Analysis of Proteomic Pattern Data for Cancer Detection

Kees Jong, Elena Marchiori, and Aad van der Vaart

Department of Mathematics and Computer Science
Vrije Universiteit Amsterdam, The Netherlands
cjong,elena,aad@cs.vu.nl

Abstract. In this paper we analyze two proteomic pattern datasets containing measurements from ovarian and prostate cancer samples. In particular, a linear and a quadratic support vector machine (SVM) are applied to the data for distinguishing between cancer and benign status. On the ovarian dataset SVM gives excellent results, while the prostate dataset seems to be a harder classification problem for SVM. The prostate dataset is futher analyzed by means of an evolutionary algorithm for feature selection (EAFS) that searches for small subsets of features in order to optimize the SVM performance. In general, the subsets of features generated by EAFS vary over different runs and over different data splitting in training and hold-out sets. Nevertheless, particular features occur more frequently over all the runs. The role of these "core" features as potential tumor biomarkers deserves further study.

1 Introduction

Surface-enhanced laser desorption/ionization time-of-flight mass spectronomy (SELDI-TOF MS) is a recent laboratory technology which offers high-throughput protein profiling. It measures the concentration of low molecular weight peptides in complex mixtures, like serum (cf. e.g. [1]). Because it is relatively inexpensive and noninvasive, it is a promising new technology for classifying disease status.

SELDI-TOF MS technology produces a graph of the relative abundance of ionized peptides (y-axis) versus their mass-to-charge (m/z) ratios (x-axis). (Cf. Figure 1) The m/z ratios are proportional to the peptide masses, but the technique is not able to identify individual peptides, because different peptides may have the same mass and because of limitations in the m/z resolution. Currently the graph is represented by 15000 measuring points. There is no obvious relation between neighbouring measurement points, apart from the fact that they refer to peptides of similar masses and that the resolution is such that the graph should be considered a smoothed version of the true mass density.

Given proteomic profiles for a sample of healthy and diseased individuals it is desired to build a classifier for tumor diagnostics and to identify the protein masses that are potentially involved in the disease. Because of the large number

of features (the m/z ratios) and the small sample size (the specimens), the second problem is tackled using heuristic algorithms for feature selection.

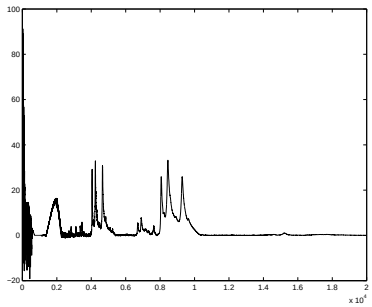


Fig. 1. A typical protein profile produced by SELDI-TOF MS.

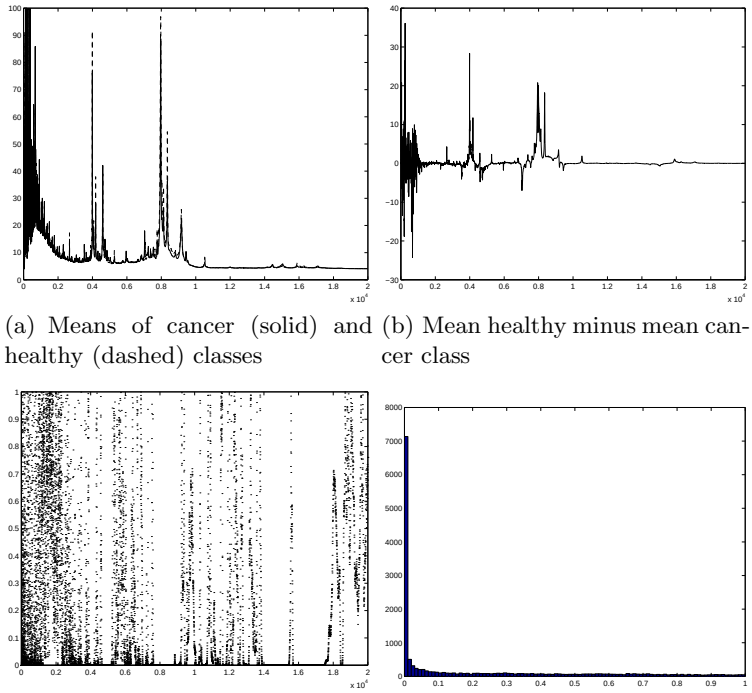
In this paper we analyze two datasets obtained by applying SELDI-TOF MS to serum samples. The first dataset concerns measurements from women with or without ovarian cancer, was previously analized in [3, 4]. The second dataset contains samples from patients with prostate cancer and patients with benign prostate conditions, was analyzed in [2]. Both datasets are publically available from the NCI/CCR and FDA/CBER Clinical Proteomics Program Databank (`http://clinicalproteomics.steem.com/`).

