MS data analysis: Preprocessing and Pattern Recognition of the Sloan-Kettering Data
CLRC Technical Report 01-02-2005,
February 2005

A.Gammerman, I.Nouretdinov, Z.Luo, A.Chervonenkis, V.Vovk
Computer Learning Research Centre, University of London
and
Paul Tempst, John Philip, Josep Villanueva
Memorial Sloan-Kettering Cancer Center, New York

This report describes pre-processing of the raw data by compressing the number of points, subtracting the baseline, normalising, calibrating and used several different ways for pattern recognitions.

1 Data

The dataset collected at Memorial Sloan-Kettering Cancer Center (MSKCC), New York, USA contains raw and processed data for each sample. Each data file has two columns: one for m/z value and the other for the corresponding intensity value. The m/z values range from 700 to 15000. There are 125,541 data points in each file. Each sample is characterised by a label:

- "Bla" - Bladder cancer
- "B" - Breast cancer
- "Mel" - Melanoma
- "Pro" - Prostate cancer
- "Ctrl" - Control group.

In addition, there is a calibrant data for each sample. 13 peaks (calibrants) in total (7 in low mass range and 6 in high mass range) are used for calibration. The m/z values for these peaks are known: 758.46, 1047.2, 1297.51, 1620.88, 2094.46, 2466.73, 3149.61, 3883.59, 4282.945, 5734.56, 6181.048, 8565.89, and 12361.088.

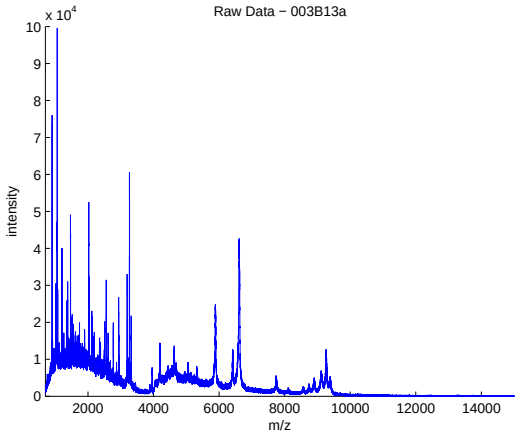

Figure 1: Raw data

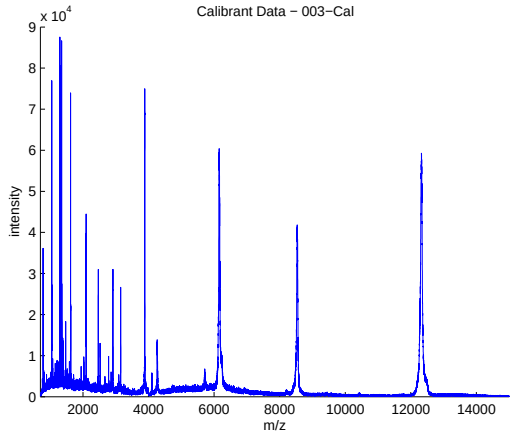

Figure 2: Calibrant data

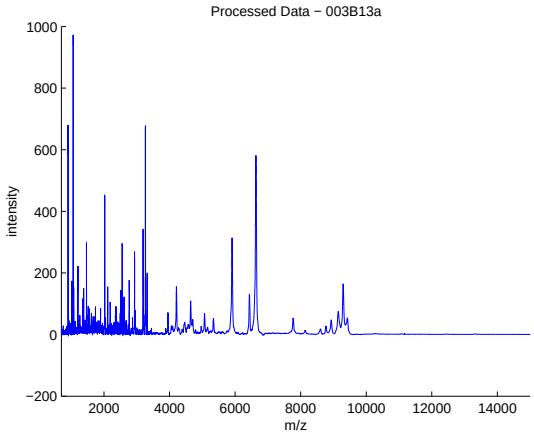

Figure 3: Processed data

Examples of raw, calibrant and processed data for the sample '003B13a' (a breast cancer sample) are shown in figures 1, 2 and 3.

