

Mass Spectrometry Diagnostic Software for Cancer Detection - addressing geographical limitations

Marcin Radlak, student

School of Computer Science
University of Birmingham
Edgbaston
B15 2TT Birmingham,GB
Email: mpr683@cs.bham.ac.uk

Ryszard Klempous, PhD

Institute of Control and Optimization
Wroclaw University of Technology
ul. Wybrzee Wyspiaskiego 27
50-370 Wroclaw, Poland
Email: ryszard.klempous@ict.pwr.wroc.pl

Abstract

This paper presents steps required to detect a cancer disease based on data obtained from SELDI-TOF-MS. Here, the full process of detection: from raw data, through pre-processing towards classification has been outlined. Importantly, methods and algorithms are presented and described in terms of their usability. Moreover, based on the analysis software developed for the purpose of this work, comparison of classifiers performance based on preprocessing methods is conducted. Finally, guidelines for further research are indicated together with suggestions of how to apply the concept of 24/7 work organization to make the process of development and research faster.

Keywords: Cancer Detection, Data Analysis, 24h Knowledge Factory, Diagnostic Software

1 Introduction

Cancer disease is the one which is very serious and every year affects millions of people all over the world. To reduce this amount, extensive work, by research centers funded by pharmaceutical companies or charity organizations is conducted to develop methods of prediction, prevention and treatment. This paper presents methods for cancer detection using data obtained from SELDI-TOF-MS. This type of Mass Spectrometer has been proposed, because of its ability to produce high resolution spectrogram of proteins content in an organic sample. Assuming, that cancerous cells consist of proteins which are usually absent in healthy tissue, there is a hope to develop a method, which will allow to distinguish between those two states, giving a solid base for final diagnosis which doctors have to make. Although, hardware is a breakthrough in the field it is still not precise enough to produce

superior results expected from diagnosis equipment. Low repeatability of results, high noise and huge amount of data are only few of the difficulties encountered. Therefore, to supplement hardware deficiency, it is important to utilize more or less intelligent data analysis algorithms. It is still unknown which of them are the best, but properly selected might, as some research show, significantly improve accuracy of the hardware. In this paper, we will suggest a process of detecting cancerous/healthy sample: from raw data, through pre-processing towards classification. Methods and algorithms, their characteristics and suggested implementation indications are presented. Analysis software has been developed using Matlab IDE. Additionally, algorithms will be compared to find the best performing method methods. Finally, the summary and future directions are proposed to create some general indicators of the future work and to point a research in the correct direction.

The paper is organized as follows: in section 2 we present a brief description of methods used in SELDI-TOF-MS data manipulation, divided into 3 groups: a) preprocessing, b) analysis and c) detection. In section 3 we provide a comparison between them and comment on results. Section presents the concept of 24/7 work organization which could improve cooperation between researchers. Finally, in section 5 we present a conclusion.

2 Analysis step by step

The concept of SELDI-TOF-MS data analysis is to classify spectrogram produced from sample tissue as cancerous or healthy. Figure ?? presents 216 samples of ovarian cancer grouped into cancer (black) and healthy (grey) - downloaded from from National Cancer Institute website <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>.

Evidently, spectrograms for cancerous and normal samples differ significantly. When looking at a population of results, it is easy to classify samples, but often single spectrograms do not contain proteins which all remaining spectrograms of the same class consists of. This may lead to misclassification. With medical software, any false classification is unacceptable.

Therefore, many issues arise when analyzing spectrograms. First of them is difficulty in detecting peaks - often high intensity for specific M/Z ratio in different samples is not high enough to be significant. There is also problem of overlapping samples, presence of noise, shifts in values (vertical and horizontal) and many more with high dimensionality at the end. To systemize process of data manipulation, following steps should be performed: preprocessing, analysis and classification. This way success - correct classification - is more likely to be achieved.

2.1 Preprocessing

Preprocessing can significantly increase performance of classifier. To be able to classify different samples, it is important to prepare them for this process. Different factors should be eliminated to prepare data for analysis and classification. If the samples were not preprocessed, classifiers may detect normal sample as cancer, or cancer as normal what, if it was medical tool, could result in very dramatic consequences. Person who is healthy could be mistakenly diagnosed as having cancer and would be referred for unnecessary treatment which is costly and very depressing. On the other side, ill person diagnosed as healthy, may loose chance of being cured , before the disease state is advanced. Most common preprocessing methods are: dimension reduction, Following are some suggested preprocessing methods, i.e. PCA, which speeds up the analysis amazingly; baseline correction, what shifts samples to the same level with respect to zero; normalization, which emphasizes differences between samples where they actually occur, rather than those which originate from different measurement conditions; denoising and peaks alignment. There is no guarantee, that applying all of them will significantly improve classifier performance, therefore, this paper will compare classifier performance with respect to applied set of preprocessing methods.

