Feature Selection on High Throughput SELDI-TOF Mass-Spectrometry Data for Identifying Biomarker Candidates in Ovarian and Prostate Cancer

Claudia Plant, Melanie Osl, Bernhard Tilg, Christian Baumgartner
Research Group for Clinical Bioinformatics, Institute of Biomedical Engeneering
University for Health Sciences, Biomedical Informatics and Technology (UMIT), Hall in Tirol, Austria
{*claudia.plant|melanie.osl|bernhard.tilg|christian.baumgartner*}*@umit.at*

Abstract

High-throughput mass-spectrometry screening has the potential of superior results in detecting early stage cancer than traditional biomarkers. Proteomic data poses novel challenges for data mining, especially concerning the curse of dimensionality. In this paper, we present a 3-step feature selection framework combining the advantages of efficient filter and effective wrapper techniques. We demonstrate the performance of our framework on two SELDI-TOF-MS data sets for identifying biomarker candidates in ovarian and prostate cancer.

1. Introduction

The identification of putative proteomic marker candidates is a big challenge in the biomarker discovery process. Pathologic states within cancer tissue may be expressed by abnormal changes in the protein and peptide abundance. By the availability of modern high throughput techniques such as SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) MS a large amount of high dimensional mass spectrometric data is produced from a single blood sample. Each spectrum is composed of peak amplitude measurements at approximately 15,200 features represented by a corresponding m/z value. Proteomic spectra potentially contain more information on the abnormal protein signaling and networking than traditional single biomarkers. The widely used cancer antigen 125 (CA125) for instance can only detect 50%-60% of patients with stage I ovarian cancer [8].

The curse of dimensionality severely affects the performance of classification algorithms in terms of efficiency and effectiveness on proteomic spectra. Feature transformation techniques can be applied before classification, e.g. as in [13]. To identify previously not discovered marker candidates, however, the transformed features are not useful. Feature selection methods, which try to find the subset of features with the highest discriminatory power, can be applied. Nevertheless, as aforementioned, the use of traditional methods is limited due to the high dimensionality of the data.

In this paper, we propose a novel 3-step feature selection framework which combines elements of existing feature selection methods and is accustomed to the special characteristics of high-throughput MS data. We present the results on two published SELDI-TOF-MS data sets on ovarian and prostate cancer. Our method identifies feature subsets with a classification accuracy between 97% and 100%.

The paper is organized as follows: In Section 2 we briefly survey related work on feature selection methods and on the data sets used and we summarize our contributions. In Section 3 we elaborate in detail our framework for a 3-step feature selection. In Section 4 we discuss the results on ovarian and prostate data and Section 5 concludes the paper.

2. Survey

Feature Selection for Classification. Numerous feature selection strategies for classification have been proposed, for a comprehensive survey see e.g. [5]. *Filter approaches* use an evaluation criterium to judge the discriminating power of the features. Rankers, e.g. information gain [12] and reliefF [7] evaluate each feature independently regarding its usefulness for classification. Rankers are very efficient, but interactions and correlations between the features are neglected. Feature subset evaluation methods, e.g. [4, 9] therefore judge the usefulness of subsets of the features. The search space of possible feature subsets expands to the size of $O(2^d)$, which also holds for the *wrapper approach*. The wrapper feature selection strategy

Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)
0-7695-2702-7/06 $20.00 © 2006 IEEE

uses a classifier to evaluate attribute subsets. Adapted to the special characteristics of the classifier, in most cases wrapper approaches identify feature subsets with a higher classification accuracy than filter approaches, cf. [5].

Data Sets. Both SELDI-TOF-MS data sets are available at the website of the US National Cancer Institute: (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). Each spectrum is composed of $d = 15,154$ features.
Ovarian Data Set. The ovarian data set 8-7-02 contains 162 instances of ovarian cancer and 91 instances of a healthy control group. The data set is an improved version of the data set published in [10], using a WCX2 protein chip and robotic automation. Trajanowski et al. [13] recently proposed an approach for ovarian cancer identification based on dimensionality reduction. They use a multi-step feature transformation technique based on wavelet transformation and present the results on this data set and on a high resolution data set on ovarian cancer. With 2-fold cross validation, an accuracy of 91.12% is reported for SVM on the wavelet reduced 8-7-02 data set. Alexe et al. [2] analyzed this data set using a combinatorics and optimization-based methodology . With a system of 41 rules they achieved a sensitivity and specificity of 100% on this data set.
Prostate Data Set. This data set consists of four classes, representing a healthy control group (c_1), patients with benign conditions and elevated PSA value (c_2) and two stages of prostate cancer (c_3 and c_4). In [11] this data set has been analyzed with with ProteinQuest, a tool combining genetic algorithms and cluster analysis. This method achieves to identify prostate cancer with 94.74% accuracy (accuracy on the class corresponding to $c3 \cup c4$) and 77.73% percent of the instances were correctly identified to have benign conditions (accuracy on $c1 \cup c2$). However, especially for class $c2$, the reported specificity is with 71 % quite low.

