**Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)**

# Extracting Efficient Fuzzy If-Then Rules from Mass Spectra of Blood Samples to Early Diagnosis of Ovarian Cancer

A.Assareh, M.H.Moradi

Faculty of Biomedical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran.

*Abstract*- Among the many applications of mass spectrometry, biomarker pattern discovery from protein mass spectra has aroused huge interest in the recent years. While research efforts have raised hopes of early and less invasive diagnosis, they have also brought to light the many issues to be tackled before mass-spectra-based proteomic patterns become routine clinical tools. Undoubtedly, biomarker selection among the high dimensional input data is the most critical part of each pattern recognition algorithm applied to this area. In this paper we pursued a new feature selection strategy that explores all data points as initial features rather than just peaks. Using the derived features in conjunction with only two intuitive fuzzy rules, we achieved a considerable accuracy over a couple of well-known ovarian cancer datasets.

*Keywords:* **Mass Spectroscopy, Ovarian Cancer, Biomarker, Data Mining, Fuzzy Linguistic Rules.**

## I. INTRODUCTION

Cancer is a major public health concern in all over the world. Cancer accounts for one of every four deaths in the US. Currently the best way of reducing the morbidity and mortality of cancer is to detect and treat it in the earliest stages [1].In particular the early diagnosis of ovarian cancer has the potential to reduce the mortality associated with this disease down to satisfactory level [2].

A biomarker is a biologically derived molecule in the body that indicates the progress or status of the disease [1].Biomarkers under investigation include genes, proteins and small molecules, which base four major bioinformatics branches: genomics, transcriptomics, proteomics and metabonomics.

In proteomics studies, the usage of mass spectrometry profiling of patient serum proteins, combined with advanced data mining algorithms, to detect protein patterns associated with malignancy, has been reported as a promising field of research to achieve the goal of early cancer detection [2].Mass spectrometry provides rapid and precise measurements of the size and relative abundance of the proteins presents in a complex biological/chemical mixture biomarkers which can distinguish between cancer and normal samples and at the meantime, provide the ability to convenient interpretation of the results for the physician.

A mass spectrum usually contains thousands of different mass/charge (m/z) ratios on the x-axis, each with corresponding signal intensity on the y-axis. For data mining purposes, each m/z ratio is represented as a distinct variable whose value is the intensity; hence each case can be seen geometrically as a single point in a very high-dimensional space. In classification for diagnosis and biomarker discovery, the problem of high dimensionality is compounded by small sample size: diseased specimens are relatively rare and difficult to collect, especially when invasive procedures are involved. This twofold pathology, called the high dimensionality- small-sample (HDSS) problem, is the main issue that plagues and propels current research on protein mass spectra classification [3].

Dimensionality reduction is crucial to biomarker discovery. First, the curse of dimensionality must be coped with if the classification problem is to be solved at all. Whatever the classification goal, the most effective way so far to get around the HDSS problem is by reducing the size of the variable set. More importantly, extracting a handful of variables from an initial set of several thousands is not a simple preprocessing expedient but the very goal of biomarker discovery. The final variables and their interaction in the learned model constitute the proteomic signature, which the biomedical researcher must then identify, validate, and interpret. In short, dimensionality reduction and classification are the co-essential goals of mass spectra mining for biomarker discovery [3].

Aspects of knowledge representation and reasoning have dominated research in fuzzy set theory (FST) for a long time, at least in that part of the theory which lends itself to intelligent systems design and applications in artificial intelligence[4]. The significance of fuzzy set theory in the realm of pattern recognition is adequately justified in

- Representing linguistically phrased input features for processing.
- Providing an estimate (representation) of missing information in terms of membership values.
- Representing multiclass membership of ambiguous patterns and in generating rules and inferences in linguistic form [5].

So far, many pattern recognition studies have done in the investigation of cancer diagnosis from mass spectra [6]. Although some of these studies utilized methods that lead to high sensitivity and specificity, but lack of intuitive insight of results make them completely "black box" approaches that are almost uninterpretable for the physician. To the best of our knowledge, the fuzzy rule based system has been rarely used in proteomics mass spectroscopy domain, while it has shown a great potential in various areas of knowledge discovery in database (KDD).

