A Brief Study of Dimensionality Reduction for Mass Spectra Classification

Pisov Maxim
IITP, MIPT
maxs987@gmail.com

Belyaev Mikhail
Skoletch, IITP
belyaevmichel@gmail.com

Abstract

Mass spectrometry is a relatively new field of study, so, for now, its data processing pipelines look fragmented and challenging. This paper covers the last two steps of such pipelines: generation of features by peak alignment and classification of spectra. A crucial machine learning issue is that, typically, the length of a feature vector significantly exceeds the number of spectra in a sample. We propose some basic ideas for dimensionality reduction based on peaks aggregation and evaluate its efficiency by intensive computational experiments.

1. Introduction

Typically the mass spectrometry pipeline is split up into separated steps [1]: denoising, normalization, peak picking, peak alignment.
The resulting data represents a set of peak intensities distributed on a two-dimensional grid having retention time (RT) and mass-to-charge ratio (M/Z) as coordinates [2].
One of many challenges on the data processing field is the lack of labelled data and a very large dimensionality of the feature space (up to several millions of features) [1, 3].
This paper is aimed at studying the feasibility of reducing the dimensionality of the feature space by projecting the data on a coarser grid, as well as the possibility of omitting the peak alignment from the pipeline.

2. Definitions

Let R, M - sets of grid knots:

$$\begin{aligned} R &= \{r_{min}, r_{min} + \delta_r, r_{min} + 2 \cdot \delta_r, ..., r_{max}\} \\ M &= \{m_{min}, m_{min} + \delta_m, r_{min} + 2 \cdot \delta_r, ..., r_{max}\} \end{aligned} \tag{1}$$

then, the function

$$MS : R \times M \mapsto \mathbb{R} \tag{2}$$

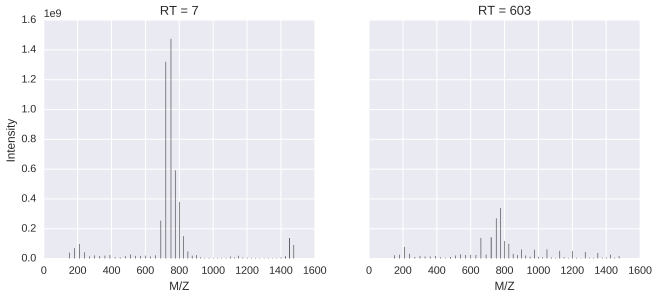

Figure 1: Example of a grid condensation

is called a ***mass spectrum***, where $MS(x, y)$ is the intensity of the peak with coordinates (x, y).
Let

$$\begin{aligned} R' &= \{r_{min}, r_{min} + n \cdot \delta_r, r_{min} + 2n \cdot \delta_r ..., r_{max}\} \subseteq R \\ M' &= \{m_{min}, m_{min} + k \cdot \delta_m, m_{min} + 2k \cdot \delta_r ..., m_{max}\} \subseteq M \\ & n, k \in \mathbb{N} \\ MS' &: R' \times M' \mapsto \mathbb{R} \end{aligned} \tag{3}$$

where

$$MS'(x, y) = \sum_{\substack{\xi \in [x, x + n \cdot \delta_r) \\ \eta \in [y, y + k \cdot \delta_m)}} MS(\xi, \eta) \tag{4}$$

Then MS' is called a ***condensation*** (or projection) of MS on a coarser grid $R' \times M'$.
Figure 1 shows an example of a such condensation on a grid with $|R| = 2, |M| = 50$, while the original grid is with $|R| \cdot |M| \sim 16$ million knots.

There are also two important particular cases:
Total Ion Spectrum (TIS): $|R'| = 1$
Total Ion Chromatogram (TIC): $|M'| = 1$

2.1. Peak Alignment

Due to measurement imperfection the peaks corresponding to the same metabolite in different samples usually are slightly misaligned.
Peak alignment is a procedure that makes the peaks

from the same metabolite to have the same values on the M/Z axis [1].

