMSA) by ProPeak achieved an average sensitivity of 93% and a specificity of 91%, for an average accuracy of 92%.

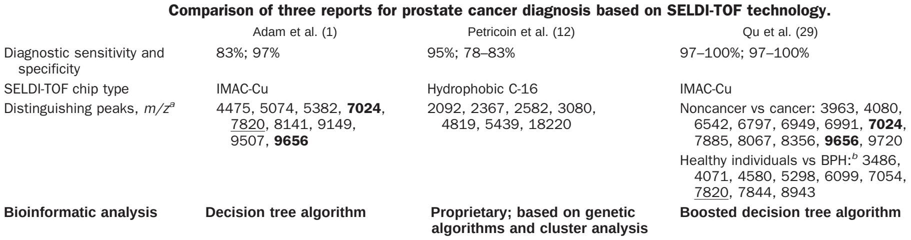
In [10], genetic algorithms together with self-organizing maps were used to distinguish between healthy women and those afflicted with ovarian cancer. Although cross-validation studies were not conducted, the approach was able to correctly classify all cancer stricken patients and 95% of healthy women, on a single test set. To improve the overall predictive accuracy the researchers turned to a higher resolution MS technique in [4]. As previously discussed this change in MS technique yielded a number of perfect predictive models when compared to the low resolution technique. Unfortunately, cross-validation studies were not presented.

In [20], the researchers used the PCA for dimensionality reduction and LDA for classification. The data set(s) was the same as in [10, 4] obtained from [8]. 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) 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 only dropped by only 0.01%. This results represents the most significant and statistically valid performance for the ovarian cancer set and over all related research literature. We can also conclude that PCA appears to be an effective way to reduce data dimensionality.

In [32], researchers compared two feature extraction algorithms together with several classification approaches. The T-statistic was used to rank features in terms of relevance. Then 2 feature subsets were greedily selected (respectively having 15 and 25 features each). Then, support vector machines, random forests, Linear Discriminant Analysis, Quadratic Discriminant Analysis, k-nearest neighbors, and bagged/boosted decision trees, were used to classify the data. In addition, random forest 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 3 results (accuracy is unclear from the graphs, but appears to be about 85%. On the other hand classification improved to approximately 92% when random forests are used as both feature extractors and classifiers (1-nearest-neighbor also had a similar performance).

2.3 Prostate Cancer Studies

In [1] the researchers used a decision tree algorithm to differentiate between healthy individuals and those having prostate cancer. This study, used the SELDI TOF to acquire the mass spectra. ROC curves were used to identify informative peaks which were subsequently used by the decision tree classification algorithm. The researchers chose not to perform cross-validation, but on a single



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. 3. Comparison of classification techniques for prostate cancer diagnosis. From [6]. Respectively, the accuracies for [1, 11, 28] are 89%, 83%, 98%. This comparison demonstrates the wide classification variance due to different MS techniques and ML approaches.

test set the classifier had 81% sensitivity and 97% specificity, yielding an accuracy of 89%.

In [28] improved performance from [1] by using ROC curves together with AdaBoost and its variant Boosted Decision Stump Feature Selection (BDSFS). AdaBoost achieved perfect accuracy on the single test set for the prostate cancer data set. However, a 10-fold cross validation performance had an average sensitivity of 98.5% and a specificity of 97.9%, for an overall performance 98%. For the BDSFS, the results were considerably 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 [20], the researchers used Principal Component Analysis for dimensionality reduction and LDA for classification. The data set was obtained from the authors of [1]. In exact same fashion as for 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) were conducted. When training sets were larger than 75% of the total sample size perfect average accuracy of 88% was achieved. Using only 50% of data for training, the performance only 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 [11, 33] researchers used Genetic Algorithms for feature extraction and Self Organizing Maps for classification of prostate cancer. This approach achieved a 95% specificity and a 71% sensitivity, for an average accuracy of 83%. Although cross validation was done, the details were not presented.

In [6], the aforementioned studies on prostate cancer raised an interesting line of questions. Why do the features and classification performance vary so drastically across the different studies. The results, summarized in Figure 3 indicate that different SELDI-TOF approaches combined with different machine learning techniques for pattern recognition produce highly variable results. This further motivates the need for comparative studies done on a regular bases using several MS techniques in conjunction with a number of ML approaches.

3 Feature Extraction and Classification Methods

Clearly, feature extraction is central to the fields of machine learning, pattern recognition and data mining. The focus of this research is on statistical and multi-resolution feature extraction approaches. The next subsection introduces general feature extraction concepts, followed by a description of algorithms used in this study.

3.1 Feature selection

Optimal[‖] feature set selection has, in general, been found to be an intractable problem [16]. However, numerous feature selection and extraction methods have been shown to perform well in practice. Traditionally, feature selection approaches have been partitioned into three categories: (i) Filter methods, (ii) Wrapper Methods, and (iii) Weighting Methods.

Filter Methods Filter based approaches attempt to select features based on simple auxiliary criteria, such as feature correlation, in order to remove redundant features. Such approaches inevitably decouple the selection process from the performance component, in order to be tractable, but may ultimately select irrelevant features as a result. An example of a filter approach is Principal Component Analysis [14], which reduces the dimensionality of the input by selecting principal components that capture a significant portion of variance within the data.

Wrapper Methods In contrast, wrapper approaches attempt to evaluate feature relevance within the context of a given task and avoid intractability by using greedy search methods. In other words, the number of possible subsets is restricted by the greedy selection procedure, and each candidate feature subset is evaluated using the performance element. Thus far a variety of greedy algorithms have been proposed to sequentially select feature sets. Sequential Forward (*resp. Backward*) selection (SFS and SBS) methods start from an empty (*resp. full*) set of features and at each step add (*resp. remove*) a single feature which produces the greatest increase in performance. One of the most effective algorithms is the Sequential Floating Forward/Backward Selection (SFFS and SFBS) algorithms [26]. This algorithm in essence combines SFS and SBS. After adding a single feature to the active set of features via SFS, the algorithm repeatedly invokes SBS to remove features from an active set. If after removing a feature from an active set performance increases, that feature can never be

[‖] In this context optimality is evaluated with respect to classification accuracy. I.e., an optimal feature subset produces the highest classification accuracy possible given the full set of features. In the worst case all 2^f feature subsets, where f is the number of features, may need to be examined in order to find the subset yielding maximal classification accuracy.

added back into the active set again. The SBS process continues until there is no improvment in performance as a result of removing a feature from the active set. Once SBS has terminated, SFS resumes its search for a single feature that produces the greatest improvement in performance as a result of being added to the active set.

Run Time Analysis of SFS and SFFS Clearly, if a feature is permanently removed from the active set, optimality has been sacrificed for a polynomial run time. To see why the algorithm is polynomial observe that worst run time of SFS occurs when the whole feature set must be added to the active feature set. Let the total number of features be f. In the worst case SFS examines f features in the first pass, $f-1$ features in the second, and so on until examining just one feature in the last iteration and adding it to the active set that contains every other feature. Hence the number of calls to the chosen classifier will be $\sum_i^f i = \frac{f(f-1)}{2}$ or more conveniently $O(f^2)$. So the worst time complexity of SFS is quadratic in the number of calls to a given classifier and the total cost is $O(f^2)O(C_f)$, where $O(C_f)$ is the worst-time complexity of the classification algorithm on f features. This is a very crucial point, that makes wrapper-methods based on SFS and SBS very costly. For the exposition of this point, consider using a least means square (LMS) algorithm as a classifier. In general LMS requires the solution of the linear system $Ax = b$, which usually costs $O(f^3)$ additions and multiplications. Hence feature extraction via SFS coupled with LMS will have a worst case running time of $O(f^6)$. Now consider the fact that a typical spectra acquired via Mass Spectrometry may contain between 5000 and 200,000 features, depending on the type of MS. Thus using SFS+LMS can easily become intractable. The same reasoning applies to sequential backward selection, which exhibits worst case behavior when all but one of the features need to be deleted from the active set (which contains all features at the beginning of the SBS procedure). This inevitably brings us to the SFFS algorithm which uses both SFS and SBS. The worst case happens when all features are added by SFS but in the very last iteration the SBS procedure deletes all but the last added feature. In that case SFFS is $O(n^4)$ in the number of calls to the base classifier. This is perhaps the key reason why wrapper methods can be impractical for large feature sets. In practice, however, this worst case behavior rarely happens.

Weighting Methods The weighting approach, simply assigns relevance weights to all possible features based on the accuracy of a simple machine learning algorithm trained on each feature individually. Examples of such approaches include weighted nearest neighbors [15] and AdaBoost [30] techniques.

Composite Methods More recently researchers have tried to combine several algorithms together on order to boost the overall performance. For example in [3], RELIEF[17] was used to remove irrelevant features. Correlation based clustering was subsequently applied to eliminate redundant features. Finally SFF(B)S algorithms were used to greedily select the required number of features from the remaining set.

In a similar fashion, we note the similarity of the wrapper and weighting approaches, and propose to combine their strengths into meta-wrappers. The major problem with wrapper approaches is the need to invoke the performance element which can be costly. Given the fact that Mass Spectrometry data is composed of thousands of features, a complex and costly performance element simply makes wrapper methods intractable. In order to make sequential selection algorithms practical an efficient classifier is needed whose run-time cost scales linearly with the number of features. Such an algorithm is presented in the next subsection.

