SVM-RFE PEAK SELECTION FOR CANCER CLASSIFICATION WITH MASS SPECTROMETRY DATA

KAIBO DUAN AND JAGATH C. RAJAPAKSE

BioInformatics Research Centre
School of Computer Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
E-mail: {askbduan, asjagath}@ntu.edu.sg

We studied two cancer classification problems with mass spectrometry data and used SVM-RFE to select a small subset of peaks as input variables for the classification. Our study shows that, SVM-RFE can select a good small subset of peaks with which the classifier achieves high prediction accuracy and the performance is much better than with the feature subset selected by T-statistics. We also found that, the best peak subset selected by SVM-RFE always have in the top ranked peaks by T-statistics while it includes some peaks that are ranked low by T-statistics. However, these peaks together give much better classification performance than the same number of most top ranked peaks by T-statistics. Our experimental comparison of the performance of Support Vector Machine classification algorithm with and without peak selection also consolidates the importance of peak selection for cancer classification with mass spectrometry data. Selecting a small subset of peaks not only improves the efficiency of the classification algorithms, but also improves the cancer classification accuracy, even for classification algorithms like Support Vector Machines, which are capable of handling large number of input variables.

1. Introduction

In the last decade or so, mass spectrometry (MS) has increasingly become the method of choice for analysis of complex protein samples. Mass spectrometry measures two properties of ion mixtures in the gas phase under a vacuum environment: the mass-to-charge ratio (m/z) of ions in the mixture; and the number of ions present at different m/z values. The output is a *mass spectrum* or chart with a series of spike peaks, each representing the ion(s) of a specific m/z value present in the sample. The heights of the peaks are related to the abundances of the ions in the sample. The heights of peaks and the m/z values of peaks are a fingerprint of the sample. For protein samples, mass spectrometry measures the mass-to-charge ratio of the ionized proteins (or protein fragments) and their abundances in the sample. The recent advances in mass spectrometry technology are starting to enable high-throughput profiling of the protein content of complex samples.

While mass spectrometry has been used intensively on purified, digested samples to identify proteins via peptide mass fingerprints,[1] recently, it has also found promising applications in cancer classification.[2−4] Proteins vary between individuals, between cell types, and in the same cell under different stimuli or different disease states. Thus, the protein

variations between cancerous samples and noncancerous samples, or between different stages of a cancer provide rich and dynamic information to discriminate cancerous samples from non-cancer samples or to discriminate between different stages of a cancer. The protein abundance changes are relatively easy to measure, especially with the recent rapid advances in mass spectrometry technology, and thus are used as feature variables for cancer classification, although the rich information contained in protein variation is not confined only to changes in abundance. For cancer classification, the protein samples from cancer patients and non-cancer patients or from different cancer stages are analyzed through mass spectrometry instruments and the mass spectrometry patterns are used to built a diagnostic classifier. However, the raw mass spectra must go through some basic preprocessing steps like baseline identification and subtraction, peak identification and extraction, intensity normalization, and peak selection etc., before them are used to build a cancer classifier. [12] Blood serum is often used as the source of protein samples for cancer classification. Blood serum constantly perfuses tissues and circulates throughout the body and thus archives rich and dynamic histological information of proteins. Besides, it also can be easily and non-invasively obtained in sufficient quantity from patient at clinics.

For cancer classification with mass spectrometry data, the peak selection step is especially important. Peak selection procedure tries to select from the original mass spectra a set of peaks that are mostly relevant to the phenotypes under study, or a subset of peaks together will form better input variables for the classification algorithm. Nowadays, for most of the MS data for cancer classification, the number of training samples (cancer or non-cancer cases) is small compared to the large number of inputs (peak intensities). When the number of input variables is significantly greater than the number of training samples, random correlation between the inputs and the phenotypes may be formed. Finding a compact small set of input variables is important as well for protecting against such spurious results. Peak selection is exactly the feature/variable selection problem commonly addressed in machine learning.[5] [6] Some statistical and machine learning methods have been used for peak selection purpose, for examples, genetic algorithm,[2] signal-to-noise ratio,[4] ROC curve criterion,[3] etc.