The main objective of this paper is to explore a simple framework of linguistic rule building to extract knowledge hidden in proteomics raw datasets which to some extent can improve the 'black box' drawback of other methods. In the other word, this approach will simplify the mutual knowledge communication between human expert and classifier that may lead to following advantages: knowledge acquisition from generated rules, extend the physician insight into the achieved results and consequently better validation of the results, and the ability of employ biological knowledge in the decision making via adding some extra rule to the rule set by the human expert.

## II. DATASET

This research takes the surface enhanced laser desorption-ionization time-of-flight (SELDI-TOF) mass spectrometry (MS) serum proteomic patterns as input.

Serum SELDI-TOF spectra data were used from patients and a healthy screening population.

Here, we utilized the two well-known MS datasets of ovarian cancer available at American National Cancer Institute (NIC) website. The first sample set, so called dataset includes 91 controls and 162 ovarian cancers and the second one consists of 100 normal, 16 benign and 100 ovarian cancer samples. Here, we dedicated roughly 70% of each dataset samples for training and the others for testing.

A mass spectrum is a curve where the x-axis indicates the ratio of the weight of a specific molecule to its electrical charge (M/Z, in Daltons per unit charge) and the y-axis is the signal intensity for the same molecule as a measure of the abundance of that molecule in the sample.

Each mass spectrum curve represents the expression profile of 15154 peptides defined by their M/z ratios with corresponding intensities. Figure 1 shows a sample of the dataset before and after preprocessing.

## III. PREPROCESSING

A typical mass spectrum consists of signals, and noise. The noise is the undesired interfering signal caused by sources unrelated to the biochemical nature of the sample being analyzed and the signal is the relative abundance of ions originating from the peptides, proteins, and contaminants present in a sample [1]. Here we have two kinds of noises: low frequency (or baseline) and high frequency. The baseline is the slowly varying trend under the spectrum; and the high frequency noise consist of chemical background, electronic noise, signal intensity fluctuations , statistical noise, warping of the signal shapes (due to overcharging in ion traps), and statistical noise in the isotopic clusters[3].

The goal of preprocessing stage is to ''clean up'' the data such that machine learning algorithms will be able to extract key information and correctly classify new samples based on a limited set of examples [1].

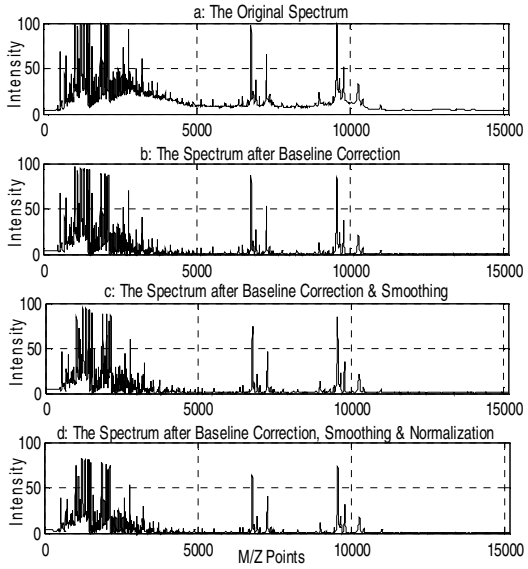


Figure 1. typical mass spectra before preprocessing(a), after baseline correction(b), smoothing(c) and normalization(d).

In analyzing mass spectra of blood samples, the preprocessing stage roughly includes three main tasks: baseline correction, smoothing and normalization. In followings, we will discuss the strategies we employed in the mentioned steps.

### *A. Baseline Correction*

Mass spectra exhibit a monotonically decreasing Baseline which can be regarded as low frequency noise because the baseline lies over a fairly long mass-to-charge ratio range.

In this study, we utilized local average within a moving window as a local estimator of the baseline and the overall baseline is estimated by sliding the window over the mass spectrum. The size of the applied window was 200 M/Z. In addition shape-preserving piecewise cubic interpolation has been applied to regress the window estimated points to a soft curve.

### *B.Smoothing*

Mass spectra of blood samples also exhibit an additive high frequency noise component. The presence of this noise influences both data mining algorithms and human observers in finding meaningful patterns in mass spectra. The heuristic high frequency noise reduction approaches employed most commonly in studies to date are smoothing filters ,the wavelet transform (WT) , or the deconvolution filter[1]. Here we employ a locally weighted linear regression method with a span of 10 M/Z to smooth the spectra. Figure 2 illustrates the smoothing effect on a section of a typical spectrum.

