

Unsupervised Classification of Treatment Resistant and Non-Resistant Cancerous Cells

Achint Oommen Thomas

Department of Computer Science and Engineering, State University of New York at Buffalo

Abstract

This report describes the work done on a real dataset for the unsupervised classification of treatment resistant and non-resistant cancerous cells. An in-depth literature study was carried out to ascertain the current work being done in this area. Various approaches were developed to tackle this problem and an analysis of the observations was done. Further work that could improve the results obtained has also been suggested.

Keywords: Mass Spectrometry, Bio-markers, Dimensionality Reduction, Multivariate Statistical Analysis, Noise Estimation and Reduction and Geospatial Statistics.

I. INTRODUCTION

Cancer is one of the leading causes of death worldwide. It is responsible for 6 million deaths or 12% of total deaths annually. Cancer is the uncontrolled growth and spread of cells, that can affect any part of the body. However, effective treatment is possible if the disease is diagnosed in its early stages. Clinical decision making would benefit from the information provided by existing molecular methods, which could reliably distinguish cancer cells that are non-responders to certain treatments from those that will respond. One such molecular method, mass spectrometry, is being used to measure the protein content of biological samples and look for molecular differences between cancer cells of different prognosis, using cancer specific molecules or bio-markers. Mass spectrometry is a method of choice over other methods, which require knowledge of what one is looking for in terms of bio-markers. [1] and [2].

II. PROBLEM AND DATASET

The problem at hand can be viewed in two ways. From the medical viewpoint, we wish

to perform statistical mining of profile mass spectra from tissue biopsies to identify specific proteins as biomarkers for the presence of cancerous cells [3]. From the pattern recognition viewpoint, we wish to perform dimensionality reduction, feature extraction and unsupervised classification of data points in a large dataset. The steps of the bio-marker identification process that we are interested in are steps 3 to 5 in Figure 1.

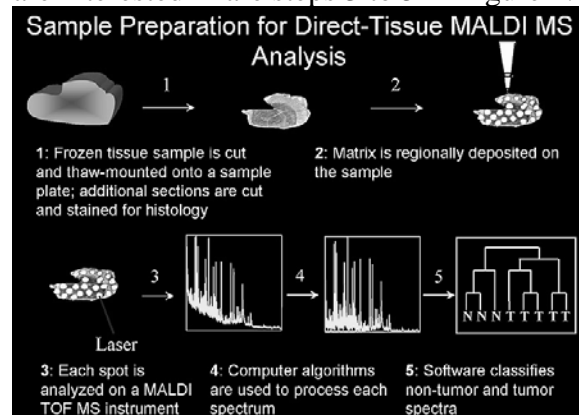


Figure 1 - Problem

The dataset that was used for this work was provided by Roswell Park Cancer Institute and the data acquisition was done by Caprioli Research Laboratory at the Vanderbilt University. Mass spectroscopy data can be stored in several formats and the dataset we have was stored in the Analyze

7.5 file format. A description of this format can be found in [4]. This particular dataset consists of multiple images of a 2D grid of pixels. Each pixel in the grid represents a point of interest. The multiple images represent different m/z (mass by charge) values and are representative of the ions of a specific molecule such as proteins etc. We consider each m/z value as a separate dimension for the pattern recognition task. Each pixel has an associated intensity value for each m/z value, which gives the intensity (count) of emitted ions in that dimension. The image is of size 53 by 120 pixels for a total of 6360 datapoints. Figure 2 shows the dataset for one m/z value. Figure 3 and Figure 4 shows plots of intensity versus m/z values for 4 random datapoints from the dataset for the resistant and non-resistant regions respectively. There are 31796 m/z values or dimensions that have been captured during data acquisition. This large dimensionality of the dataset poses a number of problems in terms of analysis that must be done on the data. The dataset represents a two class problem for this project. The left section consists of cells (represented by the pixels) that are predominantly resistant to a particular cancer treatment and the right section consists of cells that are predominantly non-resistant. The problem involves using pattern recognition techniques to perform unsupervised classification of the cells in the image into either of two classes, resistant or non-resistant.