3.2 Centroid Classification Method

A fast and simple algorithm for classification is the centroid method [23]. 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_1 is defined as the arithmetic mean:

$$P_{C_1} = \frac{1}{n}\sum_{i=1}^{n} x_i \ (\forall x_i \ \in \ C_1) \tag{1}$$

where x_i's are the training samples labeled as class C_1. Recall that the training sample is a MS spectra represented as a multi-dimensional vector. In a similar fashion we can obtain a prototypical vector (P_{C_j}) for all the other classes C_j. Classification on an unknown sample x is determined by :

$$C(x) = \arg\min_{C_j} \ d(P_{C_j}, x) \tag{2}$$

where $d(x, y)$ is some distance function or

$$C(x) = \arg\max_{C_j} \ s(P_{C_j}, x) \tag{3}$$

where $s(x, y)$ is a similarity function.

This simple classifier will form the basis of our studies. Clearly 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. This study will use simple metrics such as $L1$ and $Angle$ distances defined by:

$$L_1(x, P) = \|x - p\|_1 \tag{4}$$

with $\|y\|_1 = \sum_i^f |y(i)|$ and $y(i)$ being the value of the i^{th} feature. Similarly we define the angle similarity metric as

$$Angle(x, P) = \frac{\|x \cdot P\|}{\|x\| \cdot \|P\|} \tag{5}$$

the value $Angle(x, P)$ is actually the Cosine of the angle, but that does not matter for our purposes. Clearly both metrics have linear costs in the number of features. In this study data sets contain either 2 or 4 classes and hence the number of calls to a metric is either 2 or 4. Thus the centroid classifier is linear in the number of features as classification time. During training either 2 or 4 prototypes are computed the cost of computing each prototype is $O(fn)$, where f is the number of features and n is the number of training samples which belong to a given class. Note that n only varies from data set to data set and not during training or feature selection process. Hence we can view n as a constant and declare that the centroid classifier has $O(f)$ cost in the training phase.

3.3 Greedy and SFS Feature Selection

Using the aforementioned centroid classifier as the base classifier we can select features using the SFS technique or via the Greedy approach. The greedy approach simply ranks features according to the performance of the base classifier using each feature independently of all others. Once ranked and sorted, the greedy selection approach incrementally adds the topmost ranked feature to the active set. In total f feature subsets are tried, where s_1 contains a single top ranked feature, s_2 contains the 2 top ranked features, etc. In contrast to the SFS procedure, the Greedy approach is linear in the number of calls to the base classifier since at each stage only the topmost ranked feature is added to the data set and there are only f data sets. Unlike the SFS algorithm the Greedy approach will not stop until all f sets have been tried. The final stage of the algorithm simply selects the feature set producing the best classification accuracy.

3.4 Statistical tests

An alternative to ranking features by invoking the base classifier, is to use a filter based ranking method, such as statistical test procedures. In general, statistical tests analyze each feature independently of others just like the weighting approach of the previous section. The student-t (T) test and the Kolmogorov-Smirnov (KS) [25] tests, are no exception. Both tests compare feature values for 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 that $\mu_A = \mu_B$, meaning 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 that $cdf(A) = cdf(B)$, meaning that feature values from both classes have an identical cumulative distribution function. Both tests ask if observed differences are statistically significant and return a significance 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 ourselves to a simple multiplicative composition, whereby the T-test significance score is multiplied together with the KS-test significance score.

Both the benefits and drawbacks of these statistical tests result 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 clearly may not hold for all data sets. More technical details on statistical tests can be found in [25].

Recall that in [32] the t-statistic test and random forest were used for feature extraction together with a number of classifiers. The researchers used the T-statistic to rank each feature but chose to test classification algorithms with 15 and 25 top-ranked features. No justification was provided for selecting this specific number of features. Their line of research appears more focused on comparing classifiers rather than the two feature extractors (T-test and random forests) used in the study. In contrast, we show that feature ranking coupled with greedy selection can **automatically** find a feature subset of arbitrary size that improves performance (with respect to using either a single best feature or using the all features).

3.5 Dimensionality Reduction

While feature selection algorithms attempt to select relevant features, or conversely remove redundant or irrelevant ones, the goal in dimensionality reduction techniques is to literally reduce the number of features while preserving the information content. In proteomic pattern recognition by far the most common technique is down sampling. This technique actually filters the spectra and sub-samples it to reduce the dimensionality. The most common approach is to convolve the spectrum with a uniform filter at regular intervals (windows). In order to test the conjecture made in [4], that higher resolution data tends to improve classification performance, we will use this approach to test the merit of dimensionality reduction via down sampling.

4 Experimental Results

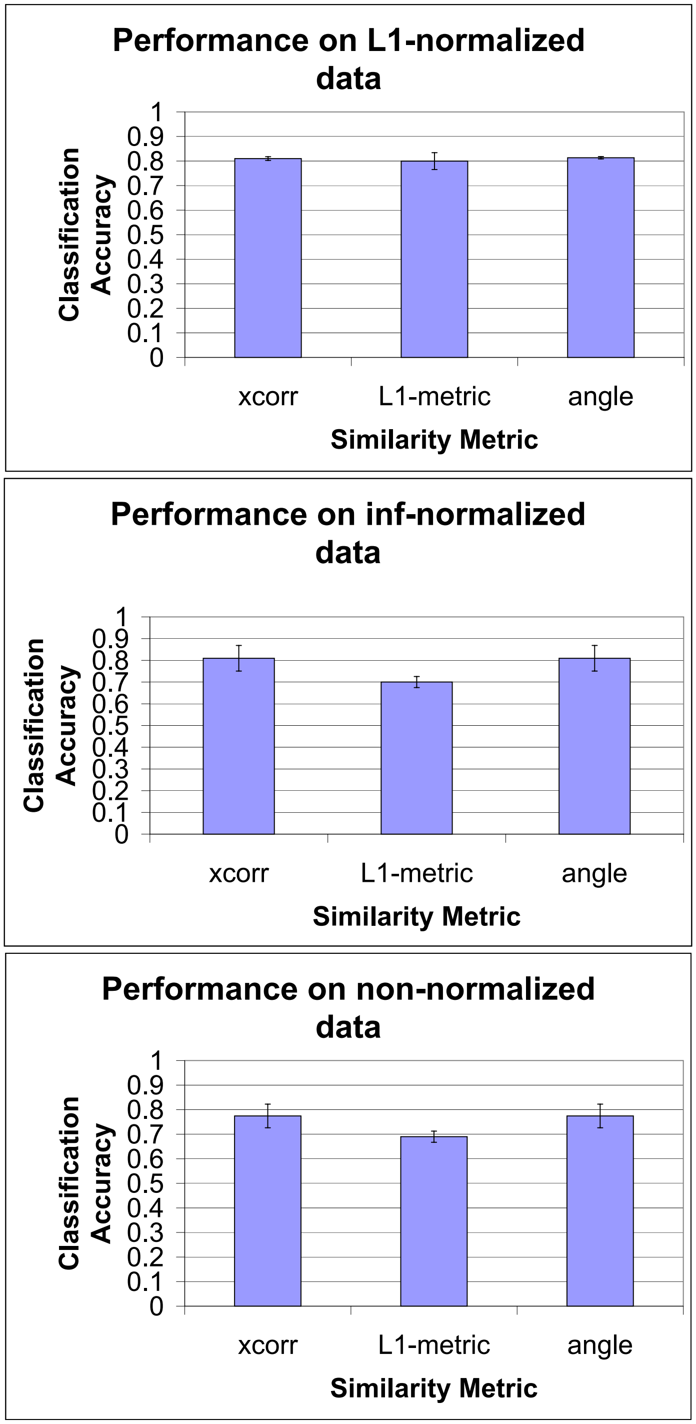
The experimental results section is split into several parts. First preliminary experiments are conducted to establish the difficulty each data set, provide a comparison of different normalization schemes, and establish which similarity metric is most appropriate for the centroid classification algorithm. The next section focuses on the dimensionality reduction by way of down sampling. The next set of experiments use statistical tests to rank each feature, then extract relevant features from the MS spectra using the Greedy selection procedure outlined in the previous section. Performance is once again reported based on the centroid classifier. After that attention is devoted to wrapper-approaches using Greedy and SFS algorithms. The final subsection provides a comparative analysis of techniques used in this and previous research endeavors.

For all experiments, each data set was split into 3 equal subsets. Each test fold used one of the three subsects with the remaining two subsets used for training and validation. Unless otherwise noted, the reported accuracy was the average classification accuracy over the three test folds and the error bars represent one standard deviation with respect to this average performance. Accuracy is taken as the arithmetic average of sensitivity and specificity.

4.1 Preliminary Experiments

The purpose of the preliminary experiments is to compare different normalization schemes, and similarity metrics. Furthermore these initial experiments enable the different data sets to be ranked in terms of classification difficulty.