In the following sections we describe our pre-processing of the raw data that includes: calibration, baseline substraction, compression, normalisation and peaks alignment. Then we describe the results of applications of several pattern recognition techniques to the transformed data to find out the number of errors made for each of the diseases. The last part describes some recent results in the attempt to establish the biomarkers by finding the most important (in some sense) peaks. These peaks could be further investigated and an attempt could be made later to identify them with the known proteins.

2 Pre-processing

We start with the raw data and perform the calibration first.

2.1 Calibration

We used the 13 peaks and calibrant file associated with each sample to perform calibration. In addition, we assumed that the relationship between m/z value (M) and time-of-flight (T) for the mass spectrometer used in the experiments can be represented as

$$M = B \times (T - A)^2 \tag{1}$$

where A and B are two constants determined by experiment setup. Our calibration algorithm is presented as follows.

As it is clear from the procedure, we calculated the constants A and B, and re-assigned to the m/z (or M) their correct values.

Algorithm 1 Calibration

Require: m/z values of 13 peaks
  fix A_s and B_s in the formula (1) (for example A_s=1.0 and B_s=0.5)
  find out TOF values for these 13 peaks (TOF_s) using the formula (1) and A_s and B_s
  for each sample's calibrant C_i **do**
    find the m/z values of these 13 peaks in C_i
    find out A_i and B_i which optimises $\sum_{k=1}^{13} ({TOF_i}^k - {TOF_s}^k)^2$
  end for
  {we now have optimal A_i and B_i for each sample}
  for each sample's raw data R_i **do**
    for each m/z value M_j in R_i **do**
      $TOF_j = \sqrt{\frac{M_j}{B_i}} + A_i$
      ${M_j}^* = B_s \times (TOF_j - A_s)^2$ {the corresponding intensity value is not changed}
    end for
  end for

2.2 Baseline Subtraction

The goal of baseline subtraction is to remove systematic artifacts, usually due to chemicals used in the experiments or to the detector overload. Ideally baseline should rest on zero. Chemical and electronic noise produces a background intensity which typically decreases when the mass m/z increases.

Our baseline estimation procedure can be described as follows.

Optionally, the raw data can be binned before baseline estimation procedure is applied. An example is shown below where the raw data was binned by a window size of 10. Figure 4 shows the data after the baseline subtraction.

Note that the results can be smoothed by moving the window averaging.

2.3 Normalisation

Due to variation in sample preparation amd deposition on the target, matrix crystallisation and ion detection, samples are not directly comparable before normalisation. The goal of normalisation is to make sure that the total amount of ions across different samples are the same. This is done by calculating the sum of all intensity values and then dividing each intensity value by the sum. We then multiple these intensity values with a constant C (for example we set C=2×10^5 in our experiments). The results are shown in Figure 5.

2.4 Comparison of Processed Data

We identify peaks as local maxima.

Algorithm 2 Baseline estimation

Require: spectrum S
Require: moving window size W
Require: threshold K
finished = false
B=S
while finished == false **do**
for each moving window of S **do**
if no of points in $B_W \leq K$ **then**
finished = true
else
find out mean μ and variance σ_2 of those points in B_W
if all points are in $[\mu - 1.5\sigma, \mu + 1.5\sigma]$ **then**
finished = true;
else
remove those points which are not in $[\mu - 1.5\sigma, \mu + 1.5\sigma]$ from B
end if
end if
end for
end while

