Predicting response to chemotherapy in AML patients

Authors: Ljubomir Buturovic (1), Damjan Krstajic (1),(2),(3), Alejandrina Pattin (1)

Affiliations:

1. Clinical Persona Inc, 932 Mouton Circle, East Palo Alto, CA 94303, USA

2. Research Centre for Cheminformatics, Jasenova 7, 11030 Beograd, Serbia

3. Laboratory for Molecular Biomedicine, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11010 Beograd, Serbia

Emails: ljubomir@clinicalpersona.com, Damjan.Krstajic@rcc.org.rs, apattin@clinicalpersona.com

Summary Sentence

We applied Principal Component Analysis and normal mixture modeling to predict which Acute Myeloid Leukemia patients will achieve complete remission after current therapy.

Background/Introduction

The goal of the AML project [5] is to build probabilistic predictor of complete remission (CR) status of AML patients who underwent chemotherapy, using clinical and proteomics features from 191 AML patients. This is a classification problem with two classes: responder and non-responder. We tried many approaches to find the best probabilistic classifier, including glmnet, Support Vector Machine, normal mixture modeling, deep learning, random forest, decision-tree adaboost and k-NN, combined with a variety of feature extraction methods, such as Pearson correlation coefficient, Mutual Information, AUC, Principal Component Analysis, Independent Component Analysis and Non-Negative Matrix Factorization.

As described in our recent publication [1], we used the grid-search cross-validation approach to select optimal model. The best cross-validation results were found for the following sequence of processing steps:

- Cleanup data
- Apply imputation and encoding of categorical variables

- Augment input data with non-linear transformations of the clinical variables
- Apply Yeo-Johnson transformation [2]
- Center/scale input variables
- Apply PCA
- Apply SpatialSign transformation [2]
- Extract relevant principal components
- Apply normal mixture model to the resulting set of variables in the PCA space
- Adjust the resulting posterior probabilities to optimize BAC measure without affecting AUC

The processing was done in R using the caret [3] and mclust [4] packages. The above steps are described in detail in the Methods.

Methods

The first step is pre-processing where we check the data and impute missing values. While inspecting the training dataset we found two issues:

- one patient is a child aged five

- one patient has FIBRINOGEN value equal to zero

Considering that there are no pediatric patients in the test set, we removed the child sample from training. As a result, we used 190 patients for building predictive models.

We concluded that the FIBRINOGEN value equal to zero was a data entry error and replaced it with NA.

We replaced all “NA”,” NotDone” and “ND” values with median or mode values from the union of the training and test sets. If an input variable was numeric then we used median, otherwise we used mode.

We augmented clinical variables with four non-linear transformations: square, square root, logarithm, and inverse. The rationale for this was generally known observation that non-linear transformations may convert difficult classification problems to linearly solvable ones. However, we observed that the application of the non-linear transformations to proteomic variables led to decrease in cross-validation AUC. Explanation of this behavior may be a subject of future research.

Following the augmentation, we applied Yeo-Johnson transformation, then centering and scaling to zero mean and unit standard deviation, followed by Principal Components Analysis and SpatialSign transformation. These steps were based on recommendations in [2], prior experience

in classification and regression analyses, and observed cross-validation performance of the models.

Following PCA, a critical step was selection of best principal components (PC). To aid in that search, we rank-ordered the top 12 PC and examined the effect of their removal on the cross-validation AUC of the predictor. The components which had noticeable effect on the AUC were considered signal, and the rest were considered noise. Using this analysis, we retained eight principal components for inclusion in the classification model.

Following identification of principal components, we performed cross-validation grid search for the best normal mixture classification model. We performed exhaustive search over 10 multivariate mixture models available in package mclust. The best AUC was achieved using ellipsoidal model with equal volume, shape and orientation ('EEE').

This model produced training and test set posterior probabilities of complete remission. The final step was optimizing the test set probabilities to maximize BAC, without affecting the AUC. Clearly, any monotone transformation of probabilities does not affect AUC. Therefore we transformed the posterior probabilities as follows:

- Compute cross-validation BAC on the training set, for each possible decision threshold
- Record proportion f of predicted positives (CR) for the threshold which maximizes BAC
- Set test set threshold to value T such that the proportion of test set examples predicted positive equals the recorded proportion f
- Monotonically transform (shift and stretch) the test set probabilities: $p > T \rightarrow (0.5, 1)$; $p <= T \rightarrow (0, 0.5)$

The resulting transformed test set probabilities were submitted to final leaderboard.

Conclusion/Discussion

The key to achieving good AUC was the combination of non-linear transformations of clinical variables and selection of informative principal components. The key to achieving good BAC was adjusting the decision threshold to the composition of the test set. The RPPA measurements did improve AUC when added to the clinical variables, but only marginally. In the AML Challenge setting, even small improvement counts, therefore we decided to use the RPPA variables. However in clinical use the value of this improvement would be questionable, considering the cost of and complexity of measuring the proteome.

References

[1] Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, *6*(1), 1-15.

[2] Kuhn, M, Johnson, K. Applied Predictive Modeling. Springer, 2013.

[3] Kuhn, M. Building Predictive Models in R Using the caret Package (2008). Journal of Statistical Software, 28(5), 1-26.

[4] Fraley C., Raftery A. E., Murphy T. B., Scrucca L. (2012). Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. University of Washington Technical Report No. 597, www.stat.washington.edu/research/reports/2012/tr597.pdf

[5] DREAM AML Outcome Prediction Challenge (syn2455683)

Authors Statement

Damjan Krstajic and Ljubomir Buturovic conceived the project approach. Damjan Krstajic and Ljubomir Buturovic developed and implemented the predictive models and pre-processing rules. Alejandrina Pattin provided software support.