

# Finding Cancer Biomarkers from Mass Spectrometry Data by Decision Lists

Jian Liu and Ming Li  
School of Computer Science,  
University of Waterloo, Waterloo, ON N2L 3G1 Canada  
jianliu@monod.uwaterloo.ca, mli@uwaterloo.ca

## Abstract

*Finding accurate biomarkers is key to early diagnosis of many otherwise incurable diseases. We study the problem of finding biomarkers for mass spectrometry (SELDI-TOF) spectra from cancerous and normal tissues. In contrast to the common practice of using vague methods, such as genetic algorithms, or un-interpretable (as biomarker) methods, such as SVM, we looked for a method that is simple, intuitive, interpretable, usable, and more accurate.*

*We introduce decision-lists to this domain. Our experiments on clinical cancer datasets show decision lists give more accurate results than other methods. More interestingly, the resulting decision lists are more interpretable, for possible causal relationship between cancer and differentially expressed proteins, and directly usable in clinical biomarker design.*

## 1 Introduction

Biomarkers are cellular, molecular or genetic alternation or patterns caused by the presence of specific diseases. Biomarkers serve as the indicators of diseases, they can also be employed to predict disease severity and monitor effectiveness of medical treatment. The discovery of biomarkers have been conducted in some major clinical areas such as cardiovascular diseases and cancers.

Combined with the protein chip technology, mass spectrometers such as SELDI-TOF have been used to diagnose cancers. In such an approach, human tissue samples are collected to produce protein spectra. Every protein spectrum consists of a sequence of peaks, each of which is characterized by its  $m/z$  ratio and intensity of a protein. The proteomic patterns are then analyzed to predict whether or not the patients have developed certain cancers. Our work focuses on this type of clinical data, while it also applies to other areas such as biomarkers in microarray data.

In recent years, a number of methods have been developed to classify mass spectra, and high sensitivity and

specificity results were reported for different cancers. In this work, we introduce decision lists [1] to this domain.

## 2 Decision Lists for Biomarker Discovery

A decision list usually expect symbolic attribute-value pairs (as Boolean variables, the value is binary). Therefore, a set of attribute-value pairs must be derived from the raw mass spectra. Given a spectrum, an attribute is a protein characterized by the  $m/z$  ratio, its value is the type of class it may belong to with maximum likelihood. For the sake of argument, let us consider the case of binary classification. In such a problem, the training data are two sets of spectra, one is generated from normal human tissues, while the other from cancer patient tissues. Hereafter, they are referred to as normals and cancers, respectively.

Each spectrum may consist of peaks for hundreds or even thousands of proteins. For each protein, its intensity is subject to individual sample and the experimental condition variability. We assume that intensity of each protein follows a normal distribution within each classes. For each individual peak, if the intensity suggests that the protein is more likely from normal sample, the value is 1, otherwise 0. Following the above steps, each spectrum is transformed into a set of attribute and binary value pairs. It has been demonstrated that 5-20 proteins would suffice to classify mass spectra [2] [3]. Inspired by such facts, we expected to identify the a small set of proteins as the discriminants. Given a protein, two groups of intensity can be collected from normal and cancer respectively.  $t$ -test is utilized to assess the its power of discriminating two groups. The proteins are sorted by their  $t$ -scores, and the top-ranked proteins are chosen to form the attribute set.

In this work, each decision list acts as a biomarker. Decision lists are constructed upon given training spectra.