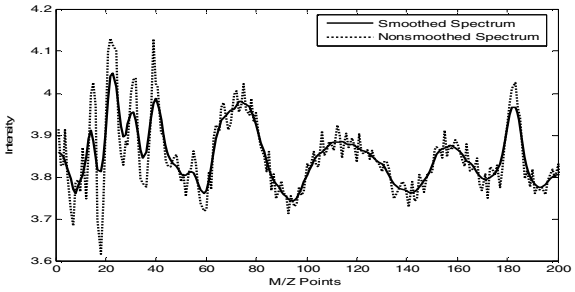


Figure 2. A section of a typical spectrum before and after smoothing.

### *C.Normalization*

A peak in mass spectra indicates the relative abundance of a protein; therefore, the magnitudes of mass spectra cannot be directly compared with each other [1].

Normalization methods scale the intensities of mass spectra to make mass spectra comparable. We normalized a group of mass spectra by standardizing the area under the curve (AUC) to the group median.

## IV. FEATURE EXTRACTION & SELECTION

Since abundance data from within the mass error rate are considered to represent the same protein, features are often extracted from mass spectra based on the properties of ‘‘peaks’’ that are comprised of multiple M/z points[1].However there is no guarantee based on which we assure that each peak represent one and only one protein. Apparently, if only peaks are taken into account, some other major discriminant features are neglected. In the present study we considered the abundance (intensity) information of every point in preprocessed mass spectra as the initial features. To avoid regional correlation between the selected features, we used regional information to outweigh the value of potential features using following factor:

$$W= (1-Exp(-(Dist/2)^2)) \qquad (1)$$

Where Dist is the distance between the candidate feature and previously selected features. A small Dist (close to 0) outweighs the significance statistics of only close features. This means that features that are close to already picked features are less likely to be included in the output list. Combining this weighting factor with the “T test” feature selection algorithm over the training set [7], we derived following 10 M/Z indices as the most discriminant features in the first dataset: 2239, 2690, 1645, 2242, 1677, 2236, 2694, 3551, 1775 and 1681.Conducting a same method, the biomarkers of the second dataset were selected as follows :3930, 4177, 3926, 4181, 6291, 3934, 6211, 4174, 4244 and 3748.

## V.CLASSIFIER

Figure 3 illustrates the discrimination power of the first 4 features (biomarkers) by showing the distribution of biomarkers values over the cancer and the normal groups in

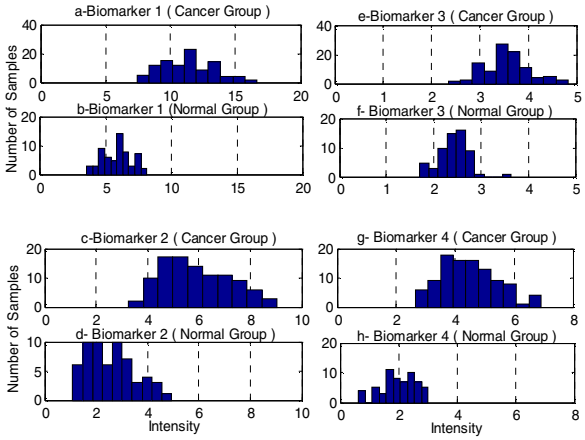


Figure 3. Histogram analysis of the first 4 derived biomarkers in the first dataset.Obviously, the distribution of biomarkers intensities over cancer and normal groups are separable.

the first dataset. Considering the histograms of the 4 biomarkers, it quickly comes to mind that the two classes are separable just by defining two simple linguistic rules over the markers. If we assign two membership functions, so called low and high, for each of the first 3 markers, the intuitively derived linguistic rules are as follows:

*In applied sample,*

*“ if biomarker 1 is low & biomarker 2 is low & biomarker 3 is low, then the sample belongs to the Normal group.”*
*“ if biomarker 1 is high & biomarker 2 is high & biomarker 3 is high, then the sample belongs to the Cancer group.”*

Consequently, we utilized fuzzy logic as the powerful tool of soft computation using linguistic variables and rules. To design two fuzzy sets, high and low, as the linguistic values of the features .We utilized Gaussian membership functions and adjusted their parameters with respect to histogram analysis of the feature value distribution over the training samples. Figure 4 illustrates the designed membership functions of biomarkers of the first dataset.