As preliminary analyses we first investigate the extent to which single m/z ratios can be used to discriminate the two classes of healthy and cancer state samples. Secondly we report the error rate of support vector machine (SVM) classifiers using the full protein profiles. It turns out that the ovarian cancer dataset is "easy" for a linear SVM classifier, whereas the prostate cancer dataset is "harder".

We perform a further analysis of the prostate cancer dataset by means of a feature selection algorithm based on EAs. We introduce an EA for feature selection, called `EAFS` (Evolutionary Algorithm for Feature Selection), in order to identify small subsets of features that discriminate the healthy and cancer groups. The results over multiple data splittings (into training, test and validation set) and multiple EA runs show that the method is slightly unstable. However, specific features occur most frequently in the solutions of multiple runs. Further study is needed in order to assess the role of these "core" features as potential tumor biomarkers.

2 Data Analysis with all Features

The "ovarian dataset" (8-7-02) consists of 253 samples, with 91 controls and 162 ovarian cancers, which include early stage cancer samples. The "prostate



(a) Means of cancer (solid) and healthy (dashed) classes

(b) Mean healthy minus mean cancer class

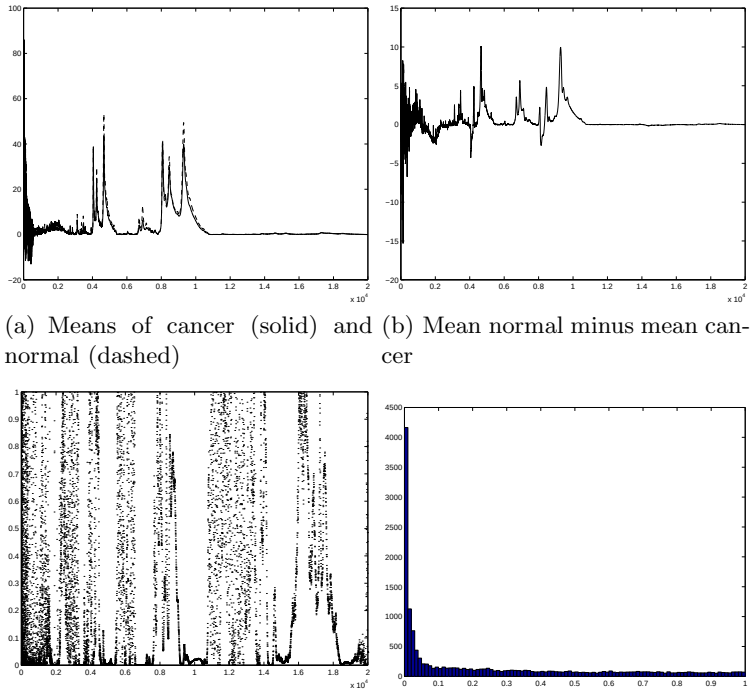
(c) p-values of t-test between groups

(d) number of p-values

Fig. 2. Mean values analysis of ovarian cancer data set

dataset" contains 322 samples, with 69 cancers and 253 healthy (or benign) samples.

We analyze the two datasets in order to assess how difficult it is to separate healthy and cancer groups. Figure 2 shows properties of the mean values of the two classes for the ovarian dataset. Parts (a) and (b) of the figure indicate that the healthy and cancer classes differ only in a few regions substantially in mean. Because the variances in the two samples vary significantly with m/z ratio, the t-test applied for each m/z ratio separately is nevertheless significant for a much larger number of m/z ratios. In fact, as shown in part (c), there are many p-values equal to zero all across the full range of m/z-values, and "most" p-values are close to zero, as shown in part (d), which is a histogram with 100 bins. This seems to suggest that it is not difficult to find a good classifier for the ovarian dataset. The same information on the prostate data set is given in Figure 3. We can see there are fewer features with significant difference in mean in the prostate than in the ovarian dataset (part (d) of the figures). Thus finding a good classifier for the prostate dataset seems to be a more difficult task.