Contributions. The main advantages of our method can be summarized as follows:

- We propose a generic framework for feature selection using a classifier C, a search strategy S and a ranker R.
- Our method is efficient and applicable on very high dimensional proteomic data sets.
- The classification results on the selected features confirm and outperform the results reported in literature on the ovarian and the prostate data set.

3. Feature Selection

An optimal feature subset for biomarker identification and diagnosis is a subset consisting of as few features as possible and achieving highest classification accuracy. We use C, R, S and special properties of proteomic data for an effective and efficient exploration of the huge search space of 2^d feature subsets to find a close to optimal solution. The classifier, the evaluation criterium and the search strategy can be arbitrarily chosen, also depending on time and memory resources.

In the following we discuss the single steps in detail. In this Section we focus on the use of linear SVM as classifier and information gain as ranker, and simulated annealing, and a novel heuristic called modified binary search as search strategies. For the classifiers we use the implementations of the WEKA machine learning package [1]. Parameterizations are $c = 1.0$ and $\gamma = 0.01$ SVM and we use 10-fold cross validation to estimate the accuracy. Some notations which are frequently used: We denote the resulting data set of step i by res_i with the classification accuracy acc_i and the dimensionality dim_i. We denote the rank of a feature f_i by $rank(f_i)$ and its quality by $quality(f_i)$. We further denote by $index(f_i)$ the index, i.e. the position of the m/z value of f_i in the original data set.

3.1 Step 1: Removing Irrelevant Features

In the first step, we want to identify and discard the irrelevant features using the ranker R and the classifier C. To get a baseline for the classification accuracy, we first determine the accuracy on the full feature space using C. For ovarian data 99.60% is achieved with linear SVM, for prostate data 90.37%. We then use the evaluation criterium to remove all irrelevant features. For information gain this means we remove all features with information gain 0 and determine the accuracy again. For the ovarian data set $dim_1 = 6,238$ attributes remain and the accuracy stays the same, i.e. $acc_1 = 99.60$. For prostate data, the reduction to $dim_1 = 9,566$ attributes improves the classification accuracy to $acc_1 = 93.16\%$.

3.2 Step 2: Selecting the Best Ranked Features

In this step, we want to further reduce the dimensionality without affecting the accuracy, i. e. our aim is to identify a feature subset res_2 with $acc_2 >= acc_1$ and $dim_2 <= dim_1$. Since dim_1 is still in the order of magnitude of several thousands of features, we restrict the search space to the ranked list generated by R in this step. This means, we reduce the search space to the size of $O(d)$ for now. Note that the features discarded now may be re-included in the following step. We use an arbitrary search strategy S and the classifier C to find a smaller attribute subset while keeping the accuracy at least constant. We discuss three different options for S: ranked

search, simulated annealing and a novel heuristic called modified binary search.

Ranked Search. Starting with res_1, ranked search removes in each step the feature with the lowest rank and evaluates the classification accuracy using C. Figure 1 shows the accuracy for SVM on both data sets for the 6,000 top ranked features. For the ovarian data set, 100% accuracy is achieved using the 38 top ranked attributes. For the prostate data the global maximum of 94.72% is reached using the 2,722 top ranked attributes. Ranked search is very inefficient because only one feature is removed in each step.

Simulated Annealing. Simulated annealing (SA) [6] has been successfully applied for solving complex global optimization problems. On the ovarian data set 40 features with an accuracy of 100% are selected with the parametrization $T = 40$, $\delta = 3,000$, $\delta \downarrow = 75$, $T \downarrow = 1$ and $S_p = 2,080$. For the prostate data set, we applied $T = 40$, $\delta = 3200$, $\delta \downarrow = 80$, $T \downarrow = 1$ and $S_p = 3198$, resulting in 2,800 features and an accuracy of 94.09 %.

