

Verison 4: Data analysis of 7 biomarkers (updated in March 2008)

I.Nouretdinov, B.Burford, Z.Luo and Alex Gammerman
Computer Learning Research Centre
Royal Holloway, University of London

March 2008

1 Abstract

The analysis of the CIPHERGEN's 7 biomarkers has been conducted using UKOPS data. The results have shown that there are two biomarkers (CTAPIII and B2M) that are useful in discriminating cases from controls using UKOPS data when they are combined with CA125. ¹

2 7 Biomarkers

The dataset consists of 7 biomarkers and CA125 (U/ml). Actually, we used only 6 biomarkers since one of the 7 - ITIH4 - did not have reliable measurements (almost all values were about 100).

- TT ($m/z = 13380$);
- ApoA1 ($m/z = 28079$);
- CTAPIII ($m/z = 9288.7$);
- TrF ($m/z = 79908$);
- ITIH4 - not used;
- HepC ($m/z = 2790$);
- B2M ($m/z = 11731$); and
- CA125.

¹This report focuses only on the data analysis - for used mass-spectrometry methods and description of UKOPS data - see various reports by CIPHERGEN as well as the reports by Usha Menon and John Timms of UCL.

Each sample was repeated 3 times (for any biomarker/mode pair) and the median of three is taken as resulting value.

This report mainly describes the most difficult problem of discriminating class Benign (B) from Malignant (M). Data were normalized, and the concentration values of each biomarker (attribute) were transformed into the logarithm scale and then linearly rescaled so that its minimum is 0 and maximum is 1. The *Support Vector Machine (SVM)* algorithm was used with the *polynomial kernel* of degree 1 and *RBF kernel* with parameters $\gamma = 100$ and $C = 1000$.

3 Results

The data consists of 3 sets: a training set (S1) with 44 benign (B) and 20 malignant (M) cases; a calibration set (S2a) with 25B and 39M, and a (blind) testing set (S2b) with 20B and 21M.

We use S1 for preliminary leave-one-out analysis, and S2a as a calibration set to improve quality of prediction.

The following table contains: (a) SVM leave-one-out results on S1; (b) SVM results trained S1 and tested on S2a; (c) SVM results trained on S1 and tested on S2b (“blind test”).

Table 1: Results on the training, calibration and testing sets

| Set: | S1 | S1 | S2a | S2a | S2b | S2b |
|--------------------------|----------|----------|----------|-----------|----------|----------|
| Total: | 44B | 20M | 25B | 39M | 20B | 21M |
| Combination | BasM | MasB | BasM | MasB | BasM | MasB |
| CA125 | 5 | 5 | 5 | 16 | 5 | 5 |
| CA125,ApoA1 | 4 | 4 | 4 | 16 | | |
| CA125,TT | 5 | 6 | 5 | 16 | | |
| CA125,HepC | 4 | 5 | 4 | 16 | | |
| CA125,B2M | 2 | 5 | 4 | 12 | 5 | 3 |
| CA125,CTAPIII | 2 | 4 | 5 | 14 | 5 | 5 |
| CA125,TRF | 4 | 5 | 5 | 14 | | |
| CA125,B2M,ApoA1 | 3 | 5 | 5 | 13 | | |
| CA125,B2M,TT | 3 | 4 | 6 | 14 | | |
| CA125,B2M,HepC | 3 | 5 | 4 | 12 | | |
| CA125,B2M,TRF | 3 | 5 | 11 | 12 | | |
| CA125,CTAPIII,B2M | 2 | 5 | 6 | 12 | 5 | 3 |

From the Table 1 it is clear that the best combination is CA125, and B2M - it gives on the testing set 8 errors (80.5% of accuracy) while CA125 alone gives 10 errors (76%). An addition of CTAPIII to the above combination gives also 8 errors. The additions of more than 2 biomarkers to CA125 leads to the overfitting.

Next table (Table 2) presents the sensitivity/specificity results. Here we combine set 2a and 2b in one joint set 2. The table also shows 95% confidence

intervals (in brackets) for our results. It was calculated using by applying formula $z\sqrt{p(1-p)/t}$ where t is the number of observed samples, $z = 1.96$, which corresponds to 95% level.