We extend a  $k$ -decision list to a  $(k, l)$ -decision list by allowing  $l$  alternative monomials at each node ( $H_i$ ), provided that all alternative monomials at each node are equivalent in the sense that they give the same classification for the training examples. Thus a  $(k, l)$ -decision list of  $L$  nodes

**Table 1. Performance comparison of DL, SVM, and C4.5 on Four Cancer Datasets**

| (a) Training results |                       |   |                  |             |
|----------------------|-----------------------|---|------------------|-------------|
| data set             | # of normals /cancers | # of correctly classified normals/cancers |                  |             |
|                      |                       | DL  | <i>SVM-light</i> | <i>C4.5</i> |
| Ovarian I            | 45/81                 | 45/80                                     | 44/79            | 44/81       |
| Ovarian II           | 58/50                 | 56/49                                     | 58/50            | 54/48       |
| Ovarian III          | 58/50                 | 57/49                                     | 51/40            | 52/42       |
| Prostate             | 127/35                | 120/28                                    | 121/30           | 125/32      |

  

| (b) Testing results |                       |   |                  |             |
|---------------------|-----------------------|---|------------------|-------------|
| data set            | # of normals /cancers | # of correctly classified normals/cancers |                  |             |
|                     |                       | DL  | <i>SVM-light</i> | <i>C4.5</i> |
| Ovarian I           | 46/81                 | 44/81                                     | 45/78            | 44/78       |
| Ovarian II          | 58/50                 | 53/48                                     | 51/48            | 46/45       |
| Ovarian III         | 58/50                 | 50/48                                     | 49/40            | 49/35       |
| Prostate            | 126/34                | 114/26                                    | 119/26           | 113/21      |

corresponds to  $l^L$  normal decision lists.

**Theorem 1**  $(k, l)$ -decision lists are pac-learnable.

**Proof.** A simple algorithm is designed to learn  $(k, l)$ -decision lists: At each step  $i$ , find as many size-up-to- $k$ -monomials as possible such that

- they all cover the same uncovered cancer (normal) examples without covering any uncovered normal (cancer) examples and
- the cover size is within  $1/2$  of the maximum cover size.

Put these monomials in the  $H_i$  node.  $O_i$  is positive (negative). Repeat for  $H_{i+1}$  until either all cancer examples are covered, then last  $O_j$  is normal, or all normal examples are covered, then the last  $O_j$  is cancer.

It is easy to show that this is a pac-learning algorithm, following similar proofs as [1]. This algorithm allows us to produce many alternative  $k$ -decision lists when there is very little data.

This algorithm allows us to produce many alternative  $k$ -decision lists when there is very little data. In practice, we need to carry out classification for multiple classes. For instance, cancers can be in different stages, while some tumors are possibly benign. In a similar manner, we can construct decision lists for multi-classification.

### 3 Experimental Results

The clinical benchmark from [4] was used in the experiment. It consists of four datasets: one is prostate cancer samples, the other three (WCXI, WCXII, Lacent) are ovarian cancer samples. These datasets were produced with different protein chip technologies. We consider 2-decision lists and 3-decision lists in our experiments, i.e., each monomial consists of at most 2 or 3 proteins.

As comparison, Support Vector Machine *SVM-light* and Decision Tree classifiers *C4.5* were also tested. Table 1 presents the empirical training and testing results for the four data sets respectively. In general, that accuracy of decision list classifiers was better than those of *SVM-light*, and *C4.5*.

### 4 Conclusions

We have proposed a new approach of using decision lists which possesses the following advantages. The concept class of decision lists captures nicely our intuition of a diagnosis process. The decision lists are interpretable and usable directly for biomarker design. Our results are more accurate than other more complicated methods. This method was tested with a set of clinical mass spectrometry proteomic data, and achieved better results than SVM and decision tree. When multiple solutions exist, our paradigm and the algorithm allow us to output a list of alternate solutions (decision lists) for the domain expert to select.

### References

- [1] R. L. Rivest. Learning decision lists. *Machine Learning*, V2:229–246, 1987.
- [2] E. F. Petricoin, A. M. Ardekani, and et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, V62:3609–3614, 2002.
- [3] T. P. Conrads, M. Zhou, and et al. Cancer diagnosis using proteomic patterns. *Expert Rev. Mol. Diagn.*, V3(4):411–420, 2003.
- [4] Clinical proteomics program databank: <http://clinicalproteomics.steem.com>, 2003.