Necrosis analyses, iteration 2

Peter Fehervari

September 25, 2019

Aims

Predict the probability of necrosis in patients with pancreatitis from blood parameters in a fashion that allows sensible in situ predictions for clinical MDs. The gold standard for necrosis diagnostics is a CT scan, which is costly and time consuming. Since only a fraction of patients develop necrosis post pancretitis, it is desirable to effectively filter low risk individuals, and require CT scans for high risk patients.

Analyses

Data description and management

Handling missing values

Two distinct datasets each with a similar set of predictors were made available for the analyses. Both datasets had unique patient identifiers, a variable indicating whether the patient suffered from necrosis and 25 blood parameters. The first dataset (n= 1435) had a total of 12022 missing values, however the second, larger dataset (n = 177) had considerably more (27). Moreover, the patterns of NA's on a single variable level are non-random (see *Figure 1* and *Figure 2*). Apparently, there are two types of variables with high propotion of missingness; 1) where NA's are in large chunks, presumably indicating a not missing at random pattern (NMAR), and 2) a pattern where NA's are seemingly independent from the index (missing at random; MAR).

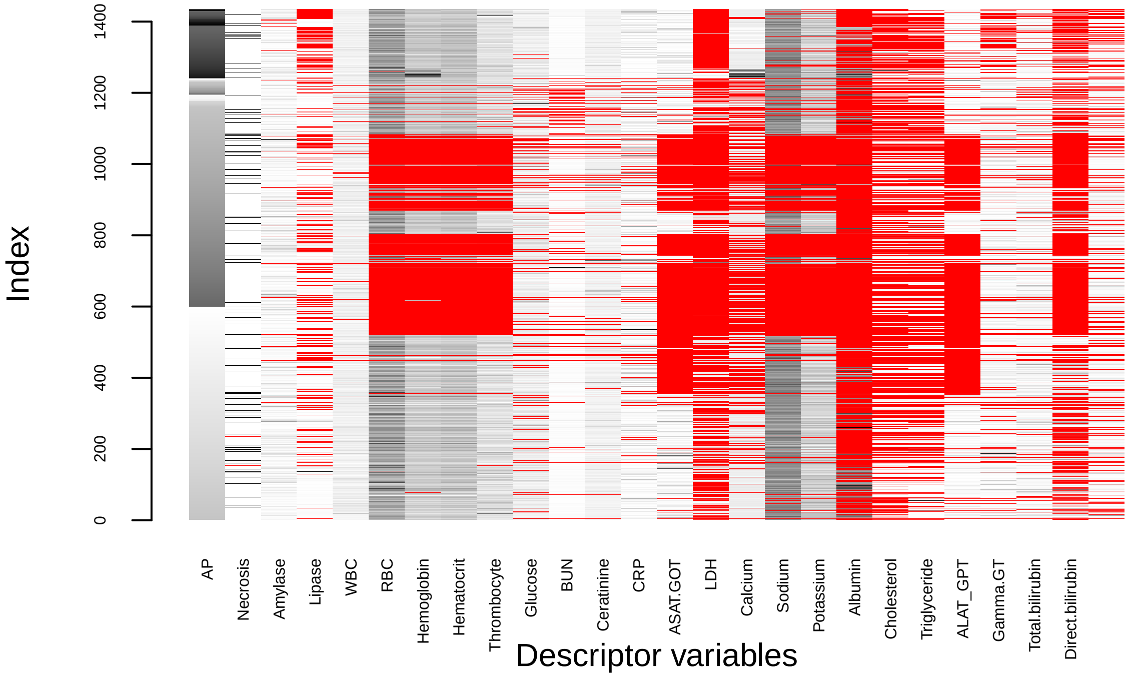

Figure 1: Missing value patterns of descriptor variables across the dataset. Red indicates missingness.

Looking at the missing data combinations, it is apparent that complete observations are not the most common record types, *Figure 2*.

