A Time Series Based Approach for Classifying Mass Spectrometry Data

F. Gullo[2], G. Ponti[2], A. Tagarelli[2]*, G. Tradigo[1], P. Veltri[1]
[1]Università di Catanzaro, Italy, veltri@unicz.it, gtradigo@libero.it
[2]DEIS-Università della Calabria, Italy, {fgullo,gponti,tagarelli}@deis.unical.it

Abstract

This paper presents a methodology to mine spectra data based on time-series analysis. MALDI-TOF spectra are modelled as time series using a compact yet feature-rich representation scheme. Experiments show that classifying mass spectrometry series is effective and can be useful for identifying peaks in spectra that can be associated to discriminant proteins.

1. Introduction

Mass spectrometry (MS) is a technique allowing to determine with high accuracy the molecular weight of chemical compounds, ranging from small molecules to large, polar biopolymers [2]. MS is able to separate gas phase ions and produce a spectrum, that is a (large) sequence of value pairs. Each pair contains a measured *intensity* and a mass to charge ratio (m/z), which depend on the quantity and the molecular mass of the detected biomolecule, respectively. Macromolecules contained in the original compounds can be identified by associating (portion of) proteins to their peak expression in a spectrum.

In this work we consider data generated from Matrix-Assisted Laser Desorption / Ionisation - Time Of Flight mass spectrometry (MALDI-TOF MS) [3]. Spectrometry output is represented as raw data containing a (large) number of value pairs (m/z, intensity). Figure 1 shows a fragment of raw spectrum coded into a text format.

Dimensions of raw data span from a few kilobytes to a few gigabytes per spectrum, thus automatic data manipulation is mandatory. The problem of dealing with large amounts of MS data arises from the need for identifying differently expressed proteins or peptides in different samples. The analysis of several spectra coming from biological samples belonging to different subjects (e.g. healthy and diseased) focuses on to identify discriminant values in spectra (m/z, intensity couples corresponding to biomarkers) that are responsible of diseases. Several approaches for knowledge discovery from spectra have been recently developed. In [1], data mining techniques have been used to identify discriminants in a female population, distinguishing ovarian cancer diseased from healthy ones. Similarly, data mining techniques have been applied in [7] for surface-enhanced laser desorption/ionization mass spectrometry (SELDI MS) data, in order to identify discriminants in rectal cancer disease. In [6] machine learning algorithms have been used to identify biomarkers in SELDI MS data generated on tens of patients to figure out cerebral accident discriminants.

The focus of this paper is on applying data mining techniques to analyze MALDI spectra data according to discriminant biomarkers which can be associated to different peaks in MALDI data. The key idea underlying our approach is to model spectra as *time series*. A

m/Z	intensity
....	...
799.976004	135.864
800.004478	156.232
800.032953	140.765
800.061429	152.13
800.089905	137.15
800.118381	132.145
800.146858	131.137
800.175336	122.761
800.203814	124.499
800.232292	125.993
...	...

Figure 1. Mass spectrum loaded in a text raw data

time series is a list of (real) numeric values upon which a total order based on timestamps is defined. Knowledge discovery and management of time series data is a fruitful area of research involving different domains, such as speech recognition, biomedical measurement, financial and market data analysis, telecommunication and telemetry, sensor networking, motion tracking, and meteorology.

We propose a framework for preprocessing, modelling and classifying MALDI spectra based on a new time series representation, which is able to realize a good trade-off between compactness and feature-richness. The ultimate objective is to automatically identify groups of spectra with similar profiles by means of a clustering task. Preliminary experimental evaluation was conducted on the ovarian cancer data set used in [1] and on a significant data set generated at the University of Catanzaro proteomics laboratory. The results obtained show that the proposed framework exhibits good effectiveness of data classification.