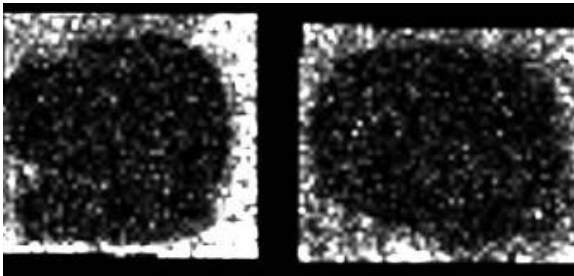


Figure 2 – The dataset in 1 dimension

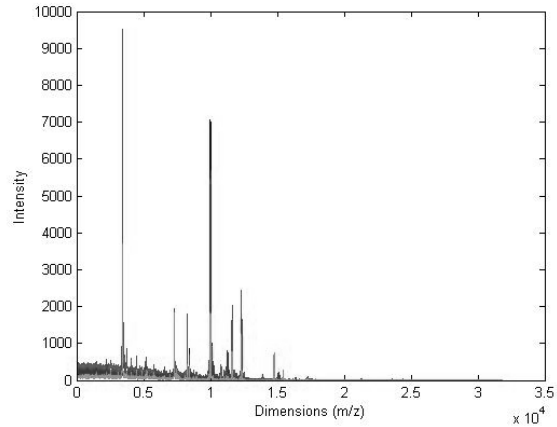


Figure 3 – Plot of 4 points in Resistant Section

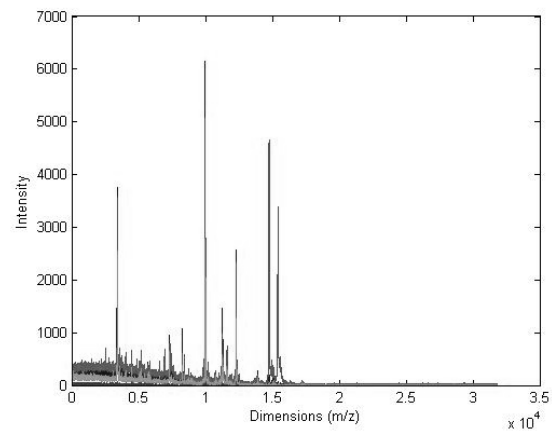


Figure 4 - Plot of 4 points in Non-Resistant Section

III. CHALLENGES

This dataset poses many challenges. The large dimensionality of the dataset makes it difficult to work on all the data at once. During tests, it was never possible to load all the data into memory and this makes it harder to have an overall view of the data. The large size makes any method chosen to work on the data, computationally expensive. Hence, workarounds need to be devised to handle this issue. The unavailability of class labels makes this problem, one of unsupervised classification. It also means we need to check with domain experts every time we get results to verify if they are correct. Also the data we have is raw; no preprocessing has been performed.

This means we need to identify useful features from the data. The data may also be noisy and noise removal needs to be done.

IV. PROGRESS OF WORK

As a first step, we tried to work on the data as it was given in the dataset. We tried a number of different approaches all designed to perform dimensionality reduction on the dataset so we could work with the most relevant dimensions. Once we get the relevant dimensions we use an unsupervised classification technique like K-Means to assign labels for the datapoints. The relevant dimensions were extracted using the algorithm presented in [5]. This algorithm, Generic Feature Extraction Algorithm (GFEA) consists of the following steps.

- *Normalize the data to prevent dominance of some features*
- *Cluster the scaled data by using the Fuzzy C-Means Clustering Algorithm*
- *Find the difference between the centers, in each dimension, for the two clusters formed*
- *Sort the differences in descending order and then select the top-k dimensions as those that contribute most to the separation between the two clusters*

Five different methods were developed for performing the dimensionality reduction. These are outlined below.

Partition Approach

- *Partition dimensions into segments of smaller size*
- *Run GFEA on partitions*
- *Get top-k dimensions for partition, cluster dataset using these dimensions*

A problem with this approach is that, the 'losing' dimensions in each segment are not considered again with those of other segments. For example, if we partition the

dataset into segments of 1000 dimensions each and select the top 100 dimensions from each segment, there may be cases where the 900 'losing' dimensions from the first segment may be more relevant than all other dimensions in the dataset. However they will not be considered again since the segments are treated as mutually exclusive.

Competing Sliding Window Approach

The problem with the Partition Approach can be avoided if we never discard any 'losing' dimension until all dimensions have been considered. The following modification incorporates this is guaranteed to find the top-k most relevant dimensions over the entire dataset.