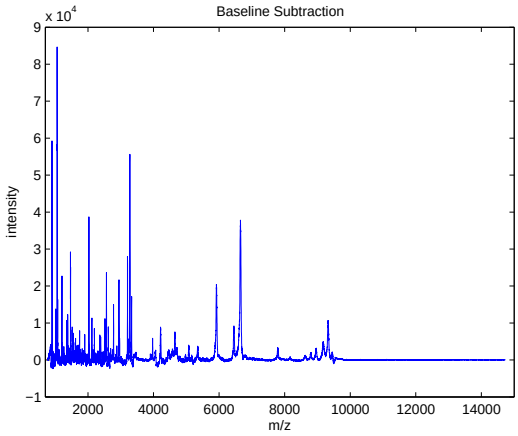

Figure 4: After baseline subtraction

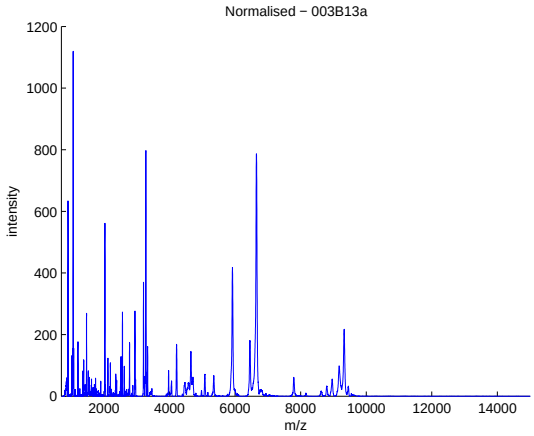

Figure 5: Normalised data

We can now compare our processed data with the one sent by MKSCC. Figure 6 shows the identified peaks in each of data where the top part of the graph shows the peaks in the MKSCC processed data and the bottom graph the peaks in the RHUL processed data. There are 1292 peaks identified in our processed data and 879 found in MKSCC processed data. This difference could be due to smoothing carried out by MKSCC.

2.5 Peak Alignment

Given a number of peak points (or features), we need to find out a unique correspondence between them. The problem is that not all peaks appear in every sample. Therefore, one-to-one correspondence does not exist between every two samples. A simple approach is to construct a super set of all peaks and use it as the anchor of alignment - every sample is aligned to this super set. For this purpose the superset is splitted to clusters. Cluster definition goes in two steps. First we find all intervals between neighbouring peak positions in the superset exceeding some fixed values d_1, where d_i is some distance metrics. These intervals split the m/z axe to clusters of order 1. Then we test if each sample has no more than 1 peak in a cluster. If so, the cluster is considered as final. Otherwise, we look whether there is an interval exceeding $d_2 < d_1$, dividing occurrences of one sample peaks within the cluster. If so we divide the cluster of order 1 to smaller ones. Otherwise we consider it as final.

Now for each sample peak we assign the number of its cluster (in case when there are more than 1 peak of the same cluster for the same sample we take into account only the largest of them).

Now all peaks are aligned to a certain cluster. Every sample is then characterised by a numerical vector, of dimension n (n is the number of final clusters) with the coordinates equal to the height of a peak corresponding to each cluster,

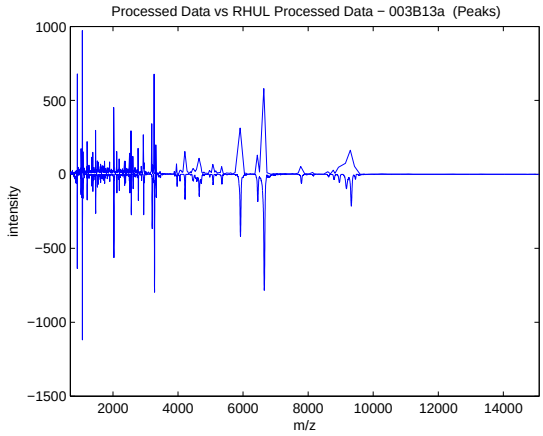

Figure 6: Comparison of processed data (top - MKSCC, bottom - RHUL)

zero if there is no such peak. These coordinates are considered as a set of sample features for pattern recognition.

3 Pattern Recongnition and Biomarkers

3.1 SVM Pattern Recongnition without Selection

SVM (with dot-product kernel) procedure was used to find decision rules for 4 dichotomy cases:

- Control group vs Bladder cancer
- Control group vs Breast cancer
- Control group vs Melanoma
- Control group vs Prostate cancer

For each case clustering procedure was carried out separately (because different samples were included). Leave one out cross validation method was used to evaluate the results. They are as follows.