2.2 Data Analysis

Principle Component Analysis (PCA) is a mathematical transformation of a set of many correlated variables into a smaller set of uncorrelated variables - principal components. This is performed as follows (Smith (2002)):

1. Prepare data set
2. Subtract the mean
3. Calculate the covariance matrix
4. Calculate the eigenvectors and eigenvalues of covariance matrix
5. Choose components and form feature vector
6. Derive new data set

Most significant advantage of PCA is that resulting data set is much smaller. However, key properties have been kept thus classifier will work faster and should produce results very similar to those, which would be obtained if raw data have been applied.

2.3 Classification

To develop a software which can be used for cancer diagnosis based on mass spectrometry, classifying algorithm is required. Previous steps were used to prepare raw data so noise and technology deficiencies were corrected. This ensures that input for classifier is always in the same range, with real rather than artificial differences

allowing to distinguish between normal or cancer sample. With this requirement in mind many classifying algorithm have been developed and all the research work is focused on developing new ones, which will be applicable to the specific data set.

Lilien et al. (2003) suggests that classification algorithms can be divided into two major groups: first, those algorithms which depend on the data and experiment conditions, return different results. The output is not determined and depend on the initial setup. Therefore, they are called heuristic. Second group of algorithms are mathematical models, which for specific input, always return the same output. Those are called deterministic or exact models.

2.3.1 Heuristic Approach to classification

Neural Networks Artificial Neural Networks (ANNs) are mathematical model of human brain. They are based on cooperation of many nonlinear processing units (neurons) connected in a network. As real neurons, artificial ones (simplified) have inputs and outputs. By inputs they collect outputs from neurons they are connected to, and if the input is strong enough to activate the neuron, it produces the output. By combining neurons in multi layer networks, it is possible to solve many complex problems, which are impossible to solve (in reasonable time) by traditional mathematics i.e they tend to approximate functions very well with reasonably small computing power requirements. Some of ANN's applications include: speech and pattern recognition, image recognition, financial prediction and many more. They also perform very well in proteomics and the results are described in following papers Ball et al. (2002), Zhou et al. (2005) or Yu & Chen (2005). Difficulty with Neural Network is that there are many parameters to set. It is thus very laborious task to adjust them correctly so the error on the output is minimized. We will use single layer perceptron and 2 layer feed forward neural network for further comparison.

2.3.2 Deterministic Approach to classification

While heuristic approach is sometimes the only way of solving problem, it is very common that results differs even if input data has not been changed. But sometimes it is possible to use exact algorithm, which will always produce determined result, for the same input data. There are many tools available, but only some of them can be applied for mass spectrometry analysis. Few of them, which have been used are described as follows.

Linear Discriminant Analysis (LDA) Implemented in so called Q5 algorithm by Lilien et al. (2003), seems to perform very well when compared to other methods. The dimension has to be reduced before LDA is applied so (PCA) is performed to accomplish it. LDA, similar to PCA, searches for linear combination of variables which best describes the data. But the difference is that LDA also models

differences between them whereas PCA does not explore it. As defined, "LDA approaches the problem by assuming that the probability density functions $p(\vec{x}|y = 1)$ and $p(\vec{x}|y = 0)$ are both normally distributed, with identical full-rank covariances $\Sigma_{y=0} = \Sigma_{y=1} = \Sigma$. It can be shown that the required probability $p(y|\vec{x})$ depends only on the dot product $\vec{w} \cdot \vec{x}$ where $\vec{w} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$. That is, the probability of an input x being in a class y is purely a function of this linear combination of the known observations." When LDA classifier is trained, mean and covariance is calculated because those parameters are not known. But training in this case is shorter than time consuming heuristic methods.

k-nearest neighbors k-Nearest Neighbors is another method of cluster analysis. The algorithm examine data and divide them into a predefined number of classes. Those classes contains categories of parameters which are derived from data during training process. After algorithm is trained, when a test sample is applied, classifier finds the k nearest neighbors and assigns their label names to the training data set. Importance of each neighbor is weighted by its rank presented in terms of the distance to test sample.

