A Time Series Based Approach for Classifying Mass Spectrometry Data

Francesco Gullo
DEIS - Università della Calabria

joint work with:

G. Ponti, A. Tagarelli (DEIS - Università della Calabria)
G. Tradigo, P. Veltri (Università di Catanzaro)
Introduction

A typical Mass Spectrum

Focus in MS-based Proteomics:
identify discriminating values in the spectra
(i.e. \(m/z\), intensity) couples corresponding to biomarkers)
that are indicators of biological states (e.g. disease).
Introduction

Mass Spectra

PROBLEMS:
HIGH dimensionality
HUGE datasets

Mandatory requirement:
AUTOMATIC DATA MANAGEMENT

DATA MINING

Clustering
Classification
Rule Discovery
Motif Discovery

...
Introduction

Traditional data mining tasks directly applied on raw spectra may not reach satisfying results.

IDEA:

Mass Spectra modelled as TIME SERIES

Need for a proper mass spectra representation

Basic Motivations:

- From spectra to time series: trivial modelling
- Several well-established and valid approaches for mining of time series
- Many proper solutions for time series dimensionality reduction
Introduction

Our idea allowed us to reach excellent results in classifying mass spectra:

- **Ovarian Cancer dataset**: classification accuracy: 87%
- **MALDI UNICZ dataset**: classification accuracy: 96%
Outline

- Introduction
- Overview of time series data management
  - The DSA model
- A Time Series based Framework for Mass Spectrometry Data
- Experimental results
- Conclusions
Traditional time series form:

\[ T = [(x_1, t_1), \ldots, (x_n, t_n)] \]

Time series form under condition of fixed sampling period:

\[ T = [x_1, \ldots, x_n] \]
Automatic management of time series data is typically accomplished by applying data mining tasks.

**Two main issues:**

- **Distance Measures**
  - One-to-one alignment (euclidean distance)
  - Warping time axis
  - String matching

- **Dimensionality Reduction**
  - Piecewise discontinuous functions
  - Low-order continuous functions
Time Series Distance Measures: Dynamic Time Warping

Euclidean

Dynamic Time Warping

Fixed Time Axis
Sequences are aligned “one to one”

“Warped” Time Axis
Nonlinear alignments are possible

figures borrowed from tutorial “Data Mining and Machine Learning in Time Series Databases” (2006), Dr. Eamonn Keogh, University of California
Time Series Dimensionality Reduction

Approximation via piecewise discontinuous functions

- DWT
  Discrete Wavelet Transform

- PLA
  Piecewise Linear Approximation

- PAA
  Piecewise Aggregate Approximation

Approximation via low-order continuous functions

- DFT
  Discrete Fourier Transform

- SVD
  Singular Value Decomposition

- Chebyshev Polynomials

figures borrowed from tutorial “Data Mining and Machine Learning in Time Series Databases” (2006), Dr. Eamonn Keogh, University of California
Time Series Dimensionality Reduction:  
**DSA model**

**DSA** *(Derivative time series Segment Approximation)* – CIKM’06

- High rate data compression
- Feature-reach representations
- Best trade-off between effectiveness and efficiency

\[
T = [(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)] \quad \rightarrow \quad \tau = [(\alpha_1, t'_1), (\alpha_2, t'_2), \ldots, (\alpha_p, t'_p)]
\]

**DSA steps:**

1. Derivation
2. Segmentation
3. Segment Approximation
Classification of MS data: our proposal

Framework for Classifying Mass Spectrometry Data

Novelty: Time Series based representation for Mass Spectra
The proposed framework

Three main parts:
1. **MS Data Preprocessing**
2. **Time Series Modelling**
3. **Classification and Evaluation**
The proposed framework: MS data preprocessing

Three main parts:
1. MS Data Preprocessing
2. Time Series Modelling
3. Classification and Evaluation
The proposed framework: 
**MS data preprocessing**

The *MS Data Preprocessing Module* performs a set of preliminary steps on the original raw spectra.

Three main steps:

- Noise reduction
- Identification of valid peaks
- Quantization
The proposed framework: *MS data preprocessing*

The *MS Data Preprocessing Module* performs a set of preliminary steps on the original raw spectra.

Three main steps:
- Noise reduction
- Identification of valid peaks
- Quantization
The proposed framework: 
**MS data preprocessing**

The **MS Data Preprocessing Module** performs a set of preliminary steps on the original raw spectra.

Three main steps:
- Noise reduction
- Identification of valid peaks
- Quantization
Three main steps:
- Noise reduction
- Identification of valid peaks
- Quantization

The proposed framework: **MS data preprocessing**

The *MS Data Preprocessing Module* performs a set of preliminary steps on the original raw spectra.
The proposed framework: time series modelling

Three main parts:
1. MS Data Preprocessing
2. Time Series Modelling
3. Classification and Evaluation
The proposed framework: time series modelling

The Time Series Modelling Module represents the preprocessed spectra into a time series based model.

\[
S = [(I_1, (m/z)_1), (I_2, (m/z)_2), \ldots, (I_n, (m/z)_n)]
\]

\[
T = [(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)]
\]

\[
\tau = [(\alpha_1, t'_1), (\alpha_2, t'_2), \ldots, (\alpha_p, t'_p)]
\]
Three main parts:

1. MS Data Preprocessing
2. Time Series Modelling
3. Classification and Evaluation
The proposed framework: classification and evaluation

The **Classification Module**
performs a task of **clustering** (i.e. **unsupervised classification**) on mass spectra
- High intra-cluster similarity
- Low inter-cluster similarity

The **Evaluation Module**
is in order to assess the accuracy of the output classification w.r.t. the desired classification.

\[
P = \frac{1}{k} \sum_{i=1}^{k} \frac{|C_i \cap \Gamma_i|}{|C_i|}\]

**precision**

\[
R = \frac{1}{k} \sum_{i=1}^{k} \frac{|C_i \cap \Gamma_i|}{|\Gamma_i|}\]

**recall**

\[
F = \frac{2PR}{P + R}\]

**f-measure**

\[
\Gamma = \{\Gamma_1, \ldots, \Gamma_k\} \quad C = \{C_1, \ldots, C_k\}\]

*desired classification*  
*output classification*
Experimental Results

Ovarian Cancer

(SELDI dataset)

50 spectra
2 classes (control - diseased)
56,384 (m/z, intensity) couples

Classification Results:

\[ P = 0.88 \]
\[ R = 0.86 \]
\[ F = 0.87 \]

(*) Available at: http://sdmc.lit.org.sg/GEDatasets/Datasets.html
Experimental Results

MALDI UNICZ

MALDI dataset

20 spectra
2 classes (healthy - diseased)
34,671 (m/z, intensity) couples

Classification Results:

P = 0.99
R = 0.93
F = 0.96
Conclusions

- Mass Spectrometry meets Time Series: a new framework for classifying MS data
- High capability in classifying and discriminating mass spectra
Thanks...