For all experiments presented in this subsection the centroid classification method was used. For each data set, the following three different normalization schemes were tested: (i) L1-norm, (ii) infinity-norm, and (iii) identity (i.e., no normalization). For each norm the following three similarity metrics were selected: (a) correlation, (b) negative L1 distance, and (c) angular distance (see [**?**] for details). Figures 4-7 show the results. From the results we can see that the Ovarian cancer set appears to be the easiest, followed by the Heart/Kidney data set. Finally the two versions of the Prostate cancer data set appear to be the most difficult sets. The reason the prostate data is split is that the set contains four groups of samples: (i) Patients with normal PSA (biomarker) levels, (ii) Patients with benign growth and slightly elevated PSA levels. The last two groups , (iii) and (iv), represent patients with prostate cancer and differing PSA levels. Clearly, the normal and benign classes can be merged into one class and similarly both cancer groups can be merged into one class. We applied the centroid algorithm to both the four class split and the two class split. The two class split produced significantly better results than the four class split. In general detecting prostate cancer appears to be a very difficult task. In contrast to the Heart/Kidney data set that has a total of $164,168$ features, both the prostate and ovarian cancer data were acquired using a low resolution MS technique that produced only $15,154$ features. From the preliminary experiments we can see that the L1 normalization together with angular similarity metric produce the best results on the Heart/Kidney and Ovarian Cancer data sets (resp. 81.3% and 86.2% accuracy). For the prostate data in the 2-class case L1 normalization together with the negative L1 distance metric produced the best accuracy of 73.0%. For the 4-class case the infinity norm together with the angular distance produced the best accuracy of 64.8%. To simplify the rest of the experiments, we selected the L1 normalization procedure together with the angular distance in all subsequent experiments (unless stated otherwise). In addition, to reduce and standardize performance comparison only the 2-class prostate cancer problem will be considered in subsequent experiments. Before moving on, we note that for any given normalization procedure the results are **not** statistically significant across the similarity metrics used. The motivation for choosing the aforementioned parameter settings is the observation that for Heart/Kidney and Ovarian cancer data sets they produced the highest accuracy along with the lowest standard deviation.


Performance on L1-normalized data

Performance on inf-normalized data

Performance on non-normalized data

Fig. 4. Performance Graphs on the Heart/Kidney data. Each graph represents a different normalization approach, and each bar within a graph represents a different similarity metric. Best performance of 81.3% (±0.47%) was achieved using L1-norm with angle distance.

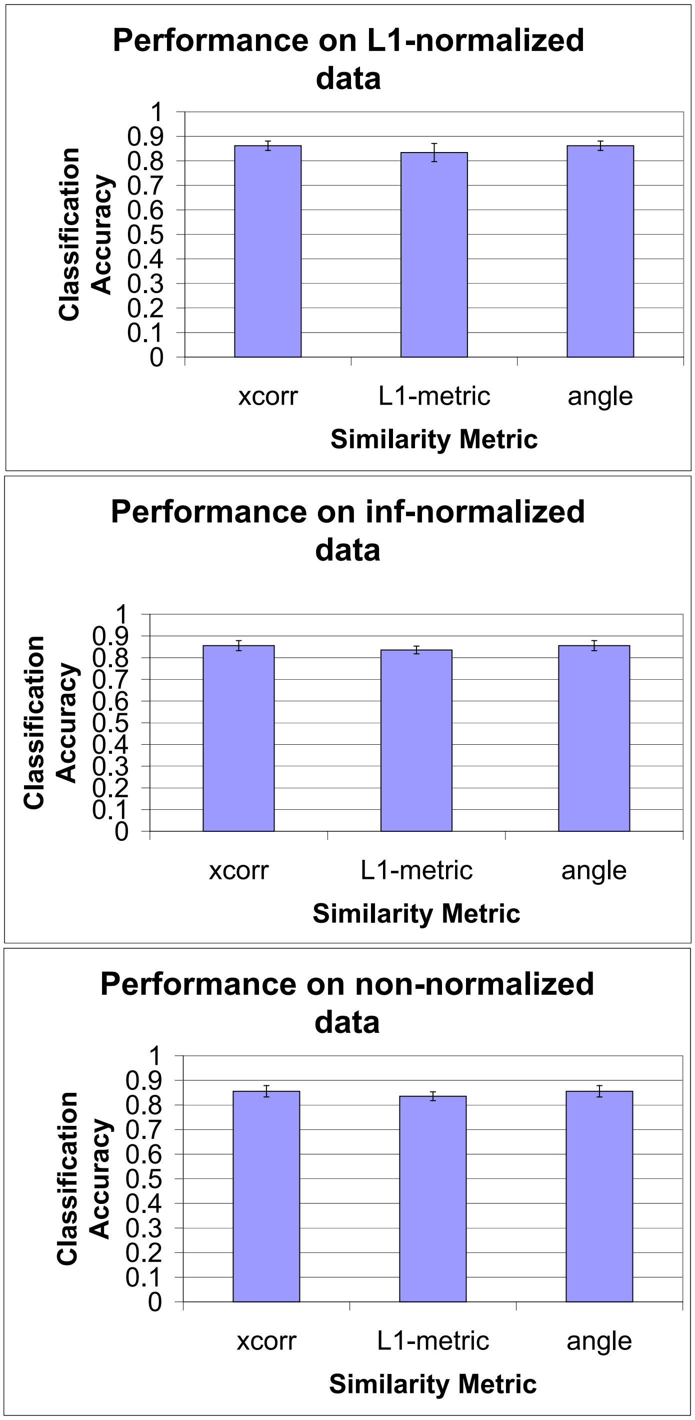


Fig. 5. Performance Graphs on the Ovarian Cancer data. Each graph represents a different normalization approach, and each bar within a graph represents a different similarity metric. Best performance of 86.2% (±1.7%) was achieved using L1-norm with angle distance.

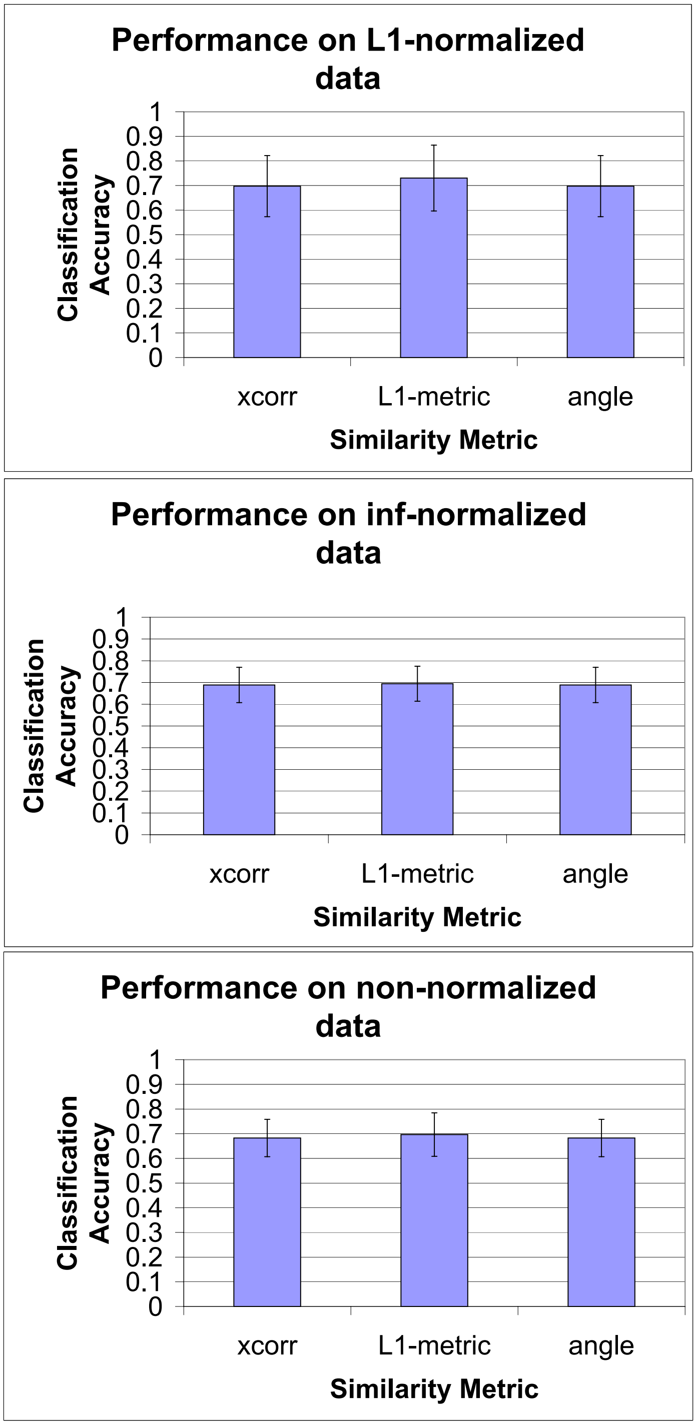
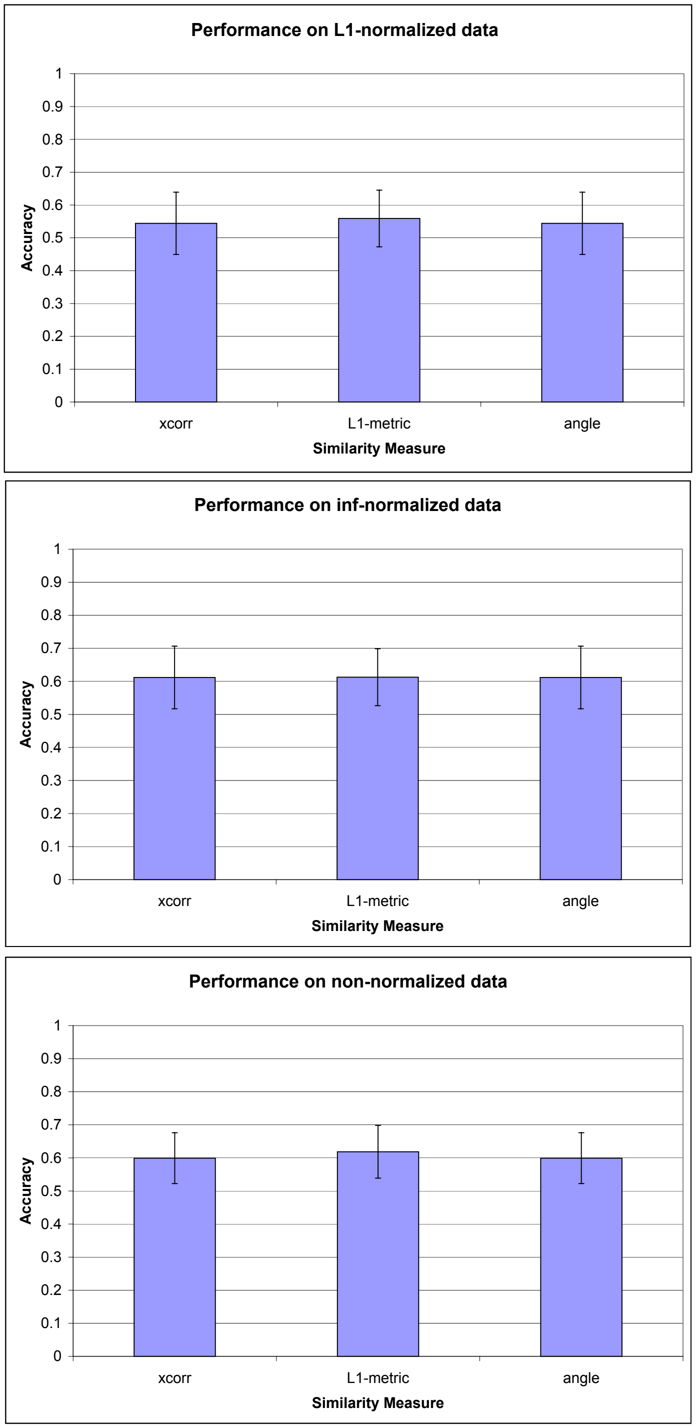