2.4 Verification

Classifier has to be verified to show how well does it perform for input data. Data set is divided into 2 subsets: training set and verification set and it is important to perform this process properly. For freely available mass spectrometry data, quantity of samples is limited. Therefore k-fold cross validation, bagging or boosting algorithms have been proposed to increase the data set. To decide about classifier's efficiency, few terms have been proposed to describe how classifier performs. Classifier performance may be described as a percentage of positive tests which correctly indicate the presence of disease (called positive predictive value - PPV), or percentage of negative tests which correctly indicate the absence of disease (negative predictive value - NPV). Further details about validation requirements were suggested by Vitzthum et al. (2005)

3 Results of comparison

Mass spectrometry data analysis software has been developed to examine efficiency of the following methods: k-Nearest Neighbors (KNN), PCA+LDA (PCA), Perceptron (PER) and Feed-Forward Neural Network (FFN) trained with Back-Propagation algorithm. Therefore, dependency of preprocessing methods in final classifier performance has been examined. Sample data *Center for Cancer Research Data Bank* (n.d.) have been equally divided into training and evaluation set. Training and classification has been repeated 50 times to create statistically significant results.

Table 1: Classifier performance based on preprocessing method

	KNN	LDA	PER	FFN
No preprocessing				
% Correct	91.74	98.28	96.44	92.33
St. Dev.	3.00	1.12	1.80	4.40
NPV (Mean)	89.24	98.66	98.02	93.32
NPV (StDev)	4.75	1.84	2.67	7.18
PPV (Mean)	92.06	98.07	95.46	92.79
PPV (StDev)	3.66	1.90	2.82	5.76
Baseline Corrected				
% Correct	91.25	98.35	96.50	93.19
St. Dev.	3.16	1.20	1.78	3.50
NPV (Mean)	90.83	98.41	97.84	95.948
NPV (StDev)	5.12	1.88	2.35	4.93
PPV (Mean)	91.89	98.37	95.68	92.22
PPV (StDev)	3.24	1.72	2.94	6.00
Normalization				
% Correct	90.81	98.68	97.27	94.24
St. Dev.	2.97	0.89	2.01	4.24
NPV (Mean)	89.08	99.10	98.20	94.47
NPV (StDev)	5.08	1.43	2.39	7.71
PPV (Mean)	92.60	98.49	96.74	95.01
PPV (StDev)	3.25	1.49	3.11	4.10
Smoothing				
% Correct	90.81	97.96	95.65	93.28
St. Dev.	2.94	1.40	2.14	3.63
NPV (Mean)	88.87	98.62	97.33	94.81
NPV (StDev)	4.74	1.42	2.93	5.78
PPV (Mean)	92.71	97.53	94.66	93.02
PPV (StDev)	3.01	2.20	3.27	5.29
Baseline Corr. + Normalization + Smoothing				
% Correct	92.09	98.28	95.95	91.27
St. Dev.	2.94	1.17	2.39	13.67
NPV (Mean)	91.19	98.11	94.66	88.17
NPV (StDev)	4.75	2.27	4.40	14.33
PPV (Mean)	93.12	98.49	97.28	94.82
PPV (StDev)	3.24	1.57	2.48	14.17

Training and evaluation data were randomly selected for each run to avoid repetition of the same set. Results are presented in Table 1. It is not straightforward to say that applying preprocessing methods will increase performance of a classifier. Some of the methods decrease performance. I.e When only normalization has been applied, performance of LDA has been increased. Also, standard deviation has been lowered indicating that classifier produce much more repeatable output. On the other side, when three different preprocessing methods were applied, Baseline Correction, Normalization and Smoothing, apart from k-Nearest Neighbors, all classifiers performed worse than if no preprocessing was applied. This leads to following conclusion: at first - preprocessing methods should be carefully chosen to not loose important information which raw data contain, secondly - performance of the classifier might be increase by applying correct preprocessing methods, but it is important to notice, that if the classifier is well constructed, it can be superior, even if no preprocessing is applied.