- Control group (33 examples) vs Bladder Cancer (20 examples): 1 false positive and 1 false negative examples (total number of features: 449).
- Control group (33 examples) vs Breast Cancer (21 examples): 2 false positive and 0 false negative examples (total number of features: 462).
- Control group (33 examples) vs Melanoma (31 examples): 5 false positive and 4 false negative examples (total number of features: 430).

- Control group (33 examples) vs Prostate Cancer (32 examples): 2 false positive and 0 false negative examples (total number of features: 456).

false positive is an example with disease wrongly classified as a control example; *false negative* is a control example classified as one with a disease.

3.2 Feature Selection

The number of features in these experiments is rather high (over 400 features or peaks), and we made an attempt to reduce the features set and to find out if such reduction sufficiently influence the results of pattern recognition.

Using t-statistics we select the top N features with the largest t-test, where N is a predefined constant.

It appeared that using only a few number of features provides the result compatible with that when we used the total feature set. The results are as follows:

- Control group (33 examples) vs Bladder Cancer (20 examples): 1 false positive and 0 false negative examples.
- Control group (33 examples) vs Breast Cancer (21 examples): 1 false positive and 1 false negative examples.
- Control group (33 examples) vs Melanoma (31 examples): 6 false positive and 5 false negative examples.
- Control group (33 examples) vs Prostate Cancer (32 examples): 2 false positive and 3 false negative examples.

The coordinates are reverted then back to the original ones, so each feature corresponds to an interval where an important peak should lie. The list of these intervals is:

Ctrl vs Bladder	Ctrl vs Breast	Ctrl vs Melanoma	Ctrl vs Prostate
[1013, 1032]	[1559, 1576]	[1190, 1203]	[1019, 1032]
[2951, 2970]	[4014, 4035]	[1345, 1362]	[1262, 1277]
		[1901, 1912]	
		[5760, 5772]	
		[5772, 5786]	

It is clear that with this significant reduction in number of features, we still almost maintain the overall accuracy of the results: the only two groups (Melanoma and Prostate Cancer) show some decrease in accuracy - two more errors in Melanoma and two more in Prostate Cancer, while Bladder cancer has actually improved (one error less). However we are operating with very small numbers here (20 or 30 examples), and it is may be pre-mature to make some important conclusions here.

Note that it is possible to reduce the width of the intervals while maintaining the prediction accuracy for biomarker discovery. For example, reduced intervals for Ctrl vs Bladder are [1015, 1019] and [2955, 2959].

3.3 Nearest Neighbours PR without Selection

To test the stability of the results we used also another approach to pattern recognition. It is different from the approach described above in the following sense:

- we reduced the number of initial features to 50 (in all cases) using appropriate clustering parameters;
- we used the nearest-neighbour procedure (instead of SVM) for pattern recognition;
- we used alternative criterion in feature reduction;
- we applied an error estimation procedure for reduced feature space. This is a less biased leave-one-out procedure since it has been moved outside of feature selection.

The number of initial features was forced to be 50 by increasing parameters d_1 and d_1 in the clustering procedure. Then applying leave-one-out procedure and nearest neighbour recognition method for the total feature set, we get the following results:

- Control group (33 examples) vs Bladder Cancer (20 examples): 1 false positive and 0 false negative examples.
- Control group (33 examples) vs Breast Cancer (21 examples): 0 false positive and 0 false negative examples.
- Control group (33 examples) vs Melanoma (31 examples): 9 false positive and 3 false negative examples.
- Control group (33 examples) vs Prostate Cancer (32 examples): 1 false positive and 1 false negative examples.

One can see that the results only slightly differ from those obtained in the above approach.