Fig. 6. Performance Graphs on the Prostate Cancer data split into 2 classes (Normal/Benign and Cancer1/Cancer2) . Each graph represents a different normalization approach, and each bar within a graph represents a different similarity metric. Best performance of 73.2% (±13.0%) was achieved using L1-norm with negative L1 distance.



Performance on L1-normalized data

Performance on inf-normalized data

Performance on non-normalized data

Fig. 7. Performance Graphs on the Prostate Cancer data split into 4 classes (Normal, Benign, Cancer1, Cancer2) . Each graph represents a different normalization approach, and each bar within a graph represents a different similarity metric. Best performance of 64.8% ($\pm 21.4\%$) was achieved using infinity-norm with angular distance.

4.2 Down Sampling Performance

To verify the conjecture made in [4] that higher resolution data produces better classification, we progressively down-sampled the spectra by averaging it out as previous studies have done. This effectively produced lower resolution data and at the same time reduced the data dimensionality. The results, presented in Figure 8 show that performance decrease as a factor of down-sampling. However, the decrease in clearly non-monotonic, and in essence noisy. This noise can be attributed to either the filtering or the sub-sampling stages of down-sampling. To see which of the two components produces the oscillations in classification accuracy a new experiment was needed.

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 high frequency components. The final stage transformed the filtered data back to spacial domain. The experiment varied the number of frequency coefficients used in reconstructing the MS spectra. In essence, this experiment considered feature selection in the frequency domain. Clearly the loss in accuracy, shown in Figure 9, is much more monotonic in comparison with the down-sampling method. This indicates that the majority of oscillations result form the sub-sampling step rather than the blurring step. However, the performance does decrease as a result of frequency filtering as in the case of down-sampling. This leads us to conclude that down sampling is in general detrimental to performance. To further validate this conclusion, we ran the centroid classifier on each individual feature for the down sampled spectra and found that performance again decreased when compared to performance using a single best feature from the non-down-sampled spectra. This further solidifies the claim that down sampling appears detrimental to classification accuracy. The conclusions drawn are in line with those in [4] where changes in resolution created by different MS techniques produced similar results. Recall that the MS spectra is really a histogram describing the ion concentrations based on the mass-to-charge ratios. The low resolution techniques effectively average together distinct ion concentrations into a single bin. Hence, down-sampling weather due to low-resolution MS hardware or or deliberately done in software appears to have the same effect of lowering performance.

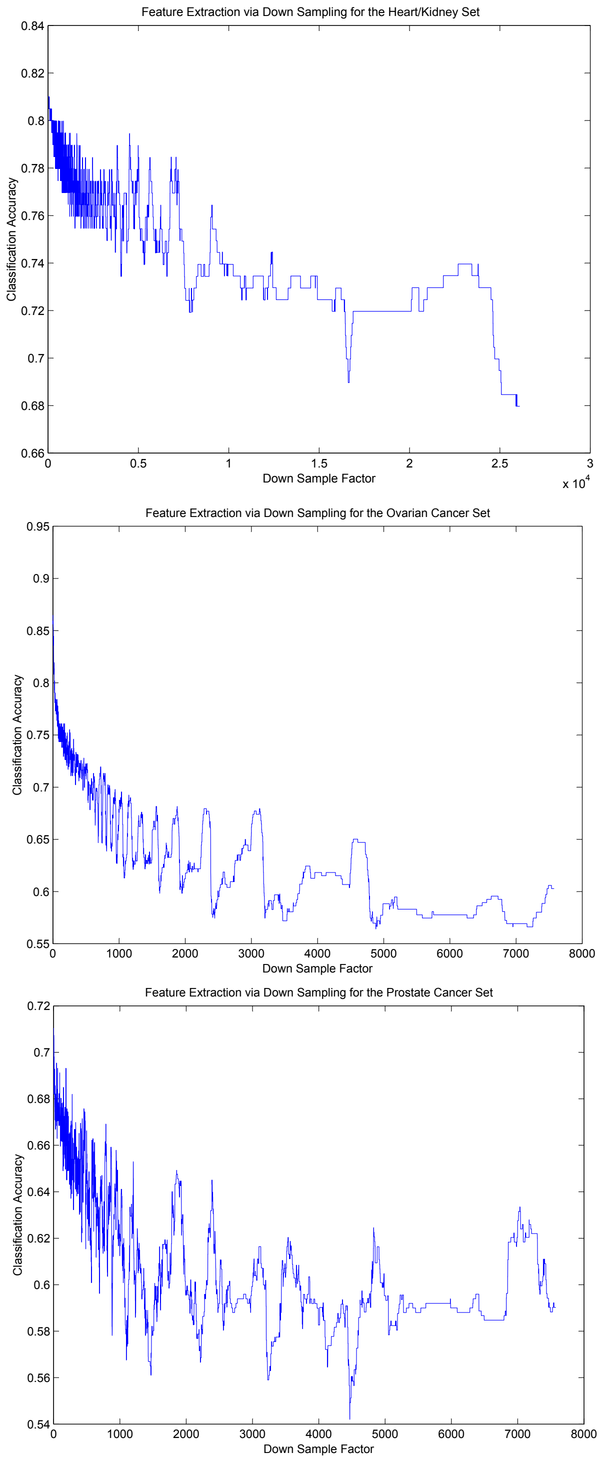


Fig. 8. Performance on progressively down-sampled data. **Top:** Heart/Kidney data set. **Middle:** Ovarian cancer data set. **Bottom:** Prostate Cancer data set. While all data sets exhibit oscillations, the performance nevertheless gradually declines as the dimensionality of the data is reduced (indicted by the increasing down-sample factor on the x-axis).