SVM-RFE (Support Vector Machine Recursive Feature Elimination) was originally proposed for gene selection,[7] where a linear version of the popular Support Vector Machine (SVM) methods [8] [9] is used as the learning algorithm in a recursive procedure to select a subset of genes for cancer classification. In this paper, we will study the usefulness of SVM-RFE for peak selection for cancer classification with mass spectrometry data. For comparison, we also include the T-statistics feature selection method, which chooses a set of features that are most relevant to the concept under study. The goodness of the selected peak subsets, in this study, are evaluated by the classification performance of a linear SVM classifier with only the selected peaks as input variables. However, ultimately, peaks in the selected subset have to be examined by biological experiments. These peaks should be further analyzed to identify the underlying proteins. The subsequent functional study of the identified proteins may help to get new biological insights into the disease pathways and may eventually lead to reliable diagnostic test methods and potential therapeutic targets.

The rest of paper is organized as follows: in Section 2, we briefly review the SVM classification methods, SVM-RFE and T-statistics feature selection methods; in Section 3, we describe the numerical experiments; in Section 4, we analyze the experiment results and make the conclusions.

2. SVM, SVM-RFE and T-Statistics

In this section we briefly review the SVM classification method, SVM-RFE and T-statistics feature selection methods.

2.1. *SVM*

Support Vector Machines [8] [9] have been very popular in solving classification problems. It constructs an optimal hyperplane decision function in a so-called *feature space* that is mapped from the original input space. The mapping Φ is usually nonlinear and the feature space is usually a much higher dimensional space than the original input space. Let us use $\mathbf{x}_i$ to denote the ith example vector in the original input space and $\mathbf{z}_i$ to denote the corresponding vector in the feature space, $\mathbf{z}_i = \Phi(\mathbf{x}_i)$. *Kernel* is one of the core concepts in SVMs and plays an very important role. Kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ computes the inner product of two vectors in the feature space and thus implicitly defines the mapping function: $k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) = \mathbf{z}_i \cdot \mathbf{z}_j$.

The following are three types of commonly used kernel functions

$$\begin{aligned} \text{Linear Kernel} \quad & k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j \\ \text{Polynomial Kernel} \quad & k(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^p \\ \text{Gaussian Kernel} \quad & k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2) \end{aligned}$$

where the order p of polynomial kernel and the spread width σ of Gaussian kernel are adjustable kernel function parameters.

For a typical classification problem with ℓ training samples $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_\ell, y_\ell)$ where y_i denotes the class label of $\mathbf{x}_i$ and $y_i \in \{+1, -1\}$, finding the discriminant function $f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b$ is formulated by SVMs into the following optimization problem

$$\min_{\mathbf{w}, b, \xi_i} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 + C \textstyle\sum_{i=1}^{\ell} \xi_i \tag{1}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{z}_i + b) \geq 1 - \xi_i \,, \quad \xi_i \geq 0 \tag{2}$$

where $C > 0$ is another predefined higher-level parameter, besides the kernel function parameters. This optimization problem is usually solved in its dual form [9]

$$\min_{\alpha_i} \quad \tfrac{1}{2} \textstyle\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{z}_i \cdot \mathbf{z}_j) - \sum_{i=1}^{\ell} \alpha_i \tag{3}$$

$$\text{subject to} \qquad 0 \leq \alpha_i \leq C, \quad \textstyle\sum_{i=1}^{\ell} \alpha_i y_i = 0 \tag{4}$$

The weight vector $\mathbf{w}$ and the hyperplane decision function can be expressed by using the dual variables α_i's:

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{z}_i \tag{5}$$

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i (\mathbf{z} \cdot \mathbf{z}_i) + b \tag{6}$$

If a nonlinear kernel is used, because of the nonlinear mapping relation between the input space and the feature space, the linear discriminant function constructed by an SVM in the feature space corresponds to a nonlinear function in the original input space. The function family richness and the discriminant power of SVMs are thus incorporated in by the mapping function and ultimately the kernel function, while problem formulation is kept in the same and neat form.