4 24h work organization - accelerating research

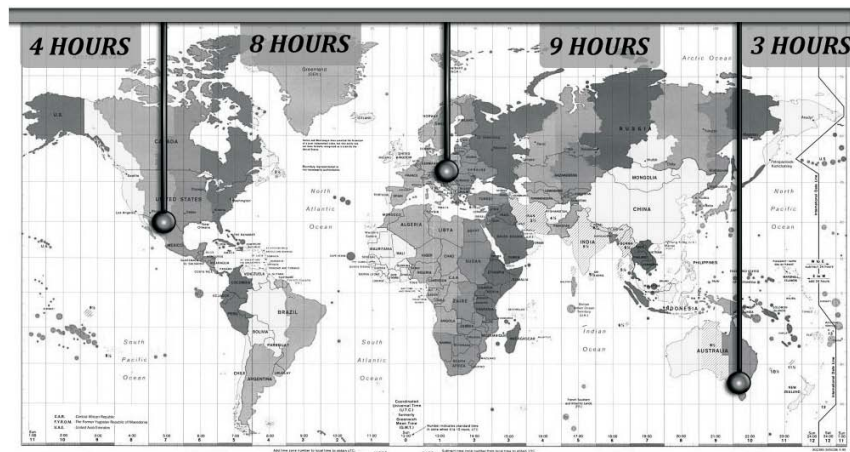


Figure 1: World Time Zones

Based on Chaczko et al. (2006) and Gupta et al. (2006) an idea of round the clock work organization could be introduced which might result in significant research acceleration. The concept is based on splitting production process in factory into 8 hours shifts. In this case, development of a product is limited by some area, where it is assembled. By utilizing time differences between different time zones, well developed communication and skilled experts living around the world - research could be done much faster, i.e. one research group based in UK could work 8 hours from 9am till 5pm. When their working day is finished, another research group based in time zone shifted 8 hours, could continue work on the same project. Afterwards,

third research group could work when the previous has finished their working day. This has been presented on figure 1. Ideal schedule of the group would be similar to the one presented on figure 3. Greatest advantage of this approach is that each team can work within their best brain activity hours. The need for work over night is eliminated. Although, the concept looks simple, it has many obstacles that need to be addressed in order to achieve successful process. Gupta et al. (2008) named the concept 24-Knowledge Factory and in his paper, proposed application of it in a software developments. Here, we will propose it to be used in medical projects. But first, some important issues have to be explained.

4.1 Knowledge sharing

Simple, at first sight idea, become more complicated, when one starts to break it into functional pieces. First obstacle is the problem of complexity of interdependencies between processes. Gupta et al. (2008) suggests that there are 3 types of dependency scenarios between members of the teams as presented on figure 2

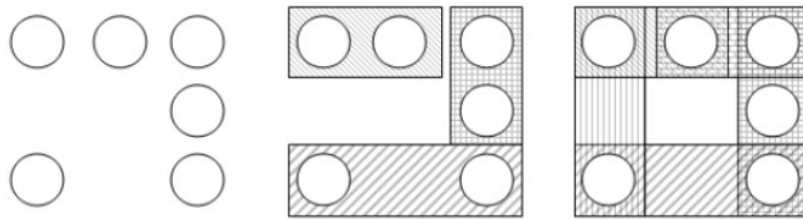


Figure 2: Interdependency between decision making scenarios

First from the left, autonomous, is when dependency is very loose, each group does not relate on the job that has to be accomplished by other, therefore they can work simultaneously, towards completion of a project or larger task. Second structure, semi-autonomous is characterized by dependencies of teams, but the one, which can be separated to form groups. These groups, if required, cooperate between each other much more often than with the whole team. Finally, tightly interdependent scenario describes a team, where each subgroup requires high degree of communication between each other.

Applying presented concept to medical research or diagnosis, one can distinguish clearly, that it is a perfect area of application for 24/7 working scheme. The reason is, that the whole research towards developing diagnosis software is semi-autonomous thus excellent for spreading research between groups. This is because of modularity of data manipulation process: preprocessing, analysis, classification. Each module can be developed and improved independently, thus there is no serial dependency at this stage. However, in terms of the whole project, collaboration at the higher level is a must.