- *Run GFEA on a window of size w dimensions of the dataset*
- *Extract top-k dimensions*
- *Slide window k dimensions to the right, discard included dimensions and extract top-k dimensions*
- *Let the previous and new k dimensions compete for inclusion in best k dimensions set using GFEA*
- *Repeat over all dimensions*
- *Perform classification using the k best dimensions*

The above two methods look at the data from a purely pattern recognition viewpoint. They do not take into account any domain knowledge. The next two approaches use some domain knowledge to select the dimensions.

Top-k Highest Peaks Approach

This approach was selected based on how most medical researchers perform dimensionality reduction.

- *Find the dimensions with the k highest values*
- *Use these dimensions to classify the datapoints*

One problem with this is, some peaks may correspond to noisy dimensions and hence will be misleading while performing the classification.

Top-k Frequent Peaks Approach

- Find the top-k dimensions which peak over all the datapoints in the dataset
- Use these k dimensions to classify the datapoints

A variation of the above approach, this approach is more robust in that it doesn't select only the dimensions which peak over the dataset, but rather those dimensions that peak frequently enough.

Hybrid Approach

- Use the top-k approach to select k dimensions
 - Either top-k highest peaks or top-k frequent peaks
- Use GFEA to search these k dimensions for k' most relevant dimensions
- Perform classification using these k' dimensions

Observations

After performing dimensionality reduction using each of the five approaches and classification using K-Means, it was found that the results obtained were not useful. As can be seen in Figure 5, this was the best separation that was possible. This result was obtained using the Competing Sliding Window Approach. Ideally, we are looking for the classification to return clumps of points of different classes in either the left or the right section. However, points belonging to the same class are distributed more or less uniformly across both sections.

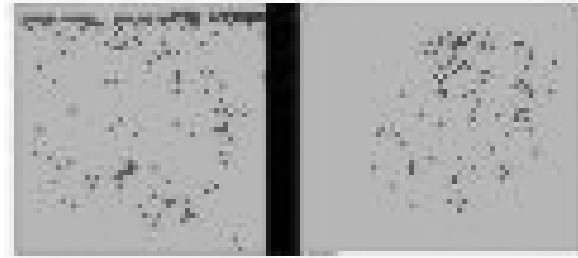


Figure 5 – Classification Results

It was also observed that using the Competing Sliding Window Approach, most of the selected dimensions were in the 3000+ range. Using Top-k Frequent Peaks, no dimensions in the 30000+ range were selected as relevant dimensions. This suggests that dimensions in the 30000+ range could be predominantly noise. They could be widely separated in the high dimensional space but not significant contributors to the separation between the two classes, resistant and non-resistant.

Based on these observations, a more in-depth literature survey was conducted. We also met with Dr. Latif Kazim and Sarah Hejaily from Roswell Park Cancer Institute and Dr. Joseph A. Gardella, Jr. from Dept. of Chemistry, SUNY-Buffalo, to gather more domain knowledge to tackle this problem.

The first change made to the approach to solving the problem, was to define regions of interest (ROIs) within the 2D grid of pixels to identify which areas correspond to potentially useful information. Figure 6 shows the mask that gives the ROIs for this dataset. The black regions represent the pixels in the 2D grid that would be considered potentially useful. Any pixel not falling within this mask would be discarded. Previously we were considering all 6360 pixels for the dimensionality reduction problem. This introduced noise in analysis. The mask reduced this number to around 4000.

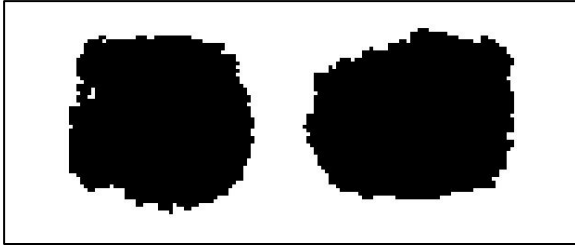


Figure 6 – Mask defining ROIs

The five proposed approaches were run again using only the pixels in the ROIs. However, the results were similar to those obtained previously. On further investigation it was concluded that the approaches were failing to find useful separations in the data due to the high degree of noise present in the dataset. The data acquisition was noisy and since no preprocessing or noise reduction was performed on the data, the features (dimensions) selected were not optimal in terms of differentiating between the two classes. Figure 7 shows the effect of noise. The high peaks correspond to relevant dimensions. However there are smaller peaks among these dimensions that just noise. Further, there are some more peaks to the right which are relevant peaks again. This suggests that any form of noise estimation and reduction to be performed on the data has to be local in nature. Global noise estimation and reduction will not identify the correct relevant dimensions.