In the dual problem of SVMs, all the computation involving the input vectors is in the form of inner products of vectors in feature space. The discriminant function also can be expressed in inner products of feature space vectors. These inner products $(\mathbf{z}_i \cdot \mathbf{z}_j)$ can be replaced by corresponding kernel computations $k(\mathbf{x}_i, \mathbf{x}_j)$, which can be executed easily in the original input space. Thus, we usually do not need to know the mapping function Φ explicitly. It is implicitly defined by the kernel function that computes the inner product in the feature space. Similarly, we do not need to explicitly compute the weight vector $\mathbf{w}$. However, if a linear kernel is used, the decision function $f(\mathbf{x})$ is simply a linear function of $\mathbf{x}$ and the weight vector of the linear function also can be explicitly computed as

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{x}_i. \tag{7}$$

SVMs with linear kernel are often referred to as linear SVMs.

2.2. *SVM-RFE*

Support Vector Machine Recursive Feature Elimination (SVM-RFE) method was originally proposed to perform gene selection for cancer classification.[7] Nested subsets of features are selected in a sequential backward elimination manner, which starts with all the features and remove one feature each time. In this way, in the end, all the feature variables are ranked. At each step, the coefficients of the weight vector $\mathbf{w}$ of a linear SVM are used as the feature ranking criterion. The recursive elimination procedure used in Ref. 7 is as follows:

(1) Start: ranked feature $R = [\ ]$; selected subset $S = [1, \cdots, d]$;
(2) Repeat until all features are ranked:
    (a) Train a linear SVM with all the training data and variables in S;
    (b) Compute the weight vector using Eq. (5);
    (c) Compute the ranking scores for features in S: $c_i = (w_i)^2$;
    (d) Find the feature with the smallest ranking score: $e = \arg\min_i c_i$;
    (e) Update R: $R = R[e, R]$;
    (f) Update S: $S = S - [e]$;
(3) Output: Ranked feature list R

For speed reasons, the algorithm can be generalized to remove more than one feature per step.[7] However, removing several features may degrade the classification performance.

Note that, in SVM-RFE,[7] the following SVM formulation is used

$$\min_{\mathbf{w},b,\xi_i} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}\xi_i^2 \tag{8}$$

$$\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{z}_i + b) \geq 1 - \xi_i\ , \quad \xi_i \geq 0 \tag{9}$$

This formulation of SVM is usually solve by the following dual problem with a slightly modified kernel function $\tilde{k}(\cdot,\cdot)$

$$\min_{\alpha_i} \ \tfrac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j \tilde{k}(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{\ell}\alpha_i \tag{10}$$

$$\text{subject to} \qquad \alpha_i \geq 0, \quad \sum_{i=1}^{\ell}\alpha_i y_i = 0 \tag{11}$$

where $\tilde{k}(\mathbf{x}_i,\mathbf{x}_j = k(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{C}\delta_{i,j}$; $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ otherwise.

Using w_i^2 as ranking score corresponds to removing the feature whose removal change the objective function least. This objective function is chosen to be $(1/2)|\mathbf{w}\|^2$ in SVM-RFE. This is explained by the OBD algorithm,[10] which approximates the change in objective function caused by removing the ith feature by expanding the objective function in Taylor series to second order

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2 \tag{12}$$

At the optimum of J, the first order term can be neglected and with $J = (1/2)\|\mathbf{w}\|^2$ Equation (12) becomes

$$\Delta J(i) = (\Delta w_i)^2 \tag{13}$$

$\Delta w_i = w_i$ corresponds to removing the ith feature.