(a) Means of cancer (solid) and normal (dashed)

(b) Mean normal minus mean cancer

(c) p-values of t-test between groups

(d) number of p-values

Fig. 3. Mean values analysis of prostate cancer data set

We classify the data using all the features and a SVM classifier. The choice of SVM is motivated by their good performance also in the presence of many features. In SVM classification [6], the samples of the two classes are mapped into a feature space where they are separated by means of a maximum margin hyperplane, that is, the hyperplane that maximizes the sum of the distances between the hyperplane and its closest points in each of the two classes (the margin). The inner product in the feature space can be computed in the input space by means of a so-called kernel function. This choice is shown to affect positively the generalization performance of the classifier. When the classes are not linearly separable in the feature space, a variant of SVM, called soft-margin SVM, is used. This SVM variant penalizes misclassification errors and employs a parameter (the soft-margin constant C) to control the cost of misclassification. In our experiments we use soft-margin SVM classifiers with C=100, and two types of kernel functions, linear and quadratic.

Table 1 contains the results obtained by applying an SVM classifier with linear and quadratic kernel to the ovarian and prostate cancer datasets. Experiments are conducted using the SVM software package LibSVM. We perform 25 runs,

where in each run the data is randomly partitioned into training (60%) and test (40 %) sets, a classifier is induced by the SVM learning algorithm applied to the training set, and its classification error rate on the test set is measured. Each entry of the table contains the average result over 25 runs (with standard deviation written between brackets) for a specific pair of SVM classifier and dataset.

	Lin., Ovar.	Poly., Ovar.	Lin., Pros.	Poly., Pros.
Test Error	**0.0028** (0.0061)	0.0040 (0.0065)	0.0600 (0.0201)	**0.0590** (0.0246)
Sens.	0.9987 (0.0044)	**0.9988** (0.0042)	0.8884 (0.0596)	**0.8936** (0.0719)
Spec.	**0.9947** (0.0153)	0.9912 (0.0169)	0.9539 (0.0205)	**0.9552** (0.0261)
Pos.Pred.Val.	**0.9969** (0.0090)	0.9951 (0.0097)	0.8361 (0.0630)	**0.8475** (0.0867)

Table 1. Average results over 25 runs of SVM with linear and quadratic kernel, on ovarian and prostate data

In all the runs the SVM classification error on the training set is zero. The table contains the average test error rate, *sensitivity* (number of cancer samples correctly classified divided by total number of cancer samples), *specificity* (number of healthy samples correcty classified divided by total number of healthy samples), and *positive predictive value* (number of cancer samples correctly classified divided by total number of samples classified as cancer).

On the ovarian dataset, the SVM classifier with linear kernel gives excellent results, being able to output the correct diagnosis in many runs, in particular cancer samples are almost always detected (pred. pos. val. and sensitivity are close to 1). Thus the ovarian dataset can be considered easy for a linear SVM classifier. Other methods applied to this dataset obtain also very good results. In [3] a commercial package that uses a genetic algorithm (GA) based feature selection method is applied. The GA searches in the space of all subsets of features. It uses a fitness function that scores a feature subset according to its ability to cluster samples in consistent groups, that is, groups containing samples with equal class. The authors report almost perfect classification for a specific data splitting in training and test sets. The paper does not mention results obtained by cross-validation. A more thorough analysis of the ovarian dataset is performed in [4], where 10-fold cross validation is applied, and different classification and feature selection methods are considered. Features are first "smoothed" by means of a discretization algorithm. Perfect classification is achieved using an SVM classifier with quadratic kernel, when all features are used but also when a small subset of 17 features is used. This feature subset is generated using a feature selection algorithm that iteratively constructs sets of features using the best-first-search and a scoring criterion, for selecting a best feature subset, which considers the correlation between pairs of features and between feature and class.

On the prostate dataset, sensitivity and predicted positive values are lower than specificity, possibly due to the unbalanced distribution of the two classes, where

cancer samples are about 1/3 of the healthy ones. The results indicate that the prostate data set is somewhat harder to classify than the ovarian one, when using SVM with a linear or quadratic kernel and all the features. In [2] the GA-based commercial package described above is applied to this dataset. The authors identify a subset of 7 features that allow their classification method to obtain 0.95 sensitivity, 0.78 specificity, and 0.1992 test error rate.