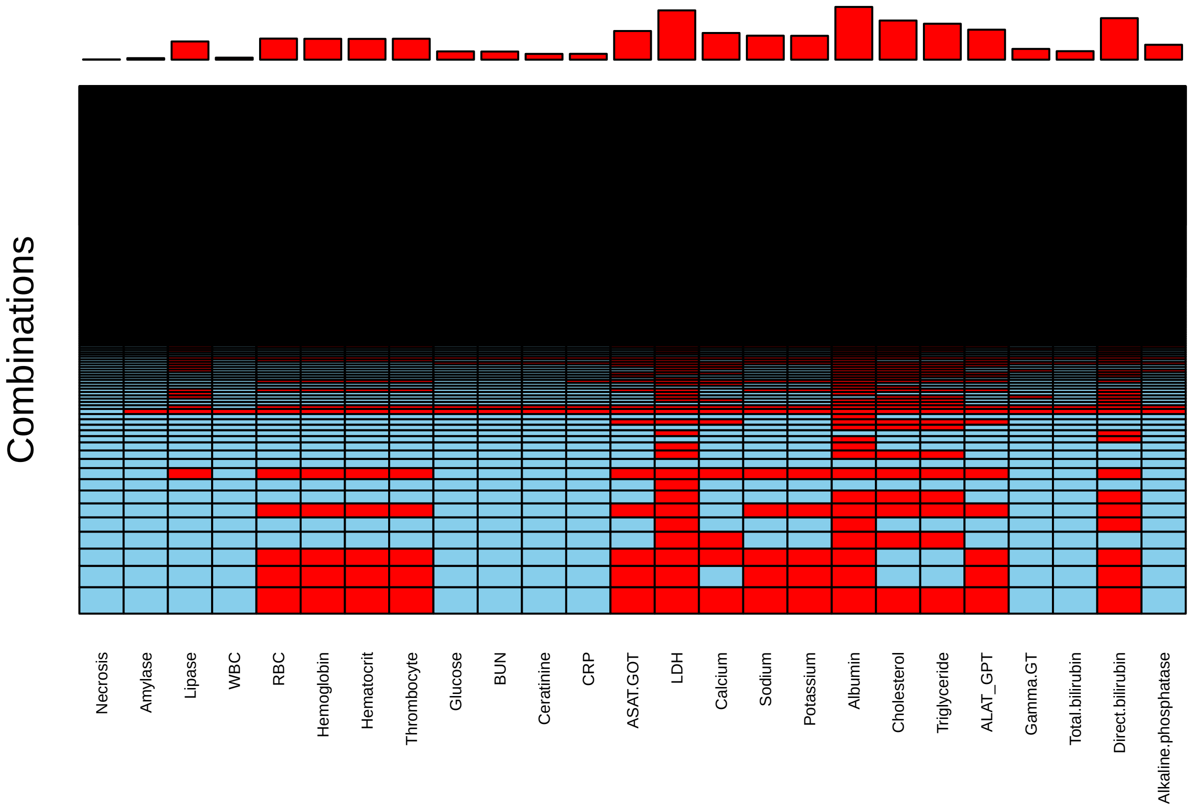

Figure 2: Missing value combinations where the height of the rows is proportional to the number of existing cominations. Note that the number of records with no missing data (all blue row) is lower than that of records with several missing predictor combinations. The bars on the top of the figure are proportional to the number of missing valeus of the variable. Red indicates missingness.

The larger dataset in its current state is probably useless for predictive modelling as most methods need complete observations. To salvage the information in the variables, I used the following approach; 1) exclude variables with more than 35% missing values in the larger dataset from both datasets, 2) impute missing values using the k-Nearest Neighbour Imputation method [1] in both datasets.

The following variables were included in all further analyses:

```
## Warning in kNN_work(as.data.table(data), variable, metric, k, dist_var, :
## Nothing to impute, because no NA are present (also after using makeNA)
```

```
## [1] "Necrosis" "Amylase"  "WBC"
```

The smaller dataset with less missing value imputations was used to test all models described below (hereafter: ***Testing Data****), while the larger dataset was used to assess model predictive performance (hereafter:* ***Training Data****)*

[1] A. Kowarik, M. Templ (2016) Imputation with R package VIM. Journal of Statistical Software, 74(7), 1-16

Unbalanced data handling

The incidence rate of necrosis is low, resulting in highly unbalanced Training and Testing Data. I used the Syntethic Minority Oversampling (SMOTE [2]) algorithm to create new examples of the minority class in the Training Data (records with necrosis) to enable more efficient model training. Briefly, this method creates synthetic samples of the minority class (individuals with necrosis in case of this study) through taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. See http://rikunert.com/SMOTE_explained%5D for a more indepth explanation. We oversampled the minority class of the training dataset by 500% and undesampled the majority class by 300% in this study.

original	SMOTE data	class
1302	1995	No_necrosis
133	798	Necrosis

Predictive modeling

I used Conditional Inference Tree[3] to model the relationship of candidate predictors and the target variable. The tree was pruned to maximum depth of 2 levels to ease the biological interpretation of results.

Results

The conditional tree fitted on the Trainig Data shows that three variables play a major role in predicting necrosis, namely Glucose, Lipase and Hemoglobin *Figure 3*. The predictive performance of this model was evaluated on the Testing Data. First we calculated the predicted probablities of all records, to later assess the threshold probability values that allow the maximum sensitivity+specificity predictions. We later used this threshold to classify observations to Necrosis or No_necrosis categories. Below is the confusion matrix of predicted and observed necrosis values and the ROC curve.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321-357.

[3] Torsten Hothorn, Kurt Hornik and Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651–674. Preprint available from http://statmath.wu-wien.ac.at/~zeileis/papers/Hothorn+Hornik+Zeileis-2006.pdf