Multiple Myeloma Challenge Approaches

## **Haibin Guan, Arpit Mehta, Joaquin Zabalo, Giri Narasimhan**

Bioinformatics Research Group (BioRG)

School of Computing and Information Sciences

Florida International University, Miami FL 33140.

# Summary

We used stacking to create an ensemble of several baseline classifiers for constructing a precise risk stratification model to identify high risk Multiple Myeloma patients using gene expression variables with ISS and age.

# Background/Introduction

We used a combination of 1) Synthetic Minority Oversampling Technique (SMOTE) & Tomek links balancing approach, 2) gene selection utilizing Support Vector Machine methods based on Recursive Feature Elimination and Minimum Redundancy and Maximum Relevance (MRMR), and 3) stacking ensemble strategy to combine several standard machine learning methods. The above process creates four predictive models, one for each training data set. Predictions on the test set is performed by first determining the most appropriate model to use and then using the model for that training set.

# Methods

Given that the four training data sets: GSE24080UAMS, HOVON65, Kryukov and MMRF were known to have distinct subject profiles, we applied the same pipeline for each of them generating separate sets of models for the four data sets. We first describe how each of the 4 datasets was processed.

**Correcting for Imbalance in the dataset:** Since the data sets were unbalanced in terms of the number of high risk versus the number of low risk subjects, we used an algorithm that creates datasets with balanced distributions by combining oversampling by SMOTE and undersampling by Tomek [Batista et al., 2003]. Synthetic Minority Oversampling Technique (SMOTE) creates new samples of the minority class by interpolating between existing ones. A pair sample is called a Tomek link if the items in the pair belong to different classes, but are the nearest neighbors to each other. Undersampling is achieved by removing majority class samples that are part of a Tomek link.

**Feature Selection:** Next, the size of the dataset was reduced by filtering out genes that were not differentially expressed. The R Bioconductor limma package was used for assessing differential gene expression and returned a ranking of the genes. To further perform feature selection, we utilized SVM-RFECV and MRMR approaches [Guyon et al., 2002]. To apply the first method, a support vector machine (SVM) classifier with a linear kernel was trained using all the features. Then, recursive feature elimination (RFE) was applied by using the weight magnitude as a ranking criterion. Smaller subsets (for backward feature elimination) of features were considered by removing lowest ranked features first and cross-validation (CV) was used to determine the optimal number of features. The idea of the second method, Minimum Redundancy and Maximum Relevance (MRMR), is to select a subset of genes so that each gene in the subset has the highest similarity with target vector (maximum relevance), while simultaneously having the lowest similarity with the rest of the genes in the subset (minimum redundancy). Mutual information was used as a measure of similarity. Note that each training data set may have different set of relevant/redundant genes. The process generated 29 genes for the GSE24080UAMS set, 638 genes for the HOVON65 set, 1314 genes for the Kryukov set, and 41 for the MMRF set.

Assuming that the feature selection process may have missed previously known relevant genes, we artificially added a collection of genes suggested by the Multiple Myeloma literature as additional features for the learning process. We added to our set of selected any genes that were absent from the 92 (EMC) and 70 genes (UAMS) selected by [Kuiper et al., 2015] as the most relevant genes for Multiple Myeloma.

**Baseline Classifiers:** For each training data set, the selected genes were used to train six baseline classifiers. The baseline classifiers included Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF), Gradient Boosting Machine (GBM), Learning Vector Quantization (LVQ) and Generalized Linear Model (GLM). The parameters of all baseline models were optimized with cross validation. Finally, we applied stacking by training ensemble classifiers to combine the results of the abovementioned baseline classifiers [Dzeroski and Zenko, 2004]. The basic idea of stacking is to train machine learning algorithms with training dataset and then generate a new dataset with these models. Then this new dataset is used as an input for the combined machine learning algorithm. Finally, a neural network learning algorithm that uses weighted voting is used as an ensemble classifier.

**Classifiers for Group Membership:** In order to figure out the right model to use, we need to first decide which group a given test data sample belongs to. We therefore trained classifiers to perform this task, where the four training sets, GSE24080UAMS, HOVON65, Kryukov and MMRF, were used to train.

**Classifying Test Data:** Any missing data in the test data sample was imputed with averaging. Next, the sample is classified into the right group using the group membership classifier mentioned above. Depending on the classification, the appropriate set of classifiers (baseline classifiers followed by the stacking method) are applied to the sample to classify its risk level.

**Software Tools:** All analysis was performed with R package caret, which in turn uses FCNN4R for NN, randomForest for RF, gbm for Gradient Boosting Machine, glm for Generalized Linear Model, e1071 for Support Vector Machines, and class for Learning Vector Quantization.

# Conclusions/Discussion

Our results show that the classifiers induced from balanced data sampled with SMOTE and Tomek links method are more accurate than those induced from the original data, and using stacking strategy to combine multiple machine learning methods produces improved results than a single model would.

# References

[G. Batista, B. Bazzan, M. Monard, “Balancing Training Data for Automated Annotation of Keywords: a Case Study,” In WOB, 10-18, 2003.](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.3904&rep=rep1&type=pdf)

[I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46, no. 1–3, pp. 389–422, 2002.](https://link.springer.com/article/10.1023%2FA%3A1012487302797?LI=true)

[Saso Dzeroski, Bernard Zenko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?" Machine Learning, vol. 54, pp. 255-273, 2004.](http://www-ai.ijs.si/SasoDzeroski/pdfs/2004/2004-DzeroskiEtAl-ML.pdf)

[Kuiper R, van Duin M, van Vliet MH, Broijl A, van der Holt B, El Jarari L, van Beers EH, Mulligan G, Avet-Loiseau H, Gregory WM, Morgan G, Goldschmidt H, Lokhorst HM, Sonneveld P, “Prediction of high- and low-risk multiple myeloma based on gene expression and the International Staging System,” Blood, 126(17):1996-2004, 2015.](https://www.ncbi.nlm.nih.gov/pubmed/26330243)

# Authors Statement

HG, AM and JZ worked on performing feature selection; HG implemented and ran the base classifiers and the stacking methods. HG implemented, tested and performed the final experiments, and prepared the leaderboard and final submissions. GN oversaw the entire training and testing process.