Modified Binary Search. It is not required to find the global maximum accuracy in the search space of the ranked features. Restricting the search space to the ranked features means rating single features only and neglecting dependencies and interactions between the features. Therefore, for complex data sets satisfactory results in terms of accuracy can not be expected from the result of this step. As a input for the next step it is sufficient to identify the smallest set of features that establish a classification accuracy which is close to the optimum.

A first idea would be to apply binary search (BS), which is very efficient (O(log(d)). For the ovarian data set, this strategy works well, since there are not many local maxima of the accuracy, cf. Figure 1. For both data sets BS is much faster than SA (cf. Table 1 showing the number of selected attributes and the runtimes on a 2.99 GHz CPU, 0.99 GB

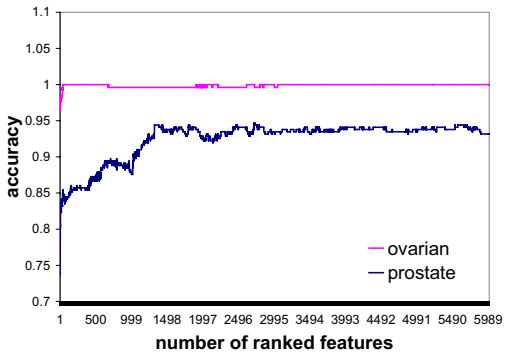


Figure 1. Search Space for Step 2.

algorithm ModifiedBinarySearch
(dataSet res_1, monotonic function f, parameter p, classifier C):
dataSet res_2

currentAccuracy: = C.accuracy(currentDim);
while currentDim $< dim_1$ **and** currentAccuracy < 1.0
  nextDim = $f(next)$;
  nextAccuracy = C.accuracy(nextDim);
  if nextAccuracy = 1.0
    res_2 = binarySeach(currentDim, nextDim);
    return res_2;
  else
    $g := \frac{nextAccuracy - currentAccuracy}{nextDim - currentDim}$
    if g < 0
    //check if plateau is reached
      check = nextDim + (nextDim - currentDim);
      checkAccuracy = C.accuracy(check)
      $g := \frac{checkAccuracy - nextAccuracy)}{check - nextDim}$
      if g $< p$
        res_2 = binarySearch(nextDim, check);
        return res_2;

Figure 2. Modified Binary Search

RAM). However, BS can get stuck at every arbitrary local maximum. To avoid this, we guide the search towards feature sets which are as small as possible.

Starting with an empty attribute subset, the accuracy shows a steep rise while the very best ranked features are added, cf. Figure 1. The accuracy then reaches a level at which it keeps at most constant while adding more and more features according their ranked order. The goal of our algorithm is to find the point where the accuracy reaches the plateau.

Our algorithm divides the search space into intervals of monotonically increasing size and determines points at which the accuracy is evaluated. We decided to use intervals of monotonically increasing size because of the steep increase of the accuracy at the beginning which is flattening later. We then determine the gradient between the accuracies of adjacent points. If the gradient is negative we are still in the region of fast increasing accuracy. We then check if we will reach the plateau soon by looking forward one step of the current interval size. If we then can observe a flattening of the accuracy, we know that we have found the desired point in the interval between the current upper bound and the upper bound of the look-ahead-interval. We then perform binary search in this region and report the found feature subset as result res_2. The algorithm also terminates searching if the maximum accuracy of 100% is reached for the current point. In this case, the algorithm tries to reduce the dimensionality by performing binary search in the interval between the current and the last point.

Pseudocode for the algorithm is given in Figure 2. Besides res_1, the algorithm takes as input a monotonically increasing function f to determine the size of the inter-

Table 1. Comparison of Search Strategies.

DS	SA	BS	MBS
ovarian	40	39	39
	100.0%	100.0%	100.0%
	6.6 min	2.6 min	0.4 min
prostate	2,800	2,539	1,331
	94.09%	94.09%	94.41%
	41.1 min	17.5 min	12.3 min

vals. In our experiments, we obtained good results using $f(x) = x^3$. The parameter p avoids that the algorithm can not detect the plateau because of random fluctuations in accuracy. It should be set as the maximal estimated contribution of the local maxima and minima to the overall accuracy. We set p to 0.1 in all experiments. On both data sets modified binary search (MBS) is the best choice for S because it determines the smallest features sets providing the highest classification accuracy in the most efficient way (cf. Table 1).