3. The Dataset

Our dataset consists of 484 mass spectra obtained from human brain tissue.
Every entry has several labels. Their repartition is the following:

- Ethnicity:
  - Asian: 148
  - Caucasian: 145
  - African American: 61
  - N/A: 130
- Sex:
  - Male: 314
  - Female: 164
  - N/A: 6
- Disease:
  - Healthy: 407
  - Schizophrenia: 44
  - Autism: 27
  - Down's syndrome: 6

The dataset is split up into two versions: spectral centroids with and without alignment.
The first one is used as a baseline, while the second one is used as data for experimentation.
Both datasets were obtained from raw experimental data with MZmine - an open-source software for MS-processing [4]. See http://mzmine.github.io/ for details.

4. The Baseline

First of all, we will estimate the accuracy of a logistic regression classifier using 5-fold cross-validation. Several problems will be analyzed:

- Sex classification
- Ethnicity classification
- Disease classification
- Autism vs Healthy binary classification
- Schizophrenia vs Healthy binary classification

The results are shown in Table 1.

Table 1: The baseline for several classification problems

	Sex	Ethnicity	Disease	Autism	Schizophrenia
Accuracy	0.650	0.714	0.820	0.926	0.880

Note, that the baseline for "Sex classification" is quite similar with the male/female repartition ratio in the dataset, which might suggest to exclude this problem from further experiments.

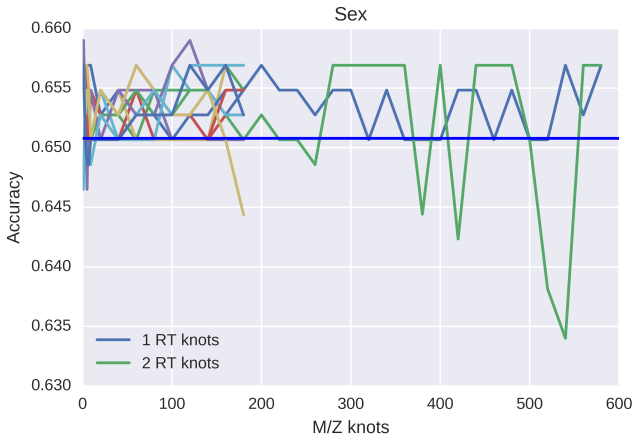

Figure 2: Condensation results for Sex

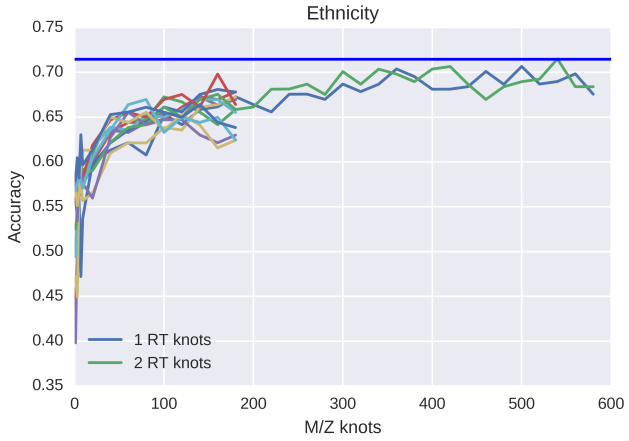

Figure 3: Condensation results for Ethnicity

5. Using a Coarser Grid

Now we will decrease the feature space dimensionality by projecting the data on a coarser grid. This procedure effectively reduces the feature space dimensionality as well as eliminates the need to align the peaks, because with a coarser grid two misaligned peaks are more likely to be projected on the same knot.
Figures 2-6 show the results of such projections for different number of grid knots. For the RT axis were used from 1 to 75 knots, while for the M/Z axis - from 1 to 180.
Each line represents a CV-curve with a fixed number of RT-knots and varying number of M/Z-knots. For a better readability not all the curves are labelled .
For the cases $|R'| = 1$ (TIS) and $|R'| = 2$ the calculations were made for a larger range of $|M'|$ values.
The blue horizontal lines are the baselines for each classification problem.
As it was stated before, the "Sex classification" problem is not reliable, which is also proved by Figure 2.
It is interesting that "Ethnicity classification" accuracy