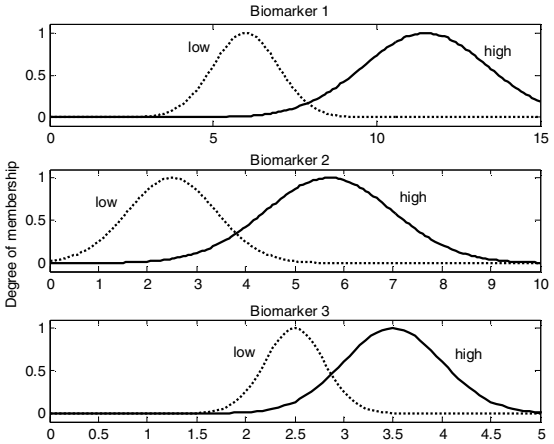


Figure 4. Membership function of the fuzzy if-then rules, designed with respect to histogram analysis of the biomarkers.

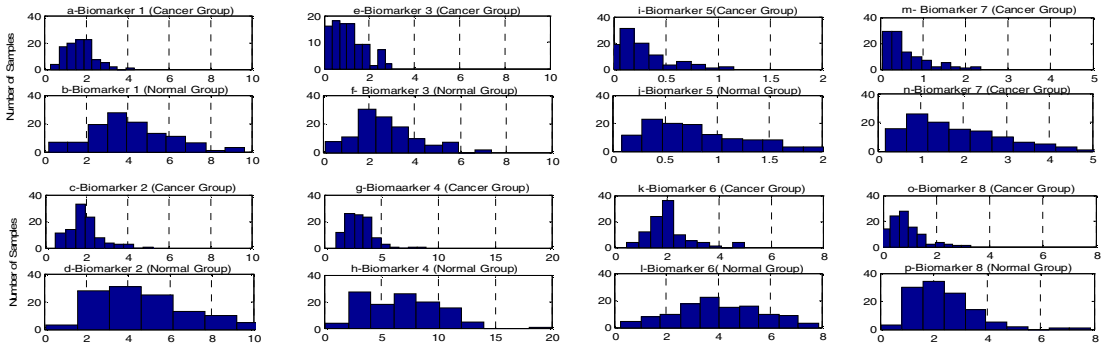


Figure 5. Histogram analysis of the first 8 derived biomarkers in the second dataset.Overlap between the biomarkers intensity distribution, makes the classification more complicated compairing to the first dataset.

Using only the two above-mentioned rules, we achieved perfect accuracy over the test dataset by following typical properties for our inference engine:
Product inference engine (comprising individual-rule base inference with union combination, Mamdani's product implication, algebraic product for all T-norm operation and maximum for all the S-norm operation), singleton fuzzifier and maximum defuzzifier[8].

But for the second dataset, the condition is not as straightforward as the first case. Figure 5 shows histogram analysis of the first 8 biomarkers over cancer and normal groups (we considered benign subjects as normal ones).Obviously the overlap of intensity distribution between cancer and normal groups, makes the classification process more complicated. This led us to utilize 3 more biomarkers to achieve a satisfactory result. Similar to the first dataset, here we defined two linguistic rules over the selected markers as follows:

*In applied sample,*
*"if biomarker 1 is low & biomarker 2 is low & …& biomarker 6 is low, then the sample belongs to the Cancer group."*
*"if biomarker 1 is high & biomarker 2 is high &…& biomarker 6 is high, then the sample belongs to the Normal group."*

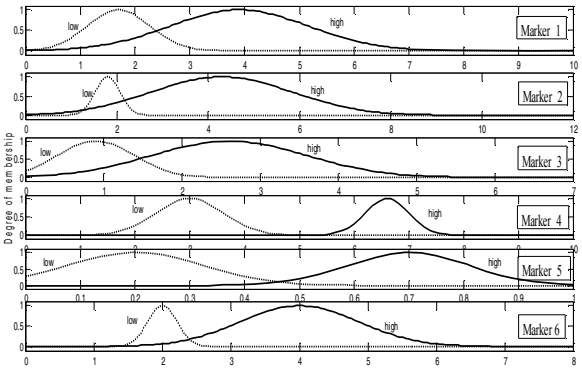


Figure 6. Membership function of the fuzzy if-then rules, designed with respect to histogram analysis of the biomarkers and adjusted by genetic algorithm.