3.3 Step 3: Selecting Region Representatives

Typical peaks in SELDI-TOF-MS data consist of continuous ranges of features (cf. Figure 4, 5). The result set res_2 contains features which have been highly ranked by R. If a region, lets say a peak, of the spectrum differentiates well among the classes, all the features of this region are highly ranked and are thus all included in res_2. However, as they are highly correlated, most of them are redundant because they represent the same information. On the other hand, there may be under-represented regions, which consist of not so highly ranked features which can contain valuable different information. In this step, we first remove the redundant features from res_2 and than add features from res_1 for under-represented regions if this leads to a further improvement in accuracy.

Removing Redundant Features. We use a forward selection method that exploits the consistency of proteomic spectra which is also assumed in binning. Spectra are often binned using a function with a linear increasing bin width. This means, in the in the area of lower m/z values fewer features are represented as one bin and in the area of high m/z values many features can be merged into one bin [3]. This is due to the fact that many different fragments of peptides with low molecular weight are causing many narrow peaks in the region of low m/z values. In the region of higher m/z values, whole peptides leading to broader peaks can be identified. We use the following simple linear increasing function b, which we call *binning function* to find a first approximation of a reasonable region size.

```
algorithm RegionRepresentatives
(dataSet res_1, dataSet res_2, classifier C): dataSet res_3
res_3 = emptySet, currentAccuracy = 0.0;
accuracy2 = C.accuracy(res_2);
while currentAccuracy < accuracy2
    res_3.add(representatives(res_2);
if currentAccuracy ¡ 1.0
   while improvement in accuracy
    res_3.add(representatives(res_1);
return res_3;

procedure representatives (dataSet DS):featureSet rep
for all features f_i in DS do
  rs: = 0.5 · b(f_i);
  if no feature f_j exists in DS with
     index(f_i) − rs < index(f_j) < index(f_i) + rs
     and quality(f_j) > quality(f_i)
     rep.add(f_i);
     DS.remove(f_i);
return rep;
```

Figure 3. Selecting Region Representatives

Definition 1 (Binning Function) *Let $s \in \mathbb{N}$. The binning function b is defined as*
$b(f_i) = max(1, index(f_i)/100 \cdot s)$

In our experiments we obtained good results for $s = 3$. For each region we now choose the best ranked feature from res_2 as representative and use C to evaluate the accuracy. For the ovarian data set we obtain 9 features and an accuracy of 100%, and we are done, since the maximum accuracy has been achieved. For the prostate data set, this results in 19 features and an accuracy of 93.48%. Since the accuracy declined from originally 94.41% on 1,331 attributes, we subsequently add in each step for each region the attribute which is best ranked among the remaining attributes and evaluate the accuracy again. For 187 attributes the accuracy of 94.41% is reached again.

Adding Missing Region Representatives. Some of the regions in our intermediate result set may be under-represented or not represented at all, since res_2 has already been a drastically reduced attribute set containing only high ranked features. Therefore, we now also use the list of ranked features of res_1. We determine for each not represented region the best representatives using the binning function and add them as long as an improvement of the accuracy can be obtained.

The pseudocode for the whole algorithm is depicted in Figure 3. The method *representatives()* selects for each region the best representative which has not been selected before. As a final step (left out in the pseudocode for simplicity), we test if there are redundant features. More precisely, we take the list of features sorted w.r.t. their

Table 2. Linear SVM and Information Gain

DS	full space	step 1	step 2	step 3
ovarian	15,154	6,238	39	9
	99.60%	99.60%	100.0%	100%
prostate	15,154	9,566	1,331	164
	90.37%	93.16%	94.41%	97.83%

Table 3. 5-NN and ReliefF

DS	full space	step 1	step 2	step 3
ovarian	15,154	15,037	66	90
	93.28%	93.20%	99.20%	99.40%
prostate	15,154	14,435	35	361
	87.27%	87.27%	85.09%	92.50%

index. We then try to leave out for each pair of neighboring features the feature which has been lower ranked by R and evaluate the accuracy again.

	c_1	c_2	c_3	c_4
c_1	63	0	0	0
c_2	0	189	1	0
c_3	0	1	22	3
c_4	0	1	1	41

For the prostate data set, we end up with 164 features and a final accuracy of 97.83%. The confusion matrix (left) for linear SVM and 10-fold cross validation shows only seven classification errors, whereas four of them are due to confusing the two different stages of prostate cancer. Our algorithm is efficient: For the prostate data set, step 3 takes 5.30 minutes, whereas for the ovarian data set we are done in 50 seconds.