In summary, on the ovarian dataset a soft-margin SVM linear classifier provides a good diagnostic tool, while for the prostate dataset the sensitivity achieved is still too low hence does not allow a direct use of this classifier in diagnostics. An early stage tumor diagnostic tool should have sensitivity equal to 1 and specificity very close to 1.

3 An EA-based Method for Feature Selection

In this section we describe a novel method for feature selection based on evolutionary algorithms (EA). Given a dataset and a learning algorithm, the goal of feature selection is to find a "small" subset of features that minimizes the generalization error (that is, the classification error on new examples) of the classifier induced by the learning algorithm when run only on the selected features.

The data is randomly partitioned into a training, a test and a validation sets (in the experiments these sets contain 60%, 30% and 10% of the data, respectively). The training and test sets are used in the feature selection algorithm and the validation set is used for assessing the performance of the resulting classifier on new data. In the standard wrapper model for feature selection, one searches for a feature subset that minimizes the test error of the classifier trained on data restricted to that feature subset. In [5], Ng shows that the main source of error in standard wrapper algorithms, when many irrelevant features are present, comes from over-fitting hold-out or cross-validation data. He proposes an exact algorithm which is more tolerant to the presence of many irrelevant features. The algorithm, called `ordered-fs`, works in two phases. First, for each feature set size $i \in [1, m]$, where m is the maximum number of features permitted, the algorithm finds a feature set of size i that minimizes the classifier training error. Next, the resulting m classifiers are run on the test set, and the one yielding minimum error is chosen.

The EA-based method we propose, called `EAFS`, is inspired by the `ordered-fs` algorithm. The core of `EAFS` (illustrated below) consists of an EA which evolves a number of populations, where each population consists of individuals representing feature subsets of a given size. The populations interact by means of highly fit individuals which are used as seed for generating new individuals of other populations. Genetic operators are used for moving in the search space in order to minimize the SVM training error. At the end of the evolutionary process, the best SVM classifier of each population is run on the test set and the one yielding minimum error on this set is selected. Thus the test set is used

only to determine the optimal size of the feature set. The selection of an optimal feature set of a given size is based only on the training set.

```
//the core of EAFS
{
  generate initial population
  while (termination criterion not satisfied)
  {
    select a population
    select to parents from that population
    generate offspring using uniform crossover
    apply mutation to offspring
    find populations with right number of features
    replace worst individuals with offspring
    determine fitness (error SVM on training set)
    if (offspring has very good fitness)
        apply migration operator
  }
}
```

Feature subsets are represented by bit strings of length equal to the total number of features. A bit value equal to 1 means the corresponding feature is considered by the learning algorithm, while a 0 means it is discarded. Individuals of each population are initialized by means of n-tournament selection which uses a feature ranking obtained from t-tests on all single features. The fitness of a feature subset F is equal to the training error of the SVM classifier restricted to the features of F. Mutation removes a feature from F and adds a new one, where both features are randomly selected. Standard GA uniform crossover is used. While mutation does not affect the size of a feature subset, this is not the case for crossover. Thus if the feature set size of an offspring is different from the one of its parents, it migrates to another population. At each iteration of the EA, a population is selected and used to generate two offspring. The EA uses tournament selection and a steady state replacement mechanism, where offspring replace the worse individuals of the population. When the EA terminates its execution, the best individual of each population is chosen and the one yielding the lowest error on the test set provides the output. In the sequel, we focus only on the application of `EAFS` to the prostate dataset, and will not compare `EAFS` to other EA-based feature selection algorithms.

4 Results

We used the following experimental setup: 20 populations, each one consisting of 10 individuals, 600 iterations, tournament selection of size 5, crossover and mutation rate of 0.95. These values have been chosen after a small number of

runs (using only training and test sets). We consider 24 random splitting of the dataset in training (60%), test (30%) and validation (10%) sets. For each "split" of the data we run `EAFS` 25 times. This amounts to a total of 600 runs. Table 2 gives the results of `EAFS` with linear kernel. The values are the averages over all the 600 runs. Standard deviation is reported between brackets. Table 3 contains the results using a quadratic kernel. `EAFS` with a quadratic kernel SVM achieves best performance, but obtains sensitivity lower than that of SVM with all the features. However, a fair comparison is not possible due to the different cross validation approaches used.