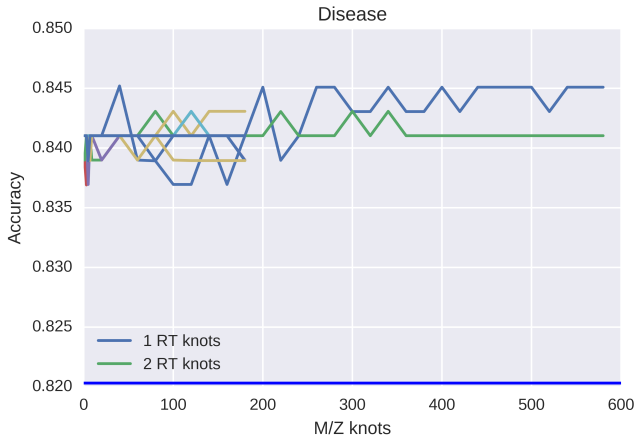

Figure 4: Condensation results for Disease

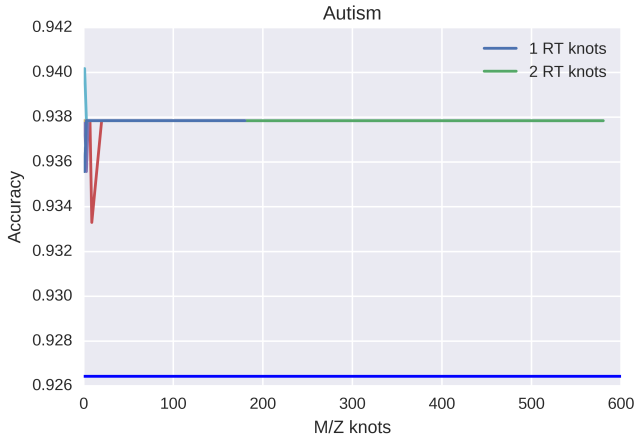

Figure 5: Condensation results for Autism

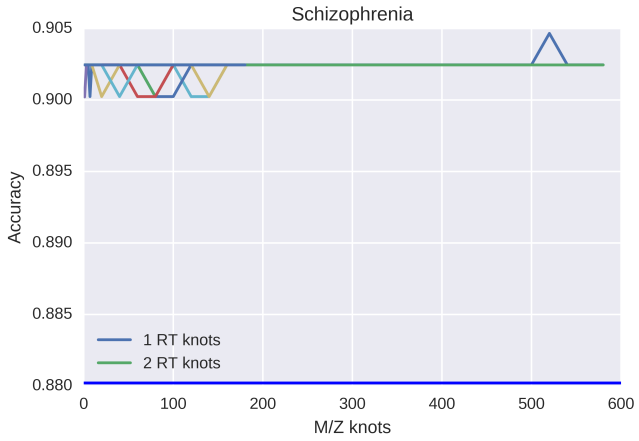

Figure 6: Condensation results for Schizophrenia

shows a positive trend when the feature space dimensionality increases. However, it does not exceed the baseline, but its deviation is quite small.
The other three problems show an increase in accuracy comparing to the baseline even for extremely small dimensionalities, yet in this case the CV-curve looks wiggly, so greater values for M/Z discretization are recommended.
Finally, for all the experiments the curves show a weak dependence on the number of RT-knots, which means that in any case the mass spectrum (MS) can be effectively replaced by its Total Ion Spectrum (TIS).

6. Conclusion

As we can see, the reduction of the feature space dimensionality can lead to an increase in the classification accuracy. Nevertheless, in some cases it might cause a slight decrease of the accuracy.
We suppose this is the case of noisy data. However further study is required, especially in the domain of low-dimensional feature generation and non-equidistant grids.
Also, we showed that excluding peak alignment (in the form it exists now) from the pipeline does not have a substantial influence on the learning accuracy, which means that this step is not essential for the problem of classification.

References

[1] Rob Smith, Andrew D Mathis, Dan Ventura, and John T Prince. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC bioinformatics*, 15(7):1, 2014.

[2] Salvatore Cappadona, Peter R Baker, Pedro R Cutillas, Albert JR Heck, and Bas van Breukelen. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino acids*, 43(3):1087–1108, 2012.

[3] Arnald Alonso, Sara Marsal, and Antonio Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3:23, 2015.

[4] Tomáš Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. Mzmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, 11(1):1, 2010.