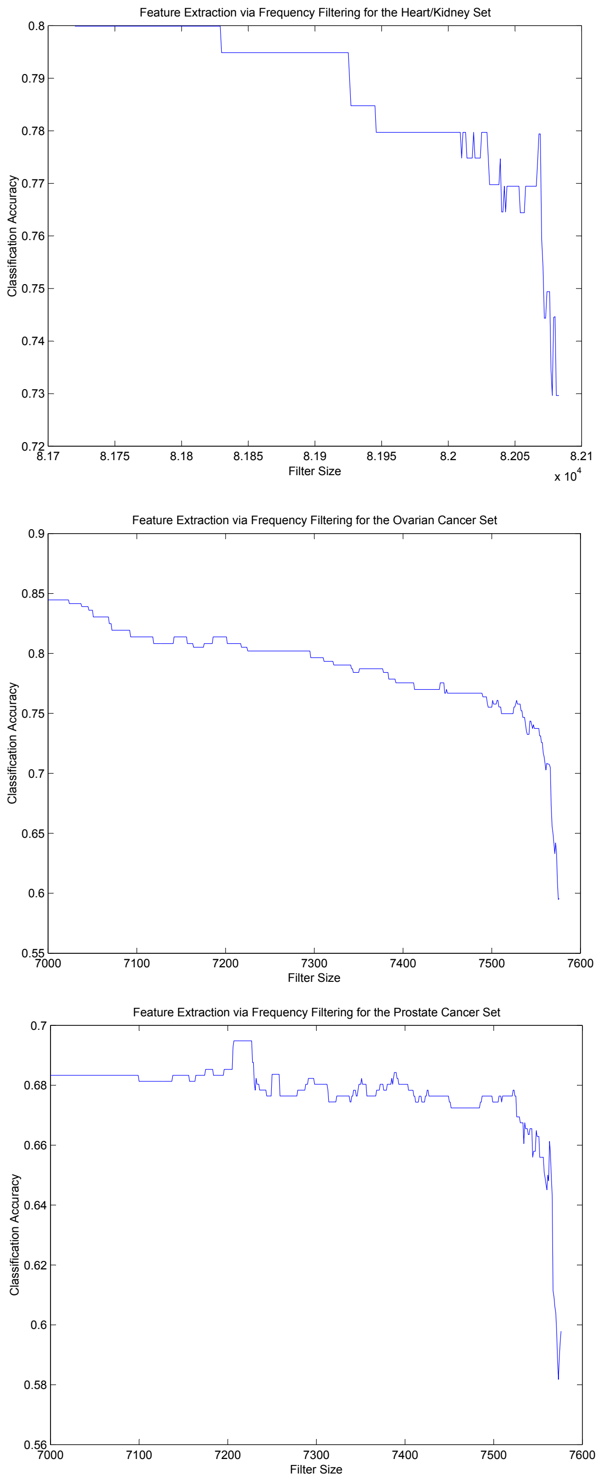


Fig. 9. Performance on frequency filtered data. **Top:** Heart/Kidney data set. **Middle:** Ovarian cancer data set. **Bottom:** Prostate Cancer data set. While all data sets exhibit oscillations, the performance nevertheless gradually declines as the lower and lower frequency components are deleted (indicted by the increasing filter size on the x-axis).

4.3 Statistical Tests with Greedy Feature Selection

As mentioned in section 3.4 on p. 14, we used the T-test, KS-test and the T*KS composite test to perform sequential forward selection (SFS) in order to establish which feature sets perform well on a given data set. In addition, each statistical technique is compared against each other. For all experiments the features were ranked using each algorithm, then the SFS procedure was used to select feature sets of increasing size and applying the centroid classification algorithm on the selected features. As usual 3-fold cross validation was performed for every feature set. The SFS procedure was stated with one top ranked feature and was run until 15,000 top ranked features were selected. After repeating the experiment for every statistical test on a given data set, we compared the best performances across different feature extraction techniques.

The experimental results, shown in Figures 10-12, demonstrate the significant improvement even using these simple statistical techniques. Unfortunately, they also indicate that there is no clearly superior statistical test for all data sets considered. On each data set the best feature extraction technique oscillates. The composite technique appears to be inferior to the two base procedures. On the positive side, for all data sets feature selection improved classification accuracy by over 10%. In addition, the number of features was reduced several orders of magnitude.

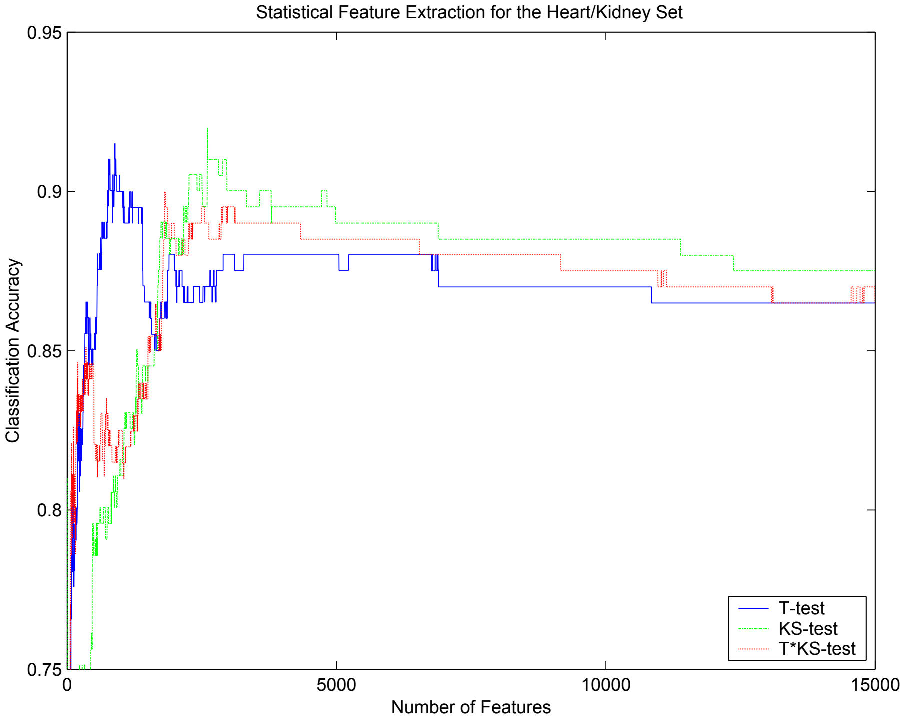


Fig. 10. Performance on the Heart/Kidney data using statistical tests for feature extraction and ranking. Sequential Forward Selection was then used on ranked features.

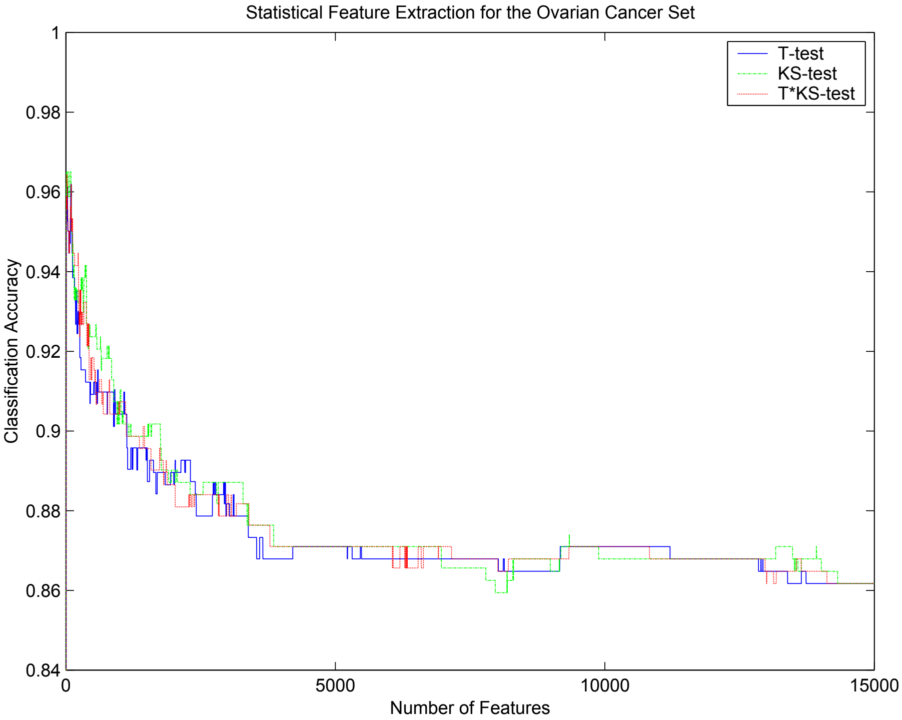


Fig. 11. Performance on the Ovarian data using statistical tests for feature extraction and ranking. Sequential Forward Selection was then used on ranked features.

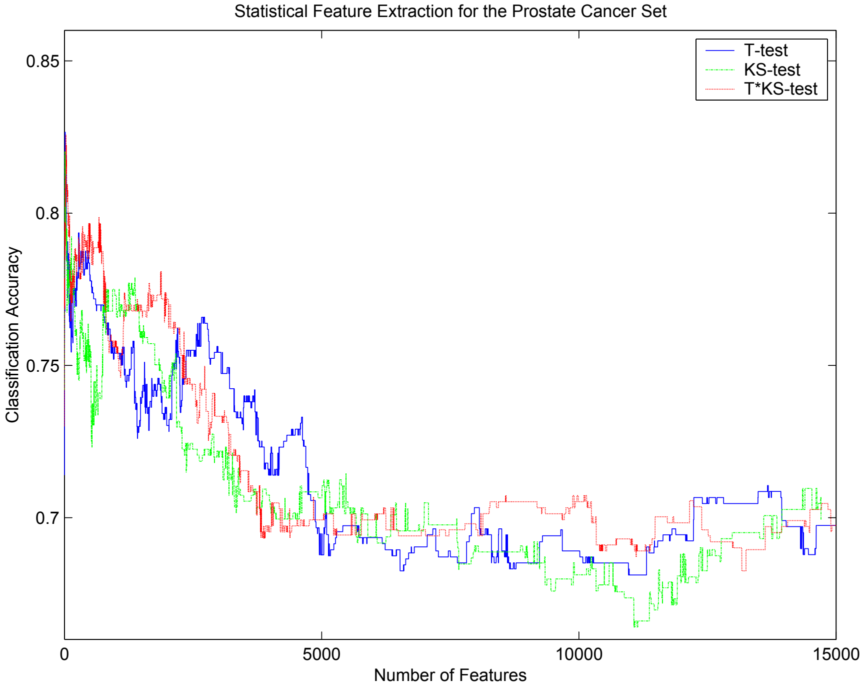


Fig. 12. Performance on the Prostate data using statistical tests for feature extraction and ranking. Sequential Forward Selection was then used on ranked features.