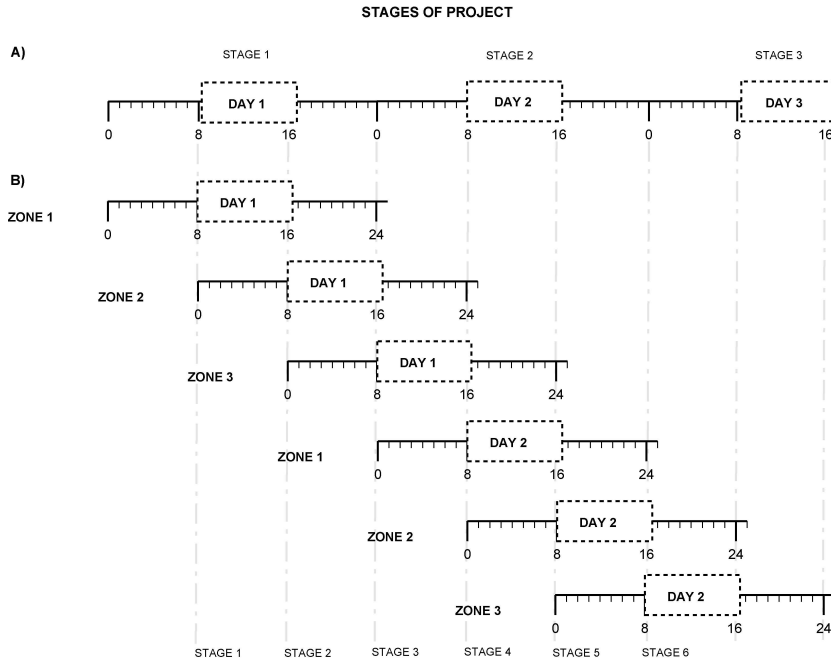


Figure 3: Example of work schedule

5 Conclusion

Decision process during diagnosis of a disease is very complicated and requires a lot of knowledge and experience from person performing it. To make the process easier, diagnosis methods are required to perform some tasks automatically, allowing larger group of less qualified specialists to be able to perform the task. This paper outlined process of samples classification and steps which should be followed. More work is required in the field as it is very promising area of research. First, hardware requirements are still very high and demand for more accurate and repeatable machines is still huge. Data produced by apparatus have to be preprocessed in order to achieve scalability, and to remove errors caused by noise or any other experimental obstacles and current methods should be improved or new developed. Furthermore, classifier is dependant on the data and preprocessing methods applied, thus deeper investigation is required to find proper methods for specific data set. It is important to remember not to apply preprocessing methods freely to any kind of classifier. As it has been shown in this paper, some preprocessing algorithms increase performance of a classifier, other decrease it. Moreover, preprocessing is required to fine tune classifier. Therefore the need for improvement in classifying methods should be similar to the need for improvement in preprocessing methods. Only joint work on at least those two algorithms can lead to the successful diagnosis software. And this can be achieved by applying concept of 24/7 work organization. It can allow to produce faster results, because of ability to organize the team members spread

around the globe removing distance and time zones limitations. And when the time is requirement, as it is in case of medical research, 24/7 concept gives opportunities to accelerate the whole work.

References

- Ball, G., Mian, S., Holding, F., Allibone, R. O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I. O., Creaser, C. & Rees, R. C. (2002), 'An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers', *Bioinformatics* **18**(3).
- Center for Cancer Research Data Bank* (n.d.).
- Chaczko, Z., Klempous, R., Nikodem, J. & Rozenblit, J. (2006), 24/7 software development in virtual student exchange groups: Redefining the work and study week, *in* 'ITHET 7th Annual International Conference, Sydney, Australia'.
- Gupta, A., Seshasai, S. & Arun, R. (2006), 'Toward the 24-hour knowledge factory a prognosis of practice and a call for concerted research'.
- Gupta, A., Seshasai, S., Crk, I. & Branson, D. (2008), 'Toward the 24-hour knowledge factory in software development', (*in print*) .
- Lilien, R. H., Farid, H. & Donald, B. R. (2003), 'Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum', *Journal of Computational Biology* **10**(6).
- Smith, L. I. (2002), *A tutorial on Principal Components Analysis*.
- Vitzthum, F., Behrens, F., Anderson, A. N. & Shaw, J. H. (2005), 'Proteomics: From basic research to diagnostic application a review of requirements and needs', *Journal of Proteome* **4**.
- Yu, J. & Chen, X.-W. (2005), 'Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data', *Bioinformatics* **1**.
- Zhou, Z.-H., IEEE, S. M. & Liu, X.-Y. (2005), 'Training cost-sensitive neural networks with methods addressing the class imbalance problem', *IEEE Transactions on Knowledge and Data Engineering* .