4 Results

Table 2 summarizes the results using linear SVM as classifier and the information gain as ranker. All three single steps of our method reduce the dimensionality, and at the same time improve the classification accuracy for this combination of C and R. For comparison, in Table 3 the results for 5-NN as C and reliefF as R are given. Also for reliefF and 5-NN our method achieves a sound reduction of features and improvement in classification accuracy. However, we obtain better results with SVM and information gain on both data sets. A standardized investigation of the huge amount of possible combinations w.r.t. their performance on proteomic data is part of our ongoing work.

Figure 4 depicts some selected regions that have been identified by our feature selection framework on the ovarian data set. A randomly selected spectrum of the control group with highlighted reagions is depicted. Below we show the highlighted regions in more detail comparing the healthy instance to a randomly selected instance with ovarian cancer. We can confirm that a majority of the

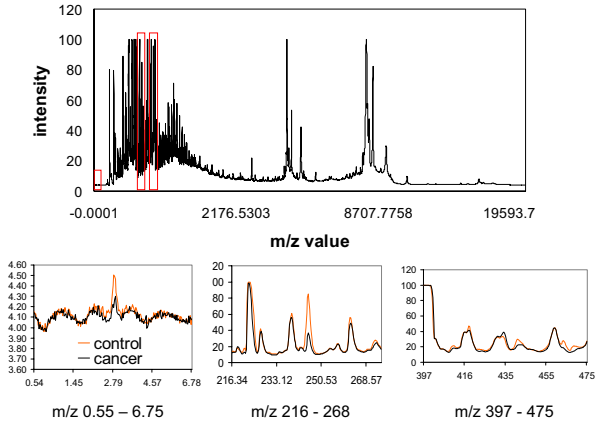


Figure 4. Results on Ovarian Data.

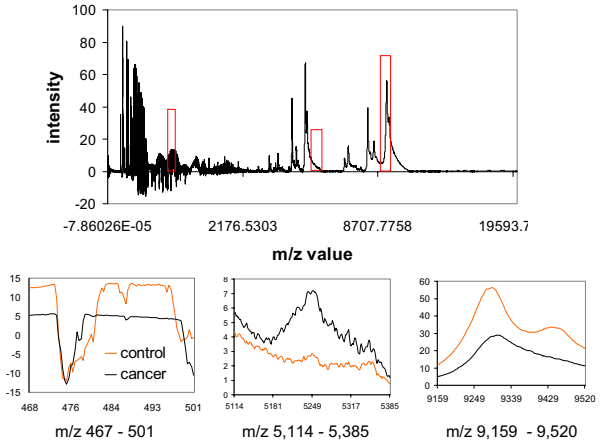
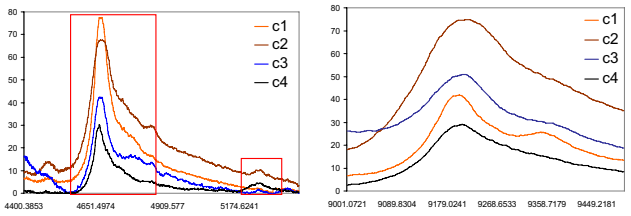


Figure 5. Results on Prostate Data.

relevant features can be found in the region of low m/z values, this has also been stated in [2]. The 9 features found using SVM and information gain are the m/z values 2.76, 25.49, 222.42, 244.66, 261.58, 417.73, 435.07, 464.36 and 4,002.46. Besides the m/z value 2.78 all these features have also been selected using reliefF and 5-NN. Among the 90 selected features with reliefF and 5-NN, 70% percent represent m/z values below 3,000.