2. A Framework for Classifying MS Data

Our framework consists of the following modules (Figure 2):

1. MS Data Preprocessing module, which filters out noise from the original raw spectra while maintaining only significant data features.
2. Time Series Modelling module, which represents the preprocessed MS data into a time series based model.
3. Classification module, which performs a task of *clustering* of time series data. Clustering is organizing a collection of objects (spectrum data), whose classification is unknown, into meaningful groups or *clusters*, based on interesting relationships discovered in the data. Objects within a cluster will be each other highly similar, but will be very dissimilar from objects in other clusters.
4. Evaluation module, which is in charge of assessing the quality of clustering results.

2.1. MS Data Preprocessing

Intensity values in MS data may be corrupted by noisy factors. Thus, MS data are usually subject to a preprocessing phase which aims at cleaning up spectrum noise and

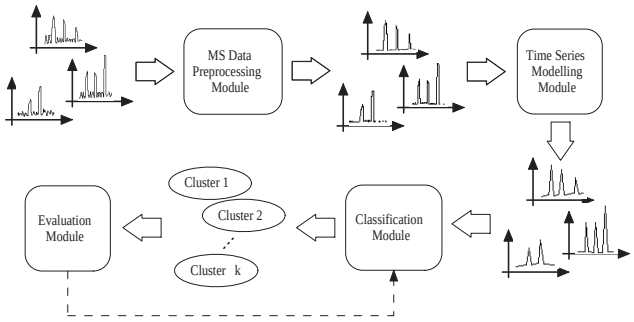

Figure 2. Conceptual architecture of the framework

contaminants without affecting biological properties. Three main steps are involved in the process: *(i)* noise reduction, *(ii)* identification of valid peaks, and *(iii)* quantization.

Noise reduction. Each mass spectrum exhibits a base intensity level (baseline) which varies from point to point across the m/z axis: intensity values that are under the baseline represent ground noise, and thus are filtered out. The baseline trend is typically approximated using a linear, logarithmic or hyperbolic model.

Identification of valid peaks. Spectra discrimination is based on the number of peaks and on the difference in their maximum intensity values. Some peaks in MS data can be due to instrumental noise. It is critical to identify only valid peaks, since up to 80% of peaks in a spectrum might be irrelevant with respect to interesting peaks. Since peaks can be approximated using a Gaussian distribution, a peak is recognized as valid if amplitude and maximum intensity value in its Gaussian representation fall within specific ranges.

Quantization. Mass spectra can be quantized to reduce the range of possible values and obtain a further noise reduction. A non-uniform quantization model is used for MS data, in which quantization step size is larger for intensity values close to ground noise mean value. A major reason to choice this model is that intensity values close to ground noise mean value are not commonly useful to identify valid peaks in the spectra.

2.2. Time Series Modelling

A preprocessed mass spectrum is a sequence $S = [((m/z)_1, I_1), \ldots, ((m/z)_n, I_n)]$, where for each pair the first value refers to the mass to charge ratio and the second one is the associated intensity value.

A mass spectrum so defined can be trivially modelled as a time series $T = [(x_1, t_1), \ldots, (x_n, t_n)]$ whose values x_i correspond to the spectrum intensity values I_i, and time steps t_i correspond to the $(m/z)_i$ values. Indeed, the notion of time implicitly lies in the sequence of mass to charge values. T can be rewritten as $T = x_1, \ldots, x_n$ when as usual the sampling period(s) is well specified.

Derivative time series Segment Approximation (DSA). Time series representing mass spectra are typically high dimensional data. Thus, it is desirable to model such time series into a compact representation which possibly synthesizes the significant variations in the time series profile.

For this purpose, we exploit a representation scheme called *DSA (Derivative time series Segment Approximation)*, proposed in [4]. Using the DSA model, a time series is transformed into a new, smaller sequence by the following main steps: *1)* computation of the first derivatives of the original series to capture its significant trends, *2)* identification of segments consisting of tight derivative points, *3)* segment approximation to finally obtain a lower dimensional still fine-grained representation of the original series.

The *derivation* step yields a sequence $\dot{T} = [\dot{x}_1, \ldots, \dot{x}_n]$, whose elements $\dot{x}_i$ are first derivative estimates. We use an estimation model that is sufficiently general (i.e. independent of the underlying data distribution model) and still enough robust to outliers [4].