Another explanation of using w_i^2 as ranking score is from the sensitivity analysis of the objective function $J = (1/2)\|\mathbf{w}\|^2$ with respect to a variable. To compute the gradient, a virtual scaling factor ν is introduced into the kernel function[11] and $k(\mathbf{x}_i,\mathbf{x}_j)$ becomes

$$k(\nu\cdot\mathbf{x}_i, \nu\cdot\mathbf{x}_j) \tag{14}$$

For a linear SVM (with a linear kernel function), using the fact that $\nu_k = 1$, the sensitivity can be computed as

$$\frac{\partial J}{\partial \nu_k} = \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j \frac{\partial k(\mathbf{x}_i,\mathbf{x}_j)}{\partial \nu_k} \tag{15}$$

$$= \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j (2\nu_k x_k^2) \tag{16}$$

$$= w_k^2 \tag{17}$$

2.3. *T-Statistics*

T-statistics basically is a filters feature selection method. It selects the feature variables that are most relevant to the concept under study. A ranking score is computed for each feature. It uses the following feature ranking criterion

$$c_i = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}}. \tag{18}$$

where μ_i^+ and μ_i^- are the mean values of the ith feature respectively over positive and negative samples; σ_i^+ and σ_i^- are the corresponding standard deviations; n^+ and n^- denote the number of positive and negative training samples. Equation (18) fundamentally measures the normalized feature value difference between two groups.

3. Numerical Experiments

We evaluate the SVM-RFE peak selection method together with T-statistics method on two cancer classification mass spectrometry datasets: Lung Cancer (Lung) and Ovarian Cancer (Ovarian). The Lung Cancer is originally from the First Annual Proteomics Datamining Conference , organized by the Department of Radiology and Biostatistics at Duke University in September 2002. We are using the version of this dataset used in Ref.12 with 229 peaks after going through some basic preprocessing steps except peak selection. We obtained the Lung Cancer dataset from Kent Ridge Bio-medical Data Set Repository.[13] All the two datasets originally have no test set. For performance validation, we still spare some samples for testing purpose. Thus, we randomly split the original dataset into a training set and a test set and keep percentages of the positive and negative samples same in the training and test sets. We summarize some basic information about the datasets, including the number of peaks, the sizes of the training and test sets, in Table 1. More detailed information about the two datasets can be found in the Refs.12 and 13, and the references there in.

In our study, for each dataset, we did the peak selection solely on the training set. The goodness of a selected peak subset is evaluated by the performance of a classifier built on the training set with only the selected set of peaks as input variables. In our study, we choose linear SVM as the classification algorithm. Linear classification algorithms are commonly used in cancer classification with mass spectrometry data, e.g. see Ref.14.

Test error on test set is usually used to assess the performance of a classifier. However, the total numbers of available samples in our mass spectrometry datasets are small. In such a case, the test error may be biased due to an "unfortunate" partition of training and test sets. Thus, instead of reporting such an test error from one division of training and test sets, we do as follows: we merge the training set and test set and then partition the total samples again into a training set and a test set randomly by stratified sampling for 100 times; for each division, we train a linear SVM classifier on the training set (hyperparameter C is to be selected by 10-fold cross-validation on the training set) and then test it on the corresponding test set; from this 100 trials we can compute the averages of performance measures.

Table 1. Number of peaks, sizes of training and test sets, of the two datasets.

Dataset	♯ peaks	♯ training samples	♯ test samples
Lung	229	29	12
Ovarian	15,154	177	76

Table 2. Performance of SVM without peak selection and the performance of SVM with peak selection by T-Statistics or SVM-RFE, on the two datasets.

Dataset	Measurement	SVM	T-Statistics	SVM-RFE
Lung	No. of Peaks	Full (229)	7	8
	Test Error (%)	21.58±9.63	10.75±8.89	8.41±5.98
	Sensitivity (%)	90.29±11.28	95.43±7.56	94.57±7.54
	Specificity (%)	61.80±21.29	80.60±22.28	87.40±13.23
Ovarian	No. of Peaks	Full (15,154)	6	11
	Test Error (%)	0.50±1.04	1.61±1.39	0±0
	Sensitivity (%)	99.85±0.52	99.31±1.86	100±0
	Specificity (%)	98.85±2.72	96.74±2.70	100±0

To speed up the feature selection procedure of SVM-RFE, when the number of features m is large in the feature subset S selected at a time, we eliminate r $(r \geq 1)$ features each time in our numerical experiments We choose $r = 1000$ if $m > 100000$, $r = 100$ if $10000 < m \leq 100000$, $r = 10$ if $1000 < m \leq 10000$ and $r = 1$ if $m \leq 1000$.