Here, in addition to histogram analysis, we used genetic algorithm [9] to adjust the input membership functions of the rules. Figure 6 illustrates the resulting membership functions.

## VI. RESULTS

Diagnostic tests are typically evaluated in terms of their sensitivity and specificity. Sensitivity is the fraction of disease cases that are correctly identified as disease. Specificity is the fraction of non-disease cases that are correctly identified as non-disease.

We also implemented two classifiers, based on two popular methods: linear discriminant analysis (LDA) and K nearest neighbor (KNN), to assess our system performance from classification view. All of the classifiers were driven using the same features. For each evaluation we used similar preprocessed dataset. As priory mentioned, we randomly hold out approximately 30% of each dataset for test. The results are illustrated in table 1 and show the excellence of proposed fuzzy system comparing to both KNN and LDA methods.

**Table 1**: Comparison of classification performance of the proposed method with LDA and KNN over the two datasets.

| Dataset 1 | | | |
|---|---|---|---|
| Classifier | Sensitivity | Specificity | Total Accuracy |
| KNN | 98.39% | 100% | 98.91% |
| LDA | 100% | 100% | 100% |
| Proposed Method | 100% | 100% | 100% |
| **Dataset2** | | | |
| Classifier | Sensitivity | Specificity | Total Accuracy |
| KNN | 83.33% | 55.56% | 68.18% |
| LDA | 100% | 52.78% | 74.24% |
| Proposed Method | 90% | 83.33% | 86.36% |

## VII. CONCLUSION

In this paper, we described a new method for feature selection which considers all of the M/Z points as potential features, rather than just considering peaks. We applied a

weight factor in 'T test' feature selection algorithm to eliminate those feature which are regionally correlated and thus likely to correspond to a same peptide. Intuitive linguistic rules are then built based on histogram analysis of the biomarkers intensities and the membership function are adjusted using genetic algorithm. Our emphasis was to extract limited number of simple rules, which can be understood by human being as knowledge and furthermore human expert can add his/her knowledge to the system by a number of rules. In the other word, the proposed method simplifies the "knowledge exchange" procedure between human being and classification system. Moreover from classifier perspective, and using only two simple linguistic rules, the proposed method achieved a satisfactory accuracy and outperformed two well-defined classification methods: LDA and KNN.

## REFERENCES

[1] H. Shin, M.K. Markey, "*A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples*", Journal of Biomedical Informatics 39, pp. 227–248, 2006.

[2] J. M. Sorace and M.Zhan, "*A data review and re-assessment of ovarian cancer serum proteomic profiling*", BMC Bioinformatics, June.2003.

[3] Melanie Hilario, Alexandros Kalousis, Christian Pellegrini, and Markus Muller, *"PROCESSING AND CLASSIFICATION OF PROTEIN MASS SPECTRA"*, Mass Spectrometry Reviews by Wiley Periodicals, 25, pp. 409– 449, 2006.

[4] Eyke Hüllermeier, *"Fuzzy methods in machine learning and data mining:Status and prospects",* Fuzzy Sets and Systems, 156 ,pp. 387–406, 2005.

[5] Sushmita Mitra, Sankar K. Pal,*"Fuzzy sets in pattern recognition and machine intelligence",* Fuzzy Sets and Systems 156 , pp.381–386, 2005.

[6] L. Li, H. Tang, Z. Wu, J.Gong, M.Gruidl,J.Zou, M. Tockmanb and R.A. Clark, " *Data mining techniques for cancer detection using serum proteomic profiling*", Artificial Intelligence in Medicine 32, pp.71—83, 2004.

[7] Robert F.Woolson, William R.Clarke, *Statistical Methods for the Analysis of Biomedical Data,* Wiley-Interscience, second edition, 2002, chapter 6.

[8] Li-Xin Wang, *A Course in Fuzzy Systems and Control*, Prentice-Hall International Limited, UK, 1997.

[9] W.Pedrycz and Z. A. Sosnowski., "*Genetically Optimized Fuzzy Decision Trees*", IEEE transactions on systems, man and cybernetics -part B: cybernetics, Vol. 35, No. 3, June 2005