The *segmentation* of a time series of length n consists in identifying $p-1$ points ($p \ll n$) to partition it into p contiguous subsequences of points, i.e. segments, having similar features. In DSA, the derivative time series $\dot{T} = [\dot{x}_1, \ldots, \dot{x}_n]$ is transformed into a sequence $S_{\dot{T}} = [s_1, \ldots, s_p]$ of variable-length segments $s_i = [s_{i,1}, \ldots, s_{i,k_i}] = [\dot{x}_{i_1}, \ldots, \dot{x}_{i_{k_i}}]$, such that: *i)* $s_{1,1} = \dot{x}_1$, *ii)* $s_{p,k_p} = \dot{x}_n$ for each $i \in [1..p\text{-}1]$, *iii)* s_{i,k_i} immediately precedes $s_{i+1,1}$ in the time axis. In order to determine the segment delimiters, a segment is grown until it exceeds an error threshold, and the process repeats starting from the next point not yet considered. Precisely, a sequence s_i, for each $i \in [1..p\text{-}1]$, is identified as a segment if $|\mu(s_i) - s_{i+1,1}| > \epsilon$, where $\mu(s_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} \dot{x}_{i_j}$ denotes the average over the points in s_i. The segment condition allows for aggregating subsequent data points having very close derivatives. Parameter ϵ can be estimated globally with respect to a given time series, by considering the variance over the points in the derivative series.

All individual segments of a derivative time series are finally *approximated* with a synthetic information capturing their respective main features. Each segment s_i is mapped to a pair formed by the timestamp t_i of the first point ($\dot{x}_{i_1}$) of s_i and an angle that explains the average slope of the portion of time series bounded by s_i. This is mathematically expressed by the notion of arctangent applied to the mean of the (derivative) points in each segment.

2.3. Classification and Evaluation

The proposed framework is parametric with respect to the clustering scheme. In this work, we use the well-known agglomerative hierarchical clustering algorithm with group-average linkage [5]. This algorithm initially forms one cluster for each individual object (time series), then repeatedly merges the most similar pairs of clusters growing a hierarchy until a stop criterion is met. In our context, a reference partitioning is available for each dataset we used. Thus, reaching the desired number of clusters at a hierarchy level can be naturally used as termination criterion for the algorithm.

Since the availability of reference classifications, evaluating the clustering effectiveness can be accomplished by assessing how well a clustering fits a predefined scheme of known classes (natural clusters). We resort to the standard Information Retrieval notions of precision, recall and F-measure [8].

3. Preliminary Experiments

Main experiments were performed on two datasets: the ovarian cancer dataset publicly available [1], and a smaller dataset consisting of MALDI data generated at the University of Catanzaro proteomics laboratory. In both datasets, spectra fall into either the healthy class or the diseased class.

Figure 3 shows raw and preprocessed spectra of the ovarian cancer dataset. Noise reduction was performed considering a linear model for the baseline. A peak was recognized as valid if its maximum intensity value is at least 2.5 times the corresponding noise intensity value. Quantization step was set to 2,000 counts for intensity values within the range [0..10,000] and to 500 counts for values greater than 10,000. Moreover, the range of (m/z) values was reduced to focus only on significant portions of the spectra. In particular, the cut involved two m/z intervals: the first interval corresponds to intensity values close to zero (i.e. with m/z ranging within [0..15,000]) and the second one corresponds to intensity values identifying spectrum contaminants such as the polymer between [43,000..56,384] m/z.