Table 2: Sensitivity/specificity with confidence intervals

| Set: | S1 | S1 | S2 | S2 |
|---------------|--------------|--------------|--------------|--------------|
| Total: | 44B | 20M | 45B | 60M |
| Combination | Sens. | Spec. | Sens. | Spec. |
| CA125 | 0.886(0.093) | 0.750(0.189) | 0.777(0.121) | 0.650(0.120) |
| CA125,CTAPIII | 0.954(0.061) | 0.800(0.175) | 0.777(0.121) | 0.683(0.117) |
| all | 0.931(0.074) | 0.700(0.201) | 0.733(0.129) | 0.766(0.107) |

4 Discussion

On the Fig.1 and Fig.2 we present joint distributions of (CA125,B2M) and (CA125, CTAPIII) On the figures the benign samples are presented by circles and malignant samples by crosses. We also added here the healthy samples (H) to demonstrate the usefulness of the two biomarkers. We can see that B2M and CTAPIII work well for the discrimination between B and M in certain intervals of CA125 - in the 'intermediate zone' where $10^{1.7} < CA125 < 10^3$. Or if we transform back from log-scale, we can see that the biomarkers are useful in a certain interval of CA125: between 50 and 1000.

This can be described as follows:

- Zone 1($CA125 < 50$) contains all Healthy samples, many Benigns and very few Malignants;
- Zone 2($50 < CA125 < 1000$) contains Benign and Malignant samples, Malignants usually have higher level of B2M;
- Zone 3($CA125 > 1000$) contains only Malignant samples

The combination (CA125,B2M,CTAPIII,TrF,HepC) is most effective on the testing set, but on the calibration set the accuracy has gone down due to over-fitting.

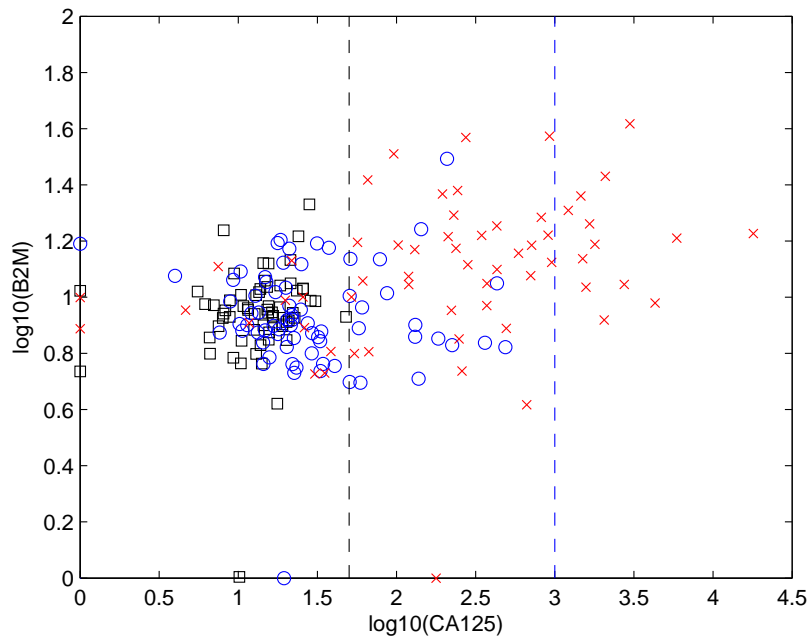


Figure 1: Distribution of healthy samples (squares), benign (circles) and malignant samples (crosses) by CA125 and B2M.

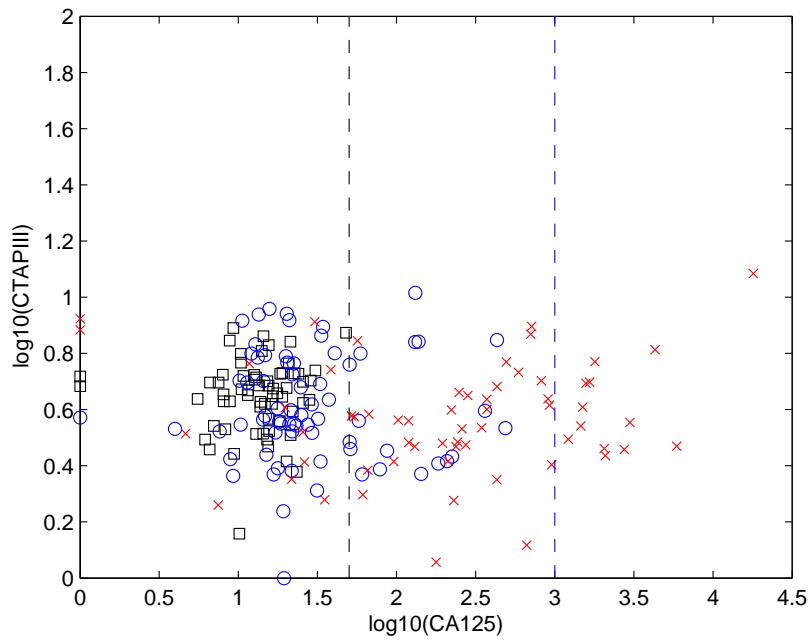


Figure 2: Distribution of healthy samples (squares), benign (circles) and malignant samples (crosses) by CA125 and CTAPIII.