For the prostate data set, also in the area of higher m/z values discriminating regions have been found. Figure 5 shows some selected regions. Out of the 164 selected features using SVM and information gain, the most evident changes between healthy and diseased instances can be observed in the regions representing the m/z values of approximately 500, 5,000 and 9,000. For clarity reasons, one randomly selected spectrum of class $c1$ (healthy, PSA $<$ 1 ng/mL) is compared to one randomly selected spectrum of class $c4$ (prostate cancer, PSA $>$ 10 ng/mL)



(a) m/z 4,500 - 5,500 (b) m/z 9,000 - 9,500

Figure 6. Selected Regions on Prostate Data.

w.r.t. the three highlighted regions in Figure 5. Most of the features selected by reliefF and 5-NN are also in these three areas. Besides this, more features in the region of very low m/z values have been selected using reliefF and 5-NN.

In Figure 6 we inspect the interesting regions from m/z 4,500 to 5,500 and m/z 9,000 to 9,500 in more detail w.r.t. all classes by depicting one randomly selected spectrum of each class. In Figure 6(a) there are two interesting regions which are highlighted: One is the peak between approximately m/z 4,550 and 4,850. The amount of the corresponding peptides is considerably lower for the instances with prostate cancer ($c3$ and $c4$) than for the instances with benign conditions (c_1 and c_2). The other interesting region is a peak of smaller intensity at approximately m/z of 5,250. Here the amount of the corresponding peptides is increased for the instances of the class $c4$ (prostate cancer, highly elevated PSA value) and $c2$ (healthy, elevated PSA value) w.r.t. class $c1$. The same region is also displayed in more detail in Figure 5.

In Figure 6(b) it can be seen that the abundance of the peptides corresponding to the m/z values around 9,200 is reduced for the instance of prostate cancer with a highly elevated PSA value (class c_4) w.r.t. the class of the healthy control group without elevation of the PSA value (class c_1). For both classes representing instances with marginally elevated PSA value (classes c_2 and c_3) the abundance of the corresponding peptides is increased w.r.t. the instance of class c_1. These interesting findings have to be systematically verified and analyzed for interpretation.

5 Conclusions

In this paper, we presented a framework for feature selection on high-throughput mass spectrometry data. We evaluated our method on two SELDI-TOF-MS data sets on cancer identification. On both data sets we found groups of features providing a very high sensitivity and specificity for cancer identification. This result can be used as an input for further data mining steps. Currently we focus on mining association rules on the selected features (after discretizing the numerical features) and clustering to identify unknown sub-classes. As cancer is a complex systemic disease with different stages, different sub-classes can be expected.

6 Acknowledgement

This work was supported by the GEN-AU project Bioinformatics Integration Network II.

References

[1] *"WEKA machine learning package, http://www.cs.waikato.ac.nz/ ml/weka"*. Universitity of Waikato.

[2] G. Alexe, S. Alexe, L. A. Liotta, E. Petricoin, M. Reiss, and P. L. Hammer. Ovarian cancer detection by logical analysis of proteomic data. *Proteomics*, 4(3):766–783, 2004.

[3] T. Conrads, V. Fusaro, S. Ross, D. Johann, V. Rajapakse, B. Hitt, S. Steinberg, E. Kohn, D. Fishman, G. Whitely, J. Barrett, L. Liotta, E. r. Petricoin, and T. Veenstra. "High-resolution serum proteomic features for ovarian cancer detection". *Endocrine-Related Cancer*, 11(2):163–178, 2004.

[4] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *ICML*, pages 359–366, 2000.

[5] M. A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447, 2003.

[6] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.

[7] I. Kononenko. Estimating attributes: Analysis and extensions of relief. In *ECML*, pages 171–182, 1994.

[8] L. A. Liotta, M. Ferrari, and E. Petricoin. Clinical proteomics: written in blood. *Nature*, 425(6961):905, Oct 2003.

[9] H. Liu and R. Setiono. A probabilistic approach to feature selection - a filter solution. In *ICML*, pages 319–327, 1996.

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

[11] E. r. Petricoin, D. Ornstein, C. Paweletz, A. Ardekani, P. Hackett, B. Hitt, A. Velassco, T. C, W. L, W. K, C. Simone, P. Levine, W. Linehan, M. Emmert-Buck, S. Steinberg, E. Kohn, and L. LA. "Serum proteomic patterns for detection of prostate cancer.". *J Natl Cancer Inst.*, 94(20):1576–1578, 2002.

[12] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[13] J. S. Yu, S. Ongarello, R. Fiedler, X. W. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics*, 21(10):2200–2209, 2005.