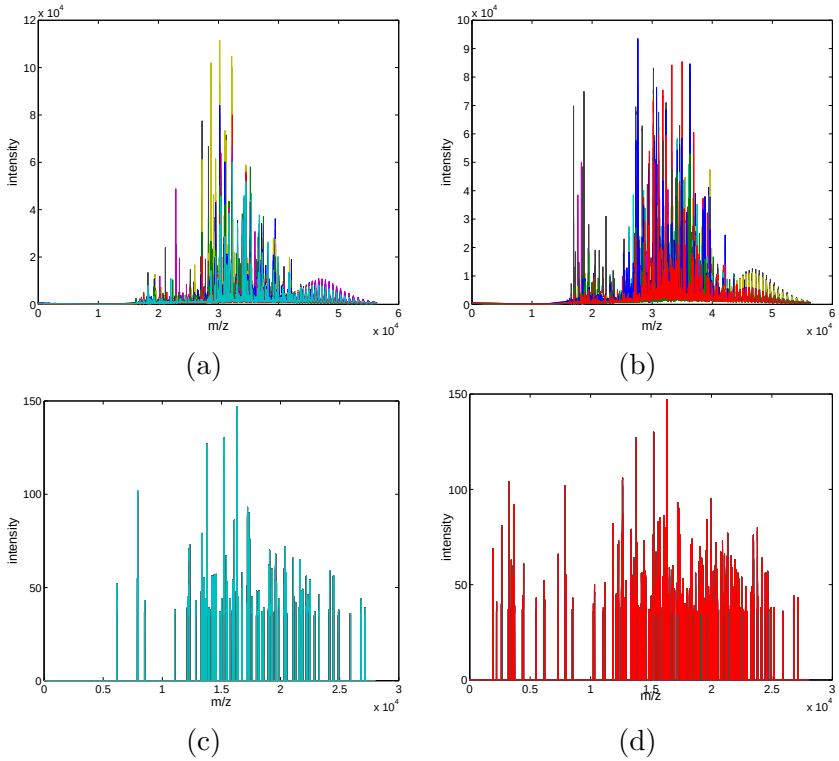

Figure 3. Ovarian Cancer MS data: on top, raw spectra of (a) control class and (b) diseased class; on bottom, preprocessed spectra of (c) control class and (d) diseased class

Clustering results on the ovarian cancer dataset highlighted high effectiveness provided by the proposed framework. In particular, the expected two classes were well-recognized with precision, recall and F-measure values measured in 0.88, 0.86 and 0.87, respectively. Note that such values are sufficiently high (i.e. they are close to 1) proving that our framework is capable both to identify homogeneous groups of MS data and to separate different data in distinct classes, according to biomarkers associated to discriminant peaks.

The framework was also tested on the MALDI dataset produced at University of Catanzaro. It is worthy to noticing that we encountered different issues in the preprocessing phase with respect to the ovarian cancer dataset, which are due to their different originating laboratories. Nevertheless, our framework obtained comparable quality results.

4. Conclusion

We presented a new framework for classifying spectrometry data based on time series analysis. The framework revealed high capability of classifying datasets and to identify discriminant peaks. We plan to test and validate the framework on more datasets that will be generated by our laboratories. Moreover, we are currently investigating how to use the discovered discriminant peaks to query publicly available protein databases and to identify biomarkers (proteins) which are potentially responsible of diseases.

Acknowledgements Authors are grateful to Marco Gaspari for advices on MALDI data, and to Mario Cannataro and Sergio Greco for their fruitful suggestions. Authors thank Andrea Grande for support in performing experiments.

References

[1] E. F. Petricoin 3rd, A. M. Ardekani, B. A. Hitt, P. Levine, V. A. Fusaro, and S. Steinberg. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 9306(359):572–577, 2002.

[2] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[3] G. L. Glish and R. W. Vachet. The basic of mass spectrometry in the twenty-first century. *Nature Reviews*, 2:140–150, 2003.

[4] S. Greco, M. Ruffolo, and A. Tagarelli. Effective and efficient similarity search in time series. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 808–809, 2006.

[5] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[6] J. Prados, A. Kalousis, J. C. Sanchez, L. Allard, O. Carrette, and M. Hilario. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, 8(4), 2004.

[7] F. M. Smith, W. M. Gallagher, E. Fox, R. B. Stephens, E. Rexhepaj, E. F. Petricoin 3rd, L. Liotta, M. J. Kennedy, and J. V. Reynolds. Combination of SELDI-TOF-MS and data mining provides early-stage response prediction for rectal tumors undergoing multimodal neoadjuvant therapy. *Ann. Surgery*, 2(245), 2007.

[8] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.