	Training	Test	Validation
Error	0.0617 (0.0254)	0.0880 (0.0313)	0.1116 (0.0515)
Sensitivity	0.7853 (0.0926)	0.7037 (0.1193)	0.6315 (0.2069)
Specificity	0.9827 (0.0118)	0.9658 (0.0238)	0.9484 (0.0456)
Pos. Pred. Value	0.9309 (0.0451)	0.8437 (0.0957)	0.7424 (0.2178)

Table 2. Results of `EAFS` with linear SVM

	Training	Test	Validation
Error	0.0463 (0.0287)	0.0774 (0.0283)	0.1096 (0.0579)
Sensitivity	0.8360 (0.1096)	0.7502 (0.1249)	0.6779 (0.2177)
Specificity	0.9874 (0.0109)	0.9674 (0.0216)	0.9441 (0.0525)
Pos. Pred. Value	0.9502 (0.0431)	0.8547 (0.0948)	0.7671 (0.1936)

Table 3. Results of `EAFS` with quadratic SVM

In order to investigate whether the data "split" influences the performance of `EAFS` significantly, we perform a one-way Analysis of Variance to compare the 24 samples of 25 validation errors resulting from the 24 "splits". The difference is statistically significant, with zero p-values for both the linear and quadratic case. The "splits" explain about 24 % of the variance in the validation errors in the linear case and about 35 % in the quadratic case, indicating that 76 % and 65 % is due to the randomness inherent in `EAFS`. A correction for the effect of "split" on the error standard deviations in Tables 2 and 3 would reduce these somewhat, but not substantially (e.g. 0.0515 becomes 0.045). Figure 4 gives a visual impression of the variation due to the splitting and `EAFS`. The three graphs show boxplots of the training, test and validation errors of the 600 runs (top to bottom), organized by "split" (left) or "run".

An important aspect of these graphs is summarized in the Figure 5, which shows boxplots of the standard deviations of the 24 samples of validation errors corresponding to the 24 "splits". The linear SVM suffers from one extreme split, but is otherwise more stable across relative to the splitting of the data.

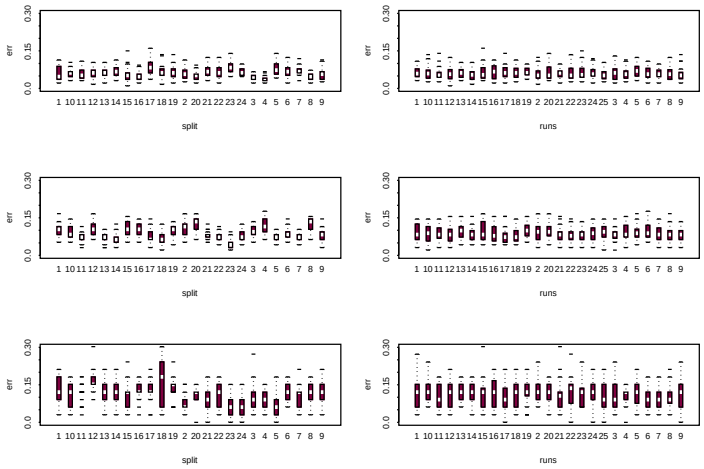


Fig. 4. Boxplot of training, test and validation errors (top to bottom), organized by split (left) and run (right)

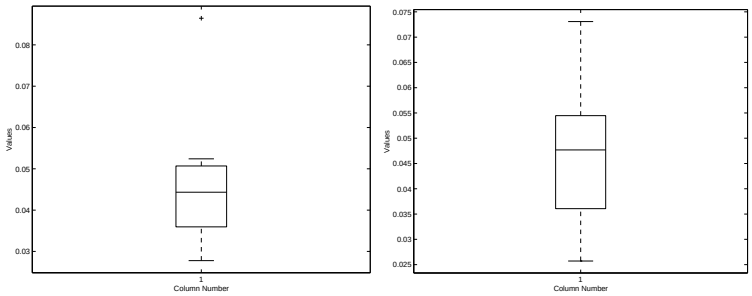
We analyze now the features obtained in all the runs.

Figure 6 shows histograms of the final feature set sizes found over the 600 runs. The algorithm shows no preference for the largest possible feature set size. We can see that `EAFS` with linear SVM has a peak on feature size 5, while `EAFS` with quadratic SVM prefers somewhat bigger feature sizes, with a peak on feature size 13.