3.3.1 SK pre-processing

Applying the same (NN) approach to the pre-processed by Sloan-Kettering data gives the following results:

- Control group (33 examples) vs Bladder Cancer (20 examples): 2 false positive and 1 false negative examples.

- Control group (33 examples) vs Breast Cancer (21 examples): 4 false positive and 0 false negative examples.
- Control group (33 examples) vs Melanoma (31 examples): 8 false positive and 1 false negative examples.
- Control group (33 examples) vs Prostate Cancer (32 examples): 3 false positive and 0 false negative examples.

There are less errors (9) in the Mel group, and less number of "false positive" (8) in the same group; the accuracies in the other 3 groups are worse.

3.4 Pattern Recognition with Nearest Neighbours and Feature Selection

On the last stage we tried to select 2 or 3 features of 50 initial ones which are most important for classification and corresponding cluster intervals. The programme looked through all possible combinations of 2 or 3 features and selected three best using leave-one-out procedure in nearest-neighbour method.

The best intervals for all dichotomies are shown below:

Ctrl vs Bladder	Ctrl vs Breast	Ctrl vs Melanoma	Ctrl vs Prostate
$[700, 800]$	$[989, 1013]$	$[889, 912]$	$[1172, 1225]$
$[3196, 3207]$	$[1013, 1026]$	$[1231, 1313]$	$[1412, 1427]$
	after 6521	$[1313, 1427]$	$[1575, 1699]$
		$[1427, 1478]$	
		$[1699, 1819]$	

To estimate recognition accuracy we used leave-one-out procedure outside feature selection. Basically, one example was left out. And the rest were used to select feature subsets and then the left out sample was classified, using selected features by nearest neighbour method. The results are as follows:

- Control group (33 examples) vs Bladder Cancer (20 examples): 1 false positive and 1 false negative examples.
- Control group (33 examples) vs Breast Cancer (21 examples): 2 false positive and 1 false negative examples.
- Control group (33 examples) vs Melanoma (31 examples): 7 false positive and 6 false negative examples.
- Control group (33 examples) vs Prostate Cancer (32 examples): 1 false positive and 1 false negative examples.

Again we see that after reduction recognition accuracy is almost the same.

To find the most important features that would help to establish the biomarkers, we again try to reduce the width of the intervals by checking the accuracy at each stage. The results are given below.

Ctrl vs Bladder	Ctrl vs Breast	Ctrl vs Melanoma	Ctrl vs Prostate
[700, 710]	[1001, 1008]	[889, 896]	[1172, 1178]
[3196, 3207]	[1013, 1020]	[1231, 1238]	[1412, 1420]
	[6521, 6537]	[1313, 1321]	[1575, 1583]
		[1448, 1456]	
		[1699, 1707]	

4 Conclusions

We have pre-processed the original raw data in several different ways, and applied the pattern recognition techniques. The following conclusions can be made:

- The original over 100,000 m/z values were compressed into 400-500 features, and then further feature selection procedures reduced them down to 4-5 features without any significant loss in accuracy of predictions.
- Three out of four diagnostic groups have shown a good accuracy of classification. The best results are: 98.1% of accuracy in case of Bla group; 100% in case of Breast group; 81.3% for Mel group; and 96.9% for Pro group. The number of the "false positives" in each of the group (apart from the Mel group) is also small. However, the results should be considered with some caution since the overall number of samples in each category is small, and the statistics of small numbers is not very reliable.
- We have made an attempt to establish very narrow intervals of m/z values in order to find out the corresponding protein biomarkers. The further experiments would require to confirm or disconfirm these results.

5 Acknowledgements

This work has been supported by MRC grant S505/65: "Proteomic analysis of the Human Serum Proteome for Population Screening, Diagnosis and Biomarker Discovery", and the Royal Society grant. We would like to thank the Sloan-Kettering team where the data were originally collected. We are also grateful to our colleagues at UCL Ian Jacobs, Usha Menon, Adam Rosental, Mike Waterfield for many useful discussions of the data and the results.