4.4 Greedy and SFS Feature Selection

To see the exact relevance of individual features, the centroid classifier was ran on individual features. Histogram plots for each data set are shown in figure 13. 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 redundant especially with the Ovarian and Prostate Cancer data sets that are skewed towards lower classification accuracies. This leads us to further question the down sampling approach which in essence aggregates individual features together. Such an approach would inevitably merge relevant and irrelevant (or redundant) features together and decrease the overall performance as evidenced by experimental results of the previous section. Interestingly, there are a number of features, within each data set that actually produce classification accuracies below 50%. These features can be definitively labeled as misleading and appear to be noise.

Once classification accuracy was established for each feature, the Greedy and SFS procedures were employed to select *relevant* feature sets. The results are presented in the next section.

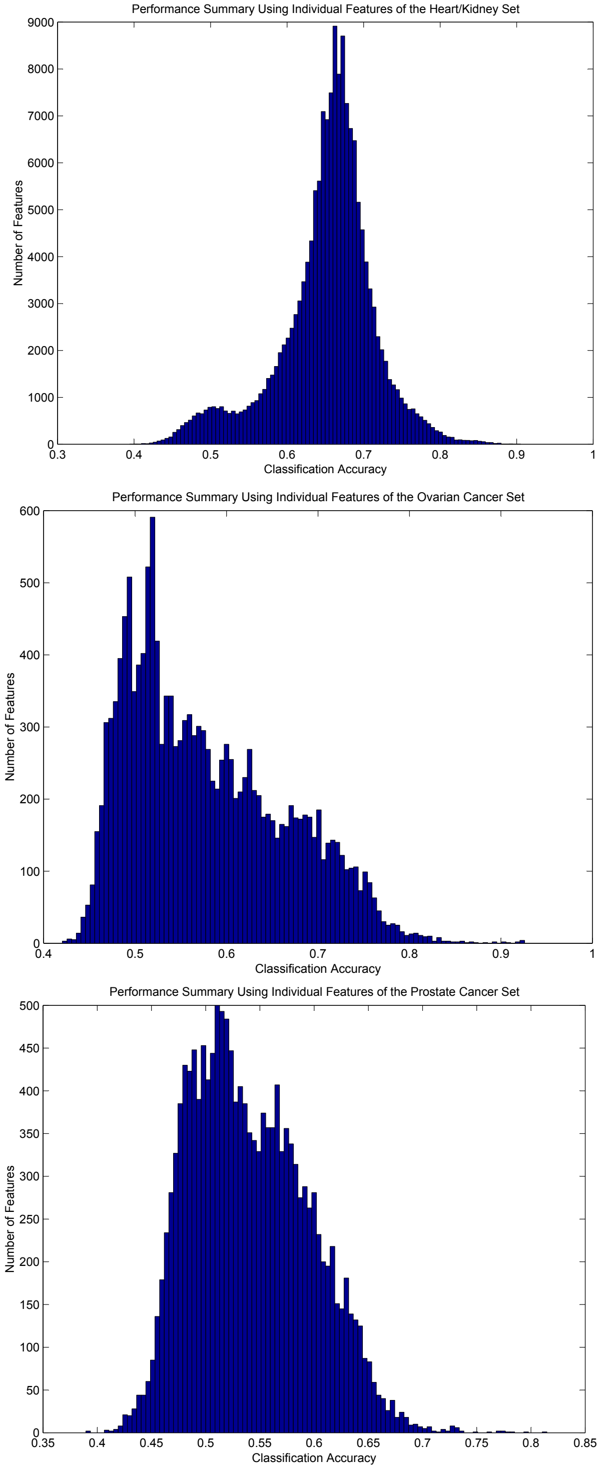


Fig. 13. Performance using individual features **Top:** Heart/Kidney data set. **Middle:** Ovarian cancer data set. **Bottom:** Prostate Cancer data set. The histograms show the number of features with a specific classification accuracy when individual features are used. Unfortunately the angle metric could not be properly used, instead the negative L1 metric was applied.

4.5 Performance Comparison

Figure 14 presents best performance for each feature extraction technique on each data set. In terms of classification accuracy, sequential forward selection clearly produces the most impressive results. In terms of producing the smallest feature sets, again the SFS feature selection procedure produced the best results. Although the graph shown in Figure 14 shows that Greedy selection on the ovarian cancer set produced set of 3 features at 94.5% accuracy, after looking through the data logs, we found that the 3 feature set selected by the SFS produced an accuracy of 96.0%. Therefore, we conclude that the **SFS algorithm produced superior results** in comparison all other algorithms tested (in terms of both smallest feature sets and highest performance). On the heart/kidney data set, SFS selected 5 features that enabled the centroid classifier to produce an accuracy of **97.5%**. On the Ovarian cancer data set the set of 4 features produced via SFS had the same accuracy of **98.0%**, tying with a 48 features set produced by the greedy procedure. Finally, 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 comparison with previous research, the SFS algorithm did reasonably well. Informally, we have been told that using genetic algorithms with self organizing maps on the Heart/Kidney data set, produced an accuracy of 92% (on a single test set). In comparison the SFS achieved a 97.5% classification accuracy. On ovarian cancer data set, PCA coupled with LDA in [20] to our knowledge produced the only perfect cross-validated (i.e., statistically valid) classification accuracy. The SFS based approach has an accuracy of 98.0% which is quite close. On the prostate cancer data set, researchers using PCA coupled with LDA in [20] produced an accuracy of 88%. In [28] the boosted decision stumps produced an impressive 98% accuracy on the same data set. Unfortunately, we were not able to get this set and used the data set from [11], where the accuracy using GA and SOM's was 83%. In contrast the SFS algorithm achieved an accuracy of 94% on this data set. Overall we believe that the SFS coupled with the centroid classifier is competitive with previously tried approaches.

Active Feature Sets To see the types of features extracted by the SFS we looked at the active set extracted by this procedure for the prostate cancer set. Table 1 shows a comparison of features extracted by our algorithms and those by algorithms used in previous research. Clearly, very few common features are present. At hypothesized in [6], it seems that different algorithms extract different relevant features based on their internal workings and biases.

Computational Complexity Analysis In terms of run time, by far the most efficient algorithm surveyed is PCA coupled with LDA. Principal Component Analysis has an $O(n^3)$ training cost, where n is the number of samples. Linear

Table 1. 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 [1, 28, 11]. Clearly the SFS procedure found a very different set of features 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**			

Discriminant Analysis has a training cost of $O(n^2 f)$, where f is the number of features. The total cost in training is then $O(n^3 + n^2 f)$. The total testing run-time, per sample, is $O(nf)$ and is dominated by the projection of the test sample onto the PCA basis. It should be noted that only n of f features can be used since LDA algorithm specifically requires the number of features to be less than the number of training samples. In fact this is why the PCA algorithm is used to reduce the dimensionality of the raw data. This in our view is not a major obstacle since in general ML algorithms will not be able to tune/learn more than n internal parameters given only n testing samples. However, we should remark that SFS/centroid combination, in contrast, can extract and use a feature set of arbitrary size independent of the training set size. To make the analysis fair, we therefore restrict the SFS feature selection algorithm to produce active set sizes less than or equal to n. With this restriction in place the training cost is the number of calls to the base classifier times the cost of each training session, given by

$$fO(n)+(f-1)O(2n)+...+(f-n-2)O(n(n-1))+(f-n-1)O(n^2) \quad (6)$$

$$= \sum_{i=1}^{n}(f-i-1)O(ni) \quad (7)$$

.

For $n << f$, the dominating term is $(f-n-1)O(n^2)$ and thus the cost is bounded by $O(fn^3)$. The testing cost is solely bounded by $O(n)$ since at most n features can now be extracted due to the artificial restriction placed on the SFS/centroid algorithm.