Over all the runs `EAFS` with linear SVM finds 3935, while with quadratic SVM finds 3797 features. Figure 7 shows histograms of the features occurring in the solutions found over the 600 runs. It is clear that `EAFS` is not stable, yet the histograms indicate the presence of few frequency peaks . By considering features occurring in these peaks, 47 features occurring in at least 10 runs are extracted. We perform 100 runs of the linear SVM restricted to the 47 features, with rando mly chosen training and test set, and obtain 0.93 sensitivity (0.058 standard deviation) and 0.98 specificity (0.016 standard deviation). These significantly better results may be due to the fact that the selection of these features implicitly uses (almost) the entire dataset.

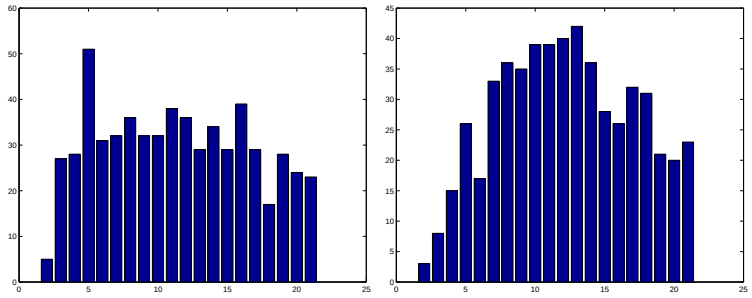
5 Conclusion

This paper analyzed two proteomic pattern datasets. We applied SVM classifiers for tumor diagnostics, and used them in `EAFS`, an EA-based feature selection algorithm for the identification of potential tumor markers identification. The



(a) Linear Kernel (b) Quadratic Kernel

Fig. 5. Boxplots of standard deviation within the 24 runs.



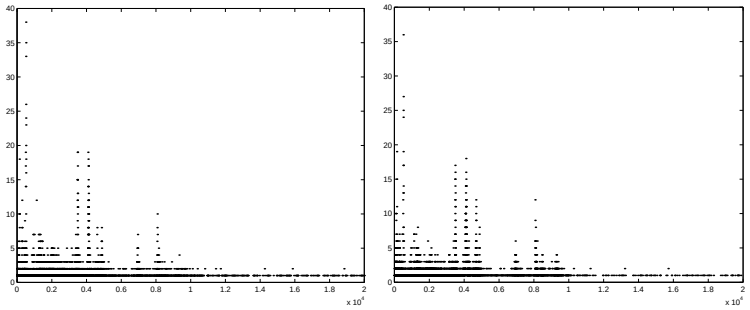
(a) Linear Kernel (b) Quadratic Kernel

Fig. 6. Histograms of obtained feature set sizes

results do not allow us to draw strong conclusions. On the ovarian dataset SVM with all the features exhibits excellent performance, while on the prostate dataset it obtains relatively low sensitivity. Results of `EAFS` show that its performance on the prostate dataset depends on the data splitting and EA run. Moreover, feature subsets generated by `EAFS` vary per run, with a small core of features occurring more often. This latter phenomenon was observed to happen also in the other methods discussed in this paper.

Future work includes: the incorporation of a pre-processing phase into `EAFS`; the investigation of other types of classifiers; the use of knowledge-based mutation operators; and the use of multiple `EAFS`'s runs with different splitting of training and test sets for extracting a "core" set of features from the resulting `EAFS`'s solutions.

Acknowledgment We would like to thank Guus Smit and Connie Jime nez from the Department of Biology of the Vrije Universiteit Amsterdam for helpful discussions on the subject of this paper.



(a) Linear Kernel (b) Quadratic Kernel

Fig. 7. Histograms of number of feature occurrence in final solutions

References

1. H.J. Issaq et al. SELDI-TOF MS for diagnostic proteomics. *Anal. Chem.*, 75(7):148A–155A, 2003.
2. Petricoin E.F. et al. Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute*, 94(20):1576–1578, 2002.
3. Petricoin E.F. et al. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359(9306):572–7, 2002.
4. H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13:51–60, 2002.
5. Andrew Y. Ng. On feature selection: learning with exponentially many irrelevant features as training examples. In *Proc. 15th International Conf. on Machine Learning*, pages 404–412. Morgan Kaufmann, San Francisco, CA, 1998.
6. V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.