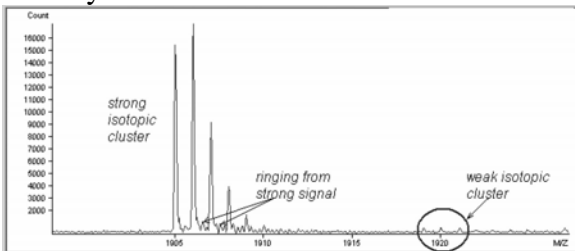


Figure 7 – Effect of Noise

The presence of background noise in turn results in low signal-to-noise ratio (SNR). Ideally we require data with high SNR so that the noise does not dominate the feature selection process. Another problem with this

dataset and mass spectroscopy data in general, is that of spectral misalignment. Figure 8 shows this effect. The top panel shows the unprocessed data where the spectra for the same m/z values are misaligned for different datapoints. This misalignment occurs mainly due to instrument error. The spectra have to be aligned to enable correct interpretation of results.

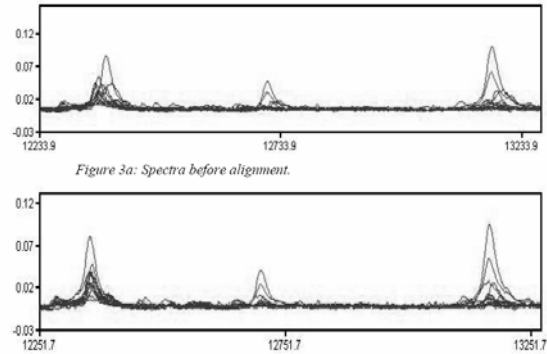


Figure 8 – Spectral Misalignment

V. AREAS OF FUTURE WORK

Data Preprocessing [1]

Going by the current trends in the research community, three steps of data preprocessing are performed. Normalization is done to ensure no dimension dominates in the feature selection process simply due to its numeric values. Possible normalization methods include mean centering and auto-scaling the data. Mean centering (1) centers the data about the mean in each dimension and auto-scaling (2) makes the standard deviation of all dimensions equal to unity.

$$x_i^D = x_i^D - \text{mean}(x^D) \text{ ----- (1)}$$

$$x_i^D = x_i^D / \text{std}(x^D) \text{ ----- (2)}$$

where D is over all dimensions
 i is over all datapoints in ROIs

Background noise estimation and removal needs to be done so that further analysis will yield more relevant information. An increase in useful information obtained from mass spectroscopy is reported in [7] and [8] after noise removal was performed. Two factors must be kept in mind while performing noise estimation and reduction. First, as mentioned earlier, the noise estimation must be done locally with respect to dimensions as opposed to globally. Second, the noise present in mass spectroscopy data can be considered to be Poisson in nature [7] and [9]. This can be explained if we consider that the spectra is formed by counting the number of ions emitted at different m/z values. The error that arises in the counting statistic is therefore dependent and directly proportional to the number of ions emitted. Such an error distribution can be modeled by the Poisson distribution. This enables us to use simple estimates for the error associated with a measurement. For instance, one estimate could be a factor times the measurement itself since the error is directly proportional to the measurement. Using the Poisson distribution makes it easier to estimate the error covariance structure and models the variance in the data more accurately [7].

Spectral alignment requires domain knowledge. One approach given in [10] is to use known standard m/z values and a tolerance window of size w . We search for peaks across the spectrum corresponding to the standard m/z values within the specified tolerance w . Given two standard m/z values, we can fit a linear regression model as in (3) and given three values, we can fit a parabolic regression model as in (4) (which tends to be more accurate). Using these models, we can shift the intensity values of the ions to coincide with their true spectral m/z locations.

$$y = ax + b \quad \text{-----} \quad (3)$$

$$y = ax + b + cx^2 \quad \text{-----} \quad (4)$$

where y is the reference m/z value
 x is the observed m/z value
 a, b, c are constants

We must be careful to note that only the m/z values that lie within the extreme standard m/z values specified can be realigned using this method.