At first glance the PCA/LDA combination appears much more efficient in training, while SFS/centroid method is much more economical in the testing phase. Although worst case comparison shows this to be so, we should take a closer look at actual performance. The PCA/LDA combination has a fixed and unchanging training cost of $O(n^3 + n^2 f)$, regardless of data content. However, on the heart/kidney data set the SFS algorithm stopped just after processing 6 features[**]. Roughly speaking, the SFS algorithm processed each feature 6 times, for a total of $6 * 164,168 = 5,910,048$ calls to the centroid classifier. In turn the centroid classifier processed 66 training samples each at most 6 features in length at any point during the SFS process. Hence we have about 400 million operations needed to extract a relevant feature set. In contrast the PCA/LDA algorithm would need about $66^3 + 66^2(164,168) \approx$ 700 million operations.

On the prostate cancer set the PCA/LDA algorithm would need about $84^3 + 84^2(15,154) \approx$ 110 million operations. On the other hand SFS/centroid combination extracted 11 features and made $12 * 15,154 = 181,848$ calls to the centroid classifier which cost at most $12 * 84$ operations. So the total cost of SFS/centoid method is about 183 million operations.

In practice we believe that the costs of both algorithms are comparable for high dimensionality data sets used in this study.

[**] The sixth feature did not improve performance so the algorithm terminated and returned the a feature set containing only 5 features.

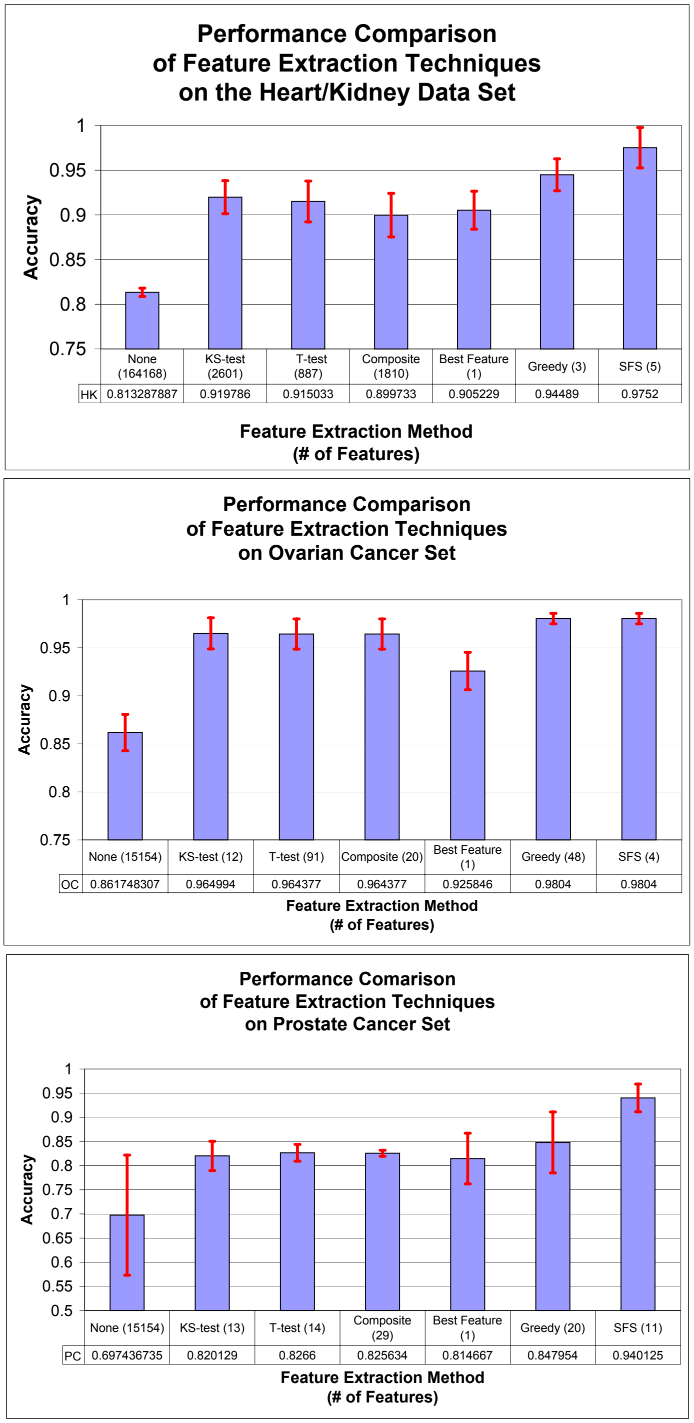


Fig. 14. Performance of Feature Extraction Algorithms **Top:** Heart/Kidney data set. **Middle:** Ovarian cancer data set. **Bottom:** Prostate Cancer data set.

5 Conclusion

Proteomic Pattern analysis is an emerging field, poised to improve the quality of medical diagnosis. However, the large dimensionality of the data samples requires the use of aggressive feature extraction techniques. This project analyzed statistical, multi-resolution, and wrapper approaches to feature extraction. Experimental results indicate that down sampling appears detrimental to classification performance. On the other hand feature selection techniques, such as sequential forward selection, can greatly reduce the dimensionality of the data while at the same time improving classification accuracy. Clearly, in the future this technique can be implemented into a real screening system, poised to revolutionize the field of early medical diagnosis.

A Mass Spectrometry: A brief Overview

Discovered by Sir J.J. Thomson in the early part of the 20^{th} century Mass spectrometry (or MS) is a technique for ’weighting’ individual molecules, fragments of molecules or individual atoms that have been ionized. In a vacuum environment, a mass spectrometer deflects the previously charged (unknown) particles in a magnetic or electric field. While different types of MS techniques exist, all use an ion source that vaporizes and changes the unknown matter. The mass spectrometer measures the molecular masses along with abundances and masses of fragments that are produced as a result of molecular breakdown. Prior to making the MS measurement a chromatograph may be used to separate the complicated mixture of compounds present in a sample into constituents.

The fundamental measurement unit of the MS is the mass-to-change ratio. For proteomic applications, Daltons (Da) are used to measure mass, with 1Da representing the atomic mass of carbon-12. The electric potential of a single electron is the measurement unit for charge (z). Thus, the mass-to-charge ratio (m/z) represents Daltons per fundamental unit of charge for each protein and/or fragment.

Depicted in Figure 15, the sample, which may be a solid, liquid, or vapor, enters the vacuum chamber through an inlet. The gas phase ions are sorted in the mass analyzer according to their mass-to-charge (m/z) ratios and then collected by a detector. In the detector the ion flux is converted to a proportional electrical current. The data system records the magnitude of these electrical signals as a function of m/z and converts this information into a mass spectrum. The spectrum is a graph of ion intensity as a function of mass-to-charge ratio and is often depicted as a histogram (see Figure 2).

A.1 Time-of-Flight (TOF)

In time-of-flight (TOF) instruments, positive ions are produced periodically by bombardment of the sample with brief pulses of electrons, secondary ions, or laser-generated photons. The ions produced by the laser are then accelerated by an electric field pulse. The accelerated particles then pass into a field-free drift tube. All ions entering the tube ideally have the same kinetic energies, their velocities in the tube must vary inversely with their masses, with lighter particles arriving at the detector earlier than the heavier ones. The ions therefore drift through a field-free path and are separated in space and time-of-flight [5]. Figure 16 shows the process. Other types of mass spectrometers include: Magnetic Sector Instruments, Quadrupole Instruments, Ion Trap, and Fourier Transform Ion Cyclotron Resonance (FT-ICR) (see [9] for more details).

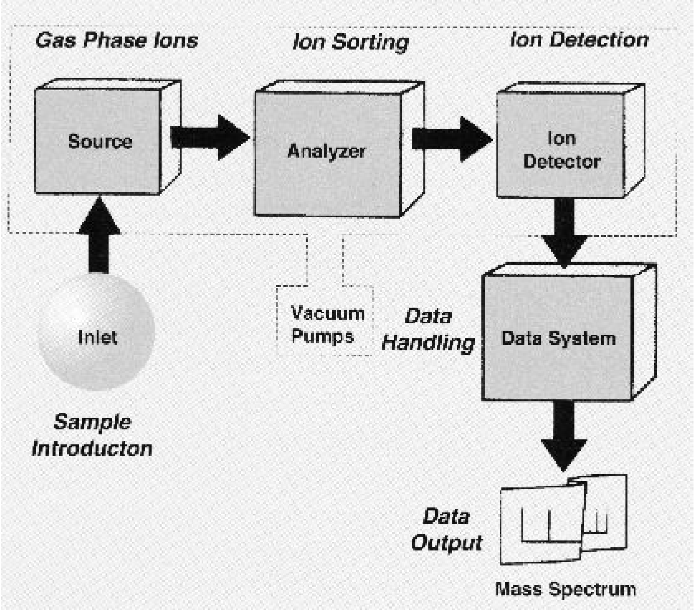


Fig. 15. Mass Spectrometry Process. From [9]

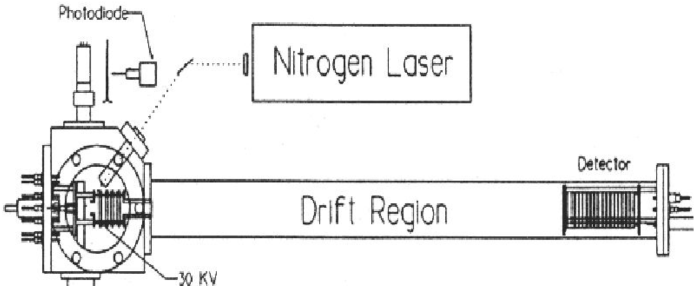


Fig. 16. Time-of-flight Ion Detection System. From [5].

A.2 MALDI

The use of the matrix assisted laser desorption/ionization (MALDI) serves several purposes. The (bio)molecules are incorporated in a large excess of matrix molecules, strong intermolecular forces are thereby reduced. The matrix molecules absorb the energy from the laser light and transfer it into excitation energy of the solid system. The effect is an instantaneous phase transition of small molecular layers of the sample into a gaseous state. Thus solid (and liquid) material can be easily analyzed by TOF MS.

A.3 SELDI

This method uses protein chip arrays with different selective surfaces such as cation or anion exchange surfaces, hydrophobic surfaces and metal binding surfaces. Also antibodies, specific proteins and DNA can be bound to chips to study protein-protein and protein-DNA interactions. **Cell lysate, plasma or urine** is applied on the selective surface and, after washing, a subset of proteins is specifically bound. The chip is analyzed in a (MALDI) TOF-MS which generates a protein spectrum of the different molecular masses present on the protein chip. This technology is therefore highly suited for research into molecular mechanisms of disease and biomarker identification.

B Acquired Data sets

For this study three data sets were acquired. Each sample in each data set is represented as a vector of real valued features forming the spectra. Each feature in turn represents the quantity (ppm[††]) of ions with a specific m/z ratio. In essence, each sample spectrum is a histogram describing the composition of the sample bio-fluid or tissue sample.

The following data are used for this study:

Heart-Kidney This data set consists of mass spectra from heart and kidney tissue samples acquired using the MALDI[‡‡]-TOF MS technique. Although this is assumed to be a relatively easy data, each sample is composed of 164,168 features and therefore provides a challenge to machine learning and feature extraction algorithm. There are 100 heart and 100 kidney samples in this set.

Ovarian Cancer This is the latest data set using the WCX2 protein array. A new set of ovarian samples were used. The sample set included 91 controls and 162 ovarian cancers. The entire process of preparing the (SELDI) chips was done using a robotic instrument. Acquired from [8], each data sample is composed of 15,156 features.

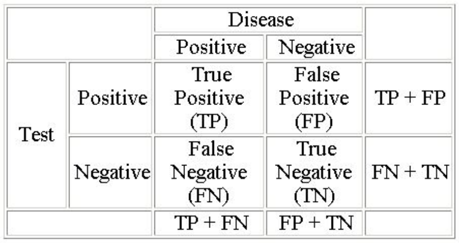
Prostate Cancer The spectra were collected using the (SELDI) H4 protein chip. The chip was prepared by hand using the recommended protocol. The spectra were exported with the baseline subtracted. This process creates negative intensities. There are 322 total samples, acquired from [8]: 190 samples with benign prostate with PSA levels greater than 4, 63 samples with no evidence of disease and PSA level less than 1, 26 samples with prostate cancer with PSA levels 4 through 10, 43 samples with prostate cancer with PSA levels greater than 10. Each sample is again a histogram composed of 15,156 features.

[††] ppm - parts per million

[‡‡] MALDI - matrix-assisted laser desorption/ionization

C Definitions

The following diagnostic definitions are used by the community.



		Disease		
		Positive	Negative	
Test	Positive	True Positive (TP)	False Positive (FP)	TP + FP
	Negative	False Negative (FN)	True Negative (TN)	FN + TN
		TP + FN	FP + TN	

Fig. 17. Diagnostic Definitions

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

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

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

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

LR+ (Likelihood Ratio Positive) $\frac{Sensitivity}{1-Specificity} = \frac{\frac{TP}{TP+FN}}{1-\frac{TN}{TN+FP}} = \frac{\frac{TP}{TP+FN}}{\frac{FP}{TN+FP}}$.

LR- (Likelihood Ratio Negative) $\frac{1-Sensitivity}{Specificity} = \frac{1-\frac{TP}{TP+FN}}{\frac{TN}{TN+FP}} = \frac{\frac{FN}{TP+FN}}{\frac{TN}{TN+FP}}$.

Accuracy In this report accuracy is defined as the average of sensitivity and specificity. That is $\frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FP}}{2}$.

References

1. Bao-Ling Adam, Yinsheng Qu, John W. Davis, Michael D. Ward, Mary Ann Clements, Lisa H. Cazares, O. John Semmes, Paul F. Schellhammer, Yutaka Yasui, Ziding Feng, and Jr. George 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.
2. Keith A. Baggerly, Jeffrey S. Morris, Jing Wang, David Gold, Lian-Chun Xiao, and Kevin R. Coombes. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, 3:1667–1672, 2003.
3. J. Bins and B. Draper. Feature selection from huge feature sets. In *Proceedings of International Conference on Computer Vision*, volume 2, pages 159–165, 2001.
4. Thomas P. Conrads, Ming Zhou, Emmanuel F. Petricoin III, Lance Liotta, and Timothy D. Veenstra. Cancer diagnosis using proteomic patterns. *Expert Reviews in Molecular Diagnostics*, 3(4):411–420, 2003.
5. R. J. Cotter. *Time-of-Flight Mass Spectrometry*. American Chemical Society, Washington, DC, 1994.
6. E. Diamandis. Proteomic patterns in biological fluinds: Do they represent the future of cancer diagnostics. *Clinical Chemistry (Point/CounterPoint)*, 48(8):1272–1278, 2003.
7. Melanie Hilario, Alexandros Kalousis, Markus Mller, and Christian Pellegrini. Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics*, 3:1716–1719, 2003.
8. http://ncifdaproteomics.com/ppatterns.php. Clinical proteomics program databank. Technical report, National Cancer Institute, Center for Cancer Research, NCI-FDA Clinical Proteomics Program, 2003.
9. http://www.bmsf.unsw.edu.au/about/index.html. Overview of mass spectrometry. Technical report, University of New South Wales/BMSF, 2002.
10. Emanuel F. Petricoin III, Ali M. Ardekani, Ben A. Hitt, Peter J. Levine, Vincent A. Fusaro, Seth M. Steinberg, Gordon B. Mills, Charles Simone, David A. Fishman, Elise C. Kohn, and Lance A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359(9306):572–577, 2002.
11. Emanuel F. Petricoin III, 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, 2002.
12. Xin Feng Jacob W. Tatay, Nancy Sobczak, Hao Jiang, Chin-Fu Chen, Roumyana Kirova, Craig Struble, Nan Jiang Wang, and Peter J. Tonellato. Multiple approaches to datamining of proteomic data based on statistical and pattern classification methods. *Proteomics*, 3:1704–1709, 2003.
13. C. Kainz. Early detection and preoperative diagnosis of ovarian carcinoma (article in german). *Wien Med Wochenschr*, 146(1–2):2–7, 1996.
14. Michael Kirby. *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. John Wiley & Sons, New York, 2001.
15. R. Kohavi, P. Langley, and Y. Yun. The utility of feature weighting in nearest-neighbor algorithms, 1997.
16. Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
17. Igor Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning*, pages 171–182, 1994.
18. Kwan R. Lee, Xiwu Lin, Daniel C. Park, and Sergio Eslava. Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics*, 3:1680–1686, 2003.

19. Jinong Li, Zhen Zhang, Jason Rosenzweig, Young Y. Wang, and Daniel W. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, 48(8):1296–1304, 2002.
20. Ryan H. Lilien, Hany Farid, and Bruce R. Donald. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *Computational Biology (In Press)*, 2003.
21. Mia K. Markey, Georgia D. Tourassi, and Carey E. Floyd Jr. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. *Proteomics*, 3:1678–1679, 2003.
22. Padraic Neville, Pei-Yi Tan, Geoffrey Mann, and Russ Wolfinger. Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum. *Proteomics*, 3:1710–1715, 2003.
23. 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.
24. S. D. Patterson and R. H. Aebersold. Proteomics: The first decade and beyond. *Nature, Genetics Supplement*, 33:311–323, 2003.
25. 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.
26. P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature-selection. *PRL*, 15(11):1119–1125, November 1994.
27. Parul V. Purohit and David M. Rocke. Discriminant models for high-throughput proteomics mass spectrometer data. *Proteomics*, 3:1699–1703, 2003.
28. Yinsheng Qu, Bao-Ling Adam, Yutaka Yasui, Michael D. Ward, Lisa H. Cazares, Paul F. Schellhammer, Ziding Feng, O. John Semmes, and Jr. George 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.
29. Douglas J. Slotta, Lenwood S. Heath, Naren Ramakrishnan, Rich Helm, and Malcolm Potts. Clustering mass spectrometry data using order statistics. *Proteomics*, 3:1687–1691, 2003.
30. Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2003.
31. Michael Wagner, Dayanand Naik, and Alex Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3:1692–1698, 2003.
32. Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor3, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classication of ovarian cancer using mass spectrometry data. *BioInformatics*, 19(13), 2003.
33. J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin. Proteomic applications for the early detection of cancer. *Nature Reviewes*, 3:267–275, 2003.
34. Hongtu Zhu, Chang-Yung Yu, and Heping Zhang. Tree-based disease classification using protein data. *Proteomics*, 3:1673–1677, 2003.