In our study, for each feature subset selected by either T-statistics method or SVM-RFE, we compute the mean and standard deviation of test error, sensitivity and specificity, from 100 times of training and testing. As we are mostly interested in small peak subsets, we evaluate the two methods only on small peak subsets with number of peaks ranging from 1 to 50. We plot the average test error versus the size of feature subsets selected by two methods on the two datasets respectively in Figs. 1 and 2.

SVMs are capable of dealing with large number of input variables with no increase in computation complexity. To see if feature selection improves the performance of SVMs, we also train and test SVMs with full number of features on the same 100 partitions of training and test sets. The means and standard deviations of the test performance of SVMs with full features are reported in Table 2, together with those of the best feature subsets selected by T-Statistics and SVM-RFE, with the number of selected peaks confined to less than 20.

4. Discussion and Conclusion

From Figs. 1 and 2, it is very clear that SVM-RFE selects better peak subsets than T-statistics feature selection method. High classification accuracy is achieved with only a small number of peaks as input variables.

Looking at the performance of SVMs without peak selection and SVMs with peaks selection in Table 2, we can see that, the classification performance of SVMs with peak selection are much better than that of SVMs with all peaks as input variables. This observation

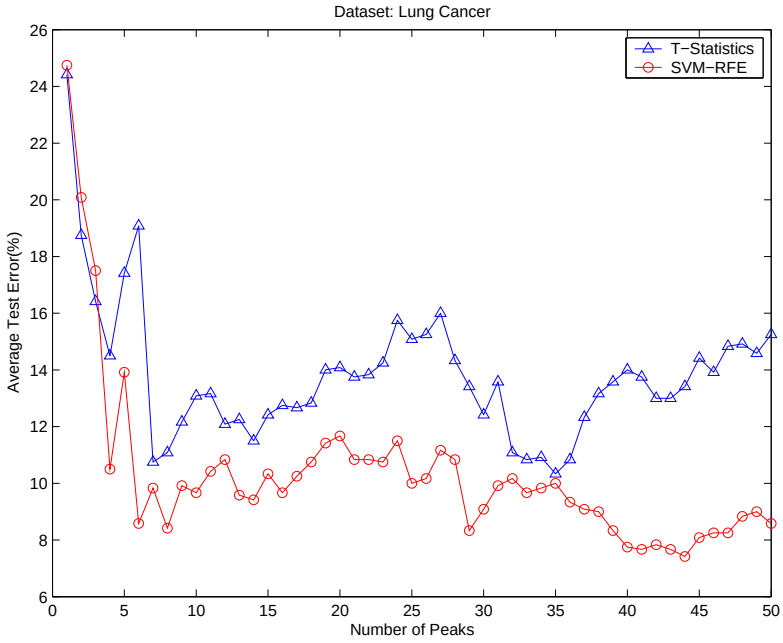

Figure 1. Average test error rates at different sizes of peak subsets, selected by T-statistics and SVM-RFE, on Lung Cancer dataset.