Apply MVSA to Data [7], [8], [10], [11], [12] and [13]

Considerable work has been done in applying Multivariate Statistical Analysis (MVSA) to mass spectrometry data to obtain useful information. The most common of these methods has been that of Principal Components Analysis (PCA). PCA is used to find the directions of highest variation in the data. The method returns principal components which can describe various features [8], [9] and can themselves be used as features to perform the classification task. An interesting option would be to incorporate the noise estimation and removal step in this statistical analysis step itself. [14] suggests that maximum likelihood PCA can be used which allows each datapoint its own uncertainty. This method is computationally expensive. [15] suggests that we transform the data into an alternate space where the uncertainty is uniform in that space. Perform normal PCA in that space and back transform the PCs into the original data space. [7] uses this approach to transform the data using the Poisson estimate of the noise present in the data.

Consider Geospatial Statistics of Data

Geospatial statistics is used to interpolate spatial distributions in data. Today, it is mainly used for image reconstruction [11].

We could consider its use for spectral analysis. Since geospatial statistics considers the neighborhood of a point as well we could extend that idea to consider the contribution of neighboring pixels when performing the spectral analysis. A further use could be to perform dimensionality reduction in terms of the number of potentially relevant pixels.

VI. CONCLUSION

This report showed how the problem can be viewed from two separate viewpoints. The pattern recognition viewpoint will enable us to use established techniques from that field. The initial work that was carried out using the algorithms developed was found to perform badly since the dataset contains raw data. It would be interesting to note the results after preprocessing has been performed. Some differences exist between the spectra for resistant and non-resistant cells. It would be an enjoyable challenge to discover these differences and use them in the automatic classification into the two classes.

VII. REFERENCES

- [1] Heinrich Roder, Julia Grigorieva and Maxim Tsybin, "The Use of Mass Spectra for Cancer Biomarker Detection"; May 2005.
- [2] Bischoff R, "Methodological Advances in the discovery of Protein and Peptide Disease Markers"; Luider TM. J Chromatogr B Analyt Technol Biomed Life Sci. 2004 April 15; 803(1):27-40.
- [3] Webpage of Mass Spectrometry Research Center, Vanderbilt Medical Center, Vanderbilt University (www.mc.vanderbilt.edu/msrc/index.php).
- [4] "ANALYZE™ Header File Format" © Copyright, 1986-1995 Biomedical Imaging Resource, Mayo Foundation.
- [5] K.G. Srinivasa, Achint O. Thomas, Amrinder Singh, K.R. Venugopal, Lalit M. Patnaik, "Generic Feature Extraction for Classification using Fuzzy C-Means Clustering"; to be presented at The Third International Conference on Intelligent Sensing and Information Processing, December 2005.
- [6] "Mass Spectrometry" by Richard Caprioli and Marc Sutter; Vanderbilt University, Mass Spectrometry Research Centre. (www.mc.vanderbilt.edu/msrc/tutorials/m s/ms.htm).
- [7] Michael R. Keenan and Paul G. Kotula, "Accounting for Poisson Noise in the multivariate analysis of TOF-SIMS Spectrum Images"; Surface and Interface Analysis 2004; 36: 203-212.
- [8] Bronwyn T. Wickes, Yongmin Kim and David G. Castner, "Denoising and multivariate analysis of TOF-SIMS images"; Surface and Interface Analysis 2003; 35: 640-648.
- [9] V.S. Smentkowski, M.R. Keenan, J.A. Ohlhausen and P.G. Kotula, "Multivariate Statistical Analysis of Time of Flight – Secondary Ion Mass Spectrometry Images: Complete Description with one Sample"; Analytical Chemistry 2005; 77: 1530 – 1536.
- [10] "ProTS Data; Reference Manual" © 2003, 2004 Efecta Technologies Corporation.
- [11] Tammy M. Milillo and Joseph A. Gardella, Jr., "Spatial Statistics and Interpolation Methods for TOF SIMS Imaging".
- [12] Daniel J. Graham, Mathew S. Wagner and David G. Castner, "Information from Complexity: Challenges of TOF-SIMS Data Interpretation".
- [13] M. von Gradowski, M. Wahl, R. Forch and H. Hilgers, "Multivariate Characterization of ultra-thin nanofunctional plasma polymer films using TOF-SIMS analysis"; Surface and Interface Analysis; 2004; 36: 1114-1118.
- [14] Wentzell P, Andrews D, Hamilton D, Faber K, Kowalski B. J., Chemomet; 1997; 11: 339.
- [15] Cochran RN, Horne FH., Analytical Chemistry 1977; 49: 846.