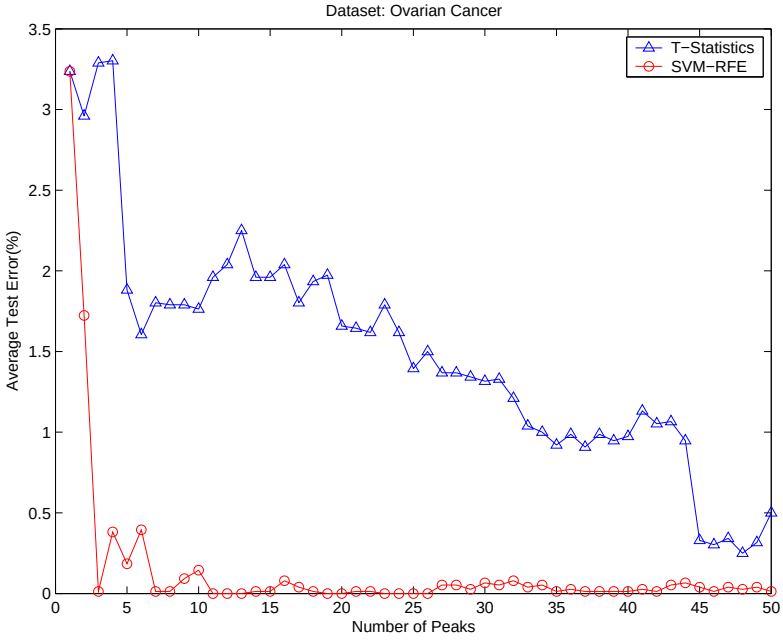

Figure 2. Average test error rates at different sizes of peak subsets, selected by T-statistics and SVM-RFE, on Ovarian Cancer dataset.

tells us that, selecting a subset of peaks not only improves the efficiency of classification algorithms, but also improves the prediction accuracy, even for classification algorithm like SVMs, which can handle large number of input variables without increase in computation complexity. Selecting a small number of peaks also prevents getting spurious results with mass spectrometry data, for which the number of training samples is usually small compared with the number of peaks in the mass spectra. The high prediction accuracy also further strengthens our belief of the promising application prospects of mass spectrometry patterns in the future cancer classification.

While we understand that T-statistics selects the peaks whose intensities differs most between the cancer and no-cancer groups, the way that SVM-RFE selects the peak subset is not well understood. Checking the T-statistics scores of the peaks selected by the SVM-RFE may helps us to get some insight into the way SVM-RFE works. On the Lung Cancer dataset, we found the T-statistics ranks of the 8 peaks in the best subset selected by SVM-RFE respectively are $\{1, 2, 3, 4, 7, 17, 29, 36\}$ (peak with rank 1 has the largest T-statistics score). On the Ovarian Cancer dataset, the T-statistics ranks of the 11 peaks in the best subset selected by SVM-RFE respectively are $\{1, 2, 3, 4, 5, 6, 7, 14, 45, 50, 73\}$. On the both datasets, the best peak subsets selected by SVM-RFE always have in the peaks top ranked by T-statistics, while they also include some peaks not top ranked by T-statistics. However, these peaks selected by SVM-RFE together achieve much smaller test error than the same number of most top ranked peaks selected by T-statistics, as we can clearly see in Figs. 1 and 2. However, to get better understanding of the way SVM-RFE works and to get a better insight into the disease pathway, ultimately we have to rely on a further investigation to identify the proteins underlying these selected peaks and a further functional study of the identified proteins.

Acknowledgment

We thank Michael Wagner at Cincinnati Children's Hospital Medical Center for sharing with us his preprocessed Lung Cancer dataset.

References

1. D.C. Liebler. *Introduction to Proteomics - Tools for the New Biology*. Humana Press, 2002.
2. 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. *Lancet*, pages 572–577, 2002.
3. B.L. 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 G.L. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishs prostate cancer from benign prostate hyperplasia and healthy men. *Can Res*, 62:3609–3614, 2002.
4. J. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang, and D.W. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem*, 48:1296–1304, 2002.
5. Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

6. A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.
7. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
8. B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pages 114–152, Pittsburgh, 1992. ACM.
9. V. Vapnik. *Statistical Learning Theory*. Wiley Interscience, 1998.
10. Y. LeCun, J. Denker, S. Solla, R.E. Howard, and L. D. Jackel. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*. Morgan Kaufmann, Mateo, CA, 1990.
11. A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research, Special Issue on Variable Selection*, 3:1357–1370, 2003.
12. M. Wagner, D. Nail, and A. Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3:1692–1698, 2003.
13. J. Li and H. Liu. Kent Ridge Bio-medical Data Set Repository, 2002. Available at: *http://sdmc.lit.org.sg/GEDatasets/Datasets.html*.
14. M. Wagner, D.N. Naik, A. Pothen, S. Kasukurti, R.R. Devineni, B.L. Adam, O.J. Semmes, and G.L. Wright Jr. Computational protein biomarker prediction: a case study for prostate cancer. *BMC Bioinformatics*